



Perceptual sensitivity to spectral properties of earlier sounds during speech categorization

Christian E. Stilp¹ · Ashley A. Assgari¹

© The Psychonomic Society, Inc. 2018

Abstract

Speech perception is heavily influenced by surrounding sounds. When spectral properties differ between earlier (context) and later (target) sounds, this can produce spectral contrast effects (SCEs) that bias perception of later sounds. For example, when context sounds have more energy in low- F_1 frequency regions, listeners report more high- F_1 responses to a target vowel, and *vice versa*. SCEs have been reported using various approaches for a wide range of stimuli, but most often, large spectral peaks were added to the context to bias speech categorization. This obscures the lower limit of perceptual sensitivity to spectral properties of earlier sounds, i.e., when SCEs begin to bias speech categorization. Listeners categorized vowels (/ɪ/-ε/, Experiment 1) or consonants (/d/-g/, Experiment 2) following a context sentence with little spectral amplification (+1 to +4 dB) in frequency regions known to produce SCEs. In both experiments, +3 and +4 dB amplification in key frequency regions of the context produced SCEs, but lesser amplification was insufficient to bias performance. This establishes a lower limit of perceptual sensitivity where spectral differences across sounds can bias subsequent speech categorization. These results are consistent with proposed adaptation-based mechanisms that potentially underlie SCEs in auditory perception.

Significance statement

Recent sounds can change what speech sounds we hear later. This can occur when the average frequency composition of earlier sounds differs from that of later sounds, biasing how they are perceived. These “spectral contrast effects” are widely observed when sounds’ frequency compositions differ substantially. We reveal the lower limit of these effects, as +3 dB amplification of key frequency regions in earlier sounds was enough to bias categorization of the following vowel or consonant sound. Speech categorization being biased by very small spectral differences across sounds suggests that spectral contrast effects occur frequently in everyday speech perception.

Keywords Perceptual categorization and identification · Psychoacoustics · Speech perception

Introduction

All perception takes place in context. Objects and events in the sensory environment are perceived relative to recent objects and events as well as the perceiver’s experiences. Speech perception is no different, as recognition of speech sounds is influenced by surrounding sounds. If the average spectral properties of earlier (context) sounds differ from those in later (target) sounds, this

difference is perceptually magnified and categorization of the target sound is biased.¹ For example, if context sounds exhibit a lower average F_1 than is present in the target sound, the latter will be perceived as having a higher F_1 by comparison, and *vice versa*. This is known as a *spectral contrast effect* (SCE) and has been widely observed in speech perception (Ladefoged & Broadbent, 1957; Watkins, 1991; Holt, 2005, 2006; Sjerps et al., 2011, 2013; Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Assgari, 2017; Sjerps et al., 2017).

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

¹ Department of Psychological and Brain Sciences, University of Louisville, 317 Life Sciences Building, Louisville, KY 40292, USA

¹ We draw the distinction between long-term SCEs (produced by average spectral properties in context sounds that are 1+ seconds in duration) and short-term SCEs (produced by spectral properties at the offset of short-duration context sounds; e.g., Lotto & Kluender, 1998). While long-term and short-term SCEs produce consistent effects in similar directions, the present investigation and subsequent citations focus on long-term SCEs.

A host of behavioral results firmly situates SCEs as a low-level phenomenon. SCEs are not limited to speech but occur quite broadly in auditory perception, being reported for non-speech contexts biasing speech categorization (signal-correlated noise contexts: Watkins, 1991; Watkins & Makin, 1994; pure tone contexts: Holt, 2005, 2006; Laing et al., 2012) and non-speech contexts biasing nonspeech categorization (musical instruments; Stilp et al., 2010). Watkins and Makin suggested that these effects are independent of sound source and precede extraction of any features from the acoustic signal (Watkins & Makin, 1994, 1996). Later, Sjerps and colleagues argued that SCEs occur before any contributions from lexical status, attention, native language background, or even category structure itself (Sjerps et al., 2012, 2013; Sjerps & Smiljanic, 2013; Sjerps & Reinisch, 2015). Altogether, SCEs appear to reflect a basic, low-level process that serves to accentuate differences between stimuli (von Bekesy, 1967; Warren, 1985; Kluender et al., 2003).

Neural mechanisms responsible for these effects have not been definitively identified, but candidate mechanisms that have been considered complement the low-level nature of SCEs described above. Simple neural adaptation has been discussed as a plausible source of SCEs (Delgutte, 1996; Delgutte et al., 1996; Holt et al., 2000; Holt & Lotto, 2002). In simple neural adaptation, neurons adapt to their characteristic frequency components that are present in earlier (context) sounds. These adapted neurons would then be less responsive to these frequency components in later (target) sounds. Conversely, neurons that encode other frequencies, particularly frequencies that are not present in earlier sounds, would not be adapted. These neurons would then be relatively more responsive to later (target) sounds. On a population level, this shift in neural responsiveness could underlie increased perceptual sensitivity to stimulus change, resulting in a contrast effect. While this is one possible underlying mechanism, effects allegedly produced by neural adaptation might instead be produced by other mechanisms, such as adaptation of inhibition (also termed adaptation of suppression; see Summerfield et al., 1984, 1987). In adaptation of inhibition, neurons respond to their characteristic frequencies in a given sound while also suppressing responses of other neurons to neighboring frequencies. Over time, this suppressive influence adapts, making neural responses to these neighboring frequencies more pronounced than they were initially. It is noteworthy that adaptation and adaptation of inhibition are frequently cited as explanations for auditory enhancement effects, where spectral changes over time are neurally and perceptually enhanced (Viemeister & Bacon, 1982; Summerfield et al., 1984, 1987; Nelson & Young, 2010; Byrne et al., 2011; Carcagno et al., 2012), a process that has been suggested to be related to SCEs (Holt & Lotto, 2002; Kluender et al., 2003).

