Predicting contrast effects following reliable spectral properties in speech perception[a)]

Christian E. Stilp[b)] and Paul W. Anderson
Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292, USA

Matthew B. Winn
Department of Surgery, Waisman Center, University of Wisconsin, Madison, Wisconsin 53706, USA

(Received 29 July 2014; revised 13 March 2015; accepted 11 May 2015)

Vowel perception is influenced by precursor sounds that are resynthesized to shift frequency regions [Ladefoged and Broadbent (1957). J. Acoust. Soc. Am. **29**(1), 98–104] or filtered to emphasize narrow [Kiefte and Kluender (2008). J. Acoust. Soc. Am. **123**(1), 366–376] or broad frequency regions [Watkins (1991). J. Acoust. Soc. Am. **90**(6), 2942–2955]. Spectral differences between filtered precursors and vowel targets are perceptually enhanced, producing spectral contrast effects (e.g., emphasizing spectral properties of /ɪ/ in the precursor elicited more /ɛ/ responses to an /ɪ/-/ɛ/ vowel continuum, and vice versa). Historically, precursors have been processed by high-gain filters, resulting in prominent stable long-term spectral properties. Perceptual sensitivity to subtler but equally reliable spectral properties is unknown. Here, precursor sentences were processed by filters of variable bandwidths and different gains, then followed by vowel sounds varying from /ɪ/-/ɛ/. Contrast effects were widely observed, including when filters had only 100-Hz bandwidth or +5 dB gain. Average filter power was a good predictor of the magnitudes of contrast effects, revealing a close linear correspondence between the prominence of a reliable spectral property and the size of shifts in perceptual responses. High sensitivity to subtle spectral regularities suggests contrast effects are not limited to high-power filters, and thus may be more pervasive in speech perception than previously thought.
[http://dx.doi.org/10.1121/1.4921600]

[MSS] Pages: 3466–3476

I. INTRODUCTION

According to the efficient coding hypothesis (Barlow, 1961), sensory systems adapt and evolve to capture reliable aspects of the sensory environment. Identifying and extracting reliable aspects of the environment allows sensory systems to be optimally sensitive to changing input that may be behaviorally relevant. Neural adaptation is an elegant example of this phenomenon. When sensory inputs are constant or predictable, neural responses generally diminish or cease altogether, as no new information is being presented to the organism. When sensory inputs change, neural firing increases, indicating new information in the environment. Extracting predictability in order to be optimally sensitive to unpredictability is a core principle by which sensory systems operate.

Auditory perception extracts and exploits acoustic properties that are predictable or reliable across time. This is particularly true for reliable aspects of the long-term average spectrum, such as spectral peaks and overall shape (e.g., Ladefoged and Broadbent, 1957; Watkins, 1991; Watkins and Makin, 1994, 1996a,b; Holt, 2005, 2006; Kiefte and Kluender, 2008; Alexander and Kluender, 2010; Stilp *et al.*, 2010; Sjerps *et al.*, 2011; Stilp and Anderson, 2014). When this spectral property changes or is no longer reliable in a subsequent sound (typically the target sound to be identified), perception magnifies this difference. Perception of the target sound is biased away from the preceding spectral regularity, resulting in spectral contrast effects.[1] For example, a precursor sound with a lower-frequency emphasis will make a neutral-frequency target stimulus sound higher-frequency by comparison, and vice versa.

Spectral contrast effects in speech perception date back at least to Ladefoged and Broadbent (1957), who examined vowel perception as a function of talker characteristics. They created multiple renditions of the sentence “Please say what this word is” by shifting frequency ranges up or down, simulating higher or lower formant frequencies as would be produced by different talkers. Ladefoged and Broadbent reported substantial changes in vowel perception: sentences with lower first formant (F_1) frequencies elicited more /ɛ/ percepts (higher F_1), while sentences with higher F_1 frequencies elicited more /ɪ/ percepts (lower F_1). These results have been replicated in unprocessed natural speech (Ladefoged, 1989), across changes in talker, spatial position, and long intervening silences (Broadbent and Ladefoged, 1960), for higher formant frequencies (F_2: Huang and Holt, 2012; F_3: Laing *et al.*, 2012), for fundamental frequency (Johnson, 1990; Huang and Holt, 2009), and in spectrally sparse sinewave speech (Remez *et al.*, 1987).

In a series of reports, Watkins and colleagues demonstrated perceptual sensitivity to other types of reliable spectral properties. Watkins (1991) processed preceding acoustic

[a)]Results were presented at the 167th Meeting of the Acoustical Society of America, Providence, Rhode Island.
[b)]Electronic mail: christian.stilp@louisville.edu

contexts using spectral envelope difference (SED) filters, where the spectrum of one endpoint of the target vowel continuum was subtracted from the spectrum of the other endpoint (e.g., /ɪ/−/ɛ/). Reliable spectral properties in the acoustic context had a broadband and more complex shape than those tested by Ladefoged and Broadbent (1957). Following this filtering, Watkins (1991) reported contrastive shifts in phoneme boundaries distinguishing /ɪ/ from /ɛ/, consistent with Ladefoged and Broadbent (1957). Contrast effects were observed even when precursor and target vowel differed in spatial location, talker, orientation in time (i.e., time-reversed precursor), and in which ear each was presented. Later investigations revealed modulation of contrast effects by varying differences between spectral peaks and valleys in the precursor versus in the vowel target (Watkins and Makin, 1994, 1996a) and whether sounds after the target sound were also processed by SED filters (Watkins and Makin, 1996b). Sjerps and colleagues (2011, 2012, 2013) and Sjerps and Smiljanic (2013) suggested that perceptual compensation for reliable SEDs is pre-categorical, independent of language background, independent of attention, and similarly evident in severely acoustically perturbed speech.

Spectral contrast effects are not restricted to speech sounds. Rather than using speech as a preceding acoustic context, Holt and colleagues (2005, 2006; Huang and Holt, 2012; Laing *et al.*, 2012) presented a sequence of short-duration sine tones ("tone history"). Tone histories were spectrally impoverished compared to speech, but they still possessed long-term average spectra sufficient to produce contrast effects in identification of speech targets. Stilp and colleagues (2010) extended this approach to perception of musical instruments. They created SED filters from endpoints of a musical instrument continuum varying from French horn to tenor saxophone. Consistent with speech studies, they reported contrastive shifts in instrument identification when the filtered acoustic context was speech or a string quintet (i.e., preceding context filtered to sound more like a French horn elicited more "tenor saxophone" responses and vice versa). Stilp *et al.* suggested that sampling reliable spectral properties and contrast effects are not specific to speech but fundamental to perception of all sounds.

