Attention, Perception, & Psychophysics (2019) 81:2037–2052
https://doi.org/10.3758/s13414-018-01659-3

Natural speech statistics shift phoneme categorization

Christian E. Stilp[1] · Ashley A. Assgari[1]

Published online: 18 March 2019

Abstract
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., more energy at lower first formant [F_1] frequencies) differs from that of a later (target) sound (e.g., vowel with intermediate F_1), the auditory system magnifies this difference, biasing target categorization (e.g., towards higher-F_1 /ɛ/). Historically, these studies used filters to force context sounds to possess desired spectral compositions. This approach is agnostic to the natural signal statistics of speech (inherent spectral compositions without any additional manipulations). The auditory system is thought to be attuned to such stimulus statistics, but this has gone untested. Here, vowel categorization was measured following unfiltered (already possessing the desired spectral composition) or filtered sentences (to match spectral characteristics of unfiltered sentences). Vowel categorization was biased in both cases, with larger biases as the spectral prominences in context sentences increased. This confirms sensitivity to natural signal statistics, extending spectral context effects in speech perception to more naturalistic listening conditions. Importantly, categorization biases were smaller and more variable following unfiltered sentences, raising important questions about how faithfully experiments using filtered contexts model everyday speech perception.

Keywords Speech perception · Perceptual categorization and identification · Hearing

Introduction

Objects and events in the environment are highly structured across space, across time, and in their physical compositions. According to the Efficient Coding Hypothesis (Attneave, 1954; Barlow, 1961), sensory systems detect and exploit these regularities in order to make processing efficient. Since the formulation of this hypothesis, considerable research has shown how neural processing and/or perceptual responses capitalize on systematicity in sensory stimuli, particularly in vision. A wide range of studies has documented the inherent statistical structure of natural images (i.e., statistical characteristics that effectively describe the composition of natural stimuli – e.g., Field, 1987; Ruderman, 1994; Olshausen & Field, 1996; Bell & Sejnowski, 1997; Simoncelli, 2003; Torralba & Oliva, 2003). This structure has been linked to neural response properties in the visual system (Field, 1987; Ruderman, Cronin, & Chiao, 1998; van Hateren & van der Schaaf, 1998; Simoncelli & Olshausen, 2001) and theories of sensory coding of natural stimuli (Field, 1994; Vinje & Gallant, 2000). Finally, these statistical regularities inform observer performance in visual perception tasks (Geisler, Perry, Super, & Gallogly, 2001; Geisler, 2008; Burge, Fowlkes, & Banks, 2010; Tkačik, Prentice, Victor, & Balasubramanian, 2010).

Given its abundant success, one might expect this efficient coding approach to be productive for studying other classes of natural stimuli, including speech. The speech signal is highly redundant, often displaying multiple acoustic properties that are correlated with each other at any given time (e.g., Lisker, 1986; Repp, 1982; McMurray & Jongman, 2011; Kluender, Stilp, & Kiefte, 2013). Yet, thus far, most investigations elected to focus

Significance statement Many studies have shown that speech sound categorization is influenced by acoustic properties of earlier sounds. Historically, these earlier sounds were carefully manipulated versions of a single sentence. However, this approach is agnostic to the inherent statistical structure of the speech signal. Here we quantify natural signal statistics for speech spectra and show that sentences possessing these regularities bias speech sound categorization. Importantly, this influence is smaller and more variable than it is for carefully manipulated stimuli. These results validate a natural signal statistics approach to speech perception, raising important questions about balancing experimental control with ecological validity.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13414-018-01659-3) contains supplementary material, which is available to authorized users.

✉ 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

not on the statistical structure of the speech signal, but on the statistics of stimulus presentation (e.g., probability density functions of stimulus presentation: Maye, Werker, & Gerken, 2002; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Toscano & McMurray, 2010; transitional probabilities between sounds: McQueen, 1998; Vitevitch & Luce, 1999; transitional probabilities between syllables: Saffran, Aslin, & Newport, 1996). Methods for analyzing the statistical structure of natural images have been directly extended to speech, but these approaches have faced considerable limitations. For example, Lewicki (2002) used independent components analysis to derive digital filters that optimally encoded the statistical structure of brief speech segments. The center frequencies and sharpness of these digital filters closely resembled those of auditory nerve fibers in the mammalian auditory system, suggesting that the statistical structure of speech was processed efficiently. However, questions abound regarding the sharpness of human auditory nerve tuning and comparisons between broadband speech stimuli and narrowband (pure tone) stimuli used to measure auditory nerve tuning (see Sayles & Winter, 2010; Stilp & Lewicki, 2014; Stilp & Assgari, 2015, for discussions). Also, it is unclear how this finding informs speech perception. So far, the clearest demonstrations of perceptual sensitivity to stimulus statistical structure utilized nonspeech sounds, including auditory textures (McDermott & Simoncelli, 2011; McDermott, Schemitsch, & Simoncelli, 2013; McWalter & McDermott, 2018) and heavily edited musical instrument sounds (Stilp, Rogers, & Kluender, 2010; Stilp & Kluender, 2011, 2012, 2016). These findings with nonspeech stimuli might not necessarily inform the natural acoustic ecology or speech and/or everyday speech perception.

Perhaps the most formidable obstacle to studying the efficient coding of speech is its extreme acoustic variability. Historically, measurements of natural signal statistics in speech have been conducted on individual phonemes. Phonemes famously lack any necessary and sufficient stimulus properties to cue their identity (the “lack of invariance”; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). However, while speech exhibits extreme acoustic variability across short (phoneme-level) timescales, this variability at least partially cedes to stable longer-term signal properties across longer timescales. For example, the long-term average spectrum of speech displays higher energy at lower frequencies (below 1 kHz) and a steady roll-off in energy at higher frequencies (Voss & Clarke, 1975; Byrne et al., 1994; ANSI, 1997). The amplitude envelope of medium-rate speech modulates most often from 2–5 Hz, and these modulations shift to lower and higher frequencies for slower and faster speaking rates, respectively (Houtgast & Steeneken, 1985; Stilp, Kiefte, Alexander, & Kluender, 2010). The speech spectrum also modulates primarily at low rates, with distinct ranges of spectral modulations for formants (<2 cycles/kHz) and talker pitch (4–6 cycles/kHz for women, 7–10 cycles/kHz for men; Singh & Theunissen, 2003; Elliot & Theunissen, 2009). This list is certainly not exhaustive nor does it dissolve the lack of invariance, but it provides relatively well known examples of statistical regularities that are evident across 1+ seconds of speech. Furthermore, when nonspeech sound textures were of sufficient duration (0.5–2 s), they were represented according to their time-averaged statistical properties (McDermott et al., 2013). This supports the examination of time-averaged statistical regularities in speech and their links to speech perception.

These longer timescales are an important context for speech perception. Speech sound recognition is accomplished using acoustic properties of the speech sounds themselves (i.e., intrinsic cues) as well as information from neighboring sounds (i.e., extrinsic cues; Ainsworth, 1975; Nearey, 1989). Extrinsic cues are of particular interest in speech research because they can help listeners perceptually adjust to different talkers’ voices (Ladefoged & Broadbent, 1957; Johnson, 1990; Sjerps, Mitterer, & McQueen, 2011) or acoustic environments in general (Watkins, 1991; Stilp, Alexander, Kiefte, & Kluender, 2010; Stilp, Anderson, Assgari, Ellis, & Zahorik, 2016). Importantly, when extrinsic and intrinsic cues to a stimulus differ, perceptual systems tend to magnify this difference, resulting in a contrast effect. For example, Ladefoged and Broadbent (1957) examined how listeners perceived vowels that followed an introductory sentence. Listeners reported whether the vowel was / / (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 / / more often. Subsequent work revealed these spectral contrast effects (SCEs)[1] to be basic, low-level phenomena that do not require talkers or even speech to occur (e.g., Watkins, 1991; Lotto & Kluender, 1998; Holt, 2005; Stilp, Alexander, et al., 2010; Stilp, Anderson, & Winn, 2015; Sjerps et al., 2011; Kingston et al., 2014; Sjerps & Reinisch, 2015; Bosker, Reinisch, & Sjerps, 2017). These effects are widespread in speech perception, occurring even for incredibly small spectral differences (adding a +3 dB spectral peak to the context; Stilp & Assgari, 2018a). SCEs bias categorization of a wide range of speech sounds (see Stilp et al., 2015 for review) and to varying degrees, depending on the size of the spectral change across earlier and target sounds (Stilp et al., 2015; Stilp & Assgari, 2017). This reveals acute sensitivity to spectral differences across sounds during speech categorization.