Historically, most investigations of SCEs in speech perception used high-gain filters to amplify relevant frequency regions in context sounds (see Stilp et al., 2015 for review). This

level of amplification produced large spectral differences between the context and target sounds. While this approach increased the probability of observing an SCE, it failed to address listeners' sensitivity to more modest spectral differences across context sounds and target sounds. Stilp, Anderson, and Winn (2015) addressed this issue by filtering context sentences using a wide range of filter gains (+5 to +20 dB amplification for narrowband spectral amplification, 25% to 100% of the total difference between broadband spectral envelopes). Across four listener groups and different filter types, as total filter power increased, the magnitudes of SCEs biasing vowel categorization increased linearly ($r = 0.74$). When one listener group was tested on a single filter type (300-Hz bandwidth) at four levels of filter gain (+5 to +20 dB in 5-dB steps), the linear relationship between filter gain and SCE magnitudes was even stronger ($r = 0.99$; Stilp & Alexander, 2016). A similarly strong linear relationship was observed for filter gains predicting SCEs that biased consonant categorization ($r = 0.99$; Stilp & Assgari, 2017).

Speech categorization exhibited a close relationship with spectral properties of earlier sounds, but the lower limit of this perceptual sensitivity remains an open question. In studies by Stilp and colleagues (Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017), the smallest spectral amplifications of context sentences were +5 dB, which was largely sufficient to produce SCEs. These results failed to identify the magnitude of change from context to target that was required to produce an SCE, or equivalently, when spectral properties of earlier sounds begin to bias speech categorization. It is likely the case that some small amount of amplification in the spectra of context sounds is insufficient to bias speech categorization. This possibility was anticipated by Holt et al. (2000), who made a prediction based on neural adaptation as a potential mechanism of SCEs: "less spectrally distinct or less intense precursors should tend to result in fewer adapted neurons. Thus, when a subsequent stimulus follows, less adaptation should result in less of a population shift in neural response. That is, there should be less neural contrast. If this change influences perception, such stimuli should exert a smaller effect of context on their neighbors" (pp 719-720). From this perspective, studies that used high-gain filters to modify context spectra invoked sufficient neural adaptation to produce SCEs. At a point, extremely modest amplification of context spectra would produce insufficient neural adaptation to bias categorization of the subsequent speech target, thus failing to produce SCEs.

Yet, it is unclear exactly when preceding context fails to influence subsequent speech categorization. Studies closely related to the present investigation revealed that +5 dB was the lower limit for spectral amplification that biased vowel categorization. Stilp and Anderson (2014) examined spectral calibration, where the context spectrum was amplified at the same frequency as a formant in the target vowel (F_2 , a key distinguishing feature for the target vowels /i/ and /u/).

Listeners decreased their reliance on F_2 and increased their reliance on other spectral cues (spectral tilt) to categorize the vowel, but only when amplification of the context spectrum was at least +5 dB. Spectral calibration effects are closely related to SCEs; the former is the deemphasis of spectral similarities and the latter is the emphasis of spectral differences (Alexander & Kluender, 2010). Given this association, because spectral amplification of less than +5 dB did not bias speech categorization for spectral calibration, it might not bias categorization through SCEs either.

While not a direct analog to context effects in speech perception, various psychophysical studies have reported sensitivity to even smaller increments in sound spectra. Normal-hearing listeners detected spectral increments as small as 1 dB in studies of profile analysis (Green, 1988) and accurately recognized vowels based on spectral increments (corresponding to formant frequencies) only 1–2 dB in magnitude (Lea & Summerfield, 1994; Leek et al., 1987; Loizou & Poroy, 2001). Other studies reported high sensitivity to small spectral differences over time. When listeners first heard a harmonic stimulus with 1–2 dB spectral notches in place of vowel formant peaks, they accurately identified a subsequent flat-spectrum stimulus as a vowel with spectral peaks at those notch frequencies (Summerfield et al., 1984, 1987). These studies demonstrate remarkable sensitivity to small spectral increments, but citing these results to predict the lower limit of SCEs in speech categorization must be done with caution. Marked differences exist across these psychophysical studies and investigations of SCEs in terms of tasks (explicitly detecting spectral increments versus categorizing a target sound), stimulus construction (presence of a spectral increment throughout the stimulus versus waxing and waning throughout the sentence context), and practice (extensive versus minimal).

The present experiments investigated perceptual sensitivity to spectral properties of earlier sounds and subsequent context effects in speech categorization. Experiment 1 explored this sensitivity in vowel categorization (extending the results of Stilp et al., 2015; Stilp & Alexander, 2016), and Experiment 2 explored this sensitivity in consonant categorization (extending Stilp & Assgari, 2017). In both experiments, spectral amplification in the context sentence ranged from +1 to +4 dB. This approach sought to establish the lower limits of perceptual sensitivity to spectral properties in earlier sounds that biased categorization of later speech sounds.

Methods

Participants

Target sample size was approximately 20 participants that met the performance-based inclusionary criteria (described below). Testing a sample of this size achieved 99% power at α

= 0.05 based on previous experiments of similar design (context sentences processed by filters with +5 dB gain produced mean SCEs of 0.5 stimulus steps and a standard deviation of 0.5; Assgari & Stilp, 2015; Assgari et al., 2016). Twenty-three undergraduate students at the University of Louisville participated in exchange for course credit. All participants were native English speakers and reported normal hearing.

Stimuli

Context

The context stimulus was a recording of the first author saying “Please say what this vowel is” (2,174 ms; mean fundamental frequency = 101.30 Hz; standard deviation [SD] = 20.88). This is the same stimulus used in previous investigations of SCEs by Stilp and colleagues (Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016). Average energy in the low- F_1 (100–400 Hz) and high- F_1 (550–850 Hz) regions was approximately equal (within 1 dB of each other). The sentence was then processed by a 300-Hz-wide finite impulse response filter near F_1 in the target vowels /ɪ/ or /ɛ/: 100–400 Hz or 550–850 Hz, respectively. The level of filter gain in the passband (with zero gain at other frequencies) varied from +1 to +4 dB in 1-dB steps. This created “low- F_1 -amplified” and “high- F_1 -amplified” versions of the context. Filters were created using the `fir2` function in MATLAB (MathWorks, Inc., Natick, MA) with 1,200 coefficients. Finally, context stimuli were low-pass filtered at a cutoff frequency of 5,000 Hz.