Reliable spectral properties have produced spectral contrast effects for a wide range of vowel contrasts, including /ɪ/-/ɛ/ (Ladefoged and Broadbent, 1957; Broadbent and Ladefoged, 1960; Sjerps *et al.*, 2011, Sjerps *et al.*, 2013), /æ/-/ɑ/ (Watkins and Makin, 1996a,b), /o/-/e/ (Mitterer, 2006), /o/-/u/ (Sjerps and Smiljanic, 2013), /ʌ/-/ɛ/ (Huang and Holt, 2012), and consonant contrasts including /d/-/g/ (Laing *et al.*, 2012) and /s/-/f/ (Watkins and Makin, 1996b). Comparable effects have been reported for musical instruments varying from French horn to tenor saxophone (Stilp *et al.*, 2010). While this phenomenon is certainly robust across these replications and extensions, the diversity of approaches obscures which aspects of the reliable spectral properties are essential for producing contrast effects. Specifically, bandwidths and amplitudes of reliable spectral properties are two properties that have varied widely across studies.

SED filters are traditionally calculated across the entire bandwidth of target vowels, which was originally 5000 Hz (Watkins, 1991; Watkins and Makin, 1994, 1996a,b). Subsequent experiments used difference filters spanning 10 000 Hz (Stilp *et al.*, 2010) or as little as 2500 Hz (Sjerps *et al.*, 2011). Tone histories that sampled frequencies across a 1000-Hz-wide region also produced contrast effects (Holt, 2005, 2006), as did later studies that spanned only 435–570 Hz (Laing *et al.*, 2012).[2] Contrast effects reported by Ladefoged and Broadbent (1957) were produced by F_1 shifts of 180–280 Hz. One might conclude that contrast effects can be produced by a spectral regularity with bandwidth spanning only a few hundred Hertz.

Recent studies of perceptual calibration demonstrate sensitivity to reliable spectral peaks that are only 100-Hz wide (Kiefte and Kluender, 2008; Alexander and Kluender, 2010; Stilp and Anderson, 2014). In these studies, the preceding acoustic context (e.g., sentence) was filtered to emphasize energy matching F_2 in the subsequent target vowel (which perceptually varied from /u/ to /i/ and acoustically varied in F_2 and spectral tilt). This filtering made the spectral peak reliable (but not constant) across all sounds on a given trial. Listeners decreased their reliance on this predictable and thus uninformative property of the acoustic environment, increasing their reliance on spectral tilt (an unpredictable and thus informative cue) to identify the target vowel. Alexander and Kluender (2010) note that perceptual calibration and spectral contrast both involve attuning to reliable spectral properties in the acoustic environment; these phenomena are distinguished by whether the reliable spectral property continues through the target sound (calibration; de-emphasis of spectral similarities) or not (contrast; emphasis of spectral differences). Both phenomena demonstrate perceptual sensitivity to reliable spectral shapes (calibration to spectral tilt; contrast to spectral envelope shapes). Finally, contrast effects have been reported when fundamental frequency of the preceding speech context varied across a range of 55–103 Hz (Johnson, 1990; Huang and Holt, 2009). These parallels suggest that spectral contrast effects might be observed when the preceding acoustic context features a 100-Hz-wide reliable spectral peak.

The relative prominence of a reliable spectral property can be characterized by filter gain. If the preceding acoustic context is processed by a high-gain filter to add a large spectral peak, this property will likely affect identification of the subsequent target sound.[3] If the context is processed by a low-gain filter that minimally amplifies a given frequency region, this spectral property might not affect target sound identification at all. The same holds true for broadband spectral regularities imposed upon the acoustic context, such as those introduced by SED filters. Large filter gains (large amplification and attenuation in the filter response) will dramatically reshape the context spectrum, increasing the likelihood of altering identification of the target sound, but small filter gains (minimal amplification and attenuation) will affect target identification minimally if at all.

To establish perceptual phenomena, it is common practice to use robust manipulations that maximize the likelihood of observing the predicted effect. Investigations of spectral contrast effects generally employed high-gain filters, but specific gains have varied widely. Watkins and Makin utilized SED filters with peak gain often +15 dB or more, in some cases as much as +30 dB (Watkins, 1991; Watkins and Makin, 1994, 1996a,b). Difference filters utilized by Sjerps and colleagues (2011) and Sjerps and Smiljanic (2013) exhibited peak gain of +13 to +25 dB. (It bears note that, due to their complex shapes, SED filters can reshape the acoustic context's spectrum in perceptually significant ways at frequencies other than where peak gain is found.) Laing and colleagues (2012) altered context spectra by 20 dB or more in key frequency regions. Huang and Holt (2012) reported contrast effects following contexts that differed by roughly 5–9 dB in key frequency regions, but spectral differences persisted across several kHz, obscuring the contributions of filter gain versus bandwidth. This variability obscures how prominent a spectral regularity must be in order to alter perception of subsequent speech sounds.

The present experiments explore spectral contrast effects in vowel identification resulting from a wide range of reliable spectral properties in the preceding acoustic context. Filter type is manipulated to compare contrast effects following narrow spectral peaks (100 Hz bandwidth, as in studies of perceptual calibration), broad spectral peaks (300 Hz bandwidth, comparable to frequency shifts by Ladefoged and Broadbent, 1957), and SED filters (Watkins, 1991). Filter gain is manipulated to investigate perceptual sensitivity to modest but still reliable spectral properties [+5 to +20 dB for narrowband (NB) and broadband (BB) spectral peaks; 25% to 100% of total spectral difference in SED filters]. Analyses move beyond evaluating the mere presence or absence of (statistically significant) contrast effects by examining perceptual sensitivity to the wide range of spectral regularities presented. Finally, key filter properties (bandwidth, peak gain, mean power) are used to predict contrast effect magnitudes.

II. EXPERIMENTS

A. METHODS

1. Participants

Fifty-six undergraduates were recruited from the Department of Psychological and Brain Sciences at the University of Louisville. All listeners reported being native English speakers with normal hearing and received course credit for their participation.