[1] There are two types of SCEs in the auditory perception literature, short term and long term. In short-term SCEs, the preceding acoustic context is typically the sound immediately preceding the target sound (e.g., Lotto & Kluender, 1998). In long-term SCEs, the preceding acoustic context is a series of sounds generally 1+ seconds in duration immediately preceding the target sound. These long-term SCEs are the focus of the present report. While short-term and long-term SCEs produce effects in consistent directions and are related to each other, short-term SCEs will not be discussed further here.

In most investigations of SCEs, the preceding acoustic context was a sentence or phrase that was filtered to produce some desired property in its long-term average spectrum. The same stimulus was filtered two slightly different ways, each time introducing a different spectral property (e.g., adding a spectral peak from 100–400 Hz or from 550–850 Hz to the long-term average spectrum). On every trial of a given experiment, listeners heard one of these two filtered contexts precede the target phoneme. These filtered contexts biased speech categorization, producing an SCE. While this approach affords good acoustic control and likely maximizes the probability of observing an SCE, it is largely agnostic to the inherent spectral composition of the speech context. Some speech stimuli might inherently possess the desired spectral properties without any additional filtering. As such, these inherent spectral characteristics (equivalently, natural signal statistics) could bias speech categorization in much the same way that highly acoustically controlled filtered contexts do. This approach would significantly advance understanding of how these context effects influence speech perception in more naturalistic listening conditions. Yet, to date, this possibility has gone unexplored.

Here, natural signal statistics and speech perception are linked through measures of SCEs that bias vowel categorization. The pervasiveness and replicability of SCEs in speech perception make them ideal for testing perceptual sensitivity to statistical regularities in earlier sounds. In each of eight experiments, unfiltered and filtered context sentences were presented before target vowels, measuring the resulting biases in vowel categorization (i.e., SCEs). Unfiltered contexts were selected based on their statistical properties, defined as the natural balance of average energy across two frequency regions (see Methods). Filtered contexts were generated to have matching spectral statistical properties, serving as a control condition. Across conditions, it was hypothesized that SCEs are not produced by an *absolute* spectral cue (an added spectral peak via filtering), but by a *relative* spectral cue (more energy in one frequency region than another, whether naturally present or introduced by filtering). Thus, unfiltered and filtered contexts should both produce SCEs in vowel categorization. More importantly, vowel categorization following unfiltered contexts provides a clearer perspective on how SCEs shape everyday speech perception.

Methods

Participants

A total of 146 undergraduate students at the University of Louisville participated in this study in exchange for course credit. All reported being native English speakers with no known hearing impairments. Eight individual experiments were conducted (n = 17, 15, 17, 17, 20, 20, 20, and 20 for Experiments 1–8, respectively), and no listener participated in multiple experiments.

Stimuli

Unfiltered contexts

Sentences were analyzed and selected according to spectral properties in the low-F_1 (100–400 Hz) or high-F_1 (550–850 Hz) frequency regions, as amplifying these regions has been highly successful in producing SCEs in previous studies (Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016; Stilp, 2017; Stilp & Assgari, 2018a). Each sentence was analyzed using two separate bandpass filters. The passband was either 100–400 Hz or 550–850 Hz, with 5-Hz transition regions between the passband and stopbands. Filters had 1,000 coefficients and were created using the fir2 command in MATLAB (MathWorks, Inc., Natick, MA, USA). The amplitude envelope in each frequency region was obtained by rectifying the signal and low-pass filtering using a second-order Butterworth filter with a cutoff frequency of 30 Hz. The root-mean-square energy for each envelope was converted into dB. The difference in energy across these two frequency regions was termed the Mean Spectral Difference (MSD; Fig. 1). MSDs were always subtracted in one direction (low-F_1 energy minus high-F_1 energy), with positive MSDs indicating more energy in the low-F_1 region and negative MSDs indicating more energy in the high-F_1 region. MSDs are likely to stem from a number of sources, including but not limited to phonemic content (particularly vowel formants) 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).[2] MSDs were calculated for every sentence in the TIMIT database (Garofolo et al., 1990), the HINT database (Nilsson, Soli, & Sullivan, 1994),[3] and 338 TIMIT sentences spoken by the first author. Long-term average spectra and distributions of MSDs for these databases are shown in Fig. 2.

Sentences were selected from these databases for use as experimental stimuli according to two criteria. First, all sentences were spoken by men. This matched the talker who produced the filtered context and the target vowels (the first author). Second, sentences were selected on the basis of having MSD magnitudes in one of two groups: large (averages of the absolute values of MSDs tested in the same block between 11 and 15 dB) or small (averages of the absolute values of MSDs tested in the same block between 5 and 8.5 dB). Sentences with

[2] These measures are also at the mercy of factors such as recording conditions and equipment, which were not controlled in the present investigation. The degree to which MSD measures are representative of everyday speech could be influenced by these factors.

[3] Several sentences in the HINT database are repeated across the Adult and Child sentence lists. Here, MSDs are presented only for the 275 unique sentences in the database.

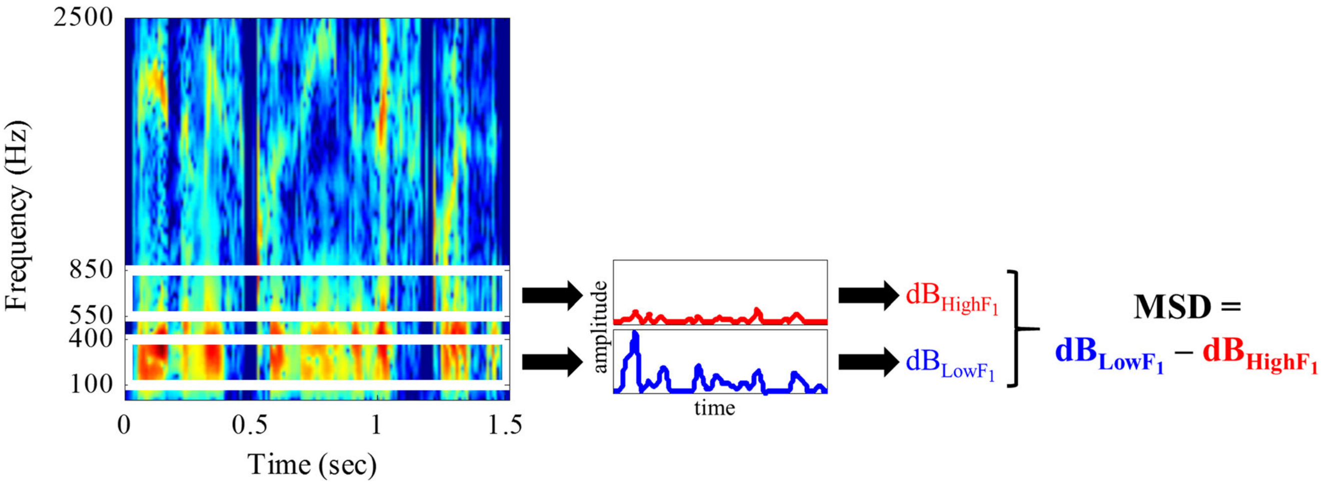

Fig. 1 Procedure for calculating Mean Spectral Differences (MSDs). Two frequency regions are excised from the sentence via bandpass filtering: low F_1 (100–400 Hz) and high F_1 (550–850 Hz). In each frequency region, the signal is rectified and low-pass filtered to produce its amplitude envelope. Energy in each frequency region is calculated using the root-mean-square amplitude of the envelope. The MSD is defined as energy in the low-F_1 region minus energy in the high-F_1 region. Here, for the sentence "Vietnamese cuisine is exquisite," MSD = 11.23 dB

large MSDs were expected to maximize the probability of observing SCEs in vowel categorization. Additionally, testing sentences with smaller and larger MSDs afforded an examination of whether SCE magnitudes scaled along with the sizes of spectral prominences in context sentences, as has been reported for filtered sentences (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017). By prioritizing talker gender and MSDs during stimulus selection, many other signal characteristics were allowed to vary freely, such as fundamental frequency, semantic and syntactic content, and duration. Sentence characteristics are listed in Table 1.

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_1 frequencies and the other had a negative MSD favoring high-F_1 frequencies (see Table 1 for details).