Targets

Target vowels were the same /ɪ/-to-/ɛ/ continuum as previously tested by Stilp and colleagues (Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016). For a detailed description of the generation procedures, see Winn and Litovsky (2015). Briefly, tokens of /ɪ/ and /ɛ/ were recorded by the first author. Formant contours from each token were extracted using PRAAT (Boersma & Weenink, 2014). In the /ɪ/ endpoint, F_1 linearly increased from 400 to 430 Hz, and F_2 linearly decreased from 2,000 to 1,800 Hz. In the /ɛ/ endpoint, F_1 linearly decreased from 580 to 550 Hz, and F_2 linearly decreased from 1,800 to 1,700 Hz. These F_1 trajectories were linearly interpolated to create a ten-step continuum of formant tracks; linear interpolations also were performed for F_2 trajectories. A single voice source was extracted from the /ɪ/ endpoint. Formant tracks were used to filter this source, producing the ten-step continuum of vowel tokens. Energy above 2,500 Hz was replaced with the energy high-pass-filtered from the original /ɪ/ token for all vowels. Final vowel stimuli were 246 ms in duration with fundamental frequency set to 100 Hz throughout the vowel.

All context sentences and vowels were set to equal root mean square (RMS) amplitude. Experimental trials were then created by concatenating one vowel to a context sentence with a 50-ms silent interstimulus interval.

Procedure

After acquisition of informed consent, participants were seated in a sound attenuating booth (Acoustic Systems, Inc., Austin, TX). Stimuli were D/A converted by RME HDSPe AIO sound cards (Audio AG, Haimhausen, Germany) on personal computers and passed through a programmable attenuator (TDT PA4, Tucker-Davis Technologies, Alachua, FL) and headphone buffer (TDT HB6). Stimuli were presented diotically at 70 dB sound pressure level (SPL) over circumaural headphones (Beyerdynamic DT-150, Beyerdynamic Inc. USA, Farmingdale, NY). A custom MATLAB script led the participants through the experiment. After each trial, participants clicked the mouse to indicate whether the target vowel sounded more like “ih (as in ‘bit’)” or “eh (as in ‘bet’)”.

Participants first completed 20 practice trials. On each practice trial, the context was a sentence from the AzBio corpus (Spahr et al., 2012) and the target was one of the two endpoints from the vowel continuum. Listeners were required to categorize vowels with at least 80% accuracy to proceed to the main experiment. If they failed to meet this criterion, they were allowed to repeat the practice session up to two more times. If participants were still unable to categorize vowels with 80% accuracy after the third practice session, they were not allowed to participate in the main experiment. A second performance criterion was implemented where participants were required to maintain 80% accuracy on endpoint vowels throughout the main experiment. If a participant met the performance criterion during practice trials but did not maintain this level of performance throughout the experiment, his or her data were not included in statistical analyses.

The experiment comprised four blocks, with each block testing 160 trials at a single filter gain (+1, +2, +3, +4 dB). Trials were presented in random order, and block order was counterbalanced across participants. The experiment was self-paced and allowed the participants the opportunity to take breaks between each block. No feedback was provided. The entire experiment lasted approximately 40 minutes.

Results

Three listeners failed to categorize vowels with 80% accuracy in the practice session and did not proceed to the main experiment. One other listener passed the practice session but failed to maintain 80% accuracy on vowel continuum endpoints throughout the experiment. This listener’s data were removed, leaving responses from the remaining 19 listeners included in analyses.

Results were analyzed using generalized linear mixed-effect models in R (R Development Core Team, 2016) using the lme4 package (Bates et al., 2014). Mixed-effects modeling allows estimation of the variables under study (fixed effects) separately from variability due primarily to individual differences of participants randomly sampled from the population (random effects). In the model, responses were transformed using the binomial logit linking function. Model architecture matched that used by Stilp et al. (2015) and Stilp and Assgari (2017). The dependent variable was modeled as binary (“ih” or “eh” coded as 0 and 1, respectively). Fixed effects in the model included vowel target (coded as a continuous variable from 1 to 10 then mean-centered, spanning -4.5 to $+4.5$), filter frequency (categorical variable with two levels: low F_1 and high F_1 , with high F_1 set as the default level), filter gain (coded as a continuous variable from 1 to 4 dB then mean-centered, spanning -1.5 to $+1.5$), and the interaction between filter frequency and filter gain. Random slopes were included for each fixed main effect and the interaction to allow the magnitudes of these factors to vary by listener, thereby capturing differential sensitivity to these manipulations. A random intercept of listener was also included to account for individual differences relative to each listener’s baseline level of performance (see Jaeger, 2008 for discussion). The final model had the following form:

$$\text{Response} \sim \text{Target} + \text{FilterFrequency} + \text{FilterGain} + \text{FilterFrequency} \times \text{FilterGain} + (1 + \text{Target} + \text{FilterFrequency} + \text{FilterGain} + \text{FilterFrequency} \times \text{FilterGain} \mid \text{Listener})$$

Model coefficients are listed in Table 1, and mean responses with model fits are shown in Figure 1. Estimates in Table 1 are relative to the default level of Filter Frequency (high F_1) and values of 0 for mean-centered variables (Target, corresponding to the hypothetical stimulus 5.5 on the 10-step continuum; Filter Gain, corresponding to the hypothetical filter gain of 2.5 dB). The significant effect of Target predicts more “eh” responses in the high- F_1 -amplified condition with each rightward step along the vowel target continuum (toward the / ϵ / endpoint). The significant positive effect of Filter Frequency predicts an increase in “eh” responses when the filtering condition is changed from high F_1 to low F_1 , consistent with the hypothesized direction of SCEs. Finally, the interaction between Filter Frequency and Filter Gain was statistically significant. This predicts that for each 1-dB increase in filter gain, listeners will respond “eh” more often when the filtering condition is changed from high F_1 to low F_1 . In other words, the model predicts that SCEs will increase in magnitude as filter gain increases, similar to Stilp et al. (2015).