2. Stimuli

a. Precursor. The precursor sentence was "Please say what this vowel is" spoken by the first author (2174 ms) (see Fig. 1). It was recorded using a Beyerdynamic M88TG microphone (Beyerdynamic, Inc., Farmingdale, NY) in a sound-treated room (Acoustic Systems, Inc., Austin, TX) onto a personal computer with a RME HDSPe AIO sound card (Audio AG, Haimhausen, Germany).

b. Vowels. Vowels were a ten-step continuum perceptually varying from /ɪ/ to /ɛ/. Exemplars of both vowels were recorded by the first author using the same setup as described above. From these recordings, the vowel continuum was created using the procedure described by Winn and Litovsky (2015). Formant contours were extracted using Praat (Boersma and Weenink, 2014) and used to create a ten-step continuum of formant tracks with F_1 and F_2 varying across the entire duration of the vowel, consistent with the natural acoustics of these vowels in English. Formant center frequencies were based on the original recordings. In the /ɪ/ endpoint, F_1 linearly increased from 400 to 430 Hz while F_2 linearly decreased from 2000 to 1800 Hz. In the /ɛ/ endpoint, F_1 linearly decreased from 580 to 550 Hz while F_2 linearly decreased from 1800 to 1700 Hz. Intermediate members of the vowel continuum linearly interpolated between these formant trajectories and center frequencies.

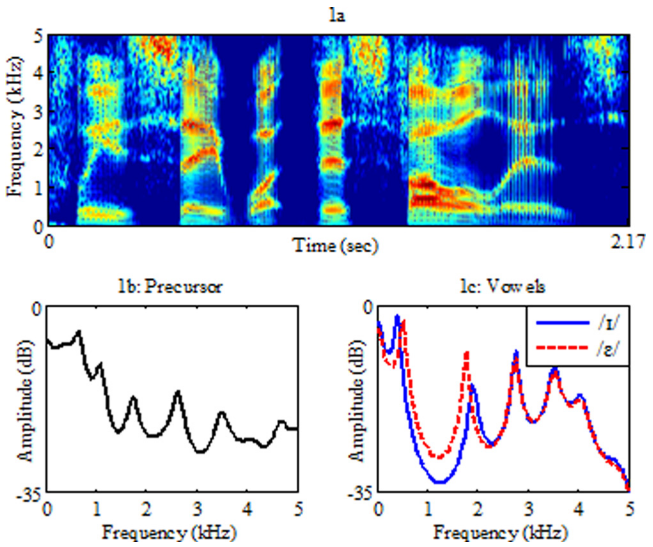

FIG. 1. (Color online) Stimulus materials. (a) Spectrogram of precursor sentence "Please say what this vowel is" (2174 ms). (b) LPC spectrum of precursor sentence. (c) LPC spectra of endpoints of the target vowel series. Solid line depicts spectrum for /ɪ/, dashed line depicts spectrum for /ɛ/.

Formant contours were used as filters to a single voice source extracted from a /ɪ/ token. The spectrum of this vowel was estimated using Burg's linear predictive coding (LPC) procedure, which was then used to inverse filter the token in order to yield the residual voice source. This method is commonly described as a way to remove formant peaks from a speech signal, in order to separate the voice "source" from the vocal tract "filter." The voice source was filtered by each member of the ten-step continuum of formant tracks. Acoustic energy above 2500 Hz in each continuum step was replaced with corresponding energy high-pass-filtered from the original /ɪ/ token. That is, frequencies above 2500 Hz were unaltered throughout the vowel continuum and thus were neutralized as cues. Final stimuli were 246 ms in duration with fundamental frequency set to 100 Hz throughout the vowel.

c. Filters. *1. NB spectral peak.* NB bandpass filters were modeled after those used in investigations of auditory perceptual calibration (Alexander and Kluender, 2010; Stilp and Anderson, 2014). Filter center frequencies were set below F_1 in the /ɪ/ endpoint and above F_1 in the /ɛ/ endpoint to avoid inducing perceptual calibration to spectral peaks that are present throughout the preceding acoustic context and target vowel. Center frequencies were 300 Hz (low F_1) and 650 Hz (high F_1), with filter bandwidths set to 100 Hz. Filter gains decreased from +20 to +5 dB in 5-dB steps [Figs. 2(a)–2(d)]. Mean power in these filters, as measured by root-mean-square (rms) amplitude (Hartmann, 1998), was 1.23 (+20 dB), 0.89 (+15 dB), 0.57 (+10 dB), and 0.27 (+5 dB). Filters were produced using the fir2 function in MATLAB with 1200 coefficients.

2. *BB spectral peak*. Rather than resynthesizing the precursor to emulate different talkers (Ladefoged and Broadbent, 1957; Laing *et al.*, 2012), low-F_1 and high-F_1 frequency regions were amplified using bandpass filters. Bandpass regions of 100–400 Hz (low F_1) and 550–850 Hz (high F_1) were amplified, again beyond F_1 values in vowel continuum endpoints as to avoid shared spectral peaks across the precursor and target sounds. These conditions were labeled BB spectral peaks to distinguish them from narrower spectral peaks (NB). Filter gains decreased from +20 to +5 dB in 5-dB steps [Figs. 2(e)–2(h)]. Mean filter power, as measured by rms amplitude, was 2.47 (+20 dB), 1.85 (+15 dB), 1.23 (+10 dB), and 0.61 (+5 dB). Filters were produced using the fir2 function in MATLAB with 1200 coefficients.

3. *SED*. SED filters were created following the methods of Stilp *et al.* (2010). Spectral envelopes for each vowel endpoint were derived from 1024-point Fourier transforms, which were smoothed using a 512-point Hamming window with 512-point overlap. Spectral envelopes of each endpoint were equated for peak power then subtracted from one another. A 500-point finite impulse response was obtained for each SED via inverse Fourier transform. Filter responses were scaled from 25% to 100% of the total spectral difference between vowel endpoints, varying in steps of 25% [Figs. 2(i)–2(l)]. Mean filter power, as measured by rms amplitude, was 2.35 (100%), 1.76 (75%), 1.17 (50%), and 0.58 (25%).

3. Procedure

Listeners were divided into four groups, each of which was tested on three experimental conditions (Group 1: NB20, BB20, SED100% [$n = 14$]; Group 2: NB15, NB10, NB5 [$n = 15$]; Group 3: BB15, BB10, BB5 [$n = 14$]; Group 4: SED75%, SED50%, SED25% [$n = 13$]). Group 1 was tested first to confirm that reliable spectral properties produced clear spectral contrast effects. Once contrast effects were established, Groups 2–4 were tested in parametric variations on a single reliable spectral property. No one participated in multiple groups.

