Variability in talkers' fundamental frequencies shapes context effects in speech perception[a)]

Ashley A. Assgari,[1] Rachel M. Theodore,[2] and Christian E. Stilp[1,b)]
[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 06828, USA*

(Received 16 April 2018; revised 15 February 2019; accepted 22 February 2019; published online 22 March 2019)

The perception of any given sound is influenced by surrounding sounds. When successive sounds differ in their spectral compositions, these differences may be perceptually magnified, resulting in spectral contrast effects (SCEs). For example, listeners are more likely to perceive /ɪ/ (low F_1) following sentences with higher F_1 frequencies; listeners are also more likely to perceive /ɛ/ (high F_1) following sentences with lower F_1 frequencies. Previous research showed that SCEs for vowel categorization were attenuated when sentence contexts were spoken by different talkers [Assgari and Stilp. (2015). J. Acoust. Soc. Am. **138**(5), 3023–3032], but the locus of this diminished contextual influence was not specified. Here, three experiments examined implications of variable talker acoustics for SCEs in the categorization of /ɪ/ and /ɛ/. The results showed that SCEs were smaller when the mean fundamental frequency ($f0$) of context sentences was highly variable across talkers compared to when mean $f0$ was more consistent, even when talker gender was held constant. In contrast, SCE magnitudes were not influenced by variability in mean $F1$. These findings suggest that talker variability attenuates SCEs due to diminished consistency of $f0$ as a contextual influence. Connections between these results and talker normalization are considered.
 https://doi.org/10.1121/1.5093638

[SHF] Pages: 1443–1454

I. INTRODUCTION

It has been established that when a listener hears different talkers, speech perception is slower and/or less accurate than when hearing speech from a single talker (e.g., Creelman, 1957; Fourcin, 1968; Assmann *et al.*, 1982; Geiselman and Bellezza, 1976; Mullenix *et al.*, 1989; Mullenix and Pisoni, 1990; Choi *et al.*, 2018). This general finding has been demonstrated in a variety of tasks including: recall of word lists (e.g., Goldinger *et al.*, 1991), word identification (e.g., Ryalls and Pisoni, 1997), vowel monitoring (e.g., Magnuson and Nusbaum, 2007), word monitoring (e.g., Magnuson and Nusbaum, 2007), consonant perception (e.g., Rand, 1971), and vowel perception (e.g., Assmann *et al.*, 1982). The decrease in accuracy and/or slower response times in different-talker compared to same-talker conditions have been attributed to the listener having to adjust to hearing a new talker on subsequent trials (e.g., Mullennix *et al.*, 1989). This effect has been referred to as talker normalization (e.g., Choi *et al.*, 2018), and we adopt the same terminology here.

Hearing different talkers influences speech perception, but precisely what it is about hearing different talkers that gives rise to slower and/or less accurate perception is not entirely clear. Several key differences exist between talkers including acoustic differences that are related to speech production. Speech production occurs through two related yet separate components: the source (related to fundamental frequency, or $f0$) and the filter (related to formants; Fant, 1973). Individuals' different vocal folds and vocal tracts lead to different combinations of $f0$ and formants across talkers, even when they are producing the same phonological content. While this combination varies for each individual, some general differences exist between men and women related to $f0$ (e.g., Peterson and Barney, 1952). In general, men have larger vocal folds that vibrate more slowly (leading to lower $f0$) and longer vocal tracts with lower resonances (leading to lower formant frequencies). Women generally have smaller vocal folds that vibrate more rapidly (higher $f0$) and shorter vocal tracts with higher resonances. While other differences exist between talkers, it is likely that the differences in $f0$ and formants could be driving the listener's need to adjust when hearing a new talker.

Goldinger (1996) offered some insight into listeners' sensitivity to low-level acoustic parameters that might distinguish talkers and thus guide talker normalization. In his first experiment, participants labeled pairs of words as the same or different. Every word pair was spoken by two different talkers (representing all possible combinations of five male and five female talkers), but listeners were instructed to ignore talker identity when responding. A multidimensional scaling analysis on response latencies to same-word trials revealed that the perceptual dimensions of gender and relative pitch influenced these judgments. In his second experiment, memory for spoken words was investigated as functions of (same/different) voice, number of talkers, and

[a)]Portions of these results were presented at the 171st (Salt Lake City, Utah) and the 173rd (Boston, Massachusetts) meetings of the Acoustical Society of America.
[b)]Electronic mail: christian.stilp@louisville.edu

study/test delay using stimuli from the scaling experiment. Different-voice trials tested both same-gender and different-gender talkers. Recognition and identification performance were each negatively correlated with intervoice distances in the scaling study. Put another way, when listeners heard words spoken by different talkers with similar voices (defined in large part by *f*0), recall was more accurate than when the words were spoken by different talkers with more dissimilar voices (Goldinger, 1996).

These results illustrate two important findings. First, fine-grained acoustic differences between talkers influence talker normalization. In many studies of talker normalization, the specific acoustic characteristics of the talkers are not explicitly reported (e.g., Creelman, 1957; Assmann *et al.*, 1982; Mullenix *et al.*, 1989; Mullenix and Pisoni, 1990; Choi *et al.*, 2018), though it is common for both male and female talkers to be used to create the different-talker conditions, likely resulting in large acoustic differences between talkers. Goldinger (1996) related the acoustic differences between talkers to the size of the talker normalization effect: as acoustic differences across talkers increased, reaction time increased and recall accuracy decreased. Second, they suggest that talker normalization may not occur every time a talker changes, but rather only when the acoustic differences across talkers are sufficiently large. Given that many talker normalization studies have not included detailed acoustic measurements of the talkers' voices, or analyses that link the magnitude of talker normalization to acoustic measurements of talkers' voices, a prevailing view is that talker normalization occurs with every change in talker. Recently, Choi and colleagues (2018) suggested that talker normalization is an obligatory aspect of speech processing. However, this study used male and female talkers (without reporting any measures of talker acoustics), so acoustic differences between the selected talkers are likely to be large. In contrast, Goldinger's (1996) findings suggest that talker normalization may be highly influenced by the acoustic differences between talkers, particularly with regard to differences in *f*0. In other words, talker normalization might only be obligatory to the degree that the acoustic input (particularly in terms of *f*0) is sufficient to cue a change in talker (or to which contextual knowledge influences listeners' expectations for multiple talkers; see Magnuson and Nusbaum, 2007).

In addition to talker information, spectral characteristics of earlier sounds also create a context for speech perception. Speech may possess spectral properties that are relatively stable or recurring over time. When these spectral properties change from a preceding context to a target sound, the change is perceptually magnified, which biases the perception of the target sound. The resulting perceptual bias is referred to as a spectral contrast effect (SCE). In speech perception, SCEs are often measured as a shift in categorization of a target phoneme following contexts with different spectral properties. In their seminal paper, Ladefoged and Broadbent (1957) observed that perception of a vowel sound could be changed based on the spectra of preceding sentences. When the first formant (F_1) of the preceding sentence was shifted down toward lower frequencies, the target vowel was perceived as the higher-F_1 /ɛ/ more often. When the F_1 of the preceding sentence was shifted up, the target vowel was perceived as the lower-F_1 /ɪ/ more often.

Ladefoged and Broadbent (1957) originally interpreted these results as a means for adjusting to differences between talkers. They suggested that if a listener learned the overall quality of a talker's speech, then that information could facilitate recognition of subsequent speech sounds from that talker. This interpretation was broadly consistent with later research on talker normalization. However, subsequent research with SCEs suggested that these effects might not be about talkers *per se* but rather stable spectral characteristics of the context (Huang and Holt, 2012). This suggestion gained support due to the pervasiveness of SCEs in speech and non-speech perception. Speech (e.g., Watkins, 1991; Watkins and Makin, 1994; Sjerps *et al.*, 2011, 2018; Stilp *et al.*, 2015; Assgari and Stilp, 2015; Stilp and Assgari, 2017, 2018, 2019) and non-speech contexts (e.g., Watkins, 1991; Holt, 2005, 2006) have been shown to bias categorization of speech targets; moreover, speech and non-speech contexts can also bias categorization of non-speech targets (Stilp *et al.*, 2010; Frazier *et al.*, 2019).