Table 1 Mixed-effects model results for Experiment 1. “Target” refers to the slope of the logistic function, defined as the change in log odds of the listener responding “eh” resulting from a change of one step along the vowel continuum. “FilterFrequency” lists the change in log odds of the listener responding “eh” resulting from changing the context filtering from the high F₁ region (550-850 Hz) to the low F₁ region (100-400 Hz). “FilterGain” lists the change in log odds of an “eh” response resulting from increasing filter gain by 1 dB. “FilterFrequency x FilterGain” indicates the change in the size of the FilterFrequency effect (*i.e.*, SCE) per dB of filter gain. SE = standard error of the mean.

	Estimate	SE	Z	p
Intercept	-0.13	0.15	-0.84	0.40
Target	1.18	0.08	14.56	<2e-16
FilterFrequency	0.28	0.07	4.18	<3e-5
FilterGain	-0.09	0.08	-1.12	0.26
FilterFrequency x FilterGain	0.21	0.06	3.19	<2e-3

A post hoc analysis was conducted to reveal whether vowel categorization was significantly biased by an SCE at each level of filter gain. Following Stilp et al. (2015) and Stilp and Assgari (2017), the mixed-effects model was reanalyzed with Filter Gain coded as a categorical factor. This tested the coefficient for Filter Frequency

against 0 using a Wald z-test at each level of filter gain. All other model parameters matched those described in the previous analysis. SCE magnitude was operationalized as the number of stimulus steps along the target continuum separating 50% points on the high-F₁-amplified and low-F₁-amplified response functions (after Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017). This is taken as an index of the shift in categorization following high-F₁-amplified context sentences versus following low-F₁-amplified context sentences. Given that the high-F₁ condition was the default level, the 50% point of this response function was calculated as -Intercept/Target. For the low-F₁ filtering condition, the 50% point of that response function was calculated as -(Intercept + Filter Frequency)/Target. SCEs biased vowel categorization at filter gains of +4 dB (mean SCE = 0.46 stimulus steps; $\beta = 0.55$, $Z = 3.69$, $p < 0.001$) and +3 dB (mean SCE = 0.49 stimulus steps; $\beta = 0.58$, $Z = 3.78$, $p < 0.001$) but did not influence vowel categorization at filter gains of +2 dB (mean SCE = -0.03 stimulus steps; $\beta = -0.03$, $Z = -0.25$, $p = 0.80$) or +1 dB (mean SCE = 0.04 stimulus steps; $\beta = 0.05$, $Z = 0.32$, $p = 0.75$).

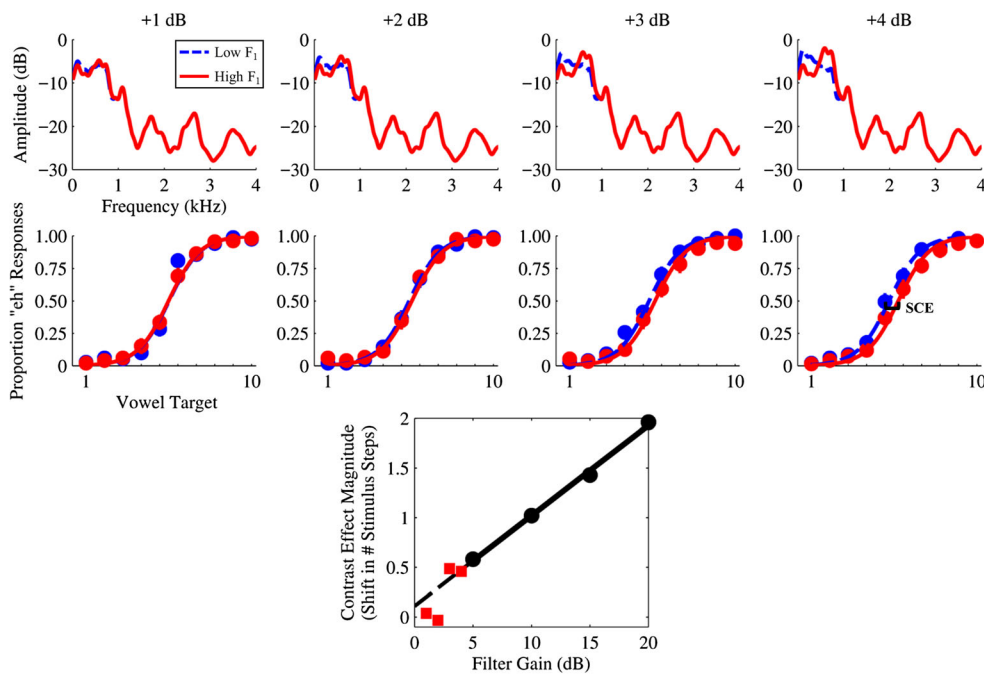


Fig. 1 Stimuli and results from Experiment 1. The top row depicts long-term average spectra of the context sentence, filtered to add +1 (leftmost panel) to +4 dB (rightmost panel) amplification in one of two frequency regions: low F₁ (100-400 Hz; dashed blue lines) or high F₁ (550-850 Hz; solid red lines). In the middle row, circles depict listeners’ mean responses as a function of each target vowel in the vowel continuum. Curves fit to the data are predicted responses generated by the generalized linear mixed-effects model. The bottom row depicts previous results from

Stilp and Alexander (2016; black circles), the linear regression fit to those data (solid black line), the same regression extrapolated down to the filter gains under investigation (dashed black line), and results from Experiment 1 (red squares). In this plot, both past and present results are from mixed-effects models where filter gains are tested categorically; *i.e.*, each condition examined individually without the assumption that SCEs scaled linearly as a function of filter gain

Discussion

Vowel categorization was biased by subtle manipulations of spectral properties in preceding sounds. Differentially amplifying key frequencies in the context sentence (100–400 Hz vs. 550–850 Hz) by +3 and +4 dB produced SCEs that biased listeners' vowel categorization, but amplifying these frequency regions by +1 or +2 dB did not influence listeners' responses. Previous studies of SCEs in vowel perception only tested listeners' sensitivity down to +5 dB amplification in key frequency regions (Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016). The present results demonstrate that these effects continue down to +3 dB and that the magnitudes of these effects appear to be in line with previous results (Figure 1). Further discussion and integration of the present results with previous findings appear in the General Discussion.