Precursors and vowel targets were low-pass filtered at 5 kHz, equated in rms-amplitude, and concatenated separated by a 50-ms inter-stimulus interval. Files were up-sampled to 44 100 Hz and presented diotically at 70 dB sound pressure level via circumaural headphones (Beyerdynamic DT-150, Beyerdynamic, Inc., Farmingdale, NY). Listeners participated individually in the same single-wall sound-isolating booths used for stimulus recording. Following acquisition of informed consent, listeners were given instructions and told to respond whether the target vowel sounded more like "ih (as in 'bit')" or "eh (as in 'bet')" on every trial. Experimental conditions were blocked and tested in random order. Each block consisted of 200 trials (10 target vowels × 2 filter conditions [low-F_1 emphasis, high-F_1 emphasis] × 10 repetitions). Each block lasted approximately 12 min, between which listeners took short breaks.

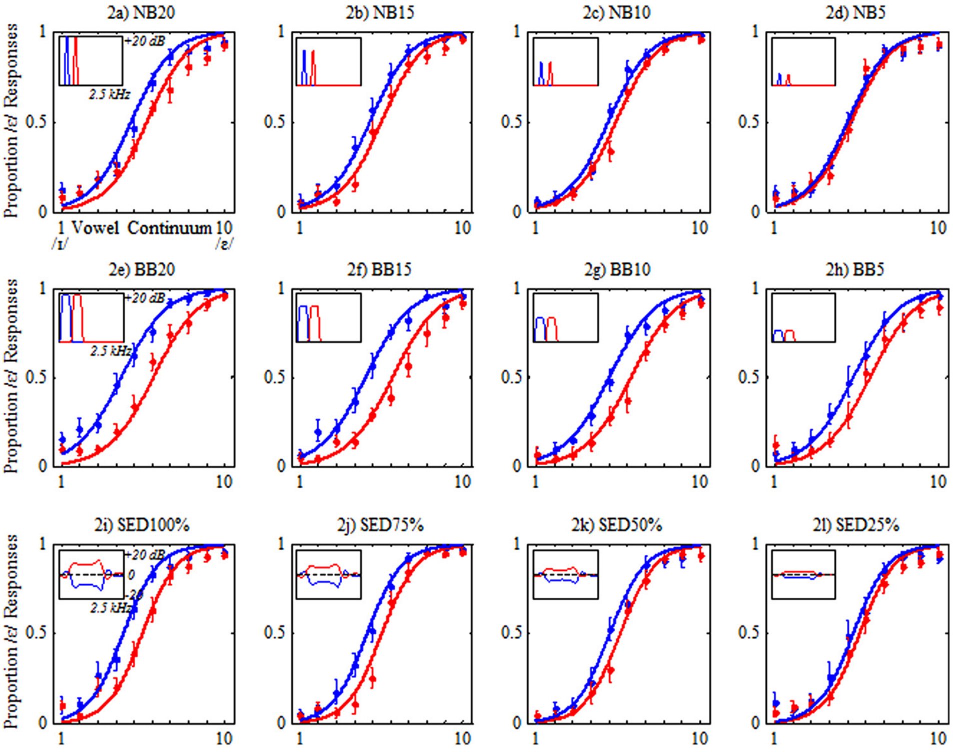

FIG. 2. (Color online) Results organized by filter type [NB spectral peak = (a)–(d), BB spectral peak = (e)–(h), SED = (i)–(l)] and filter gain (columns, arranged in decreasing order). Each ordinate shows proportion of /ε/ responses, and each abscissa shows the ten-step continuum of vowel targets (1 = /ɪ/ endpoint, 10 = /ε/ endpoint). Circles indicate mean response proportions; solid lines indicate logistic regression fits to the data. Figure insets depict filters tested in that condition [ordinates span 0 to + 20 dB filter gain in (a)–(h), ordinates span −20 to +20 dB in (i)–(l); all abscissae span 0–2.5 kHz]. Coloration is consistent across filter responses and behavioral data (i.e., responses to precursors processed by that filter). Error bars represent standard error of the mean.

4. Analysis

Responses were first analyzed on an individual listener basis to identify outliers. For every experimental condition, each listener's responses were analyzed using a generalized linear model in R (R Core Team, 2014). The binomial family call function was used to reflect responses being coded in a binary fashion (0 for "ih" response, 1 for "eh" response). The model had fixed effects of vowel (continuum step) as a continuous numerical factor (1–10), filter frequency (low-F_1 emphasis, high-F_1 emphasis) as a categorical factor, and the interaction between these two factors (denoted by the colon below):

$$\begin{aligned}\text{Response} \sim\ &\text{Vowel} + \text{Filter Frequency}\\ &+ \text{Vowel} : \text{Filter Frequency}.\end{aligned}$$

Model coefficients were used to calculate the midpoints of each psychometric function (i.e., vowel identifications following the low-F_1-filtered precursor or the high-F_1-filtered precursor). If a function midpoint fell outside the range of target vowels presented due to ill-formed functions and/or inability to reliably distinguish vowel continuum endpoints, results were deemed an outlier. In this case, all results for that listener in that condition were removed from further analyses. This occurred in 10 out of 168 data sets ($n=2$ in NB20; $n=2$ in NB15; $n=1$ in BB20; $n=2$ in BB15; $n=1$ in BB5; $n=1$ in SED100%; $n=1$ in SED75%).

Group results were sorted by filter type (NB, BB, SED) and analyzed using generalized linear mixed-effects models (Bates *et al.*, 2012). Model architectures were hypothesis-driven in order to include factors and interactions known to influence speech perception in similarly-designed experiments. Fixed effects included vowel continuum step (coded as a continuous numerical factor, 1–10), filter frequency (low-F_1 or high-F_1 emphasis; coded as a categorical factor), filter gain (peak amplitude in dB or SED in percent; coded as a continuous factor), and the interaction between filter frequency and gain. A random intercept effect of participant was included in the model, listed as (1 | *Participant*) below, and random slopes were included for each fixed effect and interaction (Barr *et al.*, 2013), listed as ({factor} | *Participant*) below. Random-effects structure allows estimation of variance attributable to the participant sample to be partitioned from the estimate of variance attributable to the fixed effects, thus providing a cleaner estimation of the effects of interest (n.b., "random" in this sense implies that the participant pool was randomly sampled from a larger population, to which these results are intended to be generalized). The final model had the following form:

$$\begin{aligned}\text{Response} \sim \text{Vowel} &+ \text{Filter Frequency} + \text{Filter Gain}\\ &+ \text{Filter Frequency} : \text{Filter Gain}\\ &+ (1 + \text{Vowel} + \text{Filter Frequency}\\ &+ \text{Filter Gain}\\ &+ \text{Filter Frequency} : \text{Filter Gain} \,|\, \text{Participant}).\end{aligned}$$

