



Contributions of natural signal statistics to spectral context effects in consonant categorization

Christian E. Stilp¹ · Ashley A. Assgari¹

Accepted: 23 March 2021 / Published online: 13 May 2021
© The Psychonomic Society, Inc. 2021

Abstract

Speech perception, like all perception, takes place in context. Recognition of a given speech sound is influenced by the acoustic properties of surrounding sounds. When the spectral composition of earlier (context) sounds (e.g., a sentence with more energy at lower third formant [F_3] frequencies) differs from that of a later (target) sound (e.g., consonant with intermediate F_3 onset frequency), the auditory system magnifies this difference, biasing target categorization (e.g., towards higher- F_3 -onset /d/). Historically, these studies used filters to force context stimuli to possess certain spectral compositions. Recently, these effects were produced using unfiltered context sounds that already possessed the desired spectral compositions (Stilp & Assgari, 2019, *Attention, Perception, & Psychophysics*, 81, 2037–2052). Here, this natural signal statistics approach is extended to consonant categorization (/g/–/d/). Context sentences were either unfiltered (already possessing the desired spectral composition) or filtered (to imbue specific spectral characteristics). Long-term spectral characteristics of unfiltered contexts were poor predictors of shifts in consonant categorization, but short-term characteristics (last 475 ms) were excellent predictors. This diverges from vowel data, where long-term and shorter-term intervals (last 1,000 ms) were equally strong predictors. Thus, time scale plays a critical role in how listeners attune to signal statistics in the acoustic environment.

Keywords Speech perception · Speech categorization · Efficient coding · Context effects · Spectral contrast

Fred Attneave (1954) elegantly declared, “The world as we know it is lawful” (p. 183). Objects and events in the sensory environment are highly structured in their compositions and across time. If sensory and perceptual processing are to be considered efficient, they ought to capitalize on this structure. These are the core tenets of the Efficient Coding Hypothesis (Attneave, 1954; Barlow, 1961).

The strongest empirical support for the Efficient Coding Hypothesis has been in the visual system. A wide range of studies has documented the statistical structure of natural images (e.g., Bell & Sejnowski, 1997; Field, 1987; Olshausen & Field, 1996; Simoncelli, 2003). This structure has been linked to neural response properties in the visual system (e.g., Field, 1987; Ruderman et al., 1998; Simoncelli & Olshausen, 2001; van Hateren & van der Schaaf, 1998) as well as informed observer performance in visual perception tasks (e.g., Burge et al., 2010; Geisler, 2008; Geisler et al., 2001; Tkačik et al., 2010). These connections have informed theories of sensory coding of natural stimuli (Field, 1994; Vinje & Gallant, 2000).

While comparatively nascent to work in the visual system, support for efficient coding in audition is growing. Much of this work can be broadly organized into two areas: sensitivity to the statistics of stimulus presentation and sensitivity to the statistics of stimulus composition. Changes in the probability and/or variance of stimulus presentation can alter neural responsiveness (through stimulus-specific adaptation; e.g., Malmierca et al., 2009; Ulanovsky et al., 2003) and tuning (Dean et al., 2005; Dean et al., 2008). Similarly, changes in the probability density and variance of speech sound presentation alters

Public significance statement

Perception of speech sounds depends critically on the sounds that precede them. This study extends our understanding of these context effects by revealing the timecourse of how earlier sounds shape the categorization of later consonant sounds. Notably, this timecourse is shorter than the one previously reported to shape vowel categorization, revealing important flexibility in how perception attunes to sounds in the listening environment. These results illuminate the ways in which preceding sounds affect everyday speech perception.

✉ 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

listeners' categorization behavior (Clayards et al., 2008; Holt & Lotto, 2006; Maye et al., 2002; Newman et al., 2001; Theodore & Monto, 2019; Toscano & McMurray, 2010).

With regard to the statistics of stimulus composition, reports suggest the statistical structure (independent components) of natural sounds reflects neural tuning in the auditory nerve and cochlear nucleus (Lewicki, 2002; Stilp & Lewicki, 2014). Additionally, neural activity in auditory cortex reflects patterns of covariance between sound properties (Lu et al., 2019). Auditory textures can be both synthesized and recognized by the statistical regularities in their respective compositions (McDermott et al., 2013; McDermott & Simoncelli, 2011; McWalter & McDermott, 2018). Finally, patterns of covariance among the acoustic properties of heavily edited musical instrument sounds can be so influential that perception abandons physical acoustics and instead represents and discriminates sounds according to these statistical properties (Stilp & Kluender, 2011, 2012, 2016; Stilp, Rogers, & Kluender, 2010).

The present experiments investigate efficient coding of the statistical composition of acoustic context. It is well established that preceding sounds (i.e., context) can influence recognition of subsequent sounds. For example, Ladefoged and Broadbent (1957) examined perception of vowels that followed an introductory sentence. Listeners reported whether the vowel was /i/ (as in “bit”; lower first formant frequency [F_1]) or /ɛ/ (as in “bet”; higher F_1). When the sentence was edited to make lower F_1 frequencies more prominent, listeners labeled the subsequent target vowel as the higher- F_1 /ɛ/ more often; when the sentence was edited to make higher- F_1 frequencies more prominent, listeners labeled the target vowel as the lower- F_1 /i/ more often. Subsequent work has revealed that these spectral contrast effects (SCEs) are quite general, influencing perception of a wide range of speech sounds (see Stilp, 2020a, for review) and nonspeech sounds as well (Kingston et al., 2014; Lanning & Stilp, 2020; Stilp, Alexander, Kiefe, & Kluender, 2010).

The vast majority of these studies utilized context stimuli that were filtered to possess the spectral characteristics necessary to produce the SCE. A particular token (e.g., a context sentence) would be filtered two slightly different ways; for example, a lower-frequency region or a higher-frequency region would be amplified. This filtering would differentially bias perception of the target sound (toward the higher-frequency vs. lower-frequency response option, respectively). Thus, trials in these experiments presented context stimuli that differed only in certain spectral properties consequent to filtering. While this approach affords high acoustic control and likely maximizes the probability of observing an SCE, it vastly underrepresents the pervasive and extreme acoustic variability in sounds encountered in everyday perception. Additionally, it ignores the fact that other sounds may inherently possess those spectral properties without any filtering necessary (i.e., natural signal statistics).

Stilp and Assgari (2019) addressed these shortcomings by expanding the experimental paradigm used to measure SCEs. They developed a simple metric, Mean Spectral Difference (MSD), to quantify the balance of energy across two frequency regions in a stimulus. Stimuli with relatively equal energy across low- F_1 (100–400 Hz) and high- F_1 frequency regions (550–850 Hz) had MSD values at or near zero; sentences with more energy in one frequency region than the other had larger MSD values. They used this metric to identify potential stimuli that already possessed more energy in either the low- F_1 region or the high- F_1 region. In their experiments, half of the blocks presented these unfiltered sentences as context stimuli. The remaining blocks presented filtered renditions of a single context stimulus (consistent with the methods of previous studies) processed to match the magnitudes of the spectral prominences in the unfiltered contexts. All trials presented a context sentence followed by a target vowel (varying from /i/ to /ɛ/ in F_1 frequency). Unfiltered and filtered contexts were both successful in producing SCEs that biased categorization of the target vowels, but SCEs magnitudes produced by unfiltered contexts were smaller and more variable than their filtered counterparts. These discrepancies were attributed in part to the greater acoustic variability in unfiltered sentences, as stimuli presented in a given block were different sentences often spoken by different talkers. Nevertheless, sensitivity to the natural signal statistics in unfiltered contexts offered keen insight as to how these context effects may be influencing everyday speech perception. These patterns of results were replicated when filtered and unfiltered musical passages biased categorization of musical instrument sounds (Lanning & Stilp, 2020), promoting the generality of this approach and supporting efficient coding in audition at large.

Here, this natural signal statistics approach is extended to context effects in consonant categorization. On each trial, listeners heard a context sentence preceding the target syllable (ranging from /ga/ to /da/, varying in the onset frequency of the F_3 transition). As in Stilp and Assgari (2019), MSDs were calculated to identify and select unfiltered sentence stimuli based on their relative amounts of energy in the low- F_3 region (1700–2700 Hz) and high- F_3 region (2700–3700 Hz). Filtered renditions of a single context sentence were created to match the balances of spectral energy observed in the unfiltered sentences. Consistent with Stilp and Assgari (2019), both approaches are predicted to produce SCEs that influence consonant categorization, with the unfiltered blocks producing smaller and more variable SCE magnitudes than the filtered blocks owing to their greater acoustic variability.

The present experiments also test the generalizability of this natural signal statistics approach. The F_3 frequency regions being queried are much higher (1700–3700 Hz) than those investigated by Stilp and Assgari (2019) (<850 Hz), and as such possess less overall energy in the long-term average spectrum of speech. Further, the target stimuli are distinguished by a short-duration

cue (63-ms formant transitions principally defined by their onset frequency), as opposed to vowel formants that are more spectrally prominent and of longer duration (246 ms in Stilp & Assgari, 2019). Replication of all results from Stilp and Assgari (2019) would strongly promote the flexibility and generalizability of this natural signal statistics approach; any differences in results would reveal important constraints on perceptual sensitivity to natural signal statistics in context spectra.