Experiment 1 identified the lower limit for perceptual sensitivity to spectral properties of preceding sounds during vowel categorization. Does this result inform perceptual sensitivity to preceding context during consonant categorization? The linear relationship between SCEs and amount of spectral amplification (*i.e.*, filter gain) was similarly strong for consonants (Stilp & Assgari, 2017) as it was for vowels (Stilp et al., 2015; Stilp & Alexander, 2016). This is despite studies differing in both context and target stimuli, spectral cues (F_1 vs. F_3), frequency regions (below 1,000 vs. 1,700–3,700 Hz), and spectral bandwidths receiving amplification (300 vs. 1,000 Hz). However, this does not guarantee that listeners will exhibit similar sensitivity to spectral increments across different bandwidths and frequency regions than those tested in Experiment 1. Experiment 2 used the same paradigm as Experiment 1 to measure perceptual sensitivity to preceding acoustic context during consonant categorization.

Methods

Participants

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

Stimuli

Context

The context stimulus was a recording of a male talker saying "Correct execution of my instructions is crucial" (2,200 ms, mean fundamental frequency = 146.90 Hz, SD = 15.58). This sentence, selected from the TIMIT database (Garofolo et al., 1990), is the same stimulus used in Stilp and Assgari (2017).

Average energy in 1,700–2,700 Hz and 2,700–3,700 Hz regions was equal in this sentence. Following Stilp and Assgari (2017), the sentence was then processed by a 1,000-Hz-wide finite impulse response filter near F_3 in the target consonants /da/ and /ga/, creating "low- F_3 -amplified" (1,700–2,700 Hz) and "high- F_3 -amplified" (2,700–3,700 Hz) renditions of the context. As in Experiment 1, filter gain in the passband (with zero gain at other frequencies) varied from +1 to +4 dB in 1-dB steps. Filters were created using the `fir2` function in MATLAB (MathWorks, Inc., Natick, MA) with 1,200 coefficients. Finally, context stimuli were low-pass filtered with a cutoff frequency of 5,000 Hz.

Targets

Target consonants were the same /da/-to-/ga/ continuum as previously tested by Stilp and Assgari (2017). These ten morphed natural tokens were taken from Stephens and Holt (2011). F_3 onset frequencies varied from 2,703 Hz (/da/ endpoint) to 2,338 Hz (/ga/ endpoint) before converging at/near 2,614 Hz for the following /a/. The duration of the consonant transition was 63 ms, and total syllable duration was 365 ms. Each context and consonant target was set to equal root mean square (RMS) amplitude. Trial sequences were then created by concatenating one consonant target to a context sentence with a 50-ms silent interstimulus interval.

Procedure

Experiment 2 used the same procedure as Experiment 1.

Results

All 20 listeners categorized endpoint vowels with 80% accuracy in the practice session and throughout the main experiment, so results from all were included in analyses. Results were again analyzed using generalized linear mixed-effect models in R, using the same model architecture as in Experiment 1 and previous experiments (Stilp et al., 2015; Stilp & Assgari, 2017)². The default level of Filter Frequency was high F_3 , and the dependent variable coded a "da" response as 0 and a "ga" response as 1.

Model coefficients are listed in Table 2, and mean responses with model fits are shown in Figure 2. Estimates in Table 2 are relative to the default level of Filter Frequency (high F_3) and values of 0 for mean-centered variables Target and Filter Gain. The significant effect of Target predicted more "ga" responses in the high- F_3 -filtered condition with each rightward step along the consonant continuum (toward the /ga/ endpoint). The significant negative effect of Filter Frequency predicted a decrease in "ga" responses when the filtering condition was changed from high F_3 to low F_3 ,

Table 2 Mixed-effects model results for Experiment 2. “Target” refers to the slope of the logistic function, defined as the change in log odds of the listener responding “ga” resulting from a change of one step along the consonant continuum. “FilterFrequency” lists the change in log odds of the listener responding “ga” resulting from changing the context filtering from the high F₃ region (2,700-3,700 Hz) to the low F₃ region (1,700–2,700 Hz). “FilterGain” lists the change in log odds of a “ga” response resulting from increasing filter gain by 1 dB. “FilterFrequency x FilterGain” indicates the change in the size of the FilterFrequency effect (*i.e.*, SCE) per dB of filter gain. SE = standard error of the mean.

	Estimate	SE	Z	p
Intercept	-0.31	0.26	-1.20	0.23
Target	1.75	0.10	17.27	<2e-16
FilterFrequency	-0.42	0.08	-5.18	<3e-7
FilterGain	0.08	0.06	1.21	0.23
FilterFrequency x FilterGain	-0.24	0.07	-3.28	<2e-3

consistent with the hypothesized direction of SCEs. Finally, the significant interaction between Filter Frequency and Filter Gain predicted that for each 1-dB increase in filter gain, listeners would respond “ga” less often when the filtering condition was changed from high F₃ to low F₃. The model predicted that SCEs would increase as filter gain increased,

similar to Experiment 1 (Table 1). This particular prediction is revisited in the Discussion.

The model was again reanalyzed with Filter Gain coded as a categorical factor to test SCEs at each level of filter gain. SCEs were calculated the same way as described in Experiment 1. SCEs were statistically significant at filter gains of +4 (mean SCE = 0.51 stimulus steps; $\beta = -0.91, Z = -4.83, p < 0.001$) and +3 dB (mean SCE = 0.24 stimulus steps; $\beta = -0.43, Z = -2.68, p < 0.01$), but were not significantly different from 0 at filter gains of +2 dB (mean SCE = 0.11 stimulus steps; $\beta = -0.20, Z = -1.28, p = 0.20$) or +1 dB (mean SCE = 0.10 stimulus steps; $\beta = -0.17, Z = -1.09, p = 0.28$).