The original interpretation that SCEs were a means to adjust for talker differences was not directly tested for quite some time. To address this question, Laing *et al.* (2012) took a single sentence and amplified different regions of F_1 or F_3 to induce the perception of different talkers. This manipulation created pairs of talkers that differed in specific frequency regions. Listeners were able to discriminate the pairs of talkers when asked if they were different "voices." Laing *et al.* (2012) predicted that if SCEs are due to the long-term average spectrum of speech in the frequency region that differentiated the targets, then only the pairs of stimuli that differ in that region would produce SCEs. However, if SCEs are a general means for compensating for different talkers, then both pairs of stimuli should produce contrast effects. Following the context sentence, participants identified consonant sounds varying from /da/ to /ga/, which are primarily differentiated by F_3 transitions. SCEs were observed following F_3-manipulated contexts but not F_1-manipulated contexts. These results were replicated using sine tone contexts, leading the authors to suggest that talker information does not play a role in SCEs. However, adding a spectral peak to different frequency regions of a single sentence from a single talker fails to reflect the considerable acoustic differences that can exist between talkers. Notably, this method did not manipulate *f*0. It is known that *f*0 is a primary cue to talker identity (e.g., Hillenbrand and Clark, 2009), and perceptual consequences of talker variability are diminished between talkers who have similar *f*0s, even when their voices are discriminable on other acoustic parameters (e.g., Goldinger, 1996). Thus, the lack of a talker effect in Laing *et al.* (2012) may reflect the lack of *f*0 cuing changes in talker identity.

To model the acoustic differences of different talkers more explicitly, Assgari and Stilp (2015) measured contrast effects after speech from 200 different talkers, each speaking a different context sentence, with *f*0 freely varying across talkers. This method provided rich sources of acoustic information to cue talker variability. Two other conditions were

also tested: a single talker producing one sentence (presented 200 times) and a single talker producing 200 different sentences. Sentences were manipulated by amplifying either low-F_1 (100–400 Hz) or high-F_1 (550–850 Hz) frequency regions by 5 dB. Listeners were asked to report whether they heard the following target vowel as /ɪ/ (low F_1) or /ɛ/ (high F_1). The 200-talker condition produced smaller SCEs than both of the single-talker conditions, which produced equivalent SCEs. Thus, SCEs were sensitive to differences between talkers. This finding was later supported by Assgari (2018), who utilized one set of acoustically variable talkers but manipulated the order of stimulus presentation. When trials were organized so that sentence mean *f*0 increased monotonically throughout the block, SCEs were observed (with equivalent results when mean *f*0 decreased monotonically in a separate block). When the same sentences were presented in random orders across trials, creating high trial-by-trial variation in sentence mean *f*0, no SCE was observed. Together, these results suggest that talker variability, particularly in terms of *f*0 information, influences SCEs in speech categorization.

While talker variability appeared to restrict the magnitudes of context effects in Assgari and Stilp (2015), the exact cause of this effect was unclear because the context sentences were chosen independently of *f*0 characteristics and talker gender, both of which freely varied. *Post hoc* analyses offered some evidence that vowel targets were categorized similarly (i.e., not influenced by context) when context sentences exhibited large *f*0 variability, and that vowels were categorized differently (i.e., biased by context) when context sentences had more consistent *f*0s. While *f*0 variability may be restricting the influence of context on perception, *f*0 was not controlled during stimulus selection or presentation. Further, effects of *f*0 variability could not be separated from variability in talker gender between the context sentence and target vowel (which was always spoken by the same male talker). The results of Assgari (2018) suggest that predictability might play a role in the effects of talker variability, but hearing all male talkers followed by all female talkers (when mean *f*0 was arranged to increase throughout the block, or vice versa in the descending-*f*0 block) was far more orderly than talker gender being randomized in Assgari and Stilp (2015)'s paradigm. In the previous literature, there are conflicting reports as to how talker gender might influence SCEs. Watkins (1991) found that talker gender had no influence on SCEs, observing similar effects when the gender of the talker producing the context and the target matched (i.e., male context and male target) or differed (i.e., female context and male target). Lotto and Kluender (1998) also observed SCEs for female contexts preceding male targets, but these shifts were smaller than when talker gender matched across context and target. Thus, while the gender of the talker producing the context does not need to match that of the target to observe a contrast effect, it might affect the size of observed shift.[1] It is important to note that these studies all presented multiple renditions of a single context stimulus throughout the experiment. This more closely resembles the single talker, one sentence condition of Assgari and Stilp (2015) than their 200 talker, 200 sentences condition, so interpretations of the influence of talker gender on SCEs should be made with caution.

The current experiments aimed to identify the locus of the talker variability effect reported in Assgari and Stilp (2015) by testing the relative influences of three sources of talker variability. In experiment 1, the *f*0 variability of sentence contexts was manipulated (low *f*0 variability vs high *f*0 variability) and talker gender was blocked in each condition. In experiment 2, *f*0 variability of sentence contexts was again manipulated (low *f*0 variability vs high *f*0 variability) while intentionally mixing talker gender within each variability condition. In experiment 3, another source of low-level acoustic variability, F_1, was manipulated to form two variability conditions (low F_1 variability vs high F_1 variability); gender was again mixed in each variability condition. In all three experiments, SCEs were examined for target vowels produced by a single male talker. If the locus of the talker variability effect reported in Assgari and Stilp (2015) reflects variability in *f*0, then diminished SCEs will be observed in high-variability conditions compared to the low-variability conditions in experiments 1 and 2. If the talker variability effect on SCEs also reflects talker gender information, then SCEs should be attenuated when the talker differs between context and target (female talker conditions in experiment 1) and when talker gender varies randomly within each block (experiments 2 and 3). If the talker variability effect on SCEs reflects acoustic variability in general, and not variability in *f*0 specifically, then SCEs will also be diminished in the high-F_1-variability condition compared to the low-F_1-variability condition in experiment 3. Results from these studies strongly imply that variability in the talkers' fundamental frequencies, not their gender nor their F_1 frequencies, influences the degree to which preceding spectral context biases speech categorization.

II. EXPERIMENT 1

A. Methods

1. Participants

Twenty undergraduate students at the University of Louisville participated in exchange for course credit. All participants reported normal hearing and were native English speakers.

2. Stimuli

a. Sentences. Context sentences were taken from the Texas Instrument and Massachusetts Institute of Technology speech corpus (TIMIT; Garofolo *et al.*, 1990). Similar to Assgari and Stilp (2015), only sentences with relatively equal energy in the low F_1 (100–400 Hz) and high F_1 (550–850 Hz) regions (within ±5 dB of each other) were selected.

The mean *f*0 in each sentence was measured in Praat (Boersma and Weenink, 2017). Unvoiced segments of sentences were identified and removed prior to analyses. In rare cases, *f*0 contours were hand edited to ensure continuity (i.e.,