TABLE I. Mixed-effects model results for NB experiments. "Vowel continuum step" refers to the slope of the psychometric function, defined as the change in log odds of the listener's response resulting from a change of one step in the vowel continuum. "Filter frequency shift" lists the change in log odds of the listener's response resulting from changing the reliable spectral peak in the preceding sentence from high F_1 (600–700 Hz) to low F_1 (250–350 Hz). "Filter gain (in dB)" lists the change in log odds of the listener's response resulting from increasing peak filter gain by 1 dB. "Filter gain (in dB): filter frequency shift" indicates the change in the size of the filter frequency shift effect (i.e., contrast effect) per dB of filter gain.

NB Model Term	Estimate	SE	*z*	*p*
Intercept	−3.787	0.190	−19.93	<0.01
Vowel continuum step	0.758	0.014	56.16	<0.01
Filter frequency shift	−0.026	0.135	−0.19	0.85
Filter gain (in dB)	−0.028	0.013	−2.22	0.03
Filter gain (in dB): filter frequency shift	0.034	0.011	3.22	<0.01

TABLE II. Mixed-effects model results for BB experiments. "Filter frequency shift" lists the change in log odds of the listener's response resulting from changing the reliable spectral peak in the preceding sentence from high F_1 (550–850 Hz) to low F_1 (100–400 Hz). All other items have the same description as in Table I.

BB Model Term	Estimate	SE	*z*	*p*
Intercept	−4.334	0.315	−13.73	<0.01
Vowel continuum step	0.695	0.013	53.47	<0.01
Filter frequency	0.385	0.175	2.21	0.03
Filter gain (in dB)	0.005	0.018	0.27	0.79
Filter gain (in dB): filter frequency shift	0.044	0.014	3.15	<0.01

B. Results

Behavioral results are presented in Fig. 2, and statistical model results are listed in Tables I–III. Table I lists model results for experiments with NB filters, Table II lists model results for experiments with BB filters, and Table III lists model results for experiments with SED filters. Intercept terms refer to the log odds of perceiving /ɛ/ at any fixed level of the other factors (starting at an extreme /ɪ/ in the default model). Slope terms correspond to the change in log odds of the listener's response attributable to a change in one stimulus step along the vowel continuum. Filter frequency shift terms indicate changes in psychometric function intercepts for low-F_1 versus high-F_1 filters. Filter gain terms indicate changes in the intercept of the high-F_1-filtered psychometric

TABLE III. Mixed-effects model results for SED experiments. "Filter frequency shift" lists the change in log odds of the listener's response resulting from changing the SED filter from /ɛ/ − /ɪ/ (higher F_1 peak) to /ɪ/ − /ɛ/ (lower F_1 peak). "Filter gain (in % of total SED)" lists the estimated change in log odds of the listener's response resulting from increasing the SED tested by 1%. "Filter gain (in % of total SED): filter frequency shift" indicates the change in the size of the filter frequency shift effect (i.e., contrast effect) per percent of total SED tested.

SED Model Term	Estimate	SE	*z*	*p*
Intercept	−4.800	0.211	−22.78	<0.01
Vowel continuum step	0.836	0.016	53.65	<0.01
Filter frequency shift	0.183	0.171	1.07	0.28
Filter gain (in % of total SED)	0.003	0.002	1.30	0.19
Filter gain (in % of total SED): filter frequency shift	0.006	0.002	2.48	0.01

function for each 1-dB (NB, BB) or 1% (SED) increase in filter gain. Of central importance is the filter gain by filter frequency shift interaction, which conveys changes in contrast effect magnitudes at different amounts of filter gain. This interaction was statistically significant in each model and is evident in the regression fits to behavioral data in Fig. 2; listeners exhibited progressively smaller contrast effects with lower amounts of filter gain. This is an important departure from previous considerations of contrast effects as being dichotomous in nature (present/absent, or statistically significant/failing to achieve statistical significance).

While model coefficients in Tables I–III can be used to estimate contrast effect magnitudes for specific conditions, they cannot indicate which contrast effects significantly differed from zero or from each other. For *post hoc* analyses, models were reanalyzed with filter gain coded as a categorical factor. Categorical coding selects one level of filter gain as the baseline condition, then uses Wald z-tests to test its model coefficient against 0 (i.e., whether the contrast effect in that condition significantly differed from 0) and against other levels of filter gain (i.e., pairwise comparisons). By rotating through each level of filter gain as the baseline condition, each level is tested against 0 and against every other level.

Table IV lists model coefficient estimates and contrast effect magnitudes for each of the 12 conditions. Contrast effect magnitudes are operationalized as the distance between psychometric function 50% points measured in stimulus steps along the vowel continuum. Fifty-percent thresholds were derived from the inverse logit function.[4] Tests of statistical significance used one-tailed Wald z-tests, as directionality was predicted *a priori* (contrast effects being significantly greater than 0, or larger contrast effects for greater amounts of filter gain as indicated by the significant interactions in Tables I–III). Contrast effects were statistically significant (i.e., greater than 0) in 11 of 12 experimental conditions; only NB5 filters failed to significantly shift listeners' responses.

TABLE IV. Model coefficients and contrast effect magnitudes for all filter types and filter gains. Mixed effects models were analyzed with filter gain coded categorically and each given level of filter gain set as the default level. This produces a Wald z-test of that model coefficient against zero, which tests contrast effect magnitude (the number of stimulus steps separating 50% points on psychometric functions) against zero. $^{**}p < 0.01$, $^{***}p < 0.001$.