Discussion

Experiments 1 and 2 utilized different context sentences where different frequency regions with different bandwidths were amplified. Target stimuli differed markedly in terms of the spectral cues and frequency regions that differentiated listeners’ responses. Nevertheless, results were highly consistent across experiments. Context sentences with +3 and +4 dB spectral amplification in key frequency regions significantly biased categorization of the subsequent vowel or consonant

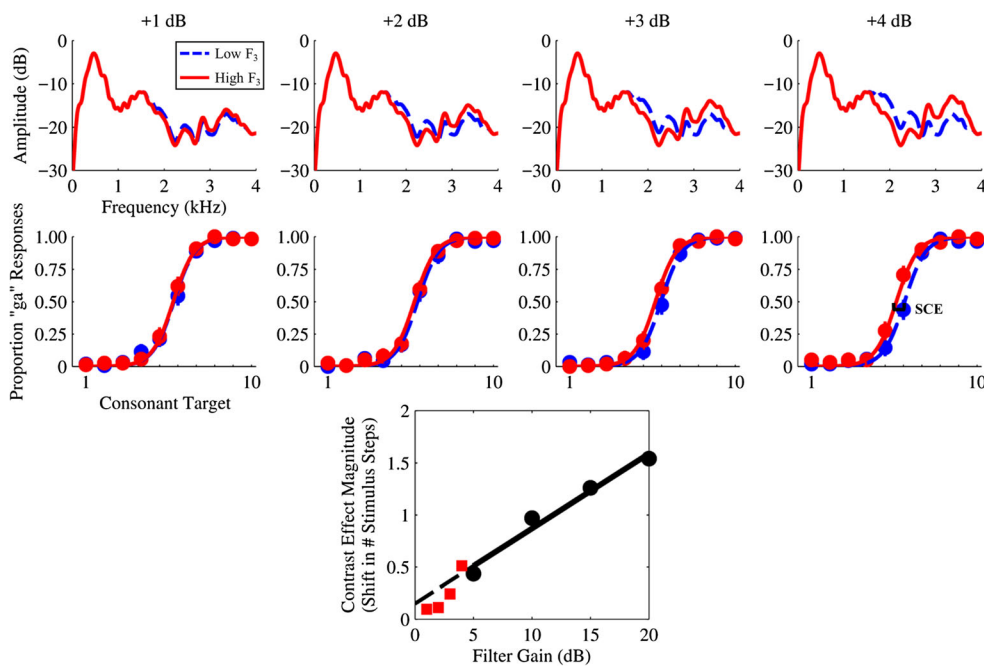


Fig. 2 Stimuli and results from Experiment 2. The top row depicts long-term average spectra of the context sentence, filtered to add +1 (leftmost panel) to +4 dB (rightmost panel) amplification in one of two frequency regions: low F₃ (1,700-2,700 Hz; dashed blue lines) or high F₃ (2,700-3,700 Hz; solid red lines). In the middle row, circles depict listeners’ mean responses as a function of each target consonant in the consonant continuum. Curves fit to the data are predicted responses generated by the generalized linear mixed-effects model. The bottom row depicts previous

results from Stilp and Assgari (2017; black circles), the linear regression fit to those data (solid black line), the same regression extrapolated down to the filter gains under investigation (dashed black line), and results from Experiment 2 (red squares). In this plot, both past and present results are from mixed-effects models where filter gains are tested categorically; *i.e.*, each condition examined individually without the assumption that SCEs scaled linearly as a function of filter gain

via SCEs, but sentences with +1 and +2 dB peaks did not bias speech categorization.

Because Experiments 1 and 2 produced the same pattern of results, one might search for common underlying acoustic properties in the context sentences. Figure 3 illustrates the amplitude envelopes in low- F_1 and high- F_1 (top) and low- F_3 and high- F_3 frequency regions (bottom) of unfiltered context sentences from each experiment using a shared y-axis. Overall energy in these frequency regions markedly differed, with more energy at lower (F_1) frequencies. Additionally, modulations in F_1 versus F_3 frequency regions also differed considerably. As such, it is difficult to identify key acoustic characteristics responsible for generating the same patterns of results across experiments.

General discussion

When spectral properties of earlier sounds differ from those in a subsequent target sound, categorization of the target sound becomes biased through spectral contrast effects (SCEs). Previous studies revealed that differentially amplifying key frequency regions in the context by +5 dB was sufficient to produce SCEs (Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017), but they did not elucidate the lower limit of perceptual sensitivity to such spectral differences. Here, filter gain was progressively decreased to reveal when the preceding sentence no longer biased categorization of a subsequent speech target. Vowel categorization (Experiment 1) and consonant categorization (Experiment 2) both showed the same pattern of results: SCEs biased

categorization performance following +4 and +3 dB amplification of key frequencies in the context, but no such biases were produced by +2 and +1 dB amplification.

Previous studies used linear regressions to relate shifts in speech categorization to the magnitudes of spectral prominences in earlier sounds (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017). When combining past and present results, these relationships maintained for vowel categorization [combining Stilp and Alexander (2016) results with Experiment 1: $r = 0.98$, $p < 0.0001$] and consonant categorization [combining Stilp and Assgari (2017) results with Experiment 2: $r = 0.99$, $p < 0.0001$]. While one might conclude that a linear relationship is an excellent descriptor of spectral amplification in earlier sounds and subsequent SCEs in categorization of later sounds, there are reasons to be circumspect about the exact nature of this relationship. First, each analysis above combines results from two different listener groups who heard different ranges of spectral amplification in the context sentences (+1 to +4 dB here; +5 to +20 dB in previous studies). Second, SCEs might increase at different rates depending on the extent of spectral amplification in the context. The mixed-effects models assumed that the Filter Frequency \times Filter Gain interaction is linear in nature, or that SCE magnitudes would increase linearly with each additional dB of filter gain. This interaction was statistically significant for each listener group in past and present studies, but predicted SCE magnitudes increased faster from +1 to +4 dB filter gain (vowels: 0.18 stimulus steps per additional dB of filter gain; consonants: 0.14 steps/dB) than from +5 to +20 dB filter gain [vowels in Stilp & Alexander (2016): 0.09 steps/dB; consonants in Stilp & Assgari (2017): 0.07 steps/dB]. Third, when the assumption of SCE magnitudes