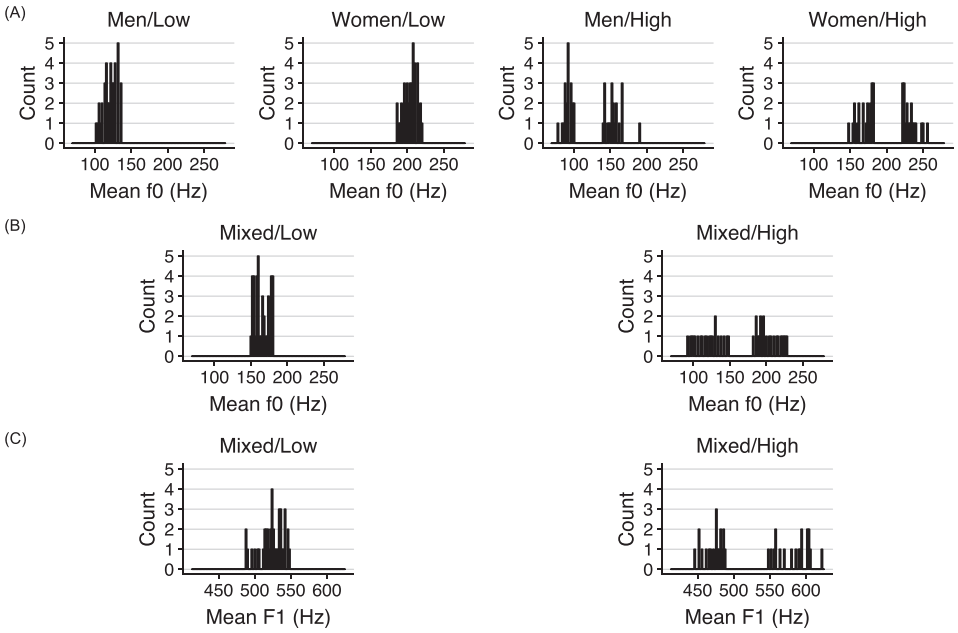

FIG. 1. Histograms depicting sentence stimuli in each condition in experiment 1 (A), experiment 2 (B), and experiment 3 (C). Stimuli were selected according to measures of mean *f*0 [(A) and (B)] or mean F_1 (C). Plot titles denote talker gender (Men, Women, Mixed) and variability condition (Low, High). As described in the main text, Low and High Variability histograms are complementary in that Low Variability stimuli fall between the tails of the High Variability stimuli for each gender in experiment 1, and for the mixed gender conditions in experiments 2 and 3.

removing spurious points that appeared highly unlikely to reflect talker pitch, such as points 1+ octave away from other pitch points). Mean *f*0 was calculated for each sentence, then a distribution of mean *f*0 values was formed for each talker gender. These distributions were used to create two experimental conditions for each talker gender, one with high *f*0 variability and one with low *f*0 variability [Fig. 1(A)]. Forty sentences were selected from the tails of each distribution for High Variability conditions (20 sentences from each tail), and 40 sentences were selected from the center of each distribution for Low Variability conditions. Thus, for each gender condition, the overall average of sentence mean *f*0 measures was well-matched across variability conditions, but the standard deviations of sentence mean *f*0s varied (see Table I). This method is distinct from that of Holt (2006), who manipulated the spectral mean and variability of the context (sequences of sine tones). Here, means and variability do not describe the spectral composition of the context sentences but rather their mean *f*0 characteristics.

Talker gender and *f*0 variability were fully crossed to form four groups (Table I). In each of these four conditions, sentences were randomly assigned to have either the low F_1 frequency region (100–400 Hz) or the high F_1 region (550–850 Hz) amplified by 5 dB using a bandpass finite impulse response filter with 1200 coefficients.

b. Vowels. Target vowels were a 10-step continuum of vowels ranging from /ɪ/ to /ɛ/. These vowels were the same as those used in previous SCE studies (Stilp *et al.*, 2015; Assgari and Stilp, 2015; Stilp and Alexander, 2016; Stilp and Assgari, 2018). Vowels were synthesized based on natural recordings from a male talker. These speech samples were resynthesized using Linear Predictive Coding in Praat (Boersma and Weenink, 2017). The /ɪ/ endpoint has an F_1 that linearly increased from 400 to 430 Hz while F_2 linearly decreased from 2000 to 1800 Hz. The /ɛ/ endpoint has 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

TABLE I. Summary of experimental conditions (Assgari and Stilp represents experiment 2 in Assgari and Stilp, 2015). Measure indicates the acoustic measure by which stimuli were sorted into high or low variability conditions. Gender indicates whether the gender of the talkers who spoke the context sentences was blocked (Men, Women) or mixed. Means indicate the overall averages of sentence mean *f*0 or mean F_1 measures in each experimental block, and SD conveys the standard deviations of these measures. The magnitude of the SCE for each block is listed in the final column, calculated as the shift in 50% points (measured in number of stimulus steps) across low-F_1-amplified and high-F_1-amplified logistic functions (see Results sections for details of linear mixed-effects models, and Sec. IV B for SCE calculation).

				Mean *f*0		Mean F_1		
Experiment	Measure	Variability	Gender	Mean	SD	Mean	SD	SCE
1	*f*0	High	Men	123.39	33.18	505.07	35.07	0.22
1	*f*0	High	Women	199.84	33.27	563.11	48.28	0.18
1	*f*0	Low	Men	121.29	9.16	504.10	28.32	0.54
1	*f*0	Low	Women	203.89	9.17	545.30	44.62	0.36
2	*f*0	High	Mixed	161.78	45.64	521.08	45.62	0.23
2	*f*0	Low	Mixed	164.81	9.78	529.43	45.26	0.51
3	F_1	High	Mixed	168.27	31.12	527.13	62.42	0.44
3	F_1	Low	Mixed	158.33	34.12	523.37	16.16	0.38
Assgari and Stilp	*f*0	High	Mixed	148.31	42.94	538.42	54.63	0.28

taking these endpoint vowels and linearly morphing their formant tracks through a script in Praat (for more detail see Winn and Litovsky, 2015). Final vowel stimuli were 246 ms in duration with an *f*0 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 stimuli were upsampled to 44 100 Hz.

3. Procedure

After obtaining informed consent, the listener was seated in a sound-attenuating booth (Acoustic Systems, Inc., Austin, TX). Stimuli were D/A converted by RME HDSPe AIO sound cards (Audio AG, Haimhausen, Germany) on personal computers and passed through a programmable attenuator (TDT PA4, Tucker-Davis Technologies, Alachua, FL) and a headphone buffer (TDT HB6). Stimuli were presented over circumaural headphones (Beyerdynamic DT-150, Beyerdynamic Inc. USA, Farmingdale, NY) at a mean presentation level of 70 dB sound pressure level. A custom script in MATLAB led the listener through the experiment, which was self-paced. The listener clicked the mouse to label the target vowel as "ih" as in "bit" or "eh" as in "bet."

Listeners first completed a set of practice trials. These trials consisted of 20 sentences from the AzBio corpus (Spahr *et al.*, 2012) paired with the endpoint vowels, as categorizing endpoints of the vowel continuum is objectively correct or incorrect. If the listener failed to reach 80% accuracy on endpoint vowels after one block, s/he repeated the practice trials up to two more times to reach 80% accuracy. If after three blocks of practice trials s/he did not achieve 80% accuracy, s/he did not continue to test.

The experiment was comprised of four blocks. Each block consisted of 160 trials (four repetitions of each unique sentence) and took about 12 min to complete. Block orders were counterbalanced across participants. Participants were allowed to take breaks in between blocks. The entire session lasted approximately 1 h.

B. Results

Participants were required to maintain 80% accuracy on vowel endpoints in each block in order for their data to be included in statistical analyses. One listener failed to achieve 80% accuracy in any of the four blocks, so his/her data were removed from subsequent analyses.

Trial-level data were analyzed in a generalized linear mixed-effects model in R (R Development Core Team, 2018) 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 model included fixed effects of Target, Filter, Gender, Variability, and all interactions between these factors. Target was entered into the model as a continuous variable (step 0–step 9), centered around the mean. Contrast-coding was used for the fixed effects of Filter (high F_1 amplification = −0.5, low F_1 amplification = 0.5), Gender (male = −0.5, female = 0.5), and Variability (low *f*0 variability = −0.5, high *f*0 variability = 0.5). The model also included random intercepts by subject and random slopes by subject for Target, Filter, Gender, and Variability. All models were run using bobyqa optimization with a maximum of 800 000 iterations.[2]

TABLE II. Beta estimate (β), SE, z, and p for the fixed effects of the mixed-effects model for experiment 1. As described in the main text, Target was entered in the model as a continuous factor, centered around the mean. Filter, Gender, and Variability were contrast-coded; the level associated with the −0.5 contrast for each factor is shown in parentheses.