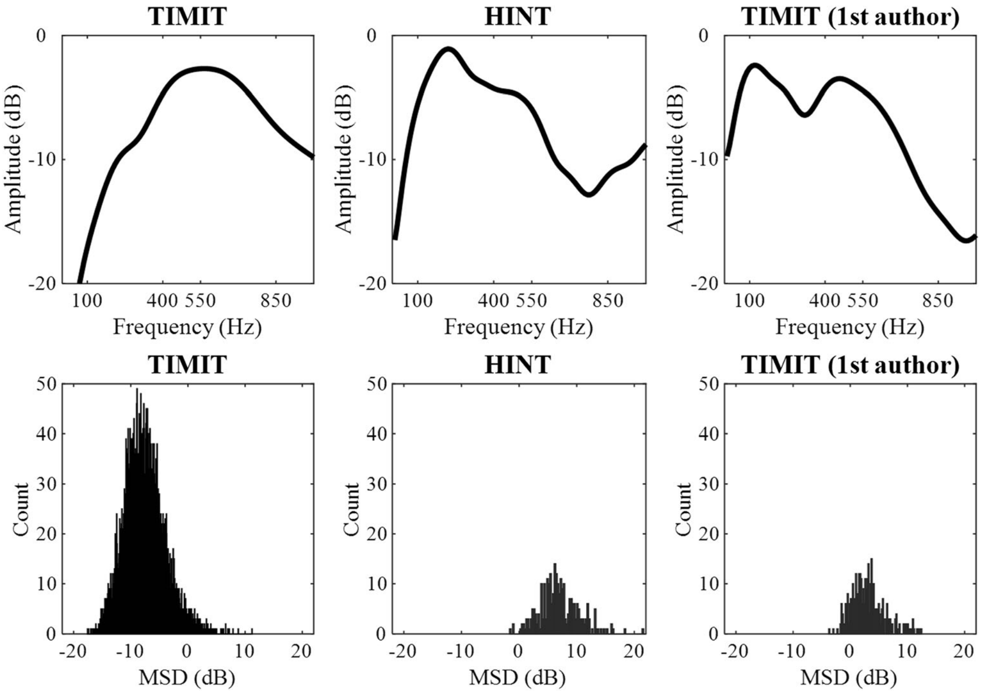

Fig. 2 (**Top row**) Long-term average spectra for 6,300 sentences in the TIMIT database (left), 275 unique sentences in the HINT database (center), and 338 sentences from the TIMIT database recorded by the first author (right). Spectra are depicted up to 1 kHz to illustrate low-F_1 (100–400 Hz) and high-F_1 (550–850 Hz) frequency regions used in calculations of MSDs. (**Bottom row**) Histograms showing the distributions of MSDs for the TIMIT database (left), HINT database (center), and sentences from the first author (right). Experimental stimuli were selected from these distributions (see Table 1)

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

Exp.	Block	Sentence	Database	Duration (ms)	MSD (dB)
1	1	Vietnamese cuisine is exquisite	TIMIT	1,516	11.23
1	1	Beg that guard for one gallon of gas	TIMIT	1,940	–13.73
1	2	Go change your shoes before you turn around	TIMIT	1,836	7.91
1	2	The drunkard is a social outcast	TIMIT	1,904	–7.09
2	1	Vietnamese cuisine is exquisite	TIMIT	1,516	11.23
2	1	A large household needs lots of appliances	TIMIT	2,700	–13.81
2	2	Go change your shoes before you turn around	TIMIT	1,836	7.91
2	2	Lots of foreign movies have subtitles	TIMIT	1,993	–7.88
3	1	She uses her spoon to eat	HINT	1,844	21.61
3	1	The ball bounced very high	HINT	2,072	–1.66
3	2	She's helping her friend move	HINT	1,673	11.64
3	2	The front yard was pretty	HINT	1,820	1.61
4	1	Vietnamese cuisine is exquisite	TIMIT	1,516	11.23
4	1	Straw hats are out of fashion this year	TIMIT	2,158	–16.68
4	2	Go change your shoes before you turn around	TIMIT	1,836	7.91
4	2	Straw hats are out of fashion this year	TIMIT	2,158	–16.68
5	1	Vietnamese cuisine is exquisite	TIMIT	1,516	11.23
5	1	A large household needs lots of appliances	TIMIT	2,700	–13.81
5	2	Go change your shoes before you turn around	TIMIT	1,836	7.91
5	2	Lots of foreign movies have subtitles	TIMIT	1,993	–7.88
6	1	Vietnamese cuisine is exquisite	TIMIT	1,516	11.23
6	1	Straw hats are out of fashion this year	TIMIT	2,158	–16.68
6	2	Go change your shoes before you turn around	TIMIT	1,836	7.91
6	2	Straw hats are out of fashion this year	TIMIT	2,158	–16.68
7	1	She uses her spoon to eat	HINT	1,844	21.61
7	1	The ball bounced very high	HINT	2,072	–1.66
7	2	The two children are laughing	HINT	1,713	12.75
7	2	The football hit the goalpost	HINT	2,135	0.59
8	1	They were pursuing him	1st author	1,089	12.52
8	1	Alfalfa is healthy for you	1st author	1,381	–3.78
8	2	Amoebas change shape constantly	1st author	1,843	11.03
8	2	"Watch it, big shot!", a hoarse voice yelled back	1st author	2,624	–2.31

Filtered contexts

Experiments also tested filtered renditions of a single sentence, a recording of the first author saying "Please say what this vowel is" (2,174 ms). This stimulus has been highly successful in biasing vowel categorization (Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016; Stilp, 2017; Stilp & Assgari, 2018a). Here, this stimulus provided an upper limit of SCE magnitudes when talker variability, duration, and all other acoustic properties were held constant from trial to trial. Filtered contexts served as control conditions in a second sense, ensuring that listeners were responding consistently in conditions where SCEs were most likely to occur based on past research.

This stimulus possessed nearly equal energy in low-F_1 and high-F_1 frequency regions before any spectral peaks were added (MSD = –0.91 dB). This stimulus was then processed by the same narrowband filters used to introduce spectral peaks in previous studies: 300-Hz-wide finite impulse response filters spanning either 100–400 Hz or 550–850 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_1 frequencies). Gain for the appropriate filter (here, amplifying low-F_1 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 0.1 dB, filter gain was increased slightly (e.g., adding slightly more energy to the low-F_1 region) and the process repeated. This continued iteratively until the MSDs for the unfiltered and filtered contexts were functionally equal (within 0.1 dB of each other). Thus, each unfiltered context sentence in an experiment was accompanied by a filtered context sentence with an equivalent MSD.

Targets

Target vowels were the same / /-to-/ɛ/ continuum as previously tested by Stilp and colleagues (Stilp et al., 2015; Assgari & Stilp, 2015; Stilp & Alexander, 2016; Stilp, 2017; Stilp & Assgari, 2018a). For a detailed description of the stimulus generation procedures, see Winn and Litovsky (2015). Briefly, tokens of / / and /ɛ/ were recorded by the first author. Formant contours from each token were extracted using PRAAT (Boersma & Weenink, 2017). In the / / endpoint, F_1 linearly increased from 400 to 430 Hz while F_2 linearly decreased from 2,000 to 1,800 Hz. In the /ɛ/ endpoint, F_1 linearly decreased from 580 to 550 Hz while F_2 linearly decreased from 1,800 to 1,700 Hz. Trajectories for each formant were linearly interpolated to create a ten-step continuum of formant tracks. A single voice source was extracted from the / / endpoint, and F_1/F_2 tracks were used to filter this source, producing the 10-step continuum of vowel tokens. Energy above 2,500 Hz was replaced with the energy high-pass filtered from the original / / token for all vowels. Final vowel stimuli were 246 ms in duration with the fundamental frequency set to 100 Hz throughout the vowel.

All context sentences and vowels were low-pass filtered at 5 kHz and set to equal root mean square (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, USA). 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, USA) 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, USA). 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 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 vowels with 80% accuracy after the third practice session, they were not allowed to participate in the main experiment.

Each experiment consisted of four blocks: two blocks presenting unfiltered contexts and two blocks presenting filtered contexts. Every block tested two context sentences. In unfiltered blocks, one sentence typically had a low-F_1-biased MSD and the other sentence had a high-F_1-biased MSD (see Table 1). In filtered blocks, the single sentence spoken by the first author was filtered two different ways in order to produce the same MSDs as those tested in the corresponding unfiltered block. Thus, each individual experiment tested two MSD values, once each in unfiltered and filtered conditions. There were 160 trials per block, with each unique context/vowel pairing tested eight times.[4] Blocks were presented in counterbalanced orders across participants, and trials within each block were randomized. The experiment was self-paced, allowing participants the opportunity to take short breaks between each block as needed. No feedback was provided. The total experimental session lasted approximately 1 h.

Results

After passing the practice session, a performance criterion was implemented such that listeners were required to achieve at least 80% accuracy identifying vowel continuum endpoints in a given experimental block. If listeners exhibited difficulty categorizing vowel endpoints, that seriously compromised the interpretability of shifts in their vowel category boundaries due to SCEs. Eighteen blocks (out of 580 blocks total) were removed from further analysis: seven listeners failed only one out of the four blocks in their respective experiments, two listeners failed two blocks, one listener failed three blocks, and one listener failed all four blocks in that experiment so his/her data were removed entirely.