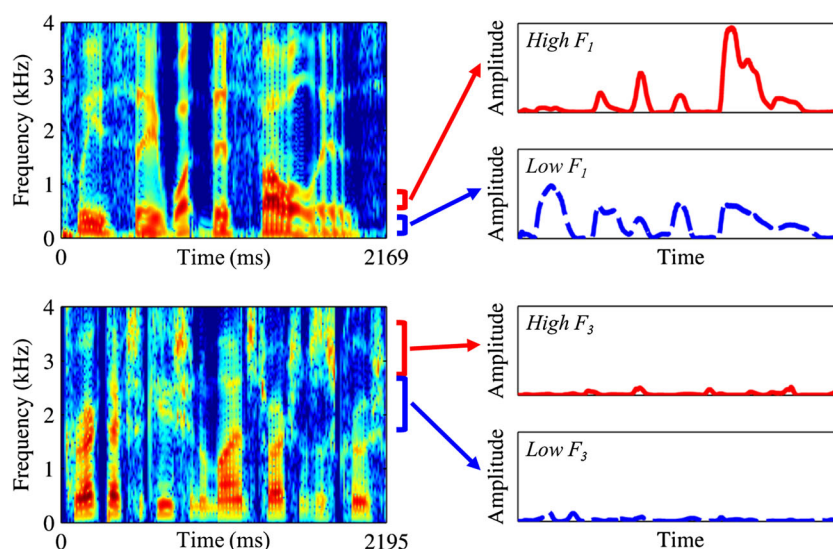


Fig. 3 Spectral characteristics of context sentences tested in Experiment 1 (“Please say what this vowel is”; top row) and Experiment 2 (“Correct execution of my instructions is crucial”; bottom row). Spectrograms are depicted at left, with low- and high- F_1 (top) or low- and high- F_3 (bottom)

frequency regions highlighted. At right, low-pass-filtered amplitude envelopes from these frequency regions are plotted using a common y-axis. Dashed blue lines depict lower-frequency (low- F_1 or low- F_3) envelopes; solid red lines depict higher-frequency (high- F_1 or high- F_3) envelopes

scaling linearly with filter gain was removed in post hoc analyses, results at smaller versus larger amounts of filter gain differed in nature. In previous studies from +5 to +20 dB, SCE magnitudes increased linearly as a function of filter gain even without the mixed-effects model assuming this to be the case. Across small filter gains in the present experiments, SCE magnitudes increased more akin to step functions, particularly in Experiment 1. These points raise questions about whether SCEs magnitudes indeed scale linearly across all amounts of spectral amplification in earlier sounds, or only over a particular range of filter gains.

Results are consistent with adaptation-related mechanisms that have been suggested to underlie SCEs in speech perception (Delgutte, 1996; Delgutte et al., 1996; Holt et al., 2000; Holt & Lotto, 2002). In earlier studies, considerable spectral amplification in the context sounds biased subsequent speech categorization through SCEs (see Stilp et al., 2015 for review). These results would be consistent with neural adaptation or adaptation of inhibition/suppression producing a greater response to unadapted or less-adapted frequencies in the target sound than adapted frequencies present in the context sounds. Here, small amounts of amplification minimally altered the spectra of context sounds; when amplification was less than +3 dB, no SCEs were observed. It is possible that +1 or +2 dB spectral amplification did not produce differential patterns of neural adaptation (following amplification of the lower- versus higher-frequency region in the context sentence); this would be consistent with no observed effects of context filtering on categorization of the subsequent speech target. This outcome was predicted by Holt et al. (2000), where less intense precursor sounds (here, key frequency regions being amplified to a smaller extent) should yield smaller context effects on categorization of subsequent speech sounds. While the present results fit within a framework where SCEs are produced by adaptation-related mechanisms, direct confirmation of these physiological substrates is still required.

Spectral context is most influential when disambiguating perceptually ambiguous speech sounds. These sounds were found toward the middle of the vowel and consonant target continua, and context effects are most evident here (Figures 1, 2). Conversely, the endpoints of these target continua are far less perceptually ambiguous and thus less influenced by context. However, naturally produced speech sounds frequently fall short of such extremes (Lindblom, 1963). Due to coarticulation, spectral properties of speech sounds are often concessions due to where speech articulators have been and where they are going next. Perception compensates for this “undershoot” by magnifying spectral differences between successive sounds, such that lower-frequency context can increase percepts of a higher-frequency target, and *vice versa* (Lindblom & Studdert-Kennedy, 1967; Holt et al., 2000). Evidence that spectral context helps disambiguate more perceptually ambiguous (here, mid-continuum) speech sounds parallels everyday speech production, where target phonemes are more acoustically ambiguous

than idealized target productions (here, continuum endpoints). The fact that spectral amplification of earlier stimuli can be extremely modest (+3 dB) yet still influence categorization of later sounds suggests that SCEs influence everyday speech perception more often than previously considered.

High perceptual sensitivity to spectral characteristics in earlier sounds was likely facilitated by extremely low acoustic variability across contexts. Each trial presented one of two filtered versions of the same context sentence, making acoustic variability from trial to trial extremely low. Assgari and colleagues have shown that acoustic variability across context sentences can diminish SCEs. When +5 dB spectral amplification was added to context sentences, hearing 200 different talkers resulted in smaller SCEs than hearing the same talker on every trial (Assgari & Stilp, 2015). Subsequent experiments revealed that variability in the talkers’ fundamental frequencies but not their gender modulated this relationship (Assgari et al., 2016). However, when large (+20 dB) spectral peaks were added to context sentences, SCE magnitudes were comparable irrespective of the number of talkers (Assgari & Stilp, 2015). Thus, if preceding contexts exhibit little acoustic variability from trial to trial, SCEs can be produced by very small spectral amplification (as in the present experiments); if contexts exhibit high acoustic variability, larger amounts of spectral amplification may be required to produce SCEs.