Methods

Participants

Ninety-eight undergraduate students at the University of Louisville participated in exchange for course credit. All reported being native English speakers with no known hearing impairments. Five individual experiments were conducted ($n = 18, 19, 21, 20,$ and 20 for Experiments 1–5, respectively), with no listener participating in more than one experiment.

Stimuli

Unfiltered contexts

Sentences were analyzed and selected according to spectral properties in the low- F_3 (1700–2700 Hz) and high- F_3 (2700–3700 Hz) frequency regions. Amplifying these frequency regions has been highly successful in producing SCEs in previous studies (Stilp, 2020b; Stilp & Assgari, 2017, 2018). Each sentence was analyzed as detailed in Stilp and Assgari (2019) using two separate bandpass filters. The passband was either 1700–2700 Hz or 2700–3700 Hz, with 5-Hz transition regions between the passband and stopbands. Filters were created using the `fir2` command in MATLAB (The MathWorks, Inc., Natick, MA) using 1,000 coefficients. The amplitude envelope for each frequency region was obtained by rectifying the signal and low-pass filtering using a second-order Butterworth filter with 30-Hz cutoff frequency. The root-mean-square (RMS) energy for each envelope was converted into dB. The Mean Spectral Difference (MSD) was defined as the difference in energy across these two frequency regions (see Fig. 1). MSDs were always subtracted in one direction (low- F_3 energy minus high- F_3 energy), with positive MSDs indicating more energy in the low- F_3 region and negative MSDs indicating more energy in the high- F_3 region. MSDs are likely to stem from a number of sources, including but not limited to phonemic content and talker size (with shorter talkers often possessing shorter vocal tracts that produce higher formants, and taller talkers often possessing longer vocal tracts that produce lower formants).¹ MSDs were calculated for every sentence in

the TIMIT database (Garofolo et al., 1990) and the 275 unique sentences in the HINT database (Nilsson et al., 1994). The results of these ecological surveys are plotted alongside the long-term average spectra for these databases in Fig. 2.

Sentences were selected from these databases for use as experimental stimuli. All sentences were spoken by men to match the sex of the talkers who produced the filtered context and the target syllables. In Stilp and Assgari (2019), MSDs were selected to be either relatively large (averages of the absolute values of MSDs tested in the same block ranged from 11–15 dB) or small (averages of the absolute values of MSDs tested in the same block spanned 5–8.5 dB). Here, absolute values of MSDs spanned a broader range (0.03 to 16.23; see Table 1) to better reflect their variability in natural speech. Other signal characteristics were allowed to vary freely across sentences (fundamental frequency, semantic and syntactic content, duration, etc.). In a given block of an experiment, one of two unfiltered sentences was presented on each trial. Generally, one sentence had a positive MSD favoring low- F_3 frequencies and the other had a negative MSD favoring high- F_3 frequencies (see Table 1).

Filtered contexts

Experiments also tested filtered renditions of a single sentence, a male talker saying “Correct execution of my instructions is crucial” (2,200 ms). This stimulus has been highly successful in biasing consonant categorization in previous studies (Stilp, 2020b; Stilp & Assgari, 2017, 2018). This stimulus served as a control in two ways. First, acoustic variability was held constant from trial to trial (talker variability, duration, and all other acoustic properties except for amplified frequencies described below). Second, given the anticipated higher variability in SCE magnitudes for unfiltered contexts, filtered contexts ensured that listeners were responding consistently in conditions where SCEs were most likely to occur. This stimulus possessed nearly equal energy in low- F_3 and high- F_3 frequency regions before any filtering was conducted (MSD = 0.08 dB). This stimulus was then processed by the same filters used to introduce spectral peaks in previous studies: 1000-Hz-wide finite impulse response filters spanning either 1700–2700 Hz or 2700–3700 Hz. Filters were created using the `fir2` function in MATLAB with 1,200 coefficients. Filter gain was determined according to the following procedure. First, the native MSD of the to-be-filtered context sentences was compared to that of an unfiltered context sentence (e.g., unfiltered sentence MSD = 15.00 dB, a strong bias toward low- F_3 frequencies). Gain for the appropriate filter (here, amplifying low- F_3 frequencies) was set to a value just below the difference between these two MSDs. The context sentence was filtered and its MSD was remeasured. If its new MSD differed from the target MSD by more than 0.10 dB, filter gain was increased slightly (e.g., adding slightly more

¹ These measures are also at the mercy of factors such as recording conditions and equipment, which were not controlled in the present investigation.

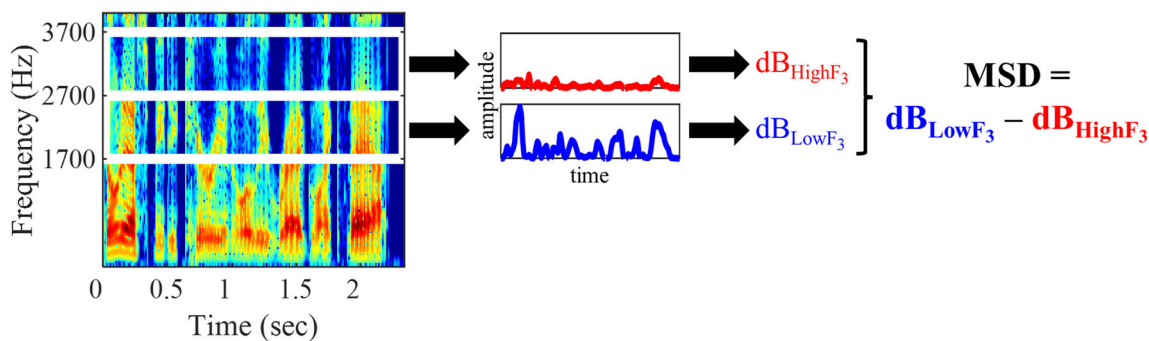


Fig. 1 Procedure for calculating Mean Spectral Differences (MSDs). Two frequency regions are excised from the sentence via bandpass filtering: low F_3 (1700–2700 Hz) and high F_3 (2700–3700 Hz). In each frequency region, the waveform is rectified and low-pass filtered to produce its amplitude envelope. Energy in each frequency region is

calculated in dB from the root-mean-square amplitude of this envelope. The MSD is defined as energy in the low- F_3 region minus energy in the high- F_3 region. Here, for the sentence “Don’t ask me to carry an oily rag like that,” $MSD = 12.38$ dB

energy to the low- F_3 region) and the process repeated. This continued iteratively until the MSDs for the unfiltered and filtered contexts were functionally equal (within 0.10 dB of each other).

Across experiments, most unfiltered context sentences were accompanied by filtered context sentences with equivalent MSDs. In Stilp and Assgari (2019), every unfiltered sentence was accompanied by a filtered sentence with an

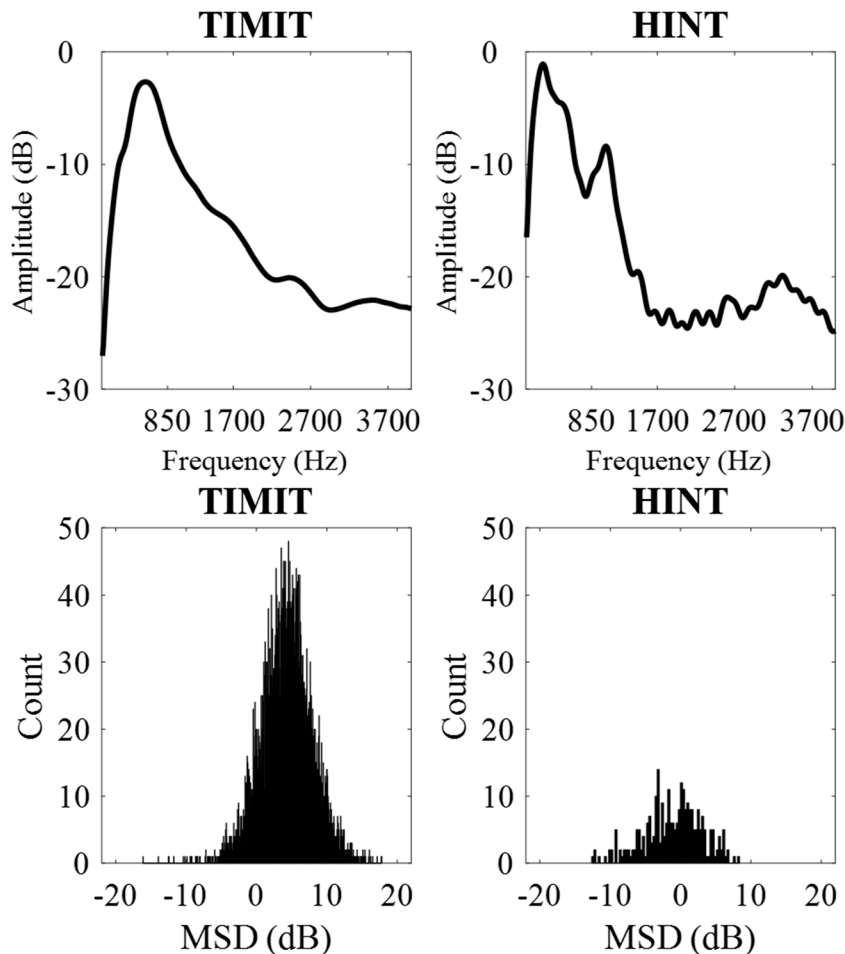


Fig. 2 (Top row) Long-term average spectra for 6,300 sentences in the TIMIT database (left) and 275 unique sentences in the HINT database (right). (Bottom row) Histograms showing the distributions of MSDs for the TIMIT database (left) and HINT database (right). Positive MSDs

indicate more energy in the low- F_3 region (1700–2700 Hz), and negative MSDs indicate more energy in the high- F_3 region (2700–3700 Hz). Experimental stimuli were selected from these distributions (see Table 1)