Omnibus analysis

Results were analyzed using mixed-effect models in R (R Development Core Team, 2016) using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014). Mixed-effects modeling allowed 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. The dependent variable was modeled as binary (“ih” or “eh” responses coded as 0 and 1, respectively). Fixed effects in the model included: Target (coded as a continuous variable from 1

[4] Some of the experiments presented a third context sentence chosen for its MSD being near zero. As this manipulation was not shared across all experiments, these trials have been removed from data analyses.

to 10 then mean-centered), Spectral Peak (contrast coded; high F_1 = –0.5, low F_1 = +0.5), Condition (contrast 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). Including the interactions between Target and other fixed effects did not significantly improve model fit [$\chi^2(7) = 3.70$, $p = 0.81$], so these interactions were excluded from the final model. Otherwise, all two-way and three-way interactions between Spectral Peak, Condition, and MSD were included in the model.

Random slopes were included for each main fixed effect to allow the magnitudes of these factors to vary by listener.[5] A random slope was also included for the Spectral Peak × Condition interaction, which allowed the differences in SCE magnitudes across filtered and unfiltered sentences to vary by listener. Finally, a random intercept of listener was included to account for individual differences relative to each listener's baseline level of performance. All models were run using bobyqa optimization with a maximum of 800,000 iterations. The final model had the following form:

$$\begin{aligned}&\text{Response}\sim\text{Target}+\text{Spectral Peak}+\text{Condition}+\text{MSD}+\text{Spectral Peak}{-}\text{by}{-}\text{Condition}+\text{Spectral}\\&\text{Peak}{-}\text{by}{-}\text{MSD}+\text{Condition}{-}\text{by}{-}\text{MSD}+\text{Spectral Peak}{-}\text{by}{-}\text{Condition}{-}\text{by}{-}\text{MSD}+\Big(1+\text{Target}+\\&\qquad\text{Spectral Peak}+\text{Condition}+\text{MSD}+\text{Spectral Peak}{-}\text{by}{-}\text{Condition}\mid\text{Listener}\Big)\end{aligned}$$

Results from this model are shown in Table 2. The model intercept was significant, indicating more "eh" responses than "ih" responses to the vowel targets. The significant effect of Target predicts more "eh" responses with each rightward step along the vowel target continuum (toward higher F_1 frequencies and the /ɛ/ endpoint), as expected. The significant positive effect of Spectral Peak predicts an increase in "eh" responses when the filtering condition is changed from high F_1 (the level coded as –0.5) to low F_1 (the level coded as +0.5), consistent with the hypothesized direction of SCEs. Listeners were also more likely to respond "eh" to vowel targets following unfiltered sentences (significant main effect of Condition) and at higher MSD values (significant main effect of MSD).

Model interactions including the Spectral Peak term (SCEs) are of particular importance. First, the significant negative interaction between Spectral Peak and Condition indicates that SCE magnitudes were significantly smaller following unfiltered sentences than following filtered sentences. Second, the significant positive interaction between Spectral Peak and MSD indicates that SCE magnitudes increased linearly as MSDs increased. This replicates and extends similar findings in studies that used only filtered sentence contexts (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017, 2018a). Finally, the three-way interaction between Spectral Peak, Condition, and MSD was not statistically significant. This indicates that the rate at which SCE magnitudes increased at larger MSDs (i.e., the observed significant interaction between Spectral Peak and MSD) did not significantly differ across filtered and unfiltered sentence conditions.

Table 2 Results from the generalized linear mixed-effects logistic regression on listeners' responses. See main text for description of factors and the model structure. SEM indicates the standard error of the mean

	Estimate	SEM	*Z*	*p*
(Intercept)	0.254	0.056	4.559	5.1e-6
Target	1.098	0.031	34.971	<2e-16
Spectral Peak	0.708	0.042	16.965	<2e-16
Condition	0.335	0.049	6.846	7.6e-12
MSD	0.011	0.005	2.229	0.026
Spectral Peak × Condition	–0.665	0.060	–11.082	<2e-16
Spectral Peak × MSD	0.061	0.009	6.962	3.4e-12
Condition × MSD	0.004	0.006	0.779	0.436
Spectral Peak × Condition × MSD	–0.006	0.013	–0.447	0.655

This omnibus analysis confirms the relationship between MSDs and SCEs in vowel categorization following filtered and unfiltered sentences, but important characteristics of each condition are not immediately evident. For example, while SCEs were significantly smaller following unfiltered sentences than following filtered sentences, it cannot speak to the amounts of variability in each condition. One might expect more variable results following different context sentences (potentially spoken by different talkers) compared to filtered renditions of a single sentence by a single talker. Also, this omnibus analysis models the probability of responding "eh" on a given trial, but the primary phenomenon of interest is the SCE, which occurs across all trials in a given block. This model offers predictions about the magnitudes of SCEs, but does not readily reveal the actual SCEs in each block of each experiment. Therefore, follow-up analyses were conducted to reveal these characteristics.

[5] Inclusion of a random slope for context sentence was explored (to reflect item variation by using different unfiltered sentences in different blocks/experiments). However, the fixed-effect model matrix was rank-deficient and the model failed to converge when this term was included, so it was omitted from analyses.

Analysis of SCEs

SCEs were calculated at the group level for each block of each experiment in the following manner. First, given their significant difference reported above, results were separated according to condition (filtered vs. unfiltered context sentences). Then, for each block in each condition, listeners' responses 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 using the following established procedures (Stilp et al., 2015; Stilp & Assgari, 2017, 2018a; see Fig. 3). The 50% points were identified on the logistic regression fits to the aggregated vowel responses following low-F_1-emphasized contexts and responses following high-F_1-emphasized contexts. These 50% points were then converted into the stimulus step number that listeners would label as /ɛ/ 50% of the time. Vowel 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 survey (Fig. 2) and the omnibus analysis reported above, sentences with positive MSDs possessed more energy in the low-F_1-frequency region, and sentences with negative MSDs possessed more energy in the high-F_1-frequency region. Each block presented two context sentences (generally one with a positive MSD and one with a negative MSD; see Table 1), which produced an SCE of some magnitude (possibly even zero magnitude, or a failure to bias vowel categorization). 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.[6] This more broadly reflects the average imbalance of energy in the spectrum of the context sentences, whether biased towards low-F_1 energy or high-F_1 energy.

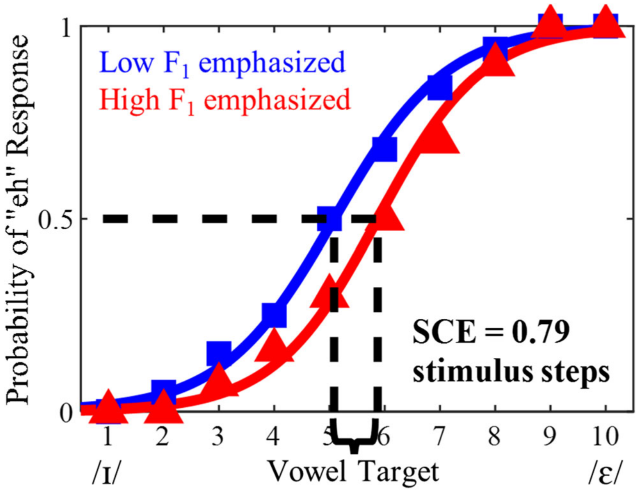

Fig. 3 Calculation of SCEs. Logistic regressions derived from the mixed-effect model are fit to group responses following low-F_1-emphasized (blue) and high-F_1-emphasized (red) context sentences. The 50% point is calculated for each regression fit, then translated into the stimulus step number (from 1 to 10, interpolated as needed; dashed vertical lines) that would be categorized as "eh" 50% of the time. The SCE is defined as the shift in 50% points, measured in stimulus steps. Here, the SCE is 0.79 stimulus steps. Logistic regressions and SCEs for each block in every experiment are illustrated in the Supplementary Figures (see online supplementary material)

SCEs in each block of each experiment are portrayed in Fig. 4 and listed in Table 3. For filtered context sentences, SCEs were significantly correlated with sentence MSDs (r = 0.83, p < .0001). A mixed-effects linear regression was fit to these data, using the fixed effect of MSD and the random effect of listener group (i.e., experiment number, from one to eight) to predict the magnitude of the resulting SCE. MSDs were indeed significant predictors of SCE magnitudes (t = 7.84, p < .0001). For unfiltered context sentences, SCEs were again significantly correlated with sentence MSDs (r = 0.65, p < .01). The same mixed-effects linear regression model was fit to these data, revealing that MSDs were again significant predictors of SCE magnitudes (t = 3.90, p < .005). Finally, the slopes of these linear regressions were highly similar (filtered sentences = 0.053 stimulus steps per additional dB filter gain, unfiltered sentences = 0.072 steps/dB), indicating that SCEs grew at similar rates as MSDs (in unfiltered sentences) or filter gain (in filtered sentences) increased in the context sentence. This is consistent with the non-significant three-way interaction between Spectral Peak, Condition, and MSD reported in the omnibus analysis above.

