



# Musical instrument categorization is highly sensitive to spectral properties of earlier sounds

Jonathan M. Frazier<sup>1</sup> · Ashley A. Assgari<sup>1</sup> · Christian E. Stilp<sup>1</sup>

Published online: 6 February 2019  
© The Psychonomic Society, Inc. 2019

## Abstract

Auditory perception is shaped by spectral properties of surrounding sounds. For example, when spectral properties differ between earlier (context) and later (target) sounds, this can produce spectral contrast effects (SCEs; i.e., categorization boundary shifts) that bias perception of later sounds. SCEs affect perception of speech and nonspeech sounds alike (Stilp Alexander, Kieffe, & Kluender in *Attention, Perception, & Psychophysics*, 72(2), 470–480, 2010). When categorizing speech sounds, SCE magnitudes increased linearly with greater spectral differences between contexts and target sounds (Stilp, Anderson, & Winn in *Journal of the Acoustical Society of America*, 137(6), 3466–3476, 2015; Stilp & Alexander in *Proceedings of Meetings on Acoustics*, 26, 2016; Stilp & Assgari in *Journal of the Acoustical Society of America*, 141(2), EL153–EL158, 2017). The present experiment tested whether this acute context sensitivity generalized to nonspeech categorization. Listeners categorized musical instrument target sounds that varied from French horn to tenor saxophone. Before each target, listeners heard a 1-second string quintet sample processed by filters that reflected part of (25%, 50%, 75%) or the full (100%) difference between horn and saxophone spectra. Larger filter gains increased spectral distinctness across context and target sounds, and resulting SCE magnitudes increased linearly, parallel to speech categorization. Thus, a highly sensitive relationship between context spectra and target categorization appears to be fundamental to auditory perception.

**Keywords** Audition · Hearing · Music cognition · Sound recognition

## Introduction

Human perception relies heavily on context for the interpretation of stimuli. Context provides a background against which

---

**Significance statement** Recently heard sounds affect recognition of later sounds. This is especially true when earlier sounds' frequencies differ from those in later sounds, making their perception more distinct from earlier sounds than they actually are. The degree to which these "spectral contrast effects" bias speech sound categorization is closely related to the magnitudes of frequency differences across sounds (larger differences produce larger contrast effects). Here, this relationship is replicated and extended to categorization of musical instrument sounds. Thus, high sensitivity to acoustic differences across sounds is not limited to speech perception, but appears to be fundamental to auditory perception most broadly.

---

✉ Christian E. Stilp  
christian.stilp@louisville.edu

<sup>1</sup> University of Louisville, 317 Life Sciences Building,  
Louisville, KY 40292, USA

a target stimulus is judged, shaping its perception. When the context and target stimulus differ, that difference may be perceptually magnified, resulting in a contrast effect. Consider the perceived brightness of a cell phone screen. For a given level of physical luminance, the screen will be perceived as less bright in well-lit conditions (such as sitting on a beach in the daytime) than in dimly lit conditions (on the same beach long after the sun has gone down). This is an example of brightness contrast in vision, but contrast effects can occur in every modality (von Békésy, 1967; Warren, 1985; Kluender, Coady, & Kieffe, 2003).

Contrast effects play an important role in speech perception. In a seminal paper, Ladefoged and Broadbent (1957) demonstrated that the spectrum of a preceding sentence context ("Please say what vowel this is") can influence vowel perception. When first formant frequencies (F1) of the context sentence were shifted up, listeners perceived the subsequent target vowel as / / (low F1) more often. When the first formant of the context sentence was shifted down, listeners perceived the target as /ε/ (high F1) more often. In this example, spectral

characteristics differed across context and target sounds, resulting in spectral contrast effects (SCEs). These effects have been widely reported in speech perception (e.g., Ladefoged & Broadbent, 1957; Watkins, 1991; Lotto & Kluender, 1998; Holt, 2005; Sjerps, Mitterer, & McQueen, 2011; Kingston et al., 2014; Sjerps & Reinisch, 2015; Assgari & Stilp, 2015; Stilp, Anderson, & Winn, 2015; Feng & Oxenham, 2018; Sjerps, Zhang, & Peng, 2018; Stilp & Assgari, *in press*; but see Mann, 1980; Mann & Repp, 1980; Fowler et al., 2000; Mitterer, 2006; Viswanathan, Fowler, & Magnuson, 2009; Viswanathan, Magnuson, & Fowler, 2010, 2013, for alternative accounts with mechanisms tied directly to speech production). SCEs are not exclusive to the use of speech stimuli. Contrast effects also bias speech categorization following nonspeech contexts, such as signal-correlated noise (Watkins, 1991) and pure tones (Lotto & Kluender, 1998; Holt, 2005; Kingston et al., 2014).