Filter	Model Estimate	Contrast Effect
NB20	0.55	0.72^{***}
NB15	0.61	0.81^{***}
NB10	0.39	0.51^{***}
NB5	0.07	0.09
BB20	1.07	1.54^{***}
BB15	1.26	1.80^{***}
BB10	0.87	1.24^{***}
BB5	0.54	0.78^{***}
SED100%	0.80	0.94^{***}
SED75%	0.79	0.93^{***}
SED50%	0.37	0.43^{**}
SED25%	0.40	0.47^{**}

Pairwise comparisons revealed the following results: for NB filters, NB5 produced smaller contrast effects than all larger filter gains ($z > 2.13$, $p < 0.025$), which did not significantly differ from each other ($z < 1.46$, $p > 0.05$). For BB filters, BB5 produced smaller contrast effects than BB20 ($z = 2.87$, $p < 0.01$), and effects grew larger across BB5, BB10, and BB15, which all differed from each other ($z > 1.95$, $p < 0.05$). For SED filters, SED100% ($z > 1.96$, $p < 0.05$) and SED75% ($z > 1.96$, $p < 0.05$) each produced larger contrast effects than SED50% and SED25%, but the two larger filter gains did not differ from each other ($z = 0.02$, $p > 0.05$) nor did the two smaller filter gains differ from each other ($z = 0.17$, $p > 0.05$). Contrast effects were slightly larger for NB15 than NB20 and for BB15 than BB20, and were comparable across SED75% and SD100%. This lack of clear monotonicity may be attributable to testing different groups of listeners across these "neighboring" conditions.

Behavioral results (Fig. 2) and significant interactions between filter gain (coded as a continuous variable) and filter frequency (Tables I–III) suggest that contrast effect magnitude varied linearly across different levels of filter gain. Pearson correlation analyses were used to assess how well different filter properties predicted contrast effect magnitudes. Filter bandwidth was a poor predictor [$r = -0.19$, $p > 0.05$; Fig. 3(a)] likely because of its restriction to only three possible values (100 Hz for NB, 300 Hz for BB, 2500 Hz for SED. Peak filter gain was positively correlated with contrast effect magnitude [$r = 0.61$, $p < 0.05$; Fig. 3(b)], suggesting that more prominent spectral regularities elicited larger contrast effects. However, this predictor is limited by its ignorance of filter type and bandwidth. Contrast effects in BB conditions are markedly larger than those in NB conditions, but these results share peak filter gain. Mean filter power was the best predictor of contrast effect magnitude [$r = 0.74$, $p < 0.01$; Fig. 3(c)]. Peak filter gain and mean filter power are clearly not independent of each other ($r = 0.63$, $p < 0.05$). However, when filter properties were entered into a multiple regression predicting contrast effect magnitude, mean filter power was a significant predictor ($t = 3.39$, $p < 0.01$) while peak gain ($t = -1.02$, $p > 0.05$) and bandwidth were not ($t = -2.22$, $p > 0.05$), confirming mean filter power as the best predictor of contrast effect magnitude.

III. GENERAL DISCUSSION

Speech perception was systematically altered when the preceding sentence featured reliable spectral properties in the form of either a narrowband spectral peak, broadband spectral peak, or a complex spectral envelope reflecting differences between spectra of target vowel continuum endpoints (SED). Contrast effects were observed in nearly every condition tested, including when the precursor was processed by filters that spanned only 100 Hz in bandwidth, were only +5 dB, or reflected only one-quarter of the difference

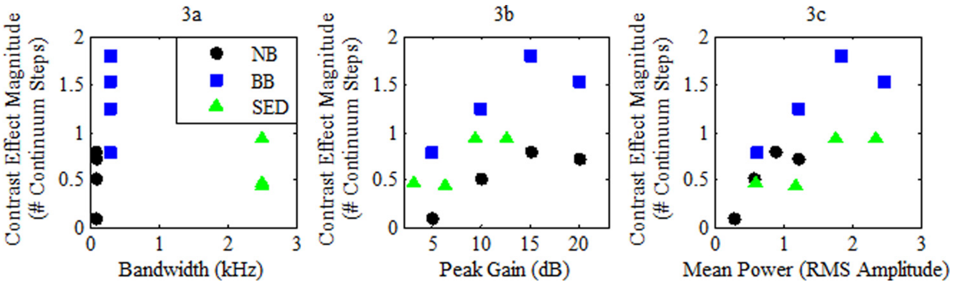

FIG. 3. (Color online) Scatterplots comparing contrast effect magnitudes (quantified as the number of stimulus steps separating 50% points on psychometric functions) to filter properties. Circles indicate contrast effect magnitudes in NB spectral peak conditions, squares for BB spectral peak conditions, triangles for SED conditions. (a) Contrast effect magnitude was not correlated with filter bandwidth ($r = -0.19$, $p > 0.05$), which was restricted to 100 (NB), 300 (BB), or 2500 Hz (SED). (b) Effects were positively correlated with peak filter gain, indicating larger contrast effects following higher-gain filters ($r = 0.61$, $p < 0.05$). (c) Mean filter power was the best predictor of contrast effect magnitude, indicating that filters with larger average power produced larger contrast effects ($r = 0.74$, $p < 0.01$).

between target vowel spectra. Reliable spectral properties need not be particularly prominent in order for the auditory system to extract and exploit them. Given this acute sensitivity, contrast effects may be more pervasive in speech perception than previously thought.

Acute sensitivity to spectral contrast relates to perception of speech amidst spectrally reflective surfaces such as walls in rooms, where amplification of resonant frequencies can create reliable spectral peaks. This also impacts speech perception over media with transfer functions that add even small spectral peaks to the signal, such as headphones, mobile phones, and loudspeakers. Altering speaking style can also add reliable spectral peaks to speech (Ladefoged, 1989). In all of these situations, it is important that the auditory system avoid mistaking predictable properties of the environment or communication medium as speech information. Accordingly, listeners are sensitive to *deviations* from reliable spectral features, both prominent and subtle.

Results extend a long history of spectral contrast effects in speech perception (see Sec. I). Across these studies, contrast effects have been operationalized by changes in response rates, error rates, phoneme boundary shifts, perceived changes in target frequency (e.g., 40 Hz change in F_1, 10 Hz change in fundamental frequency, etc.), or in cases of complex stimulus series, the number of stimulus steps separating response functions (Table IV). While such quantification puts contrast effect magnitudes in context, they have historically been treated as dichotomous: present (statistically significant differences in performance) or absent (differences in performance being absent or falling below the threshold of statistical significance). Two exceptions are those of Holt (2006), who reported larger contrast effects when the preceding acoustic context was comprised of longer tone sequences, and Holt and Lotto (2002), where contrast effect magnitudes decreased as silent inter-stimulus interval between context and target stimuli increased. These exceptions notwithstanding, no efforts attempted to predict contrast effect magnitude, addressing whether this perceptual response is dichotomous or continuous in nature. Here, contrast effects are revealed to be continuous, with magnitudes strongly correlated with total power of the reliable spectral property (i.e., mean filter power). This is an important discovery for understanding the occurrence of this phenomenon and the precise extent of its influence on perception.