While SCE magnitudes exhibited similarities across filtered and unfiltered conditions (both being significantly correlated with MSDs, both sharing comparable regression slopes), important differences were also evident. In the omnibus mixed-effects model, SCE magnitudes were significantly smaller following unfiltered sentences than filtered sentences (significant negative interaction between Spectral Peak and Condition). This point was tested directly via a paired-samples t-test on the SCEs calculated in follow-up analyses. SCEs following unfiltered context sentences (mean = 0.34 stimulus steps) were significantly smaller than SCEs following filtered context sentences (mean = 0.94; t_{15} = 10.07, p < .0001). Additionally, the magnitudes of SCEs produced by unfiltered

[6] When one MSD is positive (low-F_1 bias) and the other MSD is negative (high-F_1 bias), this formulation is mathematically equivalent to adding the absolute values of each MSD and dividing by two. This was the case for 14 out of the 16 blocks listed in Table 1. When MSDs are both positive or negative, however, averaging the absolute values overstates their expected influence on listener performance. For example, if both MSDs in an experimental block were +10, the average of absolute values is +10 but the difference divided by two is 0. The latter calculation is more appropriate given that perception is not expected to be differentially biased by two identical MSDs (in this case, both biasing responses towards high-F_1 "eh"). In Experiment 3 Block 2, the average of the absolute values of the two MSDs tested (11.64 and 1.61) is +6.63, but the difference divided by two is +5.02. The net difference between these MSDs is a better estimate of their ability to bias phoneme categorization

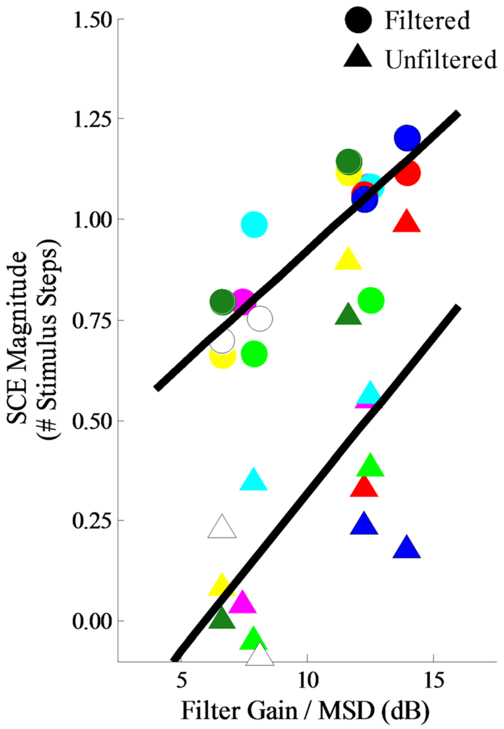

Fig. 4 Spectral contrast effect (SCE) magnitudes in each block of each experiment. SCEs produced by filtered context sentences (circles) are plotted as a function of filter gain. SCEs produced by unfiltered context sentences (triangles) are plotted as a function of relative spectral prominence (MSD). All values are obtained through mixed-effects models fit to each individual dataset (see Table 3). Icons that share color indicate results contributed by a single participant group (each completed four blocks, thus four SCEs). Solid lines represent the linear regression fits to results in each condition

context sentences were more variable compared to SCEs produced by filtered context sentences (variances = 0.110 and 0.036, respectively). This was confirmed by a statistically significant Bartlett's test, rejecting the null hypothesis of equal variances [$\chi^2(1) = 4.23$, $p < .05$].

Timecourse analysis

Unfiltered sentences were selected according to their MSDs, which reflect the long-term balance of energy across low- and high-F_1-frequency regions. While these sentences significantly biased vowel categorization (Fig. 4), these long-term averages collapse across local spectral characteristics of sentences. A sentence can produce a given MSD value an infinite number of ways, with different phonemes having energy biased toward low-F_1 or high-F_1 frequencies to varying degrees at different times throughout the sentence. Additionally, unfiltered sentences varied in their durations (see Table 1), but MSD calculations removed this information by averaging across time. Thus, the results illustrated in Fig. 4 cannot speak to any influence of more local spectrotemporal characteristics of the unfiltered context sentences.

Exploratory analyses were conducted to evaluate the predictive power of MSDs across different time-courses. 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 (which were fixed values throughout the analyses; see Table 3). The analyses proceeded as follows. First, all unfiltered sentences were aligned at their offsets. This made all stimuli uniform in terms of temporal proximity between their offsets and the onset of the vowel target 50 ms later in experimental trials, irrespective of total sentence duration. 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 1 s of silence was prepended and appended to the excerpt. The MSD of the excerpt was then calculated. Spectral analysis was conducted for both context sentences in a given experimental block; the block MSD value was calculated as the difference in MSD values divided by two. After excising t-ms excerpts from all 16 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 1,089 ms, which was the duration of the shortest sentence tested ("They were pursuing him" from Experiment 8). Exceeding this duration was undesirable because that would require removing behavioral results from the time-course analysis.

The optimal window duration for analyses was not known *a priori*, so ten different window durations were explored (t = 5–50 ms, in 5-ms steps).[7] The results of all analyses are superimposed in Fig. 5, with MSD/SCE correlation coefficient plotted as a function of the temporal analysis window (the last t ms of every sentence). For brief window durations (<500 ms), MSDs did a very poor job of predicting SCE magnitudes; correlation coefficients were almost exclusively within the range of ±0.30. This is far weaker than the correlation reported between full-sentence MSDs and SCEs (r = 0.66; denoted by the 'X' on the ordinate of Fig. 5). In general, correlation coefficient magnitude grew with longer analysis window durations. Correlation magnitudes exhibited sharp increases between 550 and 650 ms owing to the temporal patterns of phonemic content in a few stimuli. For example, "Straw hats are out of fashion this year" (high-F_1 MSD sentence in Experiments 4 and 6) exhibited many high-F_1 peaks earlier in the sentence (during "Straw hats are out of fa-") but

[7] Given the large number of correlations calculated, an inflated Type I error rate is an obvious concern. However, statistical significance is neither reported nor sought for these exploratory analyses. Instead, the primary focus is revealing the overall pattern of MSD/SCE correlation coefficients at different time-courses, including how the predictive power of longer analysis windows approaches that of full-sentence MSDs (the 'X' in Fig. 5).

Table 3 Results for each block of each experiment. Each block lists the half of the difference between the two MSDs tested in the context sentences (Block MSD), the resulting SCE measured in stimulus steps (Filtered SCE, Unfiltered SCE), the coefficient (β), and standard error for the Spectral Peak factor of the mixed-effects model fit to each dataset, and the corresponding icons in Fig. 4. Asterisks indicate model coefficients that are significantly different from zero (i.e., the presence of an SCE): * p < .05, ** p < .01, *** p < .001

Exp.	Block	Block MSD	Filtered SCE	β (se)	Icon	Unfiltered SCE	β (se)	Icon
1	1	12.48	1.08	1.33 (0.23)***	●	0.55	0.75 (0.19)***	▲
1	2	7.50	0.80	0.92 (0.18)***	●	0.04	0.05 (0.15)	▲
2	1	12.52	1.08	1.15 (0.17)***	●	0.56	0.66 (0.19)***	▲
2	2	7.90	0.99	1.10 (0.24)***	●	0.34	0.45 (0.19)*	▲
3	1	11.64	1.12	1.33 (0.16)***	●	0.89	1.02 (0.26)***	▲
3	2	5.02	0.66	0.80 (0.15)***	●	0.08	0.10 (0.15)	▲
4	1	13.96	1.12	1.18 (0.14)***	●	0.99	1.10 (0.17)***	▲
4	2	12.29	1.06	1.22 (0.21)***	●	0.33	0.37 (0.16)*	▲
5	1	12.52	0.80	0.89 (0.14)***	●	0.38	0.39 (0.13)**	▲
5	2	7.89	0.67	0.65 (0.12)***	●	-0.05	-0.06 (0.20)	▲
6	1	13.96	1.20	1.31 (0.18)***	●	0.18	0.17 (0.19)	▲
6	2	12.29	1.05	1.10 (0.19)***	●	0.24	0.26 (0.13)*	▲
7	1	11.64	1.14	1.29 (0.18)***	●	0.76	0.96 (0.18)***	▲
7	2	6.08	0.79	0.96 (0.14)***	●	0.00	0.00 (0.15)	▲
8	1	8.14	0.75	0.91 (0.15)***	○	-0.09	-0.11 (0.17)	△
8	2	6.66	0.70	0.91 (0.16)***	○	0.23	0.38 (0.24)	△

none in the last 580 ms (“-shion this year”). Similarly, “A large household needs lots of appliances” (high-F_1 MSD sentence in Experiments 2 and 5) exhibited several high-F_1 peaks but only half of one peak in the last 550 ms (during “-iances” in