Stilp, Alexander, Kiefte, and Kluender (2010) predicted that SCEs were just as important in nonspeech sound categorization as they were in speech categorization. They tested this prediction by measuring context effects in the perception of music. Music was chosen as the nonspeech stimulus because, while it is spectrotemporally complex, listeners are generally far less familiar with music than with speech. In Stilp et al. (2010), listeners categorized musical instrument sounds that varied from French horn to tenor saxophone. In one experiment, the preceding acoustic context was speech (“You will hear”). In a separate experiment, the preceding acoustic context was a brief excerpt of a string quintet. These contexts were processed by filters that emphasized the difference between horn and saxophone spectra (spectral envelope difference filters; Watkins, 1991; see *Method*). In both experiments, when the context was filtered to emphasize frequencies in the horn spectrum, listeners categorized target sounds as a saxophone more often. When the context was filtered to emphasize frequencies in the saxophone spectrum, listeners categorized target sounds as the horn more often. These results were observed irrespective of whether the context was speech or music. Thus, despite having far less experience perceiving musical instruments as compared to speech, listeners’ responses were similarly shaped by SCEs in both conditions. This supported the generality of these context effects for auditory perception at large.

While SCEs in speech perception have a long history, the underlying mechanism responsible for these effects is fiercely debated. Ladefoged and Broadbent (1957) initially interpreted their findings as a means for adjusting for talker differences. They suggested that listeners learn properties of a talker’s voice and use that information to inform perception of speech from that talker. Subsequent research interpreted these types of context effects as being rooted in speech production, reflecting the listeners’ effort to compensate for coarticulation (Mann, 1980; Mann & Repp, 1980). Lotto and Kluender

produced similar shifts in speech categorization for nonhuman animals perceiving speech (Lotto, Kluender, & Holt, 1997) and humans perceiving nonspeech (Lotto & Kluender, 1998; see also Watkins, 1991; Holt, 2005; Kingston et al., 2014), leading to the proposal that general auditory mechanisms produced these context effects and not speech production per se. This launched a decades-long debate that is still ongoing (for reviews, see Fowler et al., 2000; Diehl, Lotto, & Holt, 2004; Fowler, 2006; Lotto & Holt, 2006; Kingston et al., 2014). In this debate, it is important to acknowledge that multiple time-scales of context effects are being studied. The context preceding the categorized phoneme target can be short term (a single sound or syllable, as in Mann, 1980; Lotto & Kluender, 1998; and others) or longer term (several sounds or a sentence, as in Ladefoged & Broadbent, 1957; Holt, 2005; and others). The spectral contrast account offers the same mechanism(s) and consistent predictions across short-term and long-term context effects. While compensation for coarticulation has been proposed as the mechanism responsible for these short-term effects (e.g., Mann, 1980; Mann & Repp, 1980; Fowler et al., 2000; Viswanathan et al., 2009; Viswanathan et al., 2010; Viswanathan et al., 2013), this mechanism cannot speak directly to longer-term context effects (Viswanathan & Kely-Stephen, 2018).

The magnitudes of SCEs in auditory perception are also poorly understood. Irrespective of whether one might intuit that these effects have variable magnitudes, past studies overwhelmingly ignored this possibility and instead treated them as being merely present or absent. This was due in part to researchers processing context sounds using high-gain filters. High-gain filters introduced large spectral differences between context and target sounds, which maximized the probability of observing an SCE (if one should theoretically be present). While this approach was informative for *when* perception might be influenced by SCEs, it failed to address the question of *to what extent* perception was influenced by context.

Stilp and colleagues (Stilp et al., 2015) addressed this question by testing the effects of a variety of filters and filter gains on vowel categorization. Filter gains ranged from small (e.g., adding a +5 dB spectral peak to the context sentence spectrum, amplifying/attenuating context frequencies by only 25% of the difference between target spectral envelopes) to large (e.g., +20 dB peak, amplifying/attenuating frequencies by 100% of the difference between spectral envelopes). Not only were SCEs observed in nearly every condition tested, but their magnitudes varied continuously: As larger filter gains were tested, spectral differences between the context sentence and target vowel progressively increased, and SCE magnitudes increased in kind (see also Stilp & Alexander, 2016). This relationship was later replicated and extended in consonant categorization (Stilp & Assgari, 2017). Each of these studies reported strong linear relationships between filter gain (which introduced spectral differences between context and target

sounds) and biases in speech categorization (i.e., SCEs magnitudes), supporting acute sensitivity to context when perceiving and categorizing speech sounds.