Effect	β	SE	z	p
Intercept	0.340	0.152	2.242	0.025
Target	1.270	0.070	18.067	<0.001
Filter (High F_1)	0.426	0.075	5.687	<0.001
Gender (Male)	0.251	0.123	2.040	0.041
Variability (Low)	−0.065	0.070	−0.916	0.360
Target × Filter	0.048	0.041	1.156	0.248
Target × Gender	0.016	0.042	0.374	0.708
Filter × Gender	−0.123	0.129	−0.953	0.341
Target × Variability	−0.010	0.041	−0.231	0.817
Filter × Variability	−0.350	0.129	−2.711	0.007
Gender × Variability	−0.007	0.129	−0.057	0.954
Target × Filter × Gender	0.057	0.079	0.728	0.467
Target × Filter × Variability	−0.023	0.078	−0.290	0.772
Target × Gender × Variability	−0.224	0.079	−2.850	0.004
Filter × Gender × Variability	0.069	0.258	0.269	0.788
Target × Filter × Gender × Variability	0.107	0.156	0.682	0.495

The model results are listed in Table II. The model is visualized in Fig. 2(A), in terms of /ɛ/ responses as predicted by the fixed effects of Target, Filter, and Variability (collapsing across Gender), which was created using the jtools package in R (Long, 2018). 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 also a main effect of context Filter, indicating that changing the filtering condition from high F_1 to low F_1 increased the probability of /ɛ/ responses, confirming the presence of SCEs. Consistent with our predictions, there was a significant interaction between Filter and Variability ($p = 0.007$). The negative sign on this coefficient indicates that SCE magnitudes were larger in the Low Variability condition than the High Variability condition.

There was a significant effect of Gender ($p = 0.041$; more /ɛ/ responses following female talkers) and a significant interaction between Target, Gender, and Variability ($p = 0.004$). Figure 3 shows the results of separate mixed-effects models for each talker gender, following the structure of the primary model save for removing the fixed effect of Gender. The three-way interaction observed in the primary model reflects more /ɛ/ responses to female talkers on the lower-F_1 half of the vowel continuum in the Low Variability condition and the higher-F_1 half of the continuum in the High Variability condition. These patterns, while intriguing, do not bear on the research questions of interest, which was the degree(s) to which SCEs varied as functions of talker gender and acoustic variability, rather than the effect(s) of gender and/or variability on the

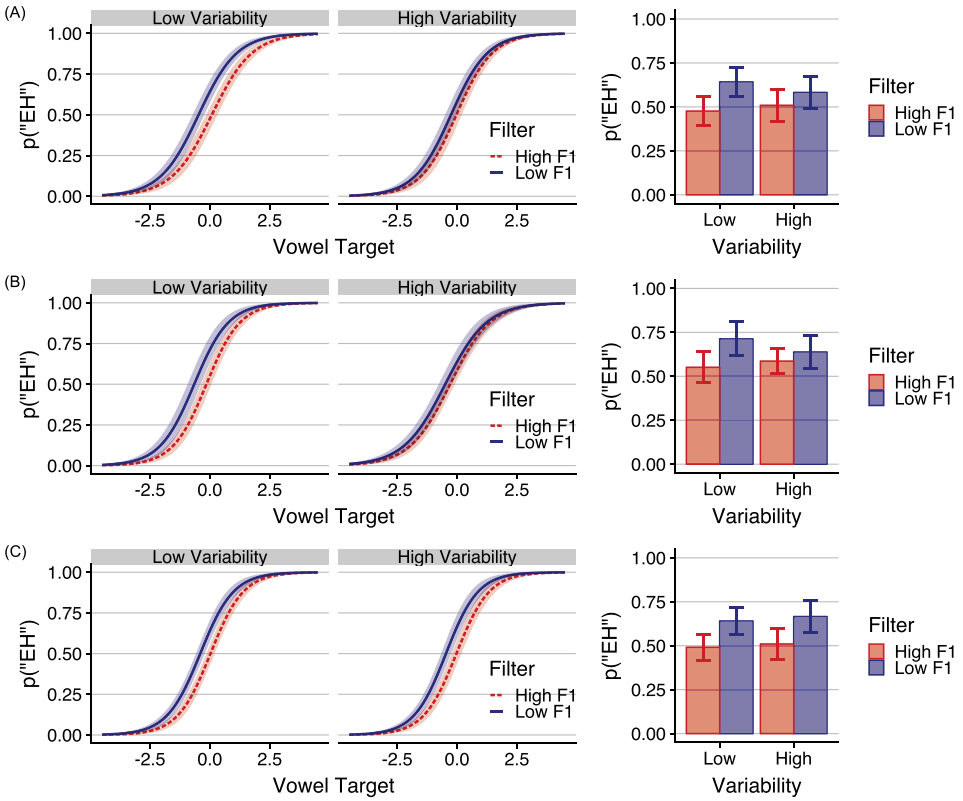

FIG. 2. (Color online) Results of the generalized linear mixed-effects models performed for experiment 1 [(A), collapsing across Gender], experiment 2 (B), and experiment 3 (C). In each panel, the plots at left show predicted /ɛ/ responses as a function of the fixed effects of Target Vowel, Variability, and Filter. Shaded regions indicate the 90% confidence interval. As described in the main text, Vowel Target was entered into the model centered around the mean continuum step, which is shown as 0 in the plots at left. The plots at right in each panel show predicted /ɛ/ responses as a function of the fixed effects of Variability and Filter (thus collapsing over Target Vowel) to illustrate the relationship between the SCE (i.e., filter manipulation) and variability condition. Error bars denote the 90% confidence interval.

probability of eliciting a particular vowel response. There was no evidence that SCE magnitudes varied as a function of talker gender, given the null interactions between Filter and Gender ($p = 0.341$) and Target, Filter, and Gender ($p = 0.467$) in the primary model.

C. Discussion

Consistent with predictions, context effect magnitudes interacted with *f*0 variability: SCEs were larger in the Low *f*0 Variability condition than in the High *f*0 Variability condition. These results parallel those of Goldinger (1996), who demonstrated that the degree to which talker normalization occurred was influenced by *f*0 similarity among talkers. In his study, when talkers were more acoustically different from each other (and thus more variable), word recall accuracy suffered to a greater degree than when talkers were more acoustically similar. Despite the fact that these tasks had very different dependent variables (i.e., recall accuracy vs categorization shifts), they were both influenced by variability in talker *f*0. This parallel is revisited in Sec. V.

Context effect magnitudes did not vary depending on whether talker gender(s) matched between the context sentence and the vowel target: SCE magnitudes were similar when context sentences were spoken by men and women. This indicates that genders of the talker producing the context (male or female) and the target vowels (male) did not have to match to observe SCEs. These results are consistent with the findings of Watkins (1991) and Lotto and Kluender (1998), where talker gender differed across (female) context

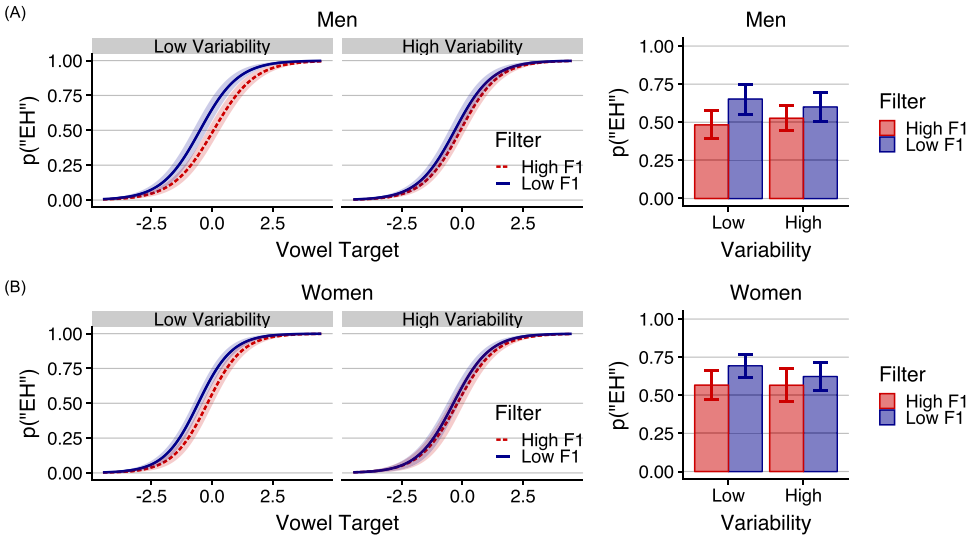

FIG. 3. (Color online) Results of the mixed-effects models performed for experiment 1 separately for the Men blocks (A) and the Women blocks (B). In each panel, the plots at left shows predicted /ɛ/ responses as a function of Target Vowel, Variability, and Filter. Shaded regions indicate the 90% confidence interval. Vowel Target was entered into the model centered around the mean continuum step, which is shown as 0 in the plots at left. The plots at right in each panel shows predicted /ɛ/ responses as a function of Variability and Filter (thus collapsing over Target Vowel) to illustrate the relationship between the SCE (i.e., filter manipulation) and Variability condition. Error bars denote the 80% confidence interval.

and (male) target but SCEs were still observed. Further, experiment 1 extends previous findings by demonstrating this pattern across 40 different female talkers, as opposed to two different renditions of a single token spoken by a woman (as tested in Watkins, 1991; Lotto and Kluender, 1998).