Table 1 Characteristics of sentences presented as unfiltered contexts in the present experiments

Expt.	Block	Sentence	Database	Duration	MSD (entire sentence)	MSD (last 475ms)
1	1	We're here to transact business.	TIMIT	1,625	16.16	6.02
1	1	Her lips, moist and parted, spoke his name.	TIMIT	2,740	-13.93	6.03
1	2	She had your dark suit in greasy wash water all year.	TIMIT	2,724	7.04	1.43
1	2	Spring Street is straight ahead.	TIMIT	1,539	-6.70	13.08
1	3	Critical equipment needs proper maintenance.	TIMIT	2,579	-0.15	-11.35
1	3	The blue jay flew over the high building.	TIMIT	2,156	-0.03	-1.93
2	1	Don't ask me to carry an oily rag like that.	TIMIT	2,773	15.45	14.69
2	1	The hallway opens into a huge chamber.	TIMIT	2,255	-12.46	1.67
2	2	The paper boy bought two apples and three ices.	TIMIT	3,075	6.68	1.63
2	2	Shipbuilding is a most fascinating process.	TIMIT	2,298	-6.30	-13.92
3	1	Don't ask me to carry an oily rag like that.	TIMIT	2,773	15.45	14.69
3	1	Those who teach values first abolish cheating.	TIMIT	2,736	-8.56	3.67
3	2	The paper boy bought two apples and three ices.	TIMIT	3,075	6.68	1.63
3	2	Shipbuilding is a most fascinating process.	TIMIT	2,298	-6.30	-13.92
4	1	The family bought a house.	HINT	1,587	8.37	-10.06
4	1	His father will come home soon.	HINT	1,890	-12.62	-1.20
4	2	They painted the wall white.	HINT	1,820	4.93	0.54
4	2	She argues with her sister.	HINT	1,659	-9.06	-9.08
5	1	Often you'll get back more than you put in.	TIMIT	2,099	13.74	15.19
5	1	Rob sat by the pond and sketched the stray geese.	TIMIT	2,860	-9.88	-19.10
5	2	Don't ask me to carry an oily rag like that.	TIMIT	2,811	13.45	18.42
5	2	The prowler wore a ski mask for disguise.	TIMIT	2,812	-6.30	-18.31
5	3	Don't ask me to carry an oily rag like that.	TIMIT	2,339	12.38	13.15
5	3	When peeling an orange, it is hard not to spray juice.	TIMIT	3,555	-3.32	-13.92

Note. Durations are listed in milliseconds; MSDs are dB differences in energy across low- F_3 and high- F_3 regions calculated across the entire sentence or only its last 475 ms. See text for details.

equivalent MSD. Given that the linear relationship between filter gain and SCE magnitude for /g/-/d/ categorization is already established (Stilp & Assgari, 2017), equal amounts of unfiltered and filtered data are not strictly necessary for the present investigation. As such, two experiments (Experiments 1 and 5) tested more unfiltered blocks than filtered blocks in order to populate the regression between unfiltered sentence MSDs and SCE magnitudes.²

Targets

Several reports have demonstrated that categorization of the /g/-/d/ contrast is influenced by SCEs (e.g., Holt, 2006; Stilp, 2020b; Stilp & Assgari, 2017, 2018), making them excellent

candidate stimuli for the present investigation. Target consonants were a series of 10 morphed natural tokens from /ga/ to /da/ (Stephens & Holt, 2011). F_3 onset frequencies varied from 2338 Hz (/ga/ endpoint) to 2703 Hz (/da/ endpoint) before converging at/near 2614 Hz for the following /a/. The duration of the consonant transition was 63 ms, and total syllable duration was 365 ms. Categorization of these targets has been shown to be influenced by SCEs (Stilp, 2020b; Stilp & Assgari, 2017, 2018).

All context sentences and vowels were low-pass filtered at 5 kHz and set to equal RMS amplitude. Experimental trials were then created by concatenating each target vowel to each context sentence with 50-ms silent interstimulus intervals.

Procedure

All experimental procedures were approved by the Institutional Review Board of the University of Louisville. 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

² Despite collecting less data in filtered blocks than unfiltered blocks here, SCE magnitudes produced by filtered sentences in Stilp and Assgari (2017) have a comparable regression slope (0.07 stimulus steps per additional dB of filter gain, compared with 0.08) and intercept (0.15, compared with 0.13) to those observed in the present study. SCE magnitudes from the previous study fit extremely well on the regression derived from the present results depicted in Fig. 3 (correlation upon adding results from Stilp & Assgari, 2017: $r = .96$, $p < .0001$).

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 an average of 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 syllable sounded more like “ga” or “da.”

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 consonant continuum. Listeners were required to categorize consonants with at least 80% accuracy in order 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 consonants with 80% accuracy after the third practice session, they did not proceed to the main experiment.

The base design for a given experiment was to test four blocks of 160 trials apiece. Two of these blocks presented unfiltered contexts and the other two blocks presented filtered contexts with matching MSDs. In each unfiltered block, one sentence typically had a low- F_3 -biased MSD and the other sentence had a high- F_3 -biased MSD (see Table 1). Experiments 2, 3, and 4 employed this base design. Experiment 1 tested three unfiltered blocks each with 200 trials (10 repetitions of each unique context/target pairing instead of eight). Experiment 5 tested three unfiltered blocks and one filtered block, each consisting of 160 trials. Blocks were presented in counterbalanced orders across participants, and trials within each block were randomized. The experiment was self-paced, with 500 ms separating the listener’s response on a given trial from the beginning of the next trial. Participants had the opportunity to take short breaks between each block as needed. No feedback was provided. The total experimental session lasted approximately 1 hour.

Results

A performance criterion was implemented such that listeners were required to achieve at least 80% accuracy identifying consonant continuum endpoints in a given experimental block. If listeners exhibited difficulty categorizing consonant endpoints, that seriously compromised the interpretability of shifts in their consonant category boundaries due to SCEs. Seven blocks (out of 373 blocks total) were removed from further analysis: one listener failed three out of four blocks, and one listener failed all four blocks in their respective experiments.

Omnibus analysis

Results were analyzed using mixed-effect models in R (R Development Core Team, 2016) using the lme4 package (Bates et al., 2014). The model architecture matched that tested in Stilp and Assgari (2019). Responses were transformed using the binomial logit linking function. The dependent variable was modeled as binary (“ga” or “da” responses coded as 0 and 1, respectively). Fixed effects in the model included: Target (coded as a continuous variable from 1 to 10, then mean-centered), Spectral Peak (sum coded; high $F_3 = -0.5$, low $F_3 = +0.5$), Condition (sum coded; filtered = -0.5 , unfiltered = $+0.5$), and the absolute value of the MSD (whether naturally occurring [unfiltered sentences] or implemented via filtering [filtered sentences]; coded as a continuous variable, then mean-centered). All possible interactions between Spectral Peak, Condition, and MSD were included in the model. Random slopes were included for each main fixed effect and for the Spectral Peak \times Condition interaction, and a random intercept of listener was also included. All models were run using bobyqa optimization with a maximum of 800,000 iterations.

Results from this model are shown in Table 2. The model intercept was significant, indicating more “da” responses than “ga” responses to the consonant targets. The significant effect of Target predicts more “da” responses with each rightward step along the target continuum (toward higher F_3 onset frequencies and the /da/ endpoint), as expected. The significant positive effect of Spectral Peak predicts an increase in “da” responses when the region of greater spectral energy is changed from high F_3 (the level coded as -0.5) to low F_3 (the level coded as $+0.5$), consistent with the hypothesized direction of SCEs. The significant positive effect of Condition indicates that listeners responded “da” more often in unfiltered blocks than filtered blocks.