The magnitudes of SCEs vary continuously in speech categorization (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017), but it is unknown whether the same is true for SCEs influencing musical instrument categorization (as in Stilp et al., 2010). One possibility is that continuous variation in SCE magnitudes only occurs for highly familiar stimuli (e.g., speech). If this is the case, SCEs for less familiar stimuli (e.g., music) would not vary gradually in their magnitudes but instead skew toward being either present or absent. On the other hand, the magnitudes of these context effects might vary continuously irrespective of this difference in listening experience. While the *occurrence* of SCEs in auditory perception generalizes across speech and nonspeech (as reviewed above), it is unclear whether the *degrees* to which SCEs shape auditory perception is equally generalizable.

To test these possibilities, in the present experiment, the variable-filter-gain approach of Stilp et al. (2015) was applied to the nonspeech stimuli used in Stilp et al. (2010). Filters introduced varying degrees of spectral differences between the context (string quintet) and the target (brass instrument) stimuli. Conditions testing high-gain filters are expected to replicate the SCEs reported in Stilp et al. (2010). The key question is whether smaller amounts of filter gain (i.e., filters that reflect less of the difference between horn and saxophone spectra) bias musical instrument categorization to progressively smaller degrees, as observed in speech categorization (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017).

## Method

### Participants

Seventeen undergraduate participants in this experiment received course credit in exchange for their participation. All self-reported no known hearing impairments.

### Stimuli

**Targets** Target stimuli were the same stimuli as used in Stilp et al. (2010). Two musical instruments, French horn and tenor saxophone, were selected from the McGill University Musical Samples database (Opolko & Wapnick, 1989). Recordings of each instrument playing the note G3 (196 Hz) were sampled at 44.1 kHz. Three consecutive pitch pulses (15.31 ms) of constant amplitude were excised at zero crossings from the center of each recording and iterated to 140-ms total duration in Praat (Boersma & Weenink, 2017). Stimuli were processed by 5-ms linear ramps at both onset and offset. Stimuli were then proportionately mixed in six steps to form a series in which the

amplitude of one instrument was +30, +18, +6, −6, −18, or −30 dB relative to the other. Stimuli with 30-dB differences between instruments served as series endpoints. Waveforms were then low-pass filtered at 10 kHz cutoff using a 10th-order, elliptical infinite impulse response filter. Instrument mixing and filtering were performed in MATLAB.

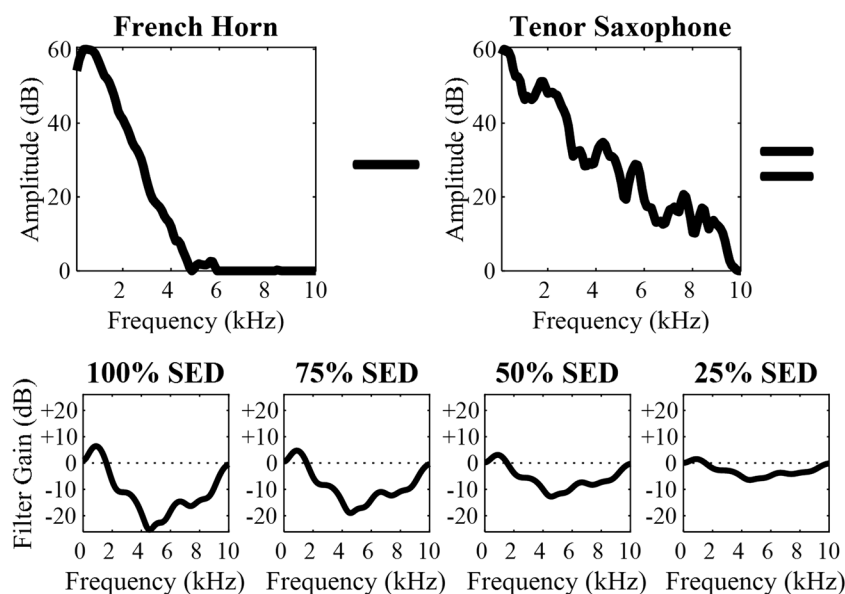
**Filters** Similar to Stilp et al. (2010), endpoint French horn and tenor saxophone stimuli were analyzed to create spectral envelope difference (SED) filters (Watkins, 1991). Spectral envelopes for each instrument were derived from 512-point Fourier transforms, and were smoothed using a 256-point Hamming window with 128-point overlap (see Fig. 1). Spectral envelopes were equated for peak power, then subtracted from one another in both directions (horn minus saxophone, saxophone minus horn). A finite impulse response was obtained for each SED using inverse Fourier transform. This impulse response reflected 100% of the difference between spectral envelopes (as in Stilp et al., 2010). Following the methods of Stilp et al. (2015) and Stilp (2017; see also Watkins & Makin, 1996), the linear amplitude values of the impulse responses were scaled down to 75%, 50%, or 25% of the original SED (see Fig. 1). This produced eight filters in all: two SEDs (horn minus saxophone, saxophone minus horn) fully crossed with four levels of filter gain (100%, 75%, 50%, 25%). If SCE magnitudes scale linearly, as observed in speech categorization (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017), then relatively large SCEs should be observed following contexts processed by 100% SED filters and progressively smaller SCEs should be observed following contexts processed by smaller amounts of filter gain.