In Assgari and Stilp (2015), SCEs in vowel categorization were smaller when talker gender and sentence mean *f*0 freely varied within a block compared to hearing a single talker throughout the block. The results of experiment 1 suggest that this result was due at least in part to variability in mean *f*0 of the context sentences. However, talker gender varied unpredictably during that block of their experiment, whereas gender was blocked in the current experiment 1. Therefore, it is unclear whether talker gender variability has a similar influence as *f*0 variability on these context effects in vowel categorization. Experiment 2 investigated whether variability in talker gender restricts SCE magnitudes beyond the observed influence of *f*0 variability. If talker gender variability does affect SCEs, then the context effects observed in the single-gender conditions of experiment 1 are predicted to be attenuated owing to the additional source of variability among the context sentences. If talker gender variability has no additional influence on SCEs, the Low Variability condition will again show larger SCEs than the High Variability condition.

III. EXPERIMENT 2

A. Methods

1. Participants

Twenty undergraduate students at the University of Louisville participated in exchange for course credit. None participated in experiment 1. All participants reported normal hearing and were native English speakers.

2. Sentences

As in experiment 1, TIMIT sentences with relatively equal energy in low-F_1 and high-F_1 frequency regions (within ±5 dB of each other) were selected as stimuli. There was no requirement that sentences from experiment 1 were not included in experiment 2, so 41 sentences were repeated across experiments. Again, *f*0 measures of candidate sentences were obtained through Praat (Boersma and Weenink, 2017). A single distribution of sentence mean *f*0 was created with talker gender intentionally mixed. Forty Low Variability sentences were pulled from the center of this distribution, and 40 High Variability sentences were pulled from the tails of the same distribution [20 sentences from each tail; see Table I and Fig. 1(B)]. In general, male talkers were pulled from the lower tail (lower *f*0) of the distribution while female talkers were pulled from the upper tail (higher *f*0). Gender was balanced within each condition (20 male sentences and 20 female sentences), making gender variability equal across conditions.

3. Procedure

The procedure was the same as in experiment 1 except that there were only two blocks: Low Variability and High Variability. Each block lasted approximately 12 min, and the entire experiment lasted about 40 min.

B. Results

All participants met the performance criterion of 80% accuracy on vowel endpoints in every block, so all data were included in analyses. Results were again analyzed in a generalized linear mixed-effects model in R (R Development Core Team, 2018) using the lme4 package (Bates *et al.*, 2014). The model had similar architecture to that detailed in experiment 1, but reflected the smaller experimental design. Responses were coded as 0 (responding /ɪ/) and 1 (responding /ɛ/), and the model tested the fixed effects of Target, Filter, Variability, and their interactions, as detailed above. Random effects included random intercepts by subject and random slopes by subject for Target, Filter, and Variability.

Model results are listed in Table III. The model is visualized in Fig. 2(B) in terms of the fixed effects of Target, Filter, and Variability as predictors of /ɛ/ responses. The intercept of the model was significant, indicating that participants responded /ɛ/ more than they responded /ɪ/ in experiment 2. As expected, there was a significant effect of Target, such that each rightward step along the vowel continuum increased the log odds of participants responding /ɛ/. There was an interaction between Target and Variability, indicating that /ɛ/ responses differed between the two variability conditions across the vowel continuum, but this result does not bear on the principal question of interest for the current study. As in experiment 1, the significant effect of Filter reflected the increased probability of /ɛ/ responses following low-F_1-amplified context sentences as compared to high-F_1-amplified sentences, confirming the presence of SCEs. Critically, there was a significant interaction between Filter and Variability, with the coefficient direction indicating that SCEs were significantly smaller in the High Variability condition than in the Low Variability condition. No other main effects or interactions were statistically significant.

In both experiments, greater variability in mean *f*0 produced smaller SCEs than blocks containing less variability

TABLE III. Beta estimate (β), SE, z, and p for the fixed effects of the mixed-effects model for 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 contrast-coded; the level associated with the −0.5 contrast for each factor is shown in parentheses.

Effect	β	SE	z	p
Intercept	0.509	0.174	2.929	0.003
Target	1.285	0.097	13.252	<0.001
Filter (High F_1)	0.462	0.114	4.056	<0.001
Variability (Low)	−0.099	0.116	−0.854	0.393
Target × Filter	−0.034	0.053	−0.645	0.519
Target × Variability	−0.237	0.054	−4.434	<0.001
Filter × Variability	−0.485	0.176	−2.765	0.006
Target × Filter × Variability	−0.012	0.100	−0.119	0.905

in mean *f*0 (Filter × Variability interactions). However, the question remains as to whether within-block gender variability (experiment 2) played any role in restraining SCE magnitudes compared to gender being constant throughout each block (experiment 1). To address this question, an additional mixed-effects model analysis was conducted on the combined results from both experiments [akin to testing for differences across Figs. 2(A) and 2(B)]. Model architecture included everything described in the analysis of experiment 2, along with a fixed effect of experiment and interactions between experiment and all other fixed effects. As expected, model terms that were significant for both experiments 1 and 2 (Tables II and III) were significant in the joint model: Intercept (overall bias toward /ɛ/ responses; $p < 0.0001$), Target (more /ɛ/ responses to higher-F_1 vowel targets; $p < 0.0001$), Filter (the presence of SCEs; $p < 0.0001$), and Filter by Variability (modulation of SCE magnitudes by Variability condition; $p = 0.0001$). Variability in talker gender did not affect SCE magnitudes across experiments [Filter × experiment: $\beta = 0.019$, standard error (SE) $= 0.126$, $z = 0.147$, $p = 0.883$] nor the relationship between SCEs and variability in mean *f*0 (Filter × Variability × Experiment: $\beta = -0.105$, SE $= 0.216$, $z = -0.484$, $p = 0.628$). There was no clear evidence that talker gender variability restrained SCE magnitudes, which are instead attributable to greater mean *f*0 variability in the context sentences.

C. Discussion

Together, the results of experiments 1 and 2 shed considerable light on why talker variability diminishes SCE magnitudes in vowel categorization. In Assgari and Stilp (2015), SCEs were diminished when context sentences were spoken by 200 different talkers, with talker gender and sentence mean *f*0 freely varying. In experiment 1, SCEs were diminished in High *f*0 Variability conditions compared to Low *f*0 Variability conditions, and talker gender was blocked. In experiment 2, talker gender freely varied but the same pattern of results was observed. Variability in talker gender appeared to have no additional influence on these spectral context effects. Thus, preceding spectral context influenced speech categorization more when hearing similar-sounding talkers (i.e., Low *f*0 Variability) than when hearing different-sounding talkers (i.e., High *f*0 Variability).