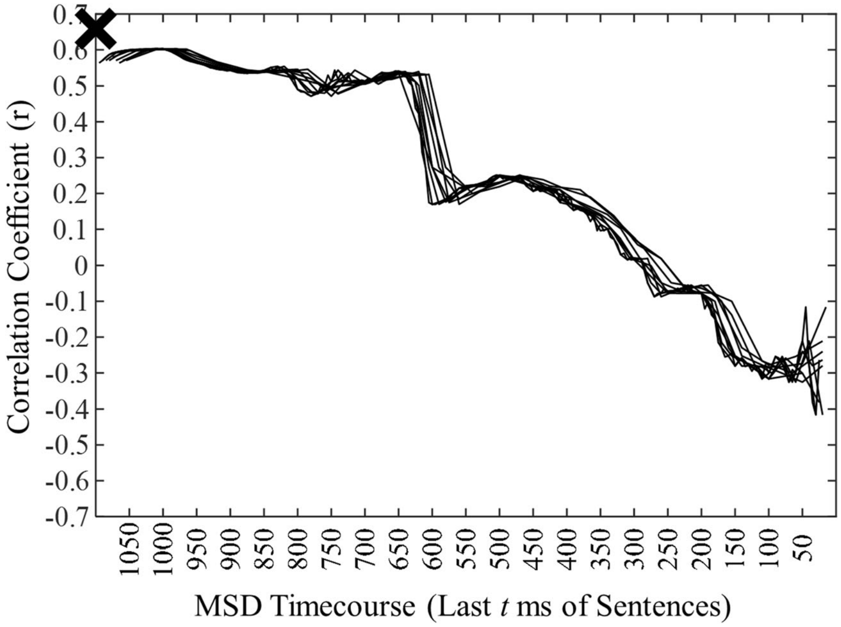

Fig. 5 Analyses of the timecourse of MSDs for predicting behavioral results. The abscissa depicts the duration of sentence spectra (relative to sentence offset) utilized for calculating MSDs. The ordinate depicts the correlation coefficient for these shorter-duration MSDs with SCEs in the unfiltered condition. Superimposed lines reflect the different window durations utilized in analyses (5–50 ms at a time, in 5-ms steps). See text for details

"appliances"). Early high-F_1 peaks in these sentences contributed significantly to their respective overall MSDs (–16.68 and –13.81), which then drove block MSDs toward their final magnitudes and improved the predictive power of time-course analyses. The longest analysis windows approached the efficacy of predicting SCEs using full-sentence MSDs (maximum correlation coefficient observed in Fig. 5: $r = 0.60$). However, this finding does not prescribe any particular temporal windows for context effects in speech perception. Unfiltered stimuli were selected without any regard for local temporal characteristics of MSDs; only their long-term (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.

Discussion

Review of the present results

Objects and events in the sensory environment are highly structured across space and time. Considerable success has been achieved in documenting this statistical structure in natural images (Field, 1987; Ruderman, 1994; Olshausen & Field, 1996; Bell & Sejnowski, 1997; Simoncelli, 2003; Torralba & Oliva, 2003; Geisler, 2008). However, efforts to reveal the statistical structure of speech have seen limited success due to its extreme acoustic variability. An alternative approach examines statistical regularities in the speech signal across longer timescales (1+ seconds) in order to reduce this variability (see *Introduction* for examples). The present report introduces a new statistical regularity in speech: the MSD, which captures the inherent balance of acoustic energy across two frequency regions. This statistical regularity is directly relevant to speech categorization, as MSDs in unfiltered sentences predicted the magnitudes of resulting SCEs in vowel categorization (Fig. 4). This validates the present approach and offers a promising direction for further research that links the statistical structure of the speech signal to perception.

Unfiltered context sentences were tested alongside filtered sentences that were tailored to test equivalent MSDs. Results across conditions shared two important similarities. First, SCE magnitudes were significantly correlated with spectral prominences in the context sentences, whether naturally present (unfiltered condition) or artificially added (filtered condition). This substantially increases the ecological validity of SCEs because they are not limited to careful laboratory manipulations that generate multiple renditions of the same token. Instead, SCEs were produced in acoustically variable conditions that more closely resemble everyday speech perception. Second, the slopes of linear regressions fit to each condition were very similar, indicating that the rates at which SCEs increased at larger MSDs were very similar for filtered and unfiltered sentences. Previous studies reported that this linear relationship existed for filtered context sentences (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017), but here the relationship is extended to unfiltered sentences as well. This increases the generality of these effects such that whatever the reason the preceding acoustic context possesses certain spectral characteristics (whether naturally present or added via filtering), subsequent categorization shifts in speech scale in orderly ways.

There are also two very important differences in the results across filtered and unfiltered conditions. First and foremost, SCEs were significantly smaller following unfiltered sentences than following filtered sentences. This is apparent in multiple ways: the visual separation between these results in Fig. 4, the significant interaction between Spectral Peak and Condition in the omnibus mixed-effects model analysis, and the significant *t*-test on filtered-context SCEs and unfiltered-context SCEs. This is particularly true at small MSDs, where filtered sentences still produced significant SCEs but unfiltered sentences failed to bias vowel categorization (SCE magnitudes ≈ 0). Second, results were three times more variable in the unfiltered condition where different sentences (sometimes spoken by different talkers) were presented. These are essential points for evaluating different approaches to measuring spectral context effects in speech categorization.

It is difficult to isolate precisely why results varied across unfiltered and filtered conditions, as several possibilities exist. First, each filtered block tested the same context sentence by the same talker on every trial, differing only in which frequency region was amplified. Listeners heard the same low-F_1 and high-F_1 patterns of energy on every trial, differing only by a matter of filter gain. Conversely, each unfiltered block tested two different sentences, each with its own patterns of low-F_1 and high-F_1 energy. Greater acoustic variability could have made listeners less sensitive to these spectral regions, slightly diminishing their effectiveness in producing SCEs. Second, conditions differed in overall uncertainty. By presenting the same talker reading the same sentence on every trial, filtered blocks contained zero uncertainty. In unfiltered blocks, listeners did not know whether the next trial would be the same talker as the previous trial or not. Even in same-talker blocks (Experiments 3, 7, and 8), listeners were uncertain which context sentence would be presented next. Third, characteristics of the context talkers' voices might influence listener performance. For example, talkers' fundamental frequencies interact with vowel categorization (Barreda & Nearey, 2012); this could conceivably affect the influence of acoustic context. Additionally, given that acoustic variability across talkers diminishes SCE magnitudes (Assgari & Stilp, 2015; Assgari, Mohiuddin, Theodore, & Stilp, 2016; under review), one might intuit that different-talker blocks (Experiments 1, 2, 4, 5, and 6) might produce smaller SCEs than single-talker blocks (Experiments 3, 7, and 8), but this was not the case (different-talker mean SCE = 0.35, same-talker mean SCE = 0.31). Fourth, conditions potentially vary in the degree to which listeners associate spectral peaks with the talker's speech or some other source. In Assgari and Stilp (2015), listeners displayed no sensitivity to changes in talker when large (+20 dB) spectral peaks were added to the context sentences, but were disrupted by talker variability when small (+5 dB) peaks were added. Spectral peaks in filtered sentences might be associated with some other external source (e.g., distortion in the communication channel; Watkins, 1991). Spectral peaks were not added to unfiltered sentences but were the result of talker and/or phoneme characteristics. This might make listeners less sensitive to these spectral characteristics, thus producing smaller SCEs overall. Finally, unfiltered sentences with negative MSDs (more energy in 550–850 Hz than 100–400 Hz) also tended to have more energy between 850 and 1,500 Hz than sentences with positive MSDs. This elevated spectral energy could have reduced the prominence of the high-F_1-frequency region through suppression and/or lateral inhibition. This inhibitory influence would produce a smaller effective spectral peak, resulting in a smaller MSD and thus a smaller SCE magnitude. Conversely, filtered contexts possessed the same energy above 850 Hz on every trial, holding any suppressive or inhibitory influences constant across all experiments. These and other factors might underlie the differences in SCE magnitudes across filtered and unfiltered context sentences, but systematic experimentation is necessary to evaluate each possibility.