**Context** The context stimulus was a 1 second excerpt of Franz Schubert's *String Quintet in C Major, Allegretto*, taken from compact disc. This was the same stimulus as used in Stilp et al. (2010). The context was processed by each of the eight SED filters detailed above.

All contexts and targets were matched in root-mean-squared (RMS) amplitude. Each of the six target instrument stimuli was concatenated to each of eight contexts, making 48 unique pairings in all. The two target instrument endpoints, absent any preceding context, were also RMS matched for use in a familiarization task. Finally, all stimuli were resampled at 44.1 kHz for presentation.

### Procedure

After obtaining informed consent, each participant was led into a sound-attenuating booth (Acoustic Systems, Inc., Austin, TX). The participant sat at a small table on top of which was a computer screen, mouse, and keyboard. All sounds were D/A converted by RME HDSPe AIO sound cards (Audio AG, Haimhausen, Germany) on a personal



**Fig. 1** Construction of spectral envelope difference (SED) filters. Top row depicts spectral envelopes for endpoints of the musical instrument target series. Bottom row depicts SED filter responses that reflect 100% (left) or progressively less of the difference between horn and saxophone

spectral envelopes. Here, filter responses reflect SEDs for horn minus saxophone; subtraction in the opposite direction (saxophone minus horn) produced complementary filter responses to those shown here

computer and passed through a programmable attenuator (TDT PA4, Tucker-Davis Technologies, Alachua, FL) and headphone buffer (TDT HB6) before being presented diotically at 70 dB sound pressure level (SPL) over circumaural headphones (Beyerdynamic DT-150, Beyerdynamic Inc. USA, Farmingdale, NY).

A custom MATLAB script guided participants through the experiment, which consisted of four phases. The first phase was Exposure, where participants heard each musical instrument endpoint played twice along with its verbal label.

The second phase was Practice, where participants categorized endpoints from the French-horn–tenor-saxophone series. Participants used the mouse to click a response button, indicating whether they heard a French horn or a tenor saxophone. Practice consisted of 120 practice trials. A performance criterion was implemented where participants were required to achieve at least 90% correct before proceeding to the next phase of the experiment. All participants met this criterion.

The third phase was the Main Experiment. On each trial, participants heard a filtered context stimulus followed by a musical instrument target. Participants clicked the mouse to indicate whether the target sounded more like a French horn or a tenor saxophone. The experiment consisted of four blocks, with each block composed of 120 trials (2 SED filters: horn minus saxophone, saxophone minus horn  $\times$  6 target instruments  $\times$  10 repetitions) at a single level of filter gain (100%, 75%, 50%, or 25%). Stimuli were randomized within each block, and blocks were tested in counterbalanced orders across participants.

The fourth and final phase was the Survey. This survey, the same as that used in Stilp et al. (2010), consisted of five questions that broadly assessed each participant's musical experience. The first question asked participants to rate their musical performing ability from 1 (*no experience*) to 5 (*virtuoso*) on a Likert-type scale. The second and third questions asked participants to report the number of years of solo or ensemble musical performance experience (with formal training/instruction) they had, respectively. The fourth question asked participants to report any other relevant musical experience they had to share. The fifth question asked participants whether they recognized or could name the musical selection used as the context stimulus. Participants clicked the mouse and typed on the keyboard to enter their responses to survey questions. In all, the entire session took approximately 50 minutes to complete.

## Results

All participants met the inclusion criterion of 90% correct in the practice block. However, two participants failed to maintain 90% accuracy on target endpoint stimuli throughout the experiment. Their results were removed from subsequent analyses. Responses from the remaining 15 participants were analyzed using a generalized linear mixed-effect logistic model in R (R Development Core Team, 2016) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). Initial model architecture matched that used by Stilp et al. (2015) and Stilp and Assgari (2017). The dependent variable was modeled as binary (“horn” or “saxophone” responses coded

**Table 1** Mixed-effects logistic model results

	Estimate	SE	Z	p
Intercept	1.10	0.14	7.86	<.0001
Target	2.21	0.28	7.87	<.0001
FilterFreq	-0.77	0.12	-6.18	<.0001
FilterGain	0.01	0.002	2.11	.035
FilterFreq × FilterGain	-0.01	0.003	-3.96	<.0001