In experiment 2, the Low Variability condition tested talkers with similar mean *f*0s, so even though trials were presented in random order, each successive trial presented a moderately-to-very similar-sounding talker (at least in terms of mean *f*0). Conversely, the High Variability condition tested talkers with highly disparate mean *f*0s, and thus the talker on each successive trial was often relatively unpredictable. These results parallel Assgari (2018), who manipulated predictability on a trial-by-trial basis. When mean *f*0 incrementally increased on each successive trial (and, in a separate block, decreasing mean *f*0), SCEs were robust; when presenting the same stimuli in random orders, SCEs were not observed. Therefore, predictability [as implemented on a block level here, or a trial-by-trial level in Assgari (2018)] plays an important role in modulating context effects in speech perception. This point is discussed further in Sec. V.

Thus far, it appears that variability in sentence mean *f*0 restricts context effects in speech categorization. Yet, across sentences and across talkers, there are many concurrent sources of acoustic variability. As previously mentioned, talkers vary on a wide variety of acoustic parameters, and *f*0 might not be the most influential acoustic property for restricting context effects in speech categorization. For example, despite there being no physiological obligation for *f*0 and formant frequencies to share a relationship, substantial covariance exists between *f*0 and formants across talkers (Kluender *et al.*, 2013). Given their covariance, it is possible that sorting stimuli into Low *f*0 Variability and High *f*0 Variability conditions in experiments 1 and 2 also sorted their formant frequencies into Low and High Variability groups. Of particular interest to the current report is variability in the region of F_1 because the target vowels /ɪ/ and /ɛ/ are primarily differentiated on F_1. F_1 information is expected to be highly variable across sentences (depending on phonemic content) and across talkers (with different vocal tract lengths), but it is an open question whether variability in mean F_1 characteristics of context sentences could similarly influence SCEs as observed for variability in mean *f*0. Experiment 3 investigated this possibility directly by manipulating F_1 variability across context sentences. If F_1 variability across context sentences has a similar influence on SCEs as does *f*0 variability, then SCE magnitudes are predicted to be smaller in a High Variability condition relative to a Low Variability condition.

IV. EXPERIMENT 3

A. Methods

1. Participants

Twenty undergraduate students at the University of Louisville participated in exchange for course credit. None participated in experiments 1 or 2. All participants reported normal hearing and were native English speakers.

2. Sentences

Sentences from experiment 2 were rearranged and presented based on measures of mean F_1. The average F_1 of each sentence was measured in Praat (Boersma and Weenink, 2017). Similar to experiment 2, unvoiced segments of sentences were removed before formant frequencies were analyzed. Formant contours were hand edited to ensure continuity (removing spurious points that did not appear to reflect the talkers' voices). A distribution was then created based on these measurements of mean F_1 [Fig. 1(C)]. Forty Low Variability sentences were selected from the center of this distribution, and 40 High Variability sentences were pulled from the tails of the same distribution (20 sentences from each tail; see Table I). While it was not possible to explicitly balance talker gender, it was almost balanced across High Variability (21 females) and Low Variability (19 females) conditions. Therefore, it does not

TABLE IV. Beta estimate (β), SE, z, and p for the fixed effects of the mixed-effects model for experiment 3. As described in the main text, Target was entered in the model as a continuous factor, centered around the mean. Filter and Variability were contrast-coded; the level associated with the −0.5 contrast for each factor is shown in parentheses.

Effect	β	SE	z	p
Intercept	0.318	0.153	2.081	0.037
Target	1.458	0.109	13.358	<0.001
Filter (High F_1)	0.632	0.115	5.515	<0.001
Variability (Low)	0.093	0.111	0.837	0.402
Target × Filter	0.005	0.060	0.086	0.932
Target × Variability	0.028	0.060	0.463	0.643
Filter × Variability	0.037	0.183	0.200	0.842
Target × Filter × Variability	0.050	0.116	0.434	0.665

present as a confound with respect to the key F_1 variability manipulation.

3. Procedure

The procedure was the same as outlined for experiment 2. The entire experiment lasted about 40 min.

B. Results

All participants met the performance criterion of 80% accuracy on vowel endpoints in every block, so all data were included in analyses. Results were analyzed in a mixed-effects model with the same architecture as that detailed in experiment 2. Model results are listed in Table IV and visualized in Fig. 2(C). The intercept of the model was significant, indicating more /ɛ/ compared to /ɪ/ responses. There was a significant effect of Target ($p < 0.001$), indicating that /ɛ/ responses increased along the vowel continuum towards the /ɛ/ endpoint. The main effect of Filter was also significant ($p < 0.001$), confirming the presence of SCEs. There was no evidence to suggest that the magnitude of the SCE was influenced by F_1 variability, given the null interactions between Filter and Variability ($p = 0.842$), and between Target, Filter, and Variability ($p = 0.665$).

Previous literature has established that SCE magnitudes are linear in nature: when stable spectral properties in context sentences are made more prominent (i.e., by using higher filter gain), thus increasing the size of the spectral difference between context and target spectra, categorization shifts increase linearly (Stilp *et al.*, 2015; Stilp and Alexander, 2016; Stilp and Assgari, 2017). This demonstrates the importance of measuring the magnitudes of categorization shifts and not just noting their presence or absence. Given this relationship, it is possible that SCE magnitudes also scale linearly with the amount of variability in mean *f*0 across context sentences. To test this possibility, a mixed-effects linear regression was conducted to predict SCE magnitude. This analysis was deemed appropriate given that different listener groups produced different numbers of SCEs; listeners completed four blocks (and thus four SCEs were measured) in experiment 1, two blocks each in experiments 2 and 3, and one block in the 200 Talkers/200 Sentences condition of Assgari and Stilp (2015). The standard deviation of *f*0 in each experimental block (listed in Table I) was entered as the fixed effect, and a random intercept was included for listener group (to match the fact that SCEs are calculated at the group level). SCEs, the dependent measure in this analysis, were calculated from mixed-effects models following the methods used in previous studies (Stilp *et al.*, 2015; Stilp and Assgari, 2017, 2018). In a given experimental block, the model fit a logistic regression to group-level responses following low-F_1-amplified-contexts, and fit a separate regression to group-level responses following high-F_1-amplified-contexts. The 50% point was derived from each regression function, then translated to the corresponding stimulus number along the vowel continuum (from 0 to 9; this number was interpolated as needed). The SCE was operationalized as the difference between the 50% points for the low-F_1 and high-F_1 context conditions, measured in stimulus steps. The results from Assgari and Stilp (2015) were reanalyzed using a mixed-effects model in order to match analyses of the present results. All SCE magnitudes are reported in the final column of Table I.

Variability in mean *f*0 was a significant predictor of SCE magnitude [$\beta = -0.008$, SE = 0.002, $t(5.210) = -4.646$, $p = 0.005$, with degrees of freedom produced using the Satterthwaite approximation as implemented in the lmerTest package (Kuznetsova *et al.*, 2013)]. As variability in mean *f*0 increased, the size of the SCE decreased (Fig. 4). It is interesting to note that SCEs observed in experiment 3 are well predicted by the regression line despite being arranged based on their F_1 variability. Thus, across different experiments with different listeners and various context sentences, a strong relationship between mean *f*0 variability and SCE magnitude is apparent. Future studies of SCEs using multiple talkers and/or

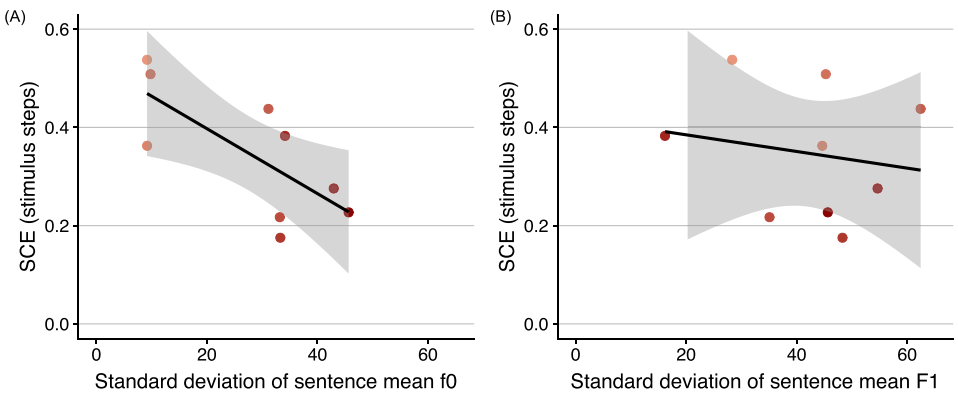

FIG. 4. (Color online) SCE magnitudes plotted as a function of standard deviation of mean *f*0 (A) and mean F_1 (B) in each corresponding experimental block. The solid line is the linear regression fit to these data, with the surrounding shaded region indicating the 95% confidence interval. All measures and corresponding listener groups are listed in Table I.

contexts should provide measures of *f*0 variability, as they may influence the magnitudes of the effects under study.