Table 2 Results from the mixed-effects logistic regression on listeners’ responses

	Estimate	SEM	Z	p
(Intercept)	0.511	0.093	5.509	3.6e-8
Target	1.715	0.045	37.889	<2e-16
Spectral Peak	0.838	0.082	10.174	<2e-16
Condition	1.498	0.117	12.772	<2e-16
MSD	-0.022	0.011	-1.914	0.056
Spectral Peak \times Condition	-1.239	0.138	-8.955	<2e-16
Spectral Peak \times MSD	0.074	0.012	6.360	2.0e-10
Condition \times MSD	0.022	0.014	1.655	0.098
Spectral Peak \times Condition \times MSD	-0.125	0.023	-5.474	4.4e-8

Note. See main text for description of factors and the model structure. SEM indicates the standard error of the mean.

Interactions including the Spectral Peak term (SCEs) are of particular importance. First, the significant negative Spectral Peak \times Condition interaction indicates that SCE magnitudes were smaller following unfiltered sentences than filtered sentences. This replicates previous studies that compared the efficacy of filtered and unfiltered contexts in eliciting SCEs (Lanning & Stilp, 2020; Stilp & Assgari, 2019). Second, the significant positive Spectral Peak \times MSD interaction indicates that SCE magnitudes increased linearly as MSDs increased. This replicates the similar significant interaction in Stilp and Assgari (2019). Finally, the three-way Spectral Peak \times Condition \times MSD interaction was significant, indicating that the rate (linear regression slope) at which SCE magnitudes increased at larger MSDs significantly differed across filtered and unfiltered conditions.

The omnibus analysis confirms the relationship between MSDs and SCEs in consonant categorization, but it is modeling the probability of responding “da” on a given trial. The primary phenomenon of interest is the SCE, which occurs across all trials in a given block. Therefore, as in Stilp and Assgari (2019), additional analyses are necessary in order to address this limitation.

Analysis of SCEs

SCEs were calculated for each block of each experiment in the same manner as Stilp and Assgari (2019). First, listeners’ responses in each block were fit with a mixed-effects logistic

regression with fixed effects of Target and Spectral Peak, random slopes for each of these fixed effects, and a random intercept for each listener. Model coefficients were used to quantify the magnitude of the SCE that occurred in that block following established procedures (Stilp & Assgari, 2017, 2018, 2019; Stilp et al., 2015). The 50% points were identified on the logistic regression fits to the aggregated responses following low- F_3 -emphasized contexts and high- F_3 -emphasized contexts. These 50% points were then converted into the stimulus step number that listeners would label as /da/ 50% of the time. Consonants targets were numbered from 1 to 10, so this stimulus number was interpolated as needed. The SCE magnitude was defined as the distance between these 50% points, measured in the number of stimulus steps.

In the ecological surveys (see Fig. 2) and the omnibus analysis reported above, sentences with positive MSDs possessed more energy in the low- F_3 frequency region and sentences with negative MSDs possessed more energy in the high- F_3 frequency region. Each block presented two context sentences (see Table 1), which produced an SCE of some magnitude (possibly even zero magnitude, a failure to bias categorization, or negative magnitude, biasing categorization in the opposite direction than that predicted by SCEs). To facilitate comparisons between MSDs and SCEs, MSD for an experimental block was calculated as the difference between the two context sentences’ MSDs divided by two. This calculation is preferable to a straight average, which

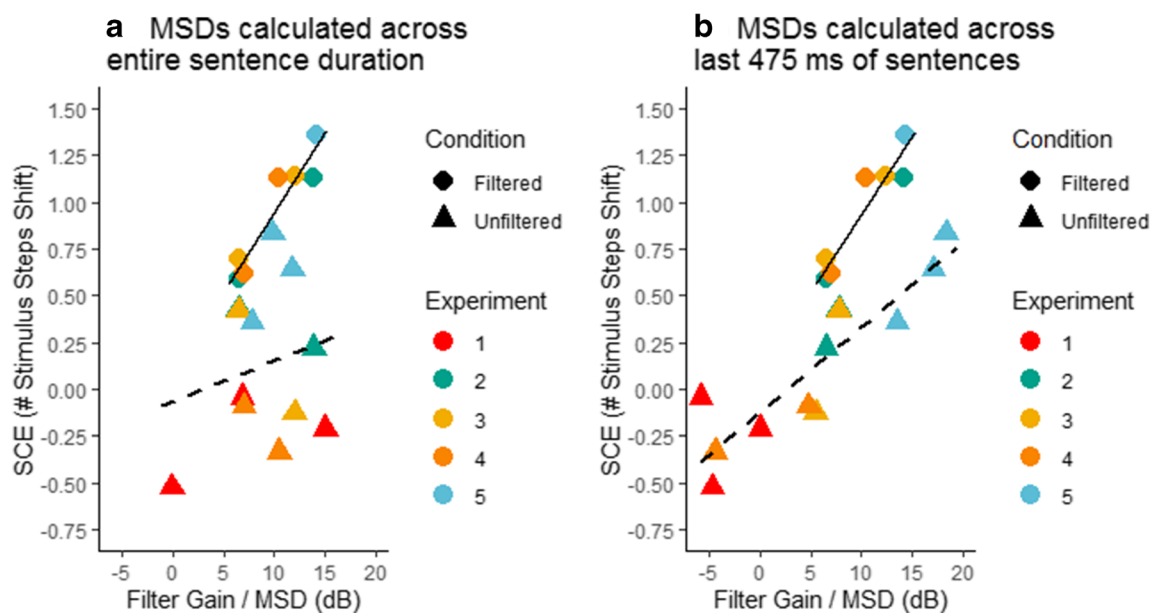


Fig. 3 Spectral contrast effect (SCE) magnitudes as calculated by mixed-effects models fit to each block of each experiment (see Table 3). Icons that share color indicate results from a given experiment contributed by a single participant group. SCEs produced by filtered context sentences (circles) are plotted as a function of filter gain. SCEs produced by unfiltered context sentences (triangles) are plotted as a function of the relative spectral prominence (MSD) calculated across the full duration

of each context sentence (a) or the last 475 ms of each context sentence (b). Ordinate values are identical across plots; abscissa values for unfiltered conditions differ depending on the timecourse of MSDs being analyzed. Solid lines represent linear regression fits to results in filtered conditions; dashed lines represent linear regression fits to results in unfiltered conditions. (Color figure online)

can produce faulty predictions when both MSDs are positive or negative (see Stilp & Assgari, 2019, for discussion).

SCEs in each block of each experiment are portrayed in Fig. 3 and listed in Table 3. Key differences in SCE magnitudes across filtered and unfiltered conditions observed in Stilp and Assgari (2019) were also observed here. SCEs following unfiltered contexts were smaller (mean = 0.13 stimulus steps) than those following filtered contexts (mean = 0.95 stimulus steps), and they were more variable (variance for unfiltered SCEs = 0.17, variance for filtered SCEs = 0.09). However, unlike Stilp and Assgari (2019), unfiltered sentences were surprisingly ineffective predictors of performance. SCEs were significantly correlated with sentence MSDs in filtered conditions ($r = .95, p < .01$) but not in unfiltered conditions ($r = .21, p = .51$). The slopes of linear regression fits to each Condition in these data sets were also discrepant (as indicated by the significant Spectral Peak \times Condition \times MSD interaction in the omnibus analysis). SCEs following filtered contexts (slope = 0.08 stimulus steps per addition dB of filter gain) grew at a markedly faster rate than SCEs following unfiltered contexts (slope = 0.02 steps/dB); slopes were comparable

across filtered and unfiltered conditions in Stilp and Assgari (2019).

Timecourse analysis

Unfiltered sentences were selected according to their MSDs, which reflect the long-term balance of energy across low-F₃ and high-F₃ frequency regions. Unlike Stilp and Assgari (2019), these MSDs did not systematically bias consonant categorization. A mixed-effects model comparable to the one reported above but limited to responses in unfiltered conditions produced only a trend toward a significant effect of Spectral Peak (i.e., SCEs; $p = .06$). The use of long-term averages to calculate stimulus statistics (MSDs) assumes that long-term averages are perceptually relevant to speech categorization; these results challenge that assumption. The current instantiation of MSDs does not consider more local spectrotemporal characteristics of context sentences; it is a distinct possibility that more local statistics may prove superior predictors of perceptual performance to more global (long-term average) ones.

Table 3 Results for each block of each experiment

Expt.	Block	MSD: Sentence	Unfiltered MSD: Last 475 ms	Unfiltered SCE	Icon	Filtered SCE	Icon
1	1	15.10	-0.01	-0.21	▲		
1	2	6.88	-5.83	-0.04	▲		
1	3	0.03	-4.71	-0.52	▲		
2	1	13.98	6.51	0.22	▲	1.13	●
2	2	6.49	7.77	0.43	▲	0.59	●
3	1	12.06	5.53	-0.12	▲	1.14	●
3	2	6.49	7.78	0.42	▲	0.70	●
4	1	10.51	-4.43	-0.33	▲	1.13	●
4	2	7.00	4.81	-0.09	▲	0.62	●
5	1	11.81	17.14	0.64	▲		
5	2	9.70	18.37	0.84	▲		
5	3	7.81	13.54	0.36	▲		
5	4	15.00				1.36	●

Note. MSDs in each block (half of the difference between the MSDs in each context sentence) were calculated across the entire sentence duration or across the last 475 ms of unfiltered context sentences. The resulting SCEs in each block are measured in the stimulus steps separating 50% points on logistic regressions fit to responses (see text for details). Icons match those illustrated in Fig. 3.