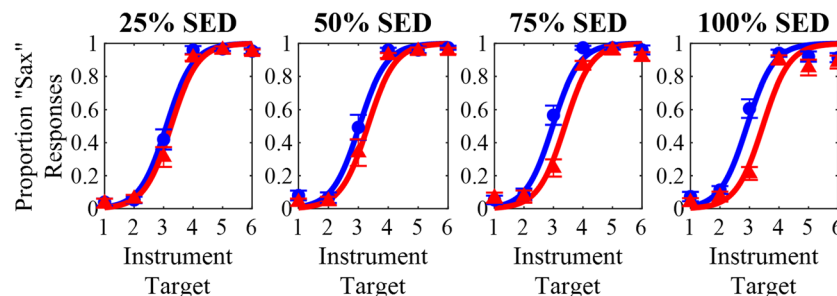
*Note.* “Target” refers to the slope of the logistic function, defined as the change in log odds of the listener responding “saxophone” resulting from one rightward step along the target instrument series (toward the saxophone endpoint). “FilterFreq” lists the change in log odds of the listener responding “saxophone” resulting from changing the context filtering from the horn-minus-saxophone SED filter to the saxophone-minus-horn SED filter. “FilterGain” lists the change in log odds of a “saxophone” response resulting from increasing filter gain by 1% of the difference between target instrument spectra. “FilterFreq × FilterGain” indicates the change in the size of the FilterFreq effect (i.e., SCE) per percentage of filter gain. *SE* = standard error of the mean

as 0 and 1, respectively). Fixed effects in the model included Target (coded as a continuous variable from 1 to 6, then mean centered), Filter Frequency (categorical variable with two levels: saxophone minus horn and horn minus saxophone, with horn minus saxophone set as the default level), Filter Gain (in percentage of SED; coded as a continuous variable from 25 to 100 in steps of 25, then mean centered), and the interaction between Filter Frequency and Filter Gain. Random slopes were included for each fixed main effect and interaction, and a random intercept of participant was also included. Preliminary analyses indicated that inclusion of the random slope for the Filter Frequency × Filter Gain interaction did not explain any additional variance,  $\chi^2(5) = 1.56, p = .91$ , so this term was omitted from the final model, which had the following form:

$$\begin{aligned} &\text{Response} \sim \text{Target} + \text{FilterFrequency} \\ &\quad + \text{FilterGain} + \text{FilterFrequency} \times \text{FilterGain} \\ &\quad + (1 + \text{Target} + \text{FilterFrequency} + \text{FilterGain} | \text{Participant}) \end{aligned}$$

Results from the mixed-effects model analysis are presented in Table 1, and behavioral results are depicted in Fig. 2. The fixed-effect Target was statistically significant, indicating that the log odds of responding “saxophone” increased with each rightward step along the musical instrument series, as expected (see Fig. 2). The fixed-effect Filter Frequency was significant, indicating that the log odds of responding “saxophone” decreased as the filtering condition was changed from horn minus saxophone to saxophone minus horn, consistent with the predicted direction of SCEs (i.e., more “horn” responses following the saxophone-minus-horn-filtered context, with which it is spectrally contrastive). The fixed effect of Filter Gain was also significant, indicating an increase in the log odds of responding “saxophone” as larger amounts of filter gain were used in processing the context musical passages. Critically, the interaction between Filter Frequency and Filter Gain was statistically significant. This indicates that SCEs increased linearly as filter gain increased, reflecting a systematic increase in SCE magnitudes as the spectral difference between context and target increased.

Following Stilp et al. (2015) and Stilp and Assgari (2017), post hoc analyses were performed by coding the interaction term Filter Frequency × Filter Gain as a categorical factor. This manipulation removes the model’s assumption that SCEs scaled linearly with different amounts of filter gain (as verified in the significant Filter Frequency × Filter Gain interaction in Table 1) and tests each SCE independently. This analysis selected one level of filter gain as the default level, then tested its model coefficient against zero using a Wald *z* test. All other model parameters matched those in the previous analysis. This process was repeated for all four levels of filter gain (25%, 50%, 75%, 100%). In each analysis, SCE magnitude was operationalized as the distance between logistic function 50% points measured in stimulus steps along the target continuum (visible as the horizontal spacing between functions in each panel of Fig. 2). For responses following the horn-minus-saxophone-filtered context, the 50% point was