Context sentences rearranged based on F_1 variability in experiment 3 did not influence SCE magnitudes in vowel categorization. However, it is possible that the measures of F_1 variability in experiment 3 alone did not constitute a large enough sample to define a potential relationship between F_1 variability and SCE magnitudes. A mixed-effects linear regression analysis was conducted parallel to the one reported for variability in mean *f*0 (Fig. 4). Variability in mean F_1 was entered as the fixed effect (listed in Table I), and a random intercept was entered for each listener group. Variability in mean F_1 was a poor predictor of SCE magnitudes [$\beta = -0.002$, SE = 0.004, $t(7) = -0.483$, $p = 0.643$].

C. Discussion

The results suggest that variability in mean F_1 does not influence SCEs in vowel categorization. Thus, not all sources of acoustic variability equally restrict the influence of context on speech perception. Interestingly, despite SCE magnitudes not varying as a function of F_1 variability, these effect magnitudes were predictable based on the *f*0 variability of these conditions arranged by their F_1 variability (Fig. 4). It is important to note, however, that the lack of a relationship F_1 variability and the magnitude of SCEs may reflect the specific range of variability examined in the current work. Recall that the range of F_1 variability used presently was constrained by that present in the stimuli used for experiment 2. It may be the case that if a large range of F_1 variability were present, then it may show an influence on SCEs.

V. GENERAL DISCUSSION

When spectral properties differ across earlier (context) and later (target) sounds, speech categorization becomes biased through SCEs. It has been demonstrated that SCEs are smaller when sentence contexts are spoken by different talkers (Assgari and Stilp, 2015). However, exactly why hearing different talkers restricted these context effects was not clear. The results from experiments 1 and 2 confirm the influence of mean *f*0 variability across context sentences on SCEs in vowel categorization. When the mean *f*0s of sentences were highly variable in a given experimental block, SCEs were significantly smaller than when mean *f*0s were less variable. These experiments also demonstrated that variability in talker gender had no influence on SCEs, whether between contexts and targets or between single- or mixed-gender talkers within a block. The stimuli from experiment 2 were rearranged into blocks based on Low or High Variability in mean F_1 across sentences, but blocking by F_1 variation did not differentially affect SCEs. Thus, not all sources of acoustic variability have equal influences on context effects; variability in mean *f*0 restricted the magnitudes of SCEs in vowel categorization but variability in mean F_1 (in the stimuli presented in experiments 2 and 3) did not.

While variability in talkers' mean *f*0s differentially affected spectral context effect magnitudes, the precise mechanism underlying these effects is unclear. Many different spectral properties of context sentences produce SCEs (see Stilp *et al.*, 2015 for review). Additionally, previous research has clearly established that preceding contexts need not be speech in order to produce SCEs in speech categorization (e.g., Watkins, 1991; Holt, 2005). So, how does talker variability change the magnitudes of these context effects? The present experiments arranged stimuli into blocks to have high or low variability in mean *f*0, but many other acoustic parameters could have been high or low variability concurrently. SCE magnitudes did not differ as a function of mean F_1 variability in experiment 3, but the stimuli tested were the same sentences as presented in experiment 2. While these results introduced and extinguished differences in SCE magnitudes depending on stimulus arrangement, it does not completely rule out influences of variability in F_1 (or other acoustic parameters). The difference between low and high F_1 variability in experiment 3 was potentially limited given the constraint of rearranging stimuli tested in experiment 2. Future research will reveal the degree to which arrangement of new stimuli into low/high *f*0 variability and even lower/even higher F_1 variability blocks mirrors the present results.

A second possible mechanism underlying the present results is the overlap between talkers' *f*0 characteristics and the frequency region designated as low F_1 (100–400 Hz). Variability in talkers' mean *f*0s might also have introduced variability in how spectral peaks in the low-F_1 region were realized. The spectral manipulations that produced SCEs in the present experiments (adding low-F_1 or high-F_1 spectral peaks to context sentences) are low-order harmonics of *f*0, so variability in talker *f*0 would result in variability in the realization of these spectral peaks (such as harmonic spacing). Research is currently underway to examine the influence of talker variability on categorization of different speech targets whose distinguishing spectral features do not overlap with talker *f*0 characteristics (e.g., /da/ - /ga/, which differ principally in F_3 onset frequencies above 2 kHz).

A third, non-exclusive possible mechanism is increased cognitive load. Listeners adapt or calibrate to a talker's speech, but hearing different talkers in succession requires repeated recalibration to new talkers. This recalibration introduces perceptual costs (longer reaction times and/or lower accuracy) frequently reported in studies comparing perception of one versus multiple talkers as reviewed in Sec. I. Diminished context effect magnitudes are another perceptual cost incurred by hearing different talkers (Assgari, 2018), and that could also underlie the present results (but see Bosker *et al.*, 2017).

In the current work, hearing talkers with highly variable mean *f*0s diminished the influence of context on subsequent vowel perception. This outcome is problematic for speech perception because SCEs serve to disambiguate perceptually ambiguous stimuli. When contrast effect magnitudes are diminished, ambiguous speech sounds are not disambiguated and speech categorization suffers and/or is slower (Stilp, 2017; Assgari, 2018). Thus, parallel to findings from talker normalization literature, talker variability—when cued by substantial variation in *f*0—leads to diminished use of systematic contextual information for optimizing speech perception. Moreover, the degree to which SCEs are attenuated is linked to the degree of *f*0 variability (Fig. 4). This

converges with findings showing that the degree to which perception is disrupted by hearing multiple talkers depends on the degree to which they differ, particularly in terms of *f*0 (Goldinger, 1996).

The broad parallel between talker normalization and SCEs deepens when considering the acoustic consequences of hearing different talkers. The current investigation of *f*0 variability restricting context effects on speech categorization closely resembles effects of *f*0 variability on talker normalization. In the different-voice trials of Goldinger's (1996) experiment 2, listeners were less accurate recalling words spoken by acoustically variable talkers than words spoken by acoustically similar talkers. Here, when talkers were more acoustically variable in terms of mean *f*0, smaller categorization shifts were observed compared to when more acoustically similar talkers were heard (Fig. 4). Assgari (2018) showed that these diminished categorization shifts for acoustically variable talkers were accompanied by slower response times. In these cases, the influence of context was mediated by the *f*0 variability in that context. These results might not necessarily be surprising considering that when talkers are more acoustically similar, they are harder to tell apart (Magnuson and Nusbaum, 2007).

We have approached talker normalization and SCEs as separate phenomena in speech perception, consistent with how they have been explicated in the literature. It is possible, however, that they reflect similar consequences of tracking systematic variability in the input. When the input is stable across trials with respect to talker, listeners are receiving processing benefits (higher accuracy and/or faster response times) that presumably reflect the ability to track that systematic structure, which is indeed consistent with the literature on talker familiarization (e.g., Nygaard *et al.*, 1994). Within this framework, the current results provide additional evidence that the ability to use contextual information to optimize speech perception is increased when context remains consistent. Namely, SCEs were larger when the context was more consistent (in terms of mean *f*0) compared to when it was more variable; consistency in terms of talker gender between context and target or mean F_1 of context did not have comparable effects. It is possible that a strict dissociation between talker normalization and SCEs is not warranted to the degree that they both reflect perceptual consequences of talker variability. Indeed, it may be preferable to consider diminished SCEs as new evidence of talker normalization as, just like increased reaction time and impaired recall accuracy, they reflect a decreased use of context to facilitate perception. Future work is needed to test this hypothesis directly.