In conclusion, the present experiments established the lower limit of perceptual sensitivity to spectral properties of earlier sounds during speech categorization. In vowel categorization and consonant categorization alike, responses were significantly biased by +3 and +4 dB amplification of key frequencies in the context sentence, but were not biased by smaller amounts of spectral amplification. At present, it is difficult to identify common acoustic characteristics that were shared across context sentences that might underlie similar patterns of results. Nevertheless, sensitivity to such modest spectral differences across sounds reinforces the fact that sensory systems are remarkably sensitive to changes in the environment, in this case even relatively small ones.

Acknowledgements The authors thank Quentin Summerfield and an anonymous reviewer for constructive comments and suggestions on an earlier draft of this manuscript. The authors also thank Ashley Batliner, Carly Newman, and Madison Rhyne for their assistance with data collection.

References

- Alexander, J. M., & Kluender, K. R. (2010). Temporal properties of perceptual calibration to local and broad spectral characteristics of a listening context. *Journal of the Acoustical Society of America*, 128(6), 3597-3613.
- Assgari, A. A., Mohiuddin, A., Theodore, R. M., & Stilp, C. E. (2016). Dissociating contributions of talker gender and acoustic variability

- for spectral contrast effects in vowel categorization. *Journal of the Acoustical Society of America*, 139(4), 2124–2124.
- Asgari, 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. Retrieved from <http://cran.r-project.org/package=lme4>
- Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer [Computer program].
- Byrne, A. J., Stellmack, M. A., & Viemeister, N. F. (2011). The enhancement effect: Evidence for adaptation of inhibition using a binaural centering task. *Journal of the Acoustical Society of America*, 129(4), 2088–2094.
- Carcagno, S., Semal, C., & Demany, L. (2012). Auditory enhancement of increments in spectral amplitude stems from more than one source. *Journal of the Association for Research in Otorhinolaryngology*, 13(5), 693–702.
- Delgutte, B. (1996). Auditory neural processing of speech. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Sciences* (pp. 507–538). Oxford: Blackwell Publishing Ltd.
- 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.), *Proceedings of Auditory Basis of Speech Perception*. European Speech Communication Association.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., & Dahlgren, N. (1990). “DARPA TIMIT acoustic-phonetic continuous speech corpus CDR0M.” *NIST Order No. PB91-505065*, National Institute of Standards and Technology, Gaithersburg, MD.
- Green, D. M. (1988). *Profile analysis: Auditory intensity discrimination*. Oxford University Press.
- Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, 16(4), 305–312.
- Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America*, 120(5), 2801–2817.
- 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.
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.
- Kluender, K. R., Coady, J. A., & Kieffe, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, 41(1), 59–69.
- Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1), 98–104.
- Laing, E. J., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology*, 3, 1–9. doi:<https://doi.org/10.3389/fpsyg.2012.00203>
- Lea, A. P., & Summerfield, Q. (1994). Minimal spectral contrast of format peaks for vowel recognition as a function of spectral slope. *Perception & Psychophysics*, 56(4), 379–391.
- Leek, M. R., Dorman, M. F., & Summerfield, Q. (1987). Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 81(1), 148–154.
- Lindblom, B. E. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, 35(11), 1773–1781.
- Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, 42(4), 830–843.
- Loizou, P. C., & Poroy, O. (2001). Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners. *Journal of the Acoustical Society of America*, 110(3), 1619–1627.
- 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.
- Nelson, P. C., & Young, E. D. (2010). Neural correlates of context-dependent perceptual enhancement in the inferior colliculus. *The Journal of Neuroscience*, 30(19), 6577–87.
- R Development Core Team. (2016). “R: A language and environment for statistical computing.” Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.r-project.org/>
- Sjerps, M. J., McQueen, J. M., & Mitterer, H. (2013). Evidence for precategorical extrinsic vowel normalization. *Attention, Perception & Psychophysics*, 75(3), 576–587.
- 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., & Smiljanic, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. *Journal of Phonetics*, 41(3–4), 145–155.
- Sjerps, M. J., Zhang, C., & Peng, G. (2017). Lexical tone is perceived relative to locally surrounding context, vowel quality to preceding context. *Journal of Experimental Psychology: Human Perception and Performance*. <https://doi.org/10.1037/xhp0000504>
- Sjerps, M., McQueen, J., & Mitterer, H. (2012). Extrinsic normalization for vocal tracts depends on the signal, not on attention. In *Proceedings of Interspeech 2012: 13th Annual Conference of the International Speech Communication Association*, 394–397.
- Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., ... Cook, S. (2012). Development and validation of the AzBio sentence lists. *Ear and Hearing*, 33(1), 112–117.
- Stephens, J. D. W., & Holt, L. L. (2011). A standard set of American-English voiced stop-consonant stimuli from morphed natural speech. *Speech Communication*, 53(6), 877–888.
- 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., Kieffe, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, and Psychophysics*, 72(2), 470–480.
- Stilp, C. E., & Anderson, P. W. (2014). Modest, reliable spectral peaks in preceding sounds influence vowel perception. *Journal of the Acoustical Society of America*, 136(5), EL383–EL389.
- 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.
- Summerfield, Q., Haggard, M., Foster, J., & Gray, S. (1984). Perceiving vowels from uniform spectra - phonetic exploration of an auditory aftereffect. *Perception & Psychophysics*, 35(3), 203–213.
- Summerfield, Q., Sidwell, A., & Nelson, T. (1987). Auditory enhancement of changes in spectral amplitude. *Journal of the Acoustical Society of America*, 81(3), 700–708.
- Viemeister, N. F., & Bacon, S. P. (1982). Forward masking by enhanced components in harmonic complexes. *Journal of the Acoustical Society of America*, 71(6), 1502–1507.

- 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. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *96*(3), 1263–1282.
- 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.
- Winn, M. B., & Litovsky, R. Y. (2015). Using speech sounds to test functional spectral resolution in listeners with cochlear implants. *Journal of the Acoustical Society of America*, *137*(3), 1430–1442.