**Fig. 2** Behavioral responses. Response proportions are on the ordinate and target instrument series is along the abscissa (1 = French horn endpoint, 6 = tenor saxophone endpoint). Symbols depict the mean proportions of “saxophone” responses across the participant sample to a given member of the target instrument series; error bars represent one standard error of the mean. Fits to these responses were generated by

the mixed-effects logistic model detailed in Table 1. Blue circles and lines depict responses following contexts processed by the horn-minus-saxophone difference filter; red triangles and lines depict responses following contexts processed by the saxophone-minus-horn difference filter. (Color figure online)

calculated as  $-\text{Intercept}/\text{Target}$ . For responses following the saxophone-minus-horn-filtered context, the 50% point was calculated as  $-(\text{Intercept} + \text{Filter Frequency})/\text{Target}$ . At every level of filter gain, SCEs were significantly greater than zero (all  $Z$ s  $> 2.03$ , all  $p$ s  $< .05$ ). Critically, SCE magnitude was again linearly related to filter gain ( $r = .96$ ,  $p < .05$ ; see Fig. 3). This replicates the linear relationships reported for vowels (Stilp et al., 2015; Stilp & Alexander, 2016) and consonants (Stilp & Assgari, 2017).

Finally, survey responses were used to predict two different outcome variables. First, responses were used to predict SCE magnitudes in the 100% SED condition. If musical experience influences SCE magnitude, then SCEs would be expected to be larger for participants with more musical experience and smaller for listeners with less musical experience. This analysis was conducted by Stilp et al. (2010), who reported no significant relationships between items. The analysis was repeated here to inform whether substantive differences existed across participant groups (if such a relationship is observed in the present report) or not (null results in both studies). Second, linear regressions were calculated on each participants' SCEs as a function of SED percentage, and survey responses were used to predict the slopes of these regressions. If musical experience promotes sensitivity to spectral differences across sounds, then regression slopes should be steeper for more experienced musicians (i.e., more differentiation across SCE magnitudes) and shallower for those with less musical experience.

The first questionnaire item, self-rated musical performance ability (median rating = 2, range: 1–4), had no predictive power for behavioral performance. Given the ordinal nature of these ratings, Spearman's correlations were calculated. This questionnaire item was not correlated with 100% SED SCE magnitudes ( $\rho = .02$ ,  $p = .93$ ) or linear regression slopes ( $\rho = .12$ ,  $p = .66$ ). Subsequent analyses used Pearson correlation coefficients. The second questionnaire item, years of solo musical performing experience (mean = 2.00 years,  $SD = 2.83$ ), was not correlated with 100% SED SCE

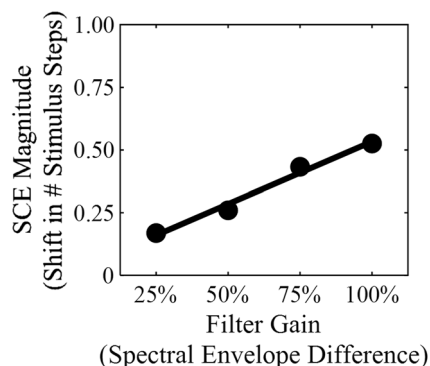
magnitude ( $r = -.22$ ,  $p = .42$ ) or linear regression slopes ( $r = -.06$ ,  $p = .82$ ). The third questionnaire item, years of ensemble musical performing experience (mean = 1.73 years,  $SD = 2.43$ ), was not correlated with 100% SED SCE magnitude ( $r = -.20$ ,  $p = .48$ ) or linear regression slopes ( $r = -.10$ ,  $p = .73$ ). No listeners had any other relevant musical experience to share, and none recognized or accurately identified the string quintet context.

## Discussion

Spectral contrast effects have a long history in speech perception research (see the introduction). Yet, much of this research focused on whether SCEs did or did not influence speech categorization, giving little to no consideration to the magnitudes of these effects. Recent work showed that SCE magnitudes scaled linearly in predictable ways: As spectral differences across context and target sounds increased, biases in target sound categorization increased in kind (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017). While SCEs have also been shown to bias categorization of musical instrument sounds (Stilp et al., 2010), whether this acute sensitivity to preceding context generalized to non-speech categorization was unclear. The current study demonstrated that SCE magnitudes also scaled linearly as a function of spectral differences in nonspeech categorization, just as they do in speech categorization. This suggests that in general, perception of complex sounds is acutely sensitive to spectral characteristics of preceding sounds.

Musical performance experience was not a reliable predictor of performance in the present study. Whether reported subjectively (rated on a Likert-type scale) or more objectively (years of performing experience), no survey items were correlated with SCE magnitudes in the 100% SED condition, nor with the rate of SCE growth as SED percentage increased (linear regression slopes fit to each listeners' SCEs). Null results were also reported in Stilp et al. (2010), who tried to predict SCE magnitudes using the same questionnaire. It bears mentioning that musicians were not explicitly recruited to participate in the present study. But, if these null results are representative, it would suggest that musical training does not modulate the proposed low-level mechanisms that are responsible for SCEs (adaptation or adaptation-like mechanisms in the early auditory system; Delgutte, 1996; Delgutte, Hammond, Kalluri, Litvak, & Cariani, 1996; Holt, Lotto, & Kluender, 2000; Holt & Lotto, 2002; Stilp & Assgari, 2018). Future research with a highly musically trained participant sample would provide a stronger test of this suggestion.