While the parallel between talker normalization and SCEs is alluring, there exists plenty of room for these research programs to converge. As previously mentioned, not all studies of talker normalization report *f*0 (or other acoustic) measurements of the talkers, though it can be inferred that *f*0 variability is relatively high in studies that use both male and female talkers. When hearing talkers with high *f*0 variability in SCE experiments, the magnitudes of the resulting categorization shifts were diminished (i.e., smaller context effects). However, when talkers are more acoustically similar in terms of their *f*0s, talker normalization effects are diminished (Goldinger, 1996) and SCEs remained intact. It is possible that these effects are more similar than previously thought, but additional research using more comparable outcome variables than accuracy and category boundary shifts would be needed to more clearly define this relationship. Assgari (2018) observed slower response times and smaller categorization shifts for acoustically variable (in terms of mean *f*0) talkers compared to acoustically similar talkers, giving this pursuit merit. Furthermore, though processing of multiple talkers' speech initially leads to decreased perceptual performance compared to single-talker conditions, research shows that listeners are able to adapt to talker variability when given increased exposure to talkers' voices, resulting in improved performance for familiar compared to unfamiliar talkers (e.g., Nygaard *et al.*, 1994). Such talker-specific adaptation emerges even with minimal exposure to talkers' voices, and has been shown to promote benefits to word intelligibility, processing time, recognition memory, and mapping to individual speech sounds (Clarke and Garrett, 2004; Nygaard *et al.*, 1994; Bradlow and Pisoni, 1999; Theodore and Miller, 2010; Theodore *et al.*, 2015). An interesting avenue for future research is to examine whether the detrimental effect of *f*0 variability on SCEs is maintained when listeners have extensive exposure to the talkers. An affirmative result would strengthen the suggestion that SCEs may be linked to talker normalization more broadly.

ACKNOWLEDGMENTS

The authors wish to thank three anonymous reviewers for their helpful comments and suggestions, and Ashley Batliner, Alexandrea Beason, Asim Mohiuddin, Madison Rhyne, Niko Sikkel, and Carly Sinclair for their assistance with data collection.

[1]In Lotto and Kluender (1998), SCEs produced by male talker and female talker contexts were not statistically compared to one another, so it is unclear whether this potential difference in context effect magnitudes is quantitative or merely qualitative.

[2]Data frames and all statistical analyses are available on the Open Science Framework (https://osf.io/s5y3m/).

Assgari, A. A. (**2018**). "Assessing the relationship between talker normalization and spectral contrast effects in speech perception," Doctoral dissertation, University of Louisville, Louisville, Kentucky.

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.

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.

Bates, D. M., Maechler, M., Bolker, B., and Walker, S. (**2014**). lme4: Linear mixed-effects models using Eigen and S4. R package version 1:1-17, https://cran.r-project.org/web/packages/lme4/index.html (Last viewed 3/6/2019).

Boersma, P., and Weenink, D. (**2017**). "Praat: Doing phonetics by computer" [Computer program]. Version 5.3.61, http://www.praat.org/ (Last viewed January 1, 2014).

Bosker, H. R., Reinisch, E., and Sjerps, M. J. (**2017**). "Cognitive load makes speech sound fast, but does not modulate acoustic context effects," J. Mem. Lang. **94**, 166–176.

Bradlow, A. R., and Pisoni, D. B. (**1999**). "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," J. Acoust. Soc. Am. **106**, 2074–2085.

Choi, J. Y., Hu, E. R., and Perrachione, T. K. (**2018**). "Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing," Attn., Percept., Psychophys **80**, 784–797.
Clarke, C. M., and Garrett, M. F. (**2004**). "Rapid adaptation to foreign-accented English," J. Acoust. Soc. Am. **116**(6), 3647–3658.
Creelman, C. D. (**1957**). "Case of the unknown talker," J. Acoust. Soc. Am. **29**, 655.
Fant, G. (**1973**). *Speech Sounds and Features* (MIT Press, Cambridge, MA).
Fourcin, A. (**1968**). "Speech source inference," IEEE Trans. Audio Electroacoust. **16**(1), 65–67.
Frazier, J. F., Assgari, A. A., and Stilp, C. E. (**2019**). "Musical instrument categorization is highly sensitive to spectral properties of earlier sounds," Attn., Percept., Psychophys. (in press).
Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (**1990**). "DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," National Institute of Standards and Technology, NIST Order No. PB91-505065.
Geiselman, R. E., and Bellezza, F. S. (**1976**). "Long-term memory for speaker's voice and source location," Memory Cognit. **4**(5), 483–489.
Goldinger, S. D. (**1996**). "Words and voices: Episodic traces in spoken word identification and recognition memory," J. Exp. Psychol. **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. **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," Attn., Percept., Psychophys. **71**(5), 1150–1166.
Holt, L. L. (**2005**). "Temporally nonadjacent nonlinguistic sounds affect speech categorization," Psychol. 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.
Huang, J., and Holt, L. L. (**2012**). "Listening for the norm: Adaptive coding in speech categorization," Front. Psychol. **3**, 10.
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. Morrison and P. Assmann (Springer, Berlin), pp. 117–151.
Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2013**). "lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)," R package version.
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. Psychol. **3**, 203.
Long, J. A. (**2018**). "jtools: Analysis and presentation of social scientific data," R package version 1.1.0, https://cran.r-project.org/web/packages/jtools/index.html (Last viewed 3/6/2019).
Lotto, A. J., and Kluender, K. R. (**1998**). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," Attn., Percept., Psychophys. **60**(4), 602–619.
Magnuson, J. S., and Nusbaum, H. C. (**2007**). "Acoustic differences, listener expectations, and the perceptual accommodation of talker variability," J. Exp. Psychol. **33**(2), 391–409.
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.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**(1), 42–46.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
R Development Core Team (**2018**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, http://www.r-project.org/ (Last viewed 3/6/2019).
Rand, T. C. (**1971**). "Vocal tract size normalization in the perception of stop consonants," J. Acoust. Soc. Am. **50**(1A), 139.
Ryalls, B. O., and Pisoni, D. B. (**1997**). "The effect of talker variability on word recognition in preschool children," Develop. Psychol. **33**(3), 441–452.
Sjerps, M. J., Mitterer, H., and McQueen, J. M. (**2011**). "Constraints on the processes responsible for the extrinsic normalization of vowels," Attn., Percept., Psychophys. **73**(4), 1195–1215.
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. Psychol. **44**(6), 914–924
Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., and Cook, S. (**2012**). "Development and validation of the AzBio sentence lists," Ear Hear. **33**(1), 112–117.
Stilp, C. E. (**2017**). "Acoustic context alters vowel categorization in perception of noise-vocoded speech," J. Assoc. Res. Otolaryngol. **18**(3), 465–481.
Stilp, C. E., and Alexander, J. M. (**2016**). "Spectral contrast effects in vowel categorization by listeners with sensorineural hearing loss," Proc. Mtgs. Acoust. **26**, 060003.
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," Attn., Percept., Psychophys. **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 Assgari, A. A. (**2018**). "Perceptual sensitivity to spectral properties in earlier sounds during speech categorization," Attn., Percept., Psychophys. **80**(5), 1300–1310.
Stilp, C. E., and Assgari, A. A. (**2019**). "Natural signal statistics shift speech sound categorization," Attn., Percept., Psychophys. (in press).
Theodore, R. M., and Miller, J. L. (**2010**). "Characteristics of listener sensitivity to talker-specific phonetic detail," J. Acoust. Soc. Am. **128**(4), 2090–2099.
Theodore, R. M., Myers, E. B., and Lomibao, J. A. (**2015**). "Talker-specific influences on phonetic category structure," J. Acoust. Soc. Am. **138**, 1068–1078.
Watkins, A. J. (**1991**). "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **90**(6), 2942–2955.
Watkins, A. J., and Makin, S. J. (**1994**). "Perceptual compensation for speaker differences and for spectral-envelope distortion," J. Acoust. Soc. Am. **96**(3), 1263–1282.
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.