Contrast effect magnitude corresponded poorly to bandwidth of the reliable spectral property, corresponded well to peak filter gain, but was best predicted by mean filter power (Fig. 3). This is certainly not an exhaustive list of properties of spectral regularities that influence speech sound identification. Duration of the preceding acoustic context, held constant in the present experiments, systematically influences speech perception. Holt (2006) reported larger contrast effects when longer-duration tone histories were presented. Alexander and Kluender (2010) reported modestly greater perceptual calibration to reliable spectral tilt in longer-duration precursors. They then revealed that it was not precursor duration *per se* that increased perceptual calibration, but opportunities to sample the precursor spectrum. Calibration to spectral tilt increased when precursor duration was held constant and rate of spectro-temporal modulations in their nonspeech precursors increased.[5] The present results are broadly consistent with these findings, as increasing the opportunities to sample the reliable spectral property (increasing its prominence vis-à-vis increased filter power) resulted in larger contrast effects. Further research is needed to understand what other characteristics of reliable spectral properties modulate perceptual responses.

Contrast effects following SED filters are qualitatively different than those following NB or BB filters. Despite mean filter power being well matched to that of BB filters, SED contrast effects are more modest and grow more slowly with increasing filter power [Fig. 3(c)]. Indeed, the correlation between mean filter power and contrast effect magnitude increases substantially when analyzing only NB and BB results ($r = 0.88$, $p < 0.005$). Two points might explain this discrepancy. First, increases in SED filter power are distributed along a wide frequency range which includes frequency regions that are less important for distinguishing /ɪ/ from /ɛ/ [Figs. 2(i)–2(l)]. This is contrary to NB and BB conditions, where increases in filter power occur only in narrow frequency regions that are sufficient for producing contrast effects [Figs. 2(a)–2(h)]. This might explain why contrast effect magnitude in the SED100% condition (where peak filter gain in F_1 regions only reaches +12.5 dB) is more comparable to those in BB10 and NB15 conditions than in BB20 and NB20 conditions [Fig. 3(b)]. Second, contrast effect magnitude may differ according to different methods of creating vowel continua. Watkins and colleagues (1991) and Watkins and Makin (1994, 1996a,b) interpolated between spectral envelopes of vowel endpoints to create their target vowel continuum, producing ambiguous mid-continuum stimuli with modest spectral peaks. Here, mid-continuum stimuli had clear formant peaks at frequencies intermediate to the canonical endpoints. Vowel identification might be more susceptible to SED filtering for ambiguous stimuli with modest formant peaks (*a la* Watkins) rather than clear peaks (as tested here).

SED50% filters produced smaller contrast effects than SED100% filters ($z = -2.13$, $p < 0.05$). This replicates the results of Watkins and Makin (1996a), who reported smaller phonemic boundary shifts when carrier contrast ratio (difference between spectral envelopes used in the difference filter) was half that of target vowel contrast ratio. Watkins and Makin concluded that compensation for spectral envelope distortion does not involve extraction of contrast-invariant spectral "features" such as formant peaks, as formant center frequencies did not change across different contrast ratios. Given the significant correlation between mean filter power and contrast effect magnitude, the auditory system does appear to be extracting some spectral feature(s) from the acoustic context, even if they are broadly defined. These features are clearly not contrast-invariant, as decreasing mean filter power resulted in smaller contrast effects. This pattern was observed for all regularities tested (NB, BB, SED), suggesting a broadly tuned process for extracting and exploiting reliable spectral properties in a listening context.

Given that target stimuli were vowels, one might situate the present results within theories of vowel normalization. Vowel normalization is traditionally viewed as relying on intrinsic (i.e., specified within the target vowel or syllable, such as fundamental and/or formant frequencies) and extrinsic factors (i.e., external to the target vowel or syllable, such as spectral properties of preceding sounds or the vowel system of a given talker). Both approaches contribute to vowel normalization, but extrinsic effects appear to influence vowel perception more than intrinsic effects (Ainsworth, 1975; Nearey, 1989). Intrinsic and extrinsic information play either direct roles in vowel normalization (information is used directly in perceptual representation of the vowel) or indirect roles (to establish a frame of reference against which other vowels are compared; see Johnson, 1990 for discussion). Reliable spectral properties in the preceding sentence may be viewed as extrinsic direct information for vowel normalization, much in the same way that Johnson (1990) classified the seminal findings of Ladefoged and Broadbent (1957).

However, it might be inappropriate to label these effects as extrinsic direct vowel normalization or perhaps vowel normalization at all. Consider the classic findings by Ladefoged and Broadbent (1957): when frequencies appropriate for F_1 were shifted downward in the precursor sentence (making it sound more /ɪ/-like), participants reported more /ɛ/ percepts in the target word, and vice versa. This effect has been reproduced by shifting key frequency ranges (Ladefoged and Broadbent, 1957; Laing *et al.*, 2012), altering speaking style (Ladefoged, 1989), and by changing talkers (Dechovitz, 1977). In BB conditions, contrast effects were observed through simple amplification of key frequency regions without any explicit consideration of the talker's formant frequencies. Remez *et al.* (1987) replicated this effect when materials were sinewave replicas of speech, leading them to question whether they falsified Ladefoged and Broadbent (1957) rather than replicated them since their results could not be attributed to talker information that is normally available in speech production. Recently, Laing and colleagues (2012) replicated this effect when the acoustic context consisted of simple tone histories (see also Holt, 2005, 2006; Huang and Holt, 2012), which contained no vocal tract characteristics to which listeners could normalize. This effect is not specific to vowel perception, also being observed in identification of stop consonants and fricatives (Watkins and Makin, 1996b; Laing *et al.*, 2012). This effect is not even specific to speech perception, also being observed for identification of musical instruments (Stilp *et al.*, 2010). Consistent patterns of results are observed due to reliable spectral properties in the precursor sounds, whether they are speech or nonspeech, and whether target sounds are speech or nonspeech.