Following Stilp and Assgari (2019), the predictive power of MSDs across different timecourses was assessed. These analyses were akin to reverse correlation: MSDs were calculated for different durations of the unfiltered sentences, then these values were correlated with SCE magnitudes calculated at the group level (as plotted in Fig. 3 and listed in Table 3; these were fixed throughout the analyses). All unfiltered sentences were aligned at their offsets, making all stimuli uniform in terms of their temporal proximity to the onset of the consonant targets (separated by the 50-ms ISI). Next, an analysis window duration was specified (e.g., $t = 20$ ms). For a given context sentence, this duration was excised from the end of the sentence (in effect, its last t ms). To facilitate spectral analysis of short-duration signals, 1-ms linear onset and offset ramps were applied, and one second of silence was prepended and appended to the excerpt. The MSD of the excerpt was then calculated for both context sentences in a given experimental block; the block MSD value was calculated as the difference in MSD values divided by two as detailed above. After excising t -ms excerpts from all context sentences, block MSDs were correlated with SCE magnitudes and the correlation coefficient was saved. This process was repeated for all integer multiples of the analysis window duration (e.g., $t = 40$ ms, 60 ms, 80 ms, etc.) until it approached 1,539 ms,

which was the duration of the shortest sentence tested (“Spring Street is straight ahead” from Experiment 1). Exceeding this duration was undesirable because that would require removing behavioral results from the timecourse analysis.

The optimal window duration for analyses was not known *a priori*, so 10 different window durations were explored (5 to 50 ms, in 5-ms steps). The results of all analyses (highly convergent across different window durations) are superimposed in Fig. 4, with the correlation between MSDs and SCEs coefficient plotted as a function of temporal analysis window (the last t ms of every sentence). This profile has a very different shape than the profile observed for MSDs influencing vowel categorization (dashed line in Fig. 4, extending out to 1,089 ms, which was the shortest context sentence duration tested in that study). In Stilp and Assgari (2019), MSDs calculated on brief window durations did a very poor job of predicting SCE magnitudes. Here, brief window durations were excellent predictors of consonant categorization shifts. Correlation coefficient magnitudes peaked for windows spanning the last 130 ms to the last 500 ms of sentences ($r = .85$ – $.91$) at values much higher than those observed in the vowel categorization timecourse analysis (maximum $r = .60$ in Stilp & Assgari, 2019). The largest correlation magnitude in this temporal vicinity of sentences was 475 ms ($r = .90$); these MSDs were fantastic

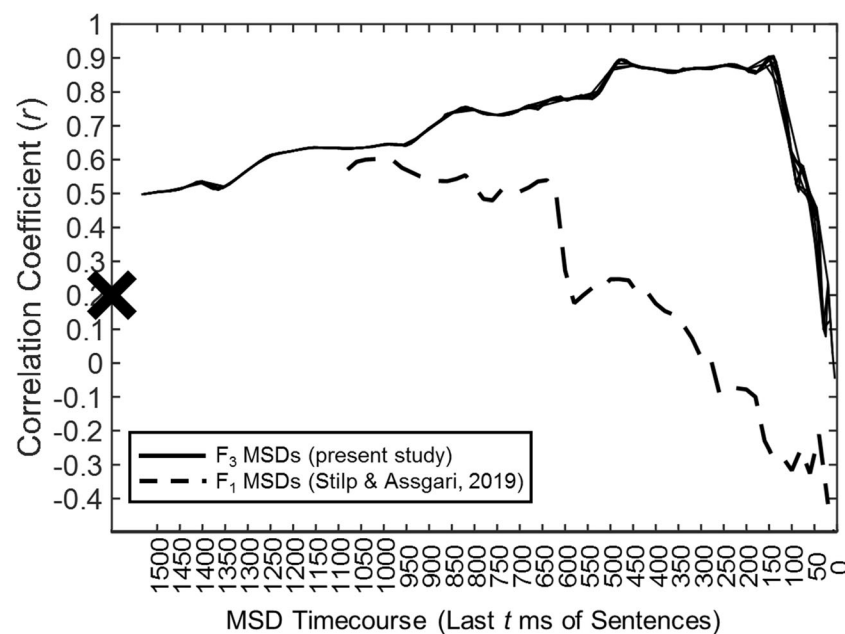


Fig. 4 Analyses of the timecourse of MSDs for predicting behavioral results (solid lines for the present results; dashed line for the timecourse analysis of different sentences tested in Stilp & Assgari, 2019). The abscissa depicts the duration of sentence spectra (relative to sentence offset) utilized for calculating MSDs. The ordinate depicts the correlation

coefficient for shorter-duration MSDs with SCEs in the unfiltered condition. Superimposed lines reflect different window durations utilized in analyses (5 to 50 ms at a time, in 5-ms steps). The black X on the ordinate represents the correlation coefficient when full-sentence MSDs were correlated with SCEs. See text for details

predictors of SCE magnitudes (see Fig. 3b).³ Beyond this point, as window duration increased, the magnitude of the correlation between MSDs and SCEs decreased. This is the reverse pattern of what was observed in Stilp and Assgari (2019): there, correlation magnitudes started small and increased as window duration increased; here, correlation magnitudes started large and decreased as window duration increased.

Aside from its peak, two other aspects of this correlation function merit discussion. Correlation coefficients are extremely modest for analysis windows spanning only the last 130 ms of sentences. In “She had your dark suit in greasy wash water all year” (Experiment 1, Block 2), the “-r” in “year” has a large low-F₃ MSD, but the 370 ms preceding it (“-ll yea-” in “all year”) has MSD values closer to 0 (all examples are illustrated in Supplementary Figures). Similarly, in “She argues with her sister” (Experiment 4, Block 2), MSD values are small and positive in the last 130 ms (“-r” in “sister”) but large and negative in the 370 ms prior to that (owing to friction noise in “siste-” of “sister”). Thus, MSDs at sentence offset (the last 130 ms) are fairly independent of those measured between the last 130–500 ms (the plateau in Fig. 4).

Correlation coefficient magnitudes decrease at analysis windows longer than the last 500 ms or so. This reflects corresponding abrupt changes in MSD measures in sentences. In “The hallway opens into a huge chamber” (Experiment 2, Block 1), affricates rich in high-F₃ energy (“-ge ch-” in “huge chamber”) convey negative MSDs immediately before the last 475 of the sentence (“-amber” in “chamber”) whose strong low-F₃ energy produces a positive MSD. A similar pattern occurs in “Shipbuilding is a most fascinating process” (Experiment 2, Block 2; Experiment 3, Block 2), where a half-second of positive MSD values (“-inating pro-” in “fascinating process”) precedes large negative MSDs in the last 475 ms owing to high-F₃ friction noise (“-cess” in “process”). This pattern also occurs in the opposite direction, as “His father will come home soon” (Experiment 4, Block 1) features a large negative MSD stemming from high-F₃ friction energy in the “s-” of “soon” before near-zero MSDs in the final 475 ms. But, it is important to note that this correlation function and its characteristics do not prescribe any specific temporal windows for context effects in speech perception. Unfiltered stimuli were selected without any regard to local temporal characteristics of MSDs; only their long-term

³ MSD values for the context sentence tested in filtered blocks differ depending on the timecourse being analyzed (MSD = 0.08 across the whole sentence, -10.63 across the last 475 ms owing to more high-F₃ friction energy in “crucial”). But this analysis utilizes the difference in MSDs across low-F₃-amplified and high-F₃-amplified renditions of the context sentence. This difference is defined by filter gains utilized and not by initial MSD value. For example, for +10 dB filter gains, the calculation becomes [(MSD + 10) - (MSD - 10)]/2 = 10. As a result, MSD values for filtered blocks are nearly identical across Figs. 3a and 3b.

(sentence-length) properties were considered. Future research using stimuli with more carefully controlled short-term MSDs (analogous to the generation of pure tone sequences with different local statistics in Holt, 2006) may shed more light on context effects in speech categorization at different temporal windows.

Given the results of the timecourse analysis, a second mixed-effects model was constructed using the same architecture as the model described at the beginning of the Results section. The key difference was that MSDs in this second model were calculated across the last 475 ms of sentences rather than their entire duration. Results from this analysis are depicted in Table 4. Key results pertaining to SCEs were replicated: SCEs occurred (positive main effect of Spectral Peak), their magnitudes increased as MSDs increased (positive interaction between Spectral Peak and MSD), and the rate of this increase was shallower for unfiltered sentences than filtered sentences (negative three-way interaction between Spectral Peak, Condition, and MSD). When this model is restricted to analyze responses in Unfiltered conditions only, the main effect of Spectral Peak is now significant ($z = 5.51, p < 4e-8$) as is its interaction with MSDs ($z = 8.64, p < 2e-16$). Neither of these results are statistically significant when MSDs were calculated across the entire durations of context sentences.