The present results are highly relevant to listeners who use hearing aids or cochlear implants to hear. Signal processing algorithms in these devices prioritize processing of the current sound, with much less emphasis on incorporating effects of preceding acoustic context. However, recent



**Fig. 3** Spectral contrast effects (SCEs) calculated independently at each level of filter gain. Solid line depicts the linear regression fit to these SCEs. A strong linear relationship exists between these variables ( $r = .96$ ,  $p < .05$ ), replicating similar results reported in studies of speech categorization

reports indicate that hearing-impaired listeners' speech perception is also influenced by SCEs. Stilp and Alexander (2016) reported that listeners with sensorineural hearing loss exhibited SCEs whose magnitudes were larger than those for normal-hearing listeners. Similarly, in Stilp (2017), SCE magnitudes were significantly larger when normal-hearing listeners categorized noise-vocoded speech compared to the spectrally intact stimuli tested in Stilp et al. (2015). This was directly confirmed by Feng and Oxenham (2018), who reported larger SCEs for cochlear implant users than for normal-hearing listeners. Finally, SCE magnitudes for hearing-impaired listeners increased in a linear fashion at higher filter gains, similar to normal hearing listeners (but with a steeper slope owing to larger SCE magnitudes; Stilp & Alexander, 2016). Thus, healthy hearing is not a prerequisite for having speech perception be influenced by SCEs. However, the magnitudes of these effects appears to differ based on hearing health. Larger-than-normal SCEs encourage miscategorization of otherwise unambiguous speech sounds, resulting in poorer speech recognition than that achieved by appropriately sized SCEs (see Stilp, 2017 for discussion). Future research should consider how finely SCE magnitudes might vary with hearing health, whether this context sensitivity generalizes to nonspeech perception by hearing-impaired listeners as well, and how signal processing in assistive listening devices might exert similar-sized context effects to those experienced by normal-hearing listeners.

Results extend the long history of replicating effects observed in speech perception using nonspeech sounds. Following several early demonstrations (Stevens & Klatt, 1974; Miller, Wier, Pastore, Kelly, & Dooling, 1976; Pisoni, 1977), strong converging evidence was offered by Diehl and colleagues, including trading relations and context effects influencing medial voicing distinctions (Parker, Diehl, & Kluender, 1986; Kluender, Diehl, & Wright, 1988), temporal cues to consonant manner (Diehl & Walsh, 1989), and related subsequent studies demonstrating the efficacy of nonspeech contexts biasing speech categorization (Lotto & Kluender, 1998; Holt, 2005). Here, not only were contrast effects observed in perception of nonspeech target sounds (as in Stilp et al., 2010), but *patterns* of contrast effects were replicated across speech and nonspeech sounds, as both scaled linearly as a function of filter gain. In other words, the larger the spectral difference between earlier and later sounds (irrespective of whether they were speech or nonspeech), the greater the extent to which categorization was biased. This reifies contrast as a fundamental mechanism that contributes substantially to perception of speech and other complex sounds (Diehl, Elman, & McCusker, 1978; Kluender, Coady, & Kieffe, 2003; Kluender & Alexander, 2007; Stilp et al., 2015; Stilp & Assgari, *in press*).

In conclusion, larger spectral differences across context and target sounds produced systematically larger biases in musical

instrument categorization. This pattern of results replicated similar reports in vowel categorization (Stilp et al., 2015; Stilp & Alexander, 2016) and consonant categorization (Stilp & Assgari, 2017). Thus, acute sensitivity to spectral differences across sounds and the contrast effects that are produced appear to be fundamental to auditory perception.

**Acknowledgments** The authors thank Samantha Cardenas, Rebecca Davis, Joshua Lanning, and Caroline Smith for their assistance with data collection. 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.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