Efficient coding of natural signal statistics in speech perception

Results from these experiments are consistent with the Efficient Coding Hypothesis (Attneave, 1954; Barlow, 1959), which states that sensory and perceptual systems adapt and evolve to efficiently capture statistical structure in the environment. Stimulus selection is a central consideration in efficient coding research. Several calls have been made to use naturalistic stimuli instead of synthetic stimuli to study sensory processing (e.g., Felsen & Dan, 2005; Einhauser & Konig, 2010). Visual and auditory psychophysics share long histories of using simple stimuli to assess sensory and perceptual system function (e.g., in vision: sinusoidal gratings, square wave gratings, white noise; in audition: pure tones, clicks, white noise of varying bandwidths). Simple stimuli afford excellent acoustic control but are sorely lacking in ecological validity (just as using ecologically valid stimuli like natural images and speech come at the cost of acoustic control; Winn & Stilp, 2018). Thus, there are limits to how much synthetic stimuli can inform our understanding of sensory systems that have adapted and evolved to process natural stimuli. For example, a wide range of studies reported more accurate predictions of neural responses and greater and more efficient information transmission when using naturalistic sounds rather than synthetic sounds (Rieke, Bodnar, & Bialek, 1995; Theunissen, Sen, & Doupe, 2000; Escabi, Miller, Read, & Schreiner, 2003; Lesica & Grothe, 2008; Theunissen & Elie, 2014). In the present report, all experiments used the naturalistic stimulus of speech, but the spectral compositions of precursor sentences varied in their naturalness. Sentences either already possessed the desired spectral characteristics (i.e., MSDs, the naturalistic approach), or were compelled to possess these spectral characteristics (i.e., through filtering, the synthetic approach). Granted, this distinction is clearly not as stark as the difference between natural (speech) sounds and synthetic (tones, clicks, noise) sounds listed above, but it is still an important one given that unfiltered and filtered sentences biased vowel categorization to different degrees. According to the Efficient Coding Hypothesis, it is advantageous to study context effects using unfiltered sentences because that models the natural acoustic ecology of speech perception far more closely than artificially manipulated sentences.

A new understanding of spectral contrast effects

The present results radically reconceptualize SCEs in four important ways. First, the longstanding view of SCEs was that they were "channel effects." Whenever the transmission channel (or equivalently, the communication channel; room, loudspeaker, telephone, etc.) introduced systematic distortion to sound spectra, SCEs were viewed as a means of perceptually compensating for this distortion (Watkins, 1991; Watkins & Makin, 1994, 1996a, 1996b; for further discussion see Assmann & Summerfield, 2004). This notion was advanced so far as to cast this perceptual process as "inverse filtering" of the transmission channel distortion (Watkins & Makin, 1994, 1996a). The present results directly challenge this view. SCEs were produced by unfiltered sentences absent from any distortion from the communication channel (i.e., absent from any filtering, which modeled systematic channel distortion in the studies of Watkins and Makin). Thus, one cannot relegate SCEs to mere "channel effects"; instead, they are better viewed as a general means of maintaining perceptual constancy irrespective of the sources or causes of spectral changes across successive sounds (Kluender, Coady, & Kiefte, 2003; Stilp, Alexander, et al., 2010; Stilp et al., 2015; Winn & Stilp, 2018).

Second, the magnitudes of SCEs require close consideration. The fact that spectral properties of earlier sounds influence speech categorization has long been known (Ladefoged & Broadbent, 1957). This and subsequent investigations generally employed a common approach, presenting multiple filtered renditions of a single context stimulus (akin to the filtered condition tested here). In the present study, unfiltered sentences were tested alongside filtered sentences for the first time, revealing smaller and more variable SCEs in speech categorization than those produced by filtered sentences. These differences are extremely important considerations for just how closely highly acoustically controlled studies (i.e., filtered conditions) model everyday speech perception and its considerable acoustic variability (i.e., unfiltered conditions). Ostensibly, the filtered-sentence approach is designed to model (unfiltered) sentences in everyday listening conditions, but highly acoustically controlled conditions may exert greater influence over speech categorization than less controlled/more acoustically variable conditions. Future research should be cognizant of which approach is under study at a given time and what conclusions or generalizations are appropriate to make. For example, one cannot accurately predict the magnitudes of context effects based solely on spectral energy in the preceding acoustic context. Other factors must be taken into consideration, such as whether the contexts are different stimuli or merely different renditions of a single token.

Third, recent research revealed that SCE magnitudes vary continuously: larger spectral differences between context and target stimuli (as introduced by larger filter gains) produced increasingly larger SCEs (Stilp et al., 2015; Stilp & Alexander, 2016; Stilp & Assgari, 2017). In some cases, this linear relationship was extremely strong ($r^2 = 0.98$ in Stilp & Alexander, 2016; Stilp & Assgari, 2017). Previous studies tested filtered context sentences exclusively, but here, strong linear relationships were replicated for filtered context sentences and extended to unfiltered sentences. Far more variability was observed in the latter case; this variability is important because it reflects acoustic variability commonly encountered in everyday speech perception. The extreme linearity in previous studies might have been facilitated by extremely low acoustic variability when presenting filtered renditions of the same context sentence on every trial. This discrepancy should guide expectations about the degrees to which speech categorization biases can be predicted from acoustic characteristics of surrounding sounds.

Finally, Stilp et al. (2015) suggested that filter power was an effective predictor of SCE magnitudes. As filter power increased (whether increasing peak gain for narrowband filters or filtering by larger percentages of the difference between target vowel spectral envelopes), SCE magnitudes increased linearly. Here, SCE magnitudes varied following sentences that inherently contained certain spectral characteristics without any filtering, rendering this particular predictor irrelevant. It is thus more effective and generalizable to predict SCE magnitudes from the size of the spectral difference between context and target, whether this difference occurs naturally (via unfiltered sentence contexts) or synthetically (via filters). In both cases, the larger the spectral prominence in context sentences (larger MSD or larger filter gain), the larger the change between that context and the subsequent target vowel, the larger the resulting shift in speech categorization will be (see Fig. 4).

Mean spectral differences

This study introduces MSDs as a statistical regularity in speech acoustics. MSD magnitudes vary widely across sentences (Fig. 2), stemming from the fact that MSD values differ as a function of phonemic content and possibly also talker. For example, several experiments utilized a recording of the sentence “Vietnamese cuisine is exquisite” because its MSD was heavily biased towards low-F_1 frequencies (MSD = 11.23; Table 1). This low-F_1 bias likely stems from the large number of high vowels with lower center frequencies for F_1. Conversely, stimuli drawn from the other end of the MSD distributions (biases toward high-F_1 frequencies) contained more low vowels (e.g., “Straw hats are out of fashion this year”, MSD = –16.68). Sentences that have more even balances of (energy in) low and high vowels are expected to have MSD values closer to 0. Thus, some sentences are expected to bias phoneme categorization based on having relatively large MSDs whereas other sentences are not. Additionally, talker characteristics might also affect MSD values. Talkers with longer vocal tracts produce formants at lower frequencies than talkers with shorter vocal tracts for a given segmental context; this could result in more low-F_1-biased MSDs. Likewise, talkers with shorter vocal tracts produce higher formant frequencies than talkers with longer vocal tracts for a given segmental context; this could result in more high-F_1-biased MSDs. Talker vocal tract lengths were not available for the present materials, but it is intriguing that the talker of the HINT database and the first author are both adult men with low-F_1-biased distributions of MSDs (Fig. 2). Further analyses are needed to identify the relative contributions of what is being said (phonemic content, particularly high-energy formant peaks) and who is talking (vocal tract length) to MSDs.

Across thousands of sentences, MSD values were constrained to a finite range (minimum value = –17.56 dB in TIMIT; maximum value = +21.61 dB in HINT). The existence of such limits (whether specifically these values or if they are modestly exceeded in future analyses) offer clear direction for future research on spectral context effects. SCEs in speech categorization have long been studied by filtering a context sentence to change its spectral composition. This approach faces no constraints as to the degree to which speech spectra can be altered; a +30-dB spectral peak can be added just as easily as a +3-dB peak can. However, there is little utility in manipulating speech spectra beyond the limits of MSD distributions. Such extreme manipulations do not occur naturally, so testing such stimuli neither models nor informs everyday speech perception.

MSDs reflect spectral energy averaged across an entire sentence, but this energy waxes and wanes over time depending on acoustic characteristics of phonemes within the sentence (see Supplementary Figures in online supplementary material). Variability in these patterns of energy modulations might have contributed to the significantly reduced SCE magnitudes following unfiltered context sentences. In unfiltered blocks, listeners heard two unfiltered sentences, each with its own patterns of energy modulation in low-F_1- and high-F_1-frequency regions. This is far more variability than listeners encountered in filtered blocks, where they heard the same sentence stimulus and thus the same patterns of energy modulations in low-F_1- and high-F_1-frequency regions. These patterns only varied as a matter of degree due to amplification in one of the two frequency regions. SCE magnitudes significantly differed despite stimuli being matched in terms of their MSDs. This finding challenges arguments that the long-term average spectrum of the preceding acoustic context fully explains subsequent biases in phoneme categorization (Huang & Holt, 2012). Other information beyond the long-term average spectrum must be considered.