The predictive power of MSDs in the last 475 of context sentences (here noted $MSD_{last475ms}$) and the poor predictive power of MSDs calculated across entire sentences (here noted $MSD_{sentence}$) for SCE magnitudes is evident at the item level. In Experiment 1, Block 1, the SCE was in the opposite direction of what was expected based on $MSD_{sentence} \cdot MSD_{last475ms}$ were equivalent, which would not be expected to differentially influence responses (i.e., extinguish the SCE). The negative SCE might reflect the last 475ms of “We’re here to transact business” containing a stretch of negative

Table 4 Results from the mixed-effects logistic regression on listeners’ responses

	Estimate	SEM	Z	p
(Intercept)	1.031	0.116	8.919	<2e-16
Target	1.707	0.044	38.560	<2e-16
Spectral Peak	0.377	0.149	2.530	0.011
Condition	0.710	0.175	4.066	4.8e-5
MSD	-0.017	0.007	-2.379	0.017
Spectral Peak × Condition	-0.017	0.300	-0.058	0.954
Spectral Peak × MSD	0.101	0.010	9.646	<2e-16
Condition × MSD	0.038	0.011	3.468	0.001
Spectral Peak × Condition × MSD	-0.102	0.021	-4.896	9.8e-7

Note. Unlike in Table 2, MSDs were calculated across the last 475 ms of context sentences. See main text for description of factors and the model structure. SEM indicates the standard error of the mean.

instantaneous MSDs due to frication energy in “business,” whereas the last 475 ms of “Her lips, moist and parted, spoke his name” contains a stretch of lower- F_3 formant energy in “name” (see [Supplementary Figures](#)). In Block 2, the SCE was extinguished altogether. “She had your dark suit in greasy wash water all year” had an extremely modest $MSD_{last475ms}$. While “Spring Street is straight ahead” had a moderate negative $MSD_{sentence}$ but a strongly positive $MSD_{last475ms}$ due to lower- F_3 formant energy in “ahead,” this was not enough to significantly shift responses. In Block 3, a negative SCE occurred despite both sentences having near-zero $MSD_{sentence}$. This is likely due to higher- F_3 frication energy in “-ance” the end of “Critical equipment needs proper maintenance” (producing a strongly negative $MSD_{last475ms}$) promoting “ga” responses to a greater degree than did the very modest negative $MSD_{last475ms}$ at the end of “The blue jay flew over the high building.”

Both blocks of Experiment 2 produced SCEs in the predicted directions. In Block 1, F_2 and F_3 peaks in the last word of “Don’t ask me to carry an oily rag like that” produced large positive instantaneous MSDs during the last 475 ms. This sentence was likely more effective in promoting “da” responses than “The hallway opens into a huge chamber,” where the instantaneous MSD values (and resulting $MSD_{last475ms}$) were more modest. A similar situation occurred in Block 2. Lower- F_3 and higher- F_3 energy in “-ces” of “The paper boy bought two apples and three ices” was relatively well balanced (resulting in a small $MSD_{last475ms}$), but the frication energy in “-cess” of “Shipbuilding is a most fascinating process” produced relatively large high- F_3 peaks. This resulted in a large negative $MSD_{last475ms}$ for this sentence, which was sufficient to shift responses in the predicted directions.

Experiments 3 and 4 were similar to Experiment 2 in that in each block, one context sentence had a large $MSD_{last475ms}$ and the other sentence had a much smaller $MSD_{last475ms}$, which might have been enough to produce an SCE. Block 1 of Experiment 3 tested the same token of “Don’t ask me to carry an oily rag like that” (with its large positive $MSD_{last475ms}$) as Block 1 of Experiment 2. This sentence was expected to be more effective in promoting “da” responses than “Those who teach values first abolish cheating,” where the raising of F_3 for the /t/ in “-eating” flipped instantaneous MSDs from positive to strongly negative, lessening the degree to which the $MSD_{last475ms}$ is positive. However, a small negative SCE was observed instead. Block 2 presented the same context sentences as those presented in Experiment 2, Block 2, and the SCE was replicated almost exactly. In Block 1 of Experiment 4, “The family bought a house” had a positive $MSD_{sentence}$ but a strong negative $MSD_{last475ms}$ due to frication energy in “house.” The last 475 ms of “His father will come home soon” exhibited small instantaneous MSD values owing to relatively balanced energy across low- F_3 and high- F_3 regions on “-oon” due to F_2 falling below the

1700-Hz lower frequency cutoff of the low- F_3 region and F_3 sitting on the border shared by low- F_3 and high- F_3 regions. Collectively, this produced more “da” responses to “The family bought a house,” resulting in a negative SCE. In Block 2, the higher- F_3 frication energy in “-ister” at the end of “She argues with her sister” (negative $MSD_{last475ms}$) should have promoted “ga” responses more effectively than “They painted the wall white,” whose small $MSD_{last475ms}$ is due to F_2 values in “white” starting out below and then transitioning into the low- F_3 frequency region. However, the resulting SCE was of small magnitude. Finally, Experiment 5 was straightforward. Sentences with positive $MSD_{sentence}$ also had positive $MSD_{last475ms}$ (lower- F_3 formant peaks in “put in” of “Often you’ll get back more than you put in” [Block 1], in “that” of “Don’t ask me to carry an oily rag like that” [Blocks 2 and 3]). Similarly, sentences with negative $MSD_{sentence}$ also had negative $MSD_{last475ms}$ (higher- F_3 frication energy in “-eese” of “Rob sat by the pond and sketched the stray geese” [Block 1], in “-ise” of “The prowler wore a ski mask for disguise” [Block 2], in “-uice” of “When peeling an orange, it is hard not to spray juice” [Block 3]). By having large $MSD_{last475ms}$ in the expected directions for both context sentences in each block, moderate-to-large SCEs were observed.

Discussion

The Efficient Coding Hypothesis (Attneave, 1954; Barlow, 1961) has long been a productive perspective for studying sensation and perception in vision (see Introduction). While nascent compared with studies in vision, applications of efficient coding to auditory perception have been fruitful, specifically for speech perception (for reviews, see Gervain & Geffen, 2019; Kluender et al., 2013; Kluender et al., 2019). The present approach further supports efficient coding perspectives of speech perception by demonstrating sensitivity to natural signal statistics in context spectra. Context sentences were selected and presented based on their Mean Spectral Differences (MSDs), the inherent balance of acoustic energy across two frequency regions. Spectra toward the ends of these sentences (particularly the last 475 ms) produced and predicted context effects that shaped listeners’ consonant categorization (SCEs). This further supports the perceptual significance of this metric and bolsters efforts to link the statistical structure of the speech signal to its perception.

The present study and Stilp and Assgari (2019) both measured the efficacy of filtered and unfiltered context sentences to produce SCEs that influenced speech sound categorization. While the previous study examined vowel categorization (/i/-/ε/) and its relevant frequency regions (<850 Hz), the present study examined consonant categorization (/g/-/d/) and its relevant frequency regions (1700–3700 Hz). In both studies, unfiltered SCE magnitudes were smaller and more variable than

filtered SCE magnitudes (see Fig. 3). The differences in results across studies, however, are more striking than the similarities. In Stilp and Assgari (2019), MSDs based on long-term average statistics of unfiltered sentences were significantly correlated with the resulting SCE magnitudes; here they were not correlated at all (see Fig. 3a). Regression slopes for predicting SCE magnitudes from MSDs were highly similar across filtered and unfiltered sentences in the previous study; here regression slopes markedly differed. In short, full-sentence MSDs were considerably less effective in biasing consonant categorization than biasing vowel categorization.

Analyses of MSD timescales offered insight as to why this was the case. While sentences were selected and presented as context stimuli owing to their long-term (i.e., full-sentence) MSDs, variation in phonemic content rapidly changed specific MSD values throughout the sentence (see [Supplementary Figures](#)). While full-sentence MSDs were poor predictors of shifts in consonant categorization, MSDs toward the ends of the context sentences were excellent predictors (see Figs. 3b and 4). These MSDs were much stronger predictors of consonant SCEs ($r \approx .90$) than those predicting vowel SCEs (maximum $r \approx .60$), which were strongest at longer timecourses and weaker for shorter timecourses. Why do MSDs exhibit such different timecourses in affecting vowel categorization and consonant categorization? Several possibilities exist that are not mutually exclusive. First, context sentences were drawn from the same databases across studies (TIMIT, HINT), but the sentences themselves differed, so item-specific factors cannot be ruled out. Second, F_1 MSDs were calculated in lower-frequency regions of the speech spectrum (<850 Hz), which have higher amplitudes than the higher-frequency regions in which F_3 MSDs were calculated (1700–3700 Hz; see long-term average spectra in Fig. 2). Thus, while specific MSD values may be similar across studies (particularly larger MSD values in the distribution tails in Fig. 2), those occurring in F_1 regions have higher overall amplitude than those occurring in F_3 regions. Third, MSDs in different frequency regions are driven by different events in the speech signal. Both low- F_1 and high- F_1 MSDs below 850 Hz are driven primarily by formant (F_1) peaks. Low- F_3 MSDs were also driven primarily by formant peaks (F_2 and F_3 peaks in the 1700–2700 Hz region), but high- F_3 MSDs were driven primarily by frication noise (see spectrograms in [Supplementary Figures](#) for examples). The durations of these events in the speech signal vary, as vowel durations often exceed those for frication noise (e.g., House, 1961; Jongman et al., 2000). As a consequence, instantaneous MSD values are generally less variable in F_1 regions than in F_3 regions.⁴ These factors bear

directly on the neural adaptation processes proposed to underlie SCEs (see Stilp, 2020a, 2020b, for discussion). In order to produce sufficient neural adaptation to result in an SCE, spectral peaks in F_3 frequency regions might need to occur closer to context sentence offset (due to being comparatively lower-amplitude, shorter-duration, and higher-variance) whereas spectral peaks in F_1 frequency regions could occur earlier in context sentences (due to being higher-amplitude, longer-duration, and lower-variance). Neural responses to speech sounds following acoustic (sentence-length) context are needed to confirm this mechanistic interpretation of MSD timecourses and patterns of SCEs across studies.