- Assgari, A. A., & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. *Journal of the Acoustical Society of America*, *138*(5), 3023–3032.
- Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using eigen and S4 (R Package Version 1.1-7) [Computer software]. Retrieved from <http://cran.r-project.org/package=lme4>
- Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer program]. Retrieved from <http://www.fon.hum.uva.nl/praat/>
- Delgutte, B. (1996). Auditory neural processing of speech. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 507–538). Oxford, UK: Blackwell.
- Delgutte, B., Hammond, B. M., Kalluri, S., Litvak, L. M., & Cariani, P. A. (1996). Neural encoding of temporal envelope and temporal interactions in speech. In W. Ainsworth & S. Greenberg (Eds.), *Auditory basis of speech perception* (pp. 1–9). Keele, UK: European Speech Communication Association.
- Diehl, R. L., Elman, J. L., & McCusker, S. B. (1978). Contrast effects on stop consonant identification. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 599–609.
- Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Reviews in Psychology*, *55*, 149–179.
- Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America*, *85*(5), 2154–2164.
- Feng, L., & Oxenham, A. J. (2018). Effects of spectral resolution on spectral contrast effects in cochlear-implant users. *The Journal of the Acoustical Society of America*, *143*(6), EL468–EL473.
- Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, *68*(2), 161–177.
- Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(3), 877–888.
- Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312.
- Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research*, *167*(1/2), 156–169.
- Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America*, *108*(2), 710–722.

- Kingston, J., Kawahara, S., Chambless, D., Key, M., Mash, D., & Watsky, S. (2014). Context effects as auditory contrast. *Attention, Perception, & Psychophysics*, *76*, 1437–1464.
- Kluender, K. R., & Alexander, J. M. (2007). Perception of speech sounds. In P. Dallos & D. Oertel (Eds.), *The senses: A comprehensive reference* (pp. 829–860). San Diego, CA: Academic.
- Kluender, K. R., Coady, J. A., & Kieft, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*(1), 59–69.
- Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics*, *16*(2), 153–169.
- Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*(1), 98–104.
- Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, *68*(2), 178–83.
- Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*(4), 602–619.
- Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America*, *102*(2), 1134–1140.
- Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.
- Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [j]–[s] distinction. *Perception & Psychophysics*, *28*(3), 213–228.
- Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise–buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, *60*(2), 410–417.
- Mitterer, H. (2006). Is vowel normalization independent of lexical processing? *Phonetica*, *63*(4), 209–229.
- Opolko, F., & Wapnick, J. (1989). *McGill University master samples user's manual*. Montreal, Canada: McGill University, Faculty of Music.
- Parker, E. M., Diehl, R. L., & Kluender, K. R. (1986). Trading relations in speech and nonspeech. *Perception & Psychophysics*, *39*(2), 129–142.
- Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, *61*(5), 1352–1361.
- R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.r-project.org/>
- Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Constraints on the processes responsible for the extrinsic normalization of vowels. *Perception & Psychophysics*, *73*(4), 1195–1215.
- Sjerps, M. J., & Reinisch, E. (2015). Divide and conquer: How perceptual contrast sensitivity and perceptual learning cooperate in reducing input variation in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 710–722.
- Sjerps, M. J., Zhang, C., & Peng, G. (2018). Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 914–924.
- Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced–voiceless distinction for stops. *The Journal of the Acoustical Society of America*, *55*(3), 653–659.
- Stilp, C. E. (2017). Acoustic context alters vowel categorization in perception of noise-vocoded speech. *Journal of the Association for Research in Otolaryngology*, *18*(3), 465–481.
- Stilp, C. E., & Alexander, J. M. (2016). Spectral contrast effects in vowel categorization by listeners with sensorineural hearing loss. *Proceedings of Meetings on Acoustics*, *26*. <https://doi.org/10.1121/2.0000233>
- Stilp, C. E., Alexander, J. M., Kieft, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics*, *72*(2), 470–480.
- Stilp, C. E., Anderson, P. W., & Winn, M. B. (2015). Predicting contrast effects following reliable spectral properties in speech perception. *Journal of the Acoustical Society of America*, *137*(6), 3466–3476.
- Stilp, C. E., & Assgari, A. A. (2017). Consonant categorization exhibits a graded influence of surrounding spectral context. *Journal of the Acoustical Society of America*, *141*(2), EL153–EL158.
- Stilp, C. E., & Assgari, A. A. (2018). Perceptual sensitivity to spectral properties of earlier sounds during speech categorization. *Attention, Perception, & Psychophysics*, *80*(5), 1300–1310.
- Stilp, C. E., & Assgari, A. A. (in press). Natural signal statistics shift speech sound categorization. *Attention, Perception, & Psychophysics*.
- Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review*, *16*(1), 74–79.
- Viswanathan, N., & Kelty-Stephen, D. G. (2018). Comparing speech and nonspeech context effects across timescales in coarticulatory contexts. *Attention, Perception, & Psychophysics*, *80*(2), 316–324.
- Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(4), 1005–1015.
- Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2013). Similar response patterns do not imply identical origins: An energetic masking account of nonspeech effects in compensation for coarticulation. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 1181–1192.
- von Békésy, G. (1967). *Sensory perception*. Princeton, NJ: Princeton University Press.
- Warren, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review*, *92*(4), 574–584.
- Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *90*(6), 2942–2955.
- Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *99*(6), 3749–3757.