MSDs in F_1 frequency regions are far from the only perceptually significant statistical regularity in speech. Each sentence can be viewed as possessing several different MSDs that could be relevant for categorization of different phonemes. The present experiments selected and tested sentences based on their energy in 100–400 Hz and 550–850 Hz frequency regions, which are relevant for categorization of / / and /ε/ but obviously not for all phoneme pairs. Analogous to the present study, sentences with skewed distributions of higher-frequency energy would be expected to bias categorization of phonemes distinguished by higher-frequency energy. Research has confirmed this expectation, as sentence MSDs in low-F_3 (1,700–2,700 Hz) or high-F_3 (2,700–3,700 Hz) frequency regions influence categorization of /d/-/g/ varying in F_3 transition onset frequency (Stilp & Assgari, 2018b). Statistics beyond MSDs also influence spectral context effects. For example, when hearing speech from many different talkers in a given block, variability in their fundamental frequencies can restrain the magnitudes of context effects in vowel categorization (Assgari & Stilp, 2015; Assgari et al., 2016, under review) or slow speech recognition (Goldinger, 1996; Assgari, 2018). Statistics such as mean speaking rate (which is tied to amplitude envelope modulations; Stilp, Kiefte, et al., 2010) can also shift phoneme categorization (e.g., Ainsworth, 1973; Repp, Liberman, Eccardt, & Pesetsky, 1978; Summerfield, 1981). Thus, efforts tying statistical characteristics of speech to perception are a fruitful direction for ongoing and future research.

Conclusion

The sensory environment is highly structured, and the present report reveals close correspondence between this structure and perception. This report linked the natural statistical structure of speech acoustics (MSDs in unfiltered sentences) to speech perception, demonstrating its significant influence on vowel categorization. Importantly, this influence was smaller and more variable than the traditional approach of using digital filters to impose spectral characteristics on context stimuli. This reveals the occurrences and magnitudes to which these statistical regularities bias categorization in everyday speech perception, while also raising important questions about how faithfully this has been modeled in past studies that used highly acoustically controlled filtered stimuli.

Acknowledgements We wish to thank Matthias Sjerps and two anonymous reviewers for their valuable comments and suggestions on an earlier draft of this manuscript. We are also grateful for the efforts of Ashley Batliner, Alexandrea Beason, Madison Rhyne, and Carly (Newman) Sinclair in data collection.

References

Ainsworth, W. (1973). Durational cues in the perception of certain consonants. *Proceedings of the British Acoustical Society*, *2*, 1–4.

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 10–113). London: Academic Press.

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.

Assgari, A. A., & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. *Journal of the Acoustical Society of America*, *138*(5), 3023–3032.

Assmann, P. F., & Summerfield, Q. (2004). *The perception of speech under adverse conditions. Speech processing in the auditory system* (Vol. 18). Springer New York.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193.

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. *The Mechanisation of Thought Processes*, *10*, 535–539.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 53–85). Cambridge, Mass.: MIT Press.

Barreda, S., & Nearey, T. M. (2012). The direct and indirect roles of fundamental frequency in vowel perception. *The Journal of the Acoustical Society of America*, *131*(1), 466–477. https://doi.org/10.1121/1.3662068

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

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer program].

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

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.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., … Ludvigsen, C. (1994). An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America, 96*(4), 2108–2120.

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.

Einhauser, W., & Konig, P. (2010). Getting real-sensory processing of natural stimuli. *Current Opinion in Neurobiology*, *20*(3), 389–395.

Elliot, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *Public Library of Science Computational Biology*, *5*(3), 1–14. https://doi.org/10.1371/journal.pcbi.1000302

Escabi, M. A., Miller, L. M., Read, H. L., & Schreiner, C. E. (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *Journal of Neuroscience*, *23*(37), 11489–11504.

Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature Reviews Neuroscience*, *8*(12), 1643–1646.

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.

Field, D. J. (1994). What is the goal of sensory coding. *Neural Computation*, *6*(4), 559–601.

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, Gaithersburg, MD.*

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Reviews in Psychology*, *59*, 167–192.

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.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183.

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.

Houtgast, T., & Steeneken, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech-intelligibility in auditoria. *Journal of the Acoustical Society of America*, *77*(3), 1069–1077.

Huang, J., & Holt, L. L. (2012). Listening for the norm: Adaptive coding in speech categorization. *Frontiers in Psychology*, *3*, 10. https://doi.org/10.3389/fpsyg.2012.00010

Institute, A. N. S. (1997). American National Standard Methods for Calculation of the Speech Intelligibility Index. In *ANSI S3.5, 1997*. New York, NY: American National Standards Institute.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, *88*(2), 642–654.

Kingston, J., Kawahara, S., Chambless, D., Key, M., Mash, D., & Watsky, S. (2014). Context effects as auditory contrast. *Attention, Perception, & Psychophysics*, *76*, 1437–1464.

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*(1), 59–69.

Kluender, K. R., Stilp, C. E., & Kiefte, 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 Berlin.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*(1), 98–104.

Lesica, N. A., & Grothe, B. (2008). Efficient temporal processing of naturalistic sounds. *Public Library of Science One*, *3*(2), e1655.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and Speech*, *29*(1), 3–11.

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.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-11.

McDermott, J. H., Schemitsch, M., & Simoncelli, E. P. (2013). Summary statistics in auditory perception. *Nature Neuroscience*, *16*(4), 493–498.

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.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–246.

McQueen, J. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*(1), 21–46.

McWalter, R., & McDermott, J. H. (2018). Adaptive and selective time averaging of auditory scenes. *Current Biology*, *28*(9), 1405–1418.e10.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*(5), 2088–2113.

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.

Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network*, *7*(2), 333–339.

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/

Repp, B. H. (1982). Phonetic trading relations and context effects - New experimental-evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 621–637.

Rieke, F., Bodnar, D. A., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society B: Biological Sciences*, *262*(1365), 259–265.

Ruderman, D. L. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, *5*(4), 517–548.

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.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Sayles, M., & Winter, I. M. (2010). Equivalent-rectangular bandwidth of single units in the anaesthetized guinea-pig ventral cochlear nucleus. *Hearing Research*, *262*(1–2), 26–33.

Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, *13*(2), 144–149.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Reviews in Neuroscience*, *24*, 1193–1216.

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *Journal of the Acoustical Society of America*, *114*(6 Pt 1), 3394–3411.

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.

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.

Stilp, C. E. (2017). Acoustic context alters vowel categorization in perception of noise-vocoded speech. *Journal of the Association for Research in Otolaryngology*, *18*(3), 465–481.

Stilp, C. E., & Alexander, J. M. (2016). Spectral contrast effects in vowel categorization by listeners with sensorineural hearing loss. *Proceedings of Meetings on Acoustics, 26*. https://doi.org/10.1121/2.0000233

Stilp, C. E., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics*, *72*(2), 470–480.

Stilp, C. E., Anderson, P. W., Assgari, A. A., Ellis, G. M., & Zahorik, P. (2016). Speech perception adjusts to stable spectrotemporal properties of the listening environment. *Hearing Research*, *341*, 168–178.

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. (2015). Languages across the world are efficiently coded by the auditory system. In *Proceedings of Meetings on Acoustics* (Vol. 23). https://doi.org/10.1121/2.0000064

Stilp, C. E., & Assgari, A. A. (2017). Consonant categorization exhibits a graded influence of surrounding spectral context. *Journal of the Acoustical Society of America*, *141*(2), EL153-EL158.

Stilp, C. E., & Assgari, A. A. (2018a). Natural signal statistics and the timecourse of spectral context effects in consonant categorization. *Journal of the Acoustical Society of America*, *143*, 1944.

Stilp, C. E., & Assgari, A. A. (2018b). Perceptual sensitivity to spectral properties of earlier sounds during speech categorization. *Attention, Perception, & Psychophysics*, *80*(5), 1300–1310.

Stilp, C. E., Kiefte, M., Alexander, J. M., & Kluender, K. R. (2010). Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences. *The Journal of the Acoustical Society of America*, *128*(4), 2112–2126.

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.

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), e30845. https://doi.org/10.1371/journal.pone.0030845

Stilp, C. E., & Kluender, K. R. (2016). Stimulus statistics change sounds from near-indiscriminable to hyperdiscriminable. *PLoS One*, *11*(8), 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*, *107*(50), 21914–21919.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095.

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355–366.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of Neuroscience*, *20*(6), 2315–31.

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*, *107*(42), 18149–18154.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391–412.

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.

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.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*(5456), 1273–1276.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Voss, R. F., & Clarke, J. (1975). “1/f noise” in music and speech. *Nature*, *258*(5533), 317–318.

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. (1996a). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *99*(6), 3749–3757.

Watkins, A. J., & Makin, S. J. (1996b). Some effects of filtered contexts on the perception of vowels and fricatives. *Journal of the Acoustical Society of America*, *99*(1), 588–594.

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.

Winn, M. B., & Stilp, C. E. (2018). Phonetics and the auditory system. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge handbook of phonetics*. Abingdon: Routledge.

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