While appealing, these interpretations are accompanied by three caveats. First, the present experiments tested 24 unfiltered context sentences (21 unique) and Stilp and Assgari (2019) tested 32 unfiltered context sentences (17 unique); this is not expected to be fully representative of all American English. Second, these different timecourses were derived in a post hoc fashion. Future research would be well-served by targeting specific timecourses in different frequency regions *a priori*. A final but nontrivial difference across studies lies in the target stimuli themselves. Stilp and Assgari (2019) measured categorization of vowels differentiated principally by F_1 frequency (/i/-/ε/). This spectral feature was relatively high-amplitude, highly prominent in the spectrum, and endured throughout the duration of the target stimulus (246 ms). This contrasts sharply with characteristics of the /g/-/d/ target consonants tested here, which are principally differentiated by the onset frequency of the F_3 transition. The entire formant transition duration is only 63 ms, of which the first few tens of milliseconds are most crucial for differentiating /g/ from /d/. As discussed above, the F_3 frequency region is considerably lower in amplitude than the F_1 frequency region. Differences in target stimuli may also contribute to discrepant patterns of results across studies, not just differences in MSD characteristics.

Acoustic context effects in speech perception can occur on different timecourses (see Stilp, 2020a, for review). Previous studies have sought to distinguish the relative influences of proximal context (i.e., temporally adjacent to the target stimulus) and distal context (i.e., temporally nonadjacent to the target stimulus) on perception of a target speech sound. Holt (2006) presented contexts comprising three successive 700-ms pure tone sequences with different local statistics (mean frequencies of 1800, 2300, or 2800 Hz in counterbalanced orders). Statistical analyses did not find consistent effects of these contexts on categorization of /ga/-/da/ targets, leading Holt (2006) to conclude that the proximal context (immediately preceding the target consonant) was less perceptually salient than the global context (grand mean of 2300 Hz across the entire 2,100-ms sequence). But, direct comparison of the present results to those from Holt (2006) are difficult. Both studies explored context spectral peaks in similar frequency regions and perception of /ga/-/da/ targets, but stimulus

⁴ This point is supported by calculating the variance on instantaneous MSD values depicted in the [Supplementary Figures](#). Variances were also calculated for instantaneous MSDs in stimuli presented in Stilp and Assgari (2019). Analyses were restricted to the last 1,000 ms or the last 500 ms to account for differing sentence durations across stimuli.

statistics were tightly controlled using pure tone contexts in Holt's study whereas they varied more naturally in sentence contexts here. While the literature on proximal versus distal spectral context effects is sparse, a host of studies have examined this question for temporal context effects (i.e., speaking rate normalization). Across studies and methodologies, the speaking rate of proximal context exerts a larger influence on perception of target speech than does the rate of distal context (Heffner et al., 2013; Kidd, 1989; Reinisch et al., 2011; Summerfield, 1981). In the present experiments, proximal and distal contexts were not explicitly pit against each other in this fashion; stimuli were selected based on their MSDs calculated across the entire duration of the sentence. Short-term MSDs in unfiltered sentences (i.e., those calculated across the last 475 ms of sentences) were not obligated to resemble long-term MSDs; they were not even correlated with each other ($r = .27, p = .40$). Yet, proximal MSDs were superior predictors of performance compared to full-sentence MSDs (see Fig. 3). Future research testing contexts whose proximal and distal statistical structure make different predictions for speech sound categorization would be highly illuminating. Given the consistent patterns of results observed in the temporal domain, the statistics of proximal contexts would be expected to bear greater influence on speech perception than the statistics of distal contexts.

Perceptual sensitivity to signal statistics on variable timescales is not limited to spectral context effects. McDermott and colleagues (McDermott et al., 2013; McDermott & Simoncelli, 2011; McWalter & McDermott, 2018) synthesized environmental sounds based on their statistical characteristics in the frequency and modulation domains. These statistics were imposed on random noise samples to create sound textures *de novo*, which listeners recognized nearly as well as their natural counterparts (McDermott & Simoncelli, 2011). Critically, perceptual processes average these statistics across time, as discrimination of different textures improved with longer durations (as statistics diverged to different long-term averages) while discrimination of different exemplars of the same texture worsened at longer durations (as statistics converged to the same long-term average; McDermott et al., 2013). Perception accommodates changes in statistics within a given texture, as the temporal averaging process lengthens for highly variable statistics but shortens for more consistent statistics (McWalter & McDermott, 2018). Aggregation of stimulus statistics over minutes of exposure can directly influence lower-level and higher-level perception. Passive exposure to a few minutes of nonsense speech is sufficient for infants to extract transitional probabilities between syllables (Saffran et al., 1996); this finding was seminal to the field exploring “statistical learning” in language development (for a recent review, see Saffran & Kirkham, 2018). After a few minutes of passive exposure or active testing, listeners extract patterns of covariance between acoustic properties in novel sounds that modulates their discriminability (Stilp & Kluender, 2012, 2016; Stilp, Rogers, &

Kluender, 2010). After a few minutes of listening or testing, listeners alter speech sound categorization in response to the probability density functions of stimulus presentation (e.g., Maye et al., 2008; McMurray et al., 2009) or changes in the variance of these distributions (Theodore & Monto, 2019). Other statistical properties may be aggregated over hours, days, or longer, such as those that promote sensitivity to the sound contrasts in one's native language at the expense of other seldom-heard languages (e.g., Werker & Tees, 1984). These and other examples exemplify perception maintaining sensitivity to statistics on different time-scales in order to make processing more efficient.

In that vein, it might not be surprising that speech perception is sensitive to stimulus statistics in preceding context. Here, consonant categorization was shaped by the natural statistical structure of context sentences (MSDs). This relationship was also observed in Stilp and Assgari (2019), where the MSDs in context sentences shaped vowel categorization. This sensitivity to natural stimulus statistics in unfiltered context sentences further supports the pervasiveness of acoustic context effects in everyday speech perception (Stilp, 2020a; Stilp & Assgari, 2019). Additionally, these results provide yet further evidence of efficient coding of structure in the sensory environment.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13414-021-02310-4>.

Acknowledgements The authors wish to thank Ashley Batliner, Alexandra Beason, Carly Newman, and Madison Rhyne for assistance with data collection.

Funding No funding was received for conducting this study.

Data availability All data and analysis scripts are available (<https://osf.io/95xpf/>).

Declarations

Ethics approval Approval was obtained from the ethics committee of the University of Louisville. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication Not applicable; no identifying information about any participants is included in this article.

Conflicts of interest/competing interests The authors have no relevant financial or nonfinancial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

- Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3), 183–193. <https://doi.org/10.1037/h0054663>

- Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 53–85). MIT Press.
- 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]. <http://cran.r-project.org/package=lme4>
- Bell, A. J., & Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. *Vision Research*, 37(23), 3327–3338. [https://doi.org/10.1016/S0042-6989\(97\)00121-1](https://doi.org/10.1016/S0042-6989(97)00121-1)
- Burge, J., Fowlkes, C. C., & Banks, M. S. (2010). Natural-scene statistics predict how the figure–ground cue of convexity affects human depth perception. *The Journal of Neuroscience*, 30(21), 7269–7280. <https://doi.org/10.1523/JNEUROSCI.5551-09.2010>
- Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804–809. <https://doi.org/10.1016/j.cognition.2008.04.004>
- Dean, I., Harper, N. S., & McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. *Nature Reviews Neuroscience*, 8(12), 1684–1689. <https://doi.org/10.1038/nrn1541>
- Dean, I., Robinson, B. L., Harper, N. S., & McAlpine, D. (2008). Rapid neural adaptation to sound level statistics. *Journal of Neuroscience*, 28(25), 6430–6438. <https://doi.org/10.1523/JNEUROSCI.0470-08.2008>
- Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12), 2379–2394. <https://doi.org/10.1364/JOSAA.4.002379>
- Field, D. J. (1994). What is the goal of sensory coding. *Neural Computation*, 6(4), 559–601. <https://doi.org/10.1162/neco.1994.6.4.559>
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., & Dahlgren, N. (1990). *DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM* (NIST Order No. PB91-505065). National Institute of Standards and Technology.
- Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Reviews in Psychology*, 59, 167–192. <https://doi.org/10.1146/annurev.psych.58.110405.085632>
- Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, 41(6), 711–724. [https://doi.org/10.1016/S0042-6989\(00\)00277-7](https://doi.org/10.1016/S0042-6989(00)00277-7)
- Gervain, J., & Geffen, M. N. (2019). Efficient neural coding in auditory and speech perception. *Trends in Neurosciences*, 42(1), 56–65. <https://doi.org/10.1016/j.tins.2018.09.004>
- Heffner, C. C., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2013). When cues combine: How distal and proximal acoustic cues are integrated in word segmentation. *Language and Cognitive Processes*, 28(9), 1275–1302.
- 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. <https://doi.org/10.1121/1.2354071>
- Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, 119(5), 3059–3071. <https://doi.org/10.1121/1.2188377>
- House, A. S. (1961). On vowel duration in English. *Journal of the Acoustical Society of America*, 33(9), 1174–1178. <https://doi.org/10.1121/1.1908941>
- Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, 108(3), 1252–1263. <https://doi.org/10.1121/1.1288413>
- Kidd, G. R. (1989). Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4), 736–748. <https://doi.org/10.1037/0096-1523.15.4.736>
- Kingston, J., Kawahara, S., Chambless, D., Key, M., Mash, D., & Watsky, S. (2014). Context effects as auditory contrast. *Attention, Perception, & Psychophysics*, 76, 1437–1464. <https://doi.org/10.3758/s13414-013-0593-z>
- Kluender, K. R., Stilp, C. E., & Kieffe, M. (2013). Perception of vowel sounds within a biologically realistic model of efficient coding. In G. S. Morrison & P. F. Assmann (Eds.), *Vowel inherent spectral change* (pp. 117–151). Springer.
- Kluender, K. R., Stilp, C. E., & Llanos, F. (2019). Longstanding problems in speech perception dissolve within an information-theoretic perspective. *Attention, Perception, & Psychophysics*, 81(4), 861–883. <https://doi.org/10.3758/s13414-019-01702-x>
- Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, 29(1), 98–104. <https://doi.org/10.1121/1.1908694>
- Lanning, J. M., & Stilp, C. E. (2020). Natural music context biases musical instrument categorization. *Attention, Perception, and Psychophysics*, 82, 2209–2214. <https://doi.org/10.3758/s13414-020-01980-w>
- Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, 5(4), 356–363. <https://doi.org/10.1038/nrn831>
- Lu, K., Liu, W., Dutta, K., Zan, P., Fritz, J. B., & Shamma, S. A. (2019). Adaptive efficient coding of correlated acoustic properties. *The Journal of Neuroscience*, 39(44), 8664–8678. <https://doi.org/10.1523/JNEUROSCI.0141-19.2019>
- Malmierca, M. S., Cristaudo, S., Perez-Gonzalez, D., & Covey, E. (2009). Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *The Journal of Neuroscience*, 29(17), 5483–5493. <https://doi.org/10.1523/JNEUROSCI.4153-08.2009>
- Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, 11(1), 122–134. <https://doi.org/10.1111/j.1467-7687.2007.00653.x>
- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101–B111. [https://doi.org/10.1016/S0010-0277\(01\)00157-3](https://doi.org/10.1016/S0010-0277(01)00157-3)
- McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, 16(4), 493–498. <https://doi.org/10.1038/nrn.3347>
- McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5), 926–940. <https://doi.org/10.1016/j.neuron.2011.06.032>
- McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, 12(3), 369–378. <https://doi.org/10.1111/j.1467-7687.2009.00822.x>
- McWalter, R., & McDermott, J. H. (2018). Adaptive and selective time averaging of auditory scenes. *Current Biology*, 28(9), 1405–1418.e10. <https://doi.org/10.1016/j.cub.2018.03.049>
- Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, 109(3), 1181–1196. <https://doi.org/10.1016/j.cub.2018.03.049>
- Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, 95(2), 1085–1099. <https://doi.org/10.1121/1.408469>
- Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network*, 7(2), 333–339. https://doi.org/10.1088/0954-898X_7_2_014

- R Development Core Team. (2016). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing. <http://www.r-project.org/>
- Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3), 978–996. <https://doi.org/10.1037/a0021923>
- Ruderman, D. L., Cronin, T. W., & Chiao, C. C. (1998). Statistics of cone responses to natural images: Implications for visual coding. *Journal of the Optical Society of America*, 15(8), 2036–2045. <https://doi.org/10.1364/JOSAA.15.002036>
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>
- Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, 69, 181–203. <https://doi.org/10.1146/annurev-psych-122216-011805>
- Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13(2), 144–149. [https://doi.org/10.1016/S0959-4388\(03\)00047-3](https://doi.org/10.1016/S0959-4388(03)00047-3)
- Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Reviews in Neuroscience*, 24, 1193–1216. <https://doi.org/10.1146/annurev.neuro.24.1.1193>
- Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L., Oakes, T., & Cook, S. (2012). Development and validation of the AzBio sentence lists. *Ear and Hearing*, 33(1), 112–117. <https://doi.org/10.1097/AUD.0b013e31822c2549>
- 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. <https://doi.org/10.1016/j.specom.2011.02.007>
- Stilp, C. E. (2020a). Acoustic context effects in speech perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(1/2), 1–18. <https://doi.org/10.1002/wics.1517>
- Stilp, C. E. (2020b). Evaluating peripheral versus central contributions to spectral context effects in speech perception. *Hearing Research*, 392, 1–12. <https://doi.org/10.1016/j.heares.2020.107983>
- 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. <https://doi.org/10.3758/APP.72.2.470>
- Stilp, C. E., Anderson, P. W., & Winn, M. B. (2015). Predicting contrast effects following reliable spectral properties in speech perception. *The Journal of the Acoustical Society of America*, 137(6), 3466–3476. <https://doi.org/10.1121/1.4921600>
- 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. <https://doi.org/10.1121/1.4974769>
- 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. <https://doi.org/10.3758/s13414-018-1488-9>
- Stilp, C. E., & Assgari, A. A. (2019). Natural speech statistics shift phoneme categorization. *Attention, Perception, & Psychophysics*, 81(6), 2037–2052. <https://doi.org/10.3758/s13414-018-01659-3>
- Stilp, C. E., & Kluender, K. R. (2011). Non-isomorphism in efficient coding of complex sound properties. *Journal of the Acoustical Society of America*, 130(5), EL352–EL357. <https://doi.org/10.1121/1.3647264>
- Stilp, C. E., & Kluender, K. R. (2012). Efficient coding and statistically optimal weighting of covariance among acoustic attributes in novel sounds. *PLOS ONE*, 7(1), Article e30845. <https://doi.org/10.1371/journal.pone.0030845>
- Stilp, C. E., & Kluender, K. R. (2016). Stimulus statistics change sounds from near-indiscernible to hyperdiscriminable. *PLOS One*, 11(8), Article e0161001. <https://doi.org/10.1371/journal.pone.0161001>
- Stilp, C. E., & Lewicki, M. S. (2014). Statistical structure of speech sound classes is congruent with cochlear nucleus response properties. In *Proceedings of Meetings on Acoustics* (Vol. 20). <https://doi.org/10.1121/1.4865250>
- Stilp, C. E., Rogers, T. T., & Kluender, K. R. (2010). Rapid efficient coding of correlated complex acoustic properties. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), 21914–21919. <https://doi.org/10.1073/pnas.1009020107>
- Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 7(5), 1074–1095. <https://doi.org/10.1037/0096-1523.7.5.1074>
- Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker’s phonetic distributions. *Psychonomic Bulletin & Review*, 26(3), 985–992. <https://doi.org/10.3758/s13423-018-1551-5>
- Tkačik, G., Prentice, J. S., Victor, J. D., & Balasubramanian, V. (2010). Local statistics in natural scenes predict the saliency of synthetic textures. *Proceedings of the National Academy of Sciences of the United States of America*, 107(42), 18149–18154. <https://doi.org/10.1073/pnas.0914916107>
- Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34(3), 434–464. <https://doi.org/10.1111/j.1551-6709.2009.01077.x>
- Ulanovsky, N., Las, L., & Nelken, I. (2003). Processing of low-probability sounds by cortical neurons. *Nature Neuroscience*, 6(4), 391–398. <https://doi.org/10.1038/nn1032>
- van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Academy B: Biological Sciences*, 265(1394), 359–366. <https://doi.org/10.1098/rspb.1998.0303>
- Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456), 1273–1276. <https://doi.org/10.1126/science.287.5456.1273>
- Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1), 49–63. [https://doi.org/10.1016/S0163-6383\(84\)80022-3](https://doi.org/10.1016/S0163-6383(84)80022-3)

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