In similar experiments, Laing *et al.* (2012) proposed that the auditory system normalizes to stable signal properties (e.g., long-term average spectrum) rather than talker characteristics such as vocal tracts. This rekindles the question of how much of talker-specific or environment-specific adaptation can be explained by low-level phenomena, and whether those effects require higher-level processes such as talker identification or perception of physical gestures. Talker information is not necessary to produce spectral contrast effects in phoneme identification (see contrast effects following nonspeech precursors in Watkins, 1991; Holt, 2005, 2006; Sjerps *et al.*, 2011; Huang and Holt, 2012; Laing *et al.*, 2012). Yet, many experimental paradigms (including the present one) are designed to reveal spectral contrast effects that arise from low-level signal properties but are *not* designed to assess higher-level influences on performance either within or across trials. Speech perception is influenced by a wide variety of top-down factors, including talker familiarity (Creelman, 1957; Mullenix *et al.*, 1989; Nygaard and Pisoni, 1998), lexicality (Ganong, 1980; Samuel, 2001; Norris *et al.*, 2003), and expectations of talker acoustics (Johnson *et al.*, 1999; Sohoglu *et al.*, 2012), the last of which can be moderated by hearing impairment (Winn *et al.*, 2013). None of these effects can be explained by low-level auditory processes alone, and few studies are designed to test both low- and high-level processes concurrently. Bottom-up and top-down influences on speech perception are more likely interactive than they are exclusive (McClelland *et al.*, 2006). It is possible (or perhaps likely) that the effects identified in the current study supplement and interact with those that occur in the process of identifying talkers, predicting words, and relying on linguistic experience (e.g., Elman and McClelland, 1998).

Sensory systems capture reliable aspects of the sensory environment in order to be optimally sensitive to changing, more informative inputs. On short timescales, neural adaptation maximizes information transmission for the organism, indicating predictability in the environment while conserving neural resources for when inputs change and convey new information (Wainwright, 1999; Clifford *et al.*, 2007; Kohn, 2007). On longer timescales, neural response properties evolve to capture stable statistical structure in the environment. This approach has been highly fruitful for understanding vision (see Schwartz and Simoncelli, 2001; Simoncelli, 2003; Geisler, 2008 for reviews), and recent research shows congruence between statistical regularities in natural sounds and auditory processing and/or perception. Perception of sound textures is mediated by their time-averaged statistical properties (McDermott *et al.*, 2013). Amplitude modulations in natural sounds such as speech and music follow a $1/f$ distribution (Voss and Clarke, 1975), and neural sensitivity to this distribution increases along the ascending auditory pathway (Garcia-Lazaro *et al.*, 2011). The statistical structure of the acoustics of human speech is well captured by response properties at the mammalian auditory nerve (Lewicki, 2002; Stilp and Assgari, 2015), and this congruence appears to continue when comparing the statistics of speech sound classes to response properties in the cochlear nucleus (Stilp and Lewicki, 2014). Means and variances of stimulus distributions are captured by response properties of auditory midbrain neurons, shifting to capture changes in these distributional properties (Dean *et al.*, 2005). Finally, diminished neural responses to repeated sounds and enhanced responses to unexpected sounds have been thoroughly documented in stimulus-specific adaptation (SSA; Ulanovsky *et al.*, 2003). SSA occurs throughout the central auditory system (Ulanovsky *et al.*, 2003; Pérez-González *et al.*, 2005; Anderson *et al.*, 2009; Malmierca *et al.*, 2009) and is of particular relevance to the present results, as Holt (2006) suggested SSA may underlie the types of effects observed here (but see Kingston *et al.*, 2014). Precursors conveyed a reliable property in their long-term average spectra (spectral peak or global spectral shape). The auditory system extracted this stable property, maintaining maximal sensitivity for when this property *changed*, as it did upon introduction of the vowel target. This difference (akin to "deviant" trials in SSA experiments) was perceptually magnified, resulting in contrast effects. The above examples and present experiments operate on varying timescales but share a common process: identifying and exploiting stable properties of the acoustic environment.

The present results reveal remarkable perceptual sensitivity to reliable spectral properties in a listening context. This process has had many names in the literature: normalization (Dechovitz, 1977; Remez *et al.*, 1987; Johnson, 1990; Mitterer, 2006; Huang and Holt, 2009, 2012; Laing *et al.*, 2012; Sjerps *et al.*, 2011; Sjerps *et al.*, 2012, 2013), calibration (Kiefte and Kluender, 2008; Alexander and Kluender, 2010; Stilp and Anderson, 2014), compensation (Watkins, 1991; Watkins and Makin, 1994, 1996a,b; Sjerps *et al.*, 2011; Sjerps and Smiljanic, 2013), perceptual constancy (Ladefoged and Broadbent, 1957; Holt, 2006; Stilp *et al.*, 2010), adaptive coding (Huang and Holt, 2012), and inverse filtering (Watkins, 1991; Watkins and Makin, 1994, 1996a,b). Except for the case of perceptual calibration where reliable spectral properties are shared across context and target, all other monikers describe the same phenomenon: differences between reliable spectral properties in the preceding acoustic context and the spectrum of the subsequent target sound are perceptually enhanced, resulting in contrast effects. This phenomenon has been observed using a wide range of materials, and in the present studies, for a wide range of spectral regularities, including spectral prominences as narrow as 100 Hz or as modest as +5 dB. Emergence of contrast effects in the presence of such subtle spectral regularities suggests that such effects are likely to occur in a wide variety of listening situations, and therefore might influence speech perception more frequently than previously considered.

ACKNOWLEDGMENTS

The authors thank Kara Hendrix, Elizabeth Niehaus, McKenzie Sexton, and Lillian Slaughter for assistance with data collection.

[1] Throughout this paper, spectral contrast effects refer to those following a reliable spectral property in the preceding acoustic context. This is distinct from similarly-named effects that occur only between the offset of a preceding sound and onset of the target sound (e.g., Lotto and Kluender, 1998). As no reliable spectral property is present in those studies, that class of contrast effects is not discussed further. The present effects are also distinct from investigations of perceptual sensitivity to level differences between spectral peaks and valleys, which are also termed spectral contrast (e.g., Leek *et al.*, 1987; Baer *et al.*, 1993). Formally speaking, Watkins and Makin (1996a) examined the influence of spectral contrast (peak-to-valley differences) on spectral contrast effects (phonemic boundary shifts), but we resist using this description to avoid confusion.

[2] Experiment 1 of Holt (2006) manipulated the variability of frequencies sampled in tone histories, concurrently varying total bandwidth of the tone history. Comparable contrast effects were reported when tone histories spanned 1000, 300, 100, or 1 Hz (repetition of one frequency). However, as frequency variability/bandwidth decreased, density of sampling that frequency region increased dramatically. Acoustic energy in any narrow frequency region of speech waxes and wanes across time, and is not consistent for such extended periods (2100 ms) as in these tone histories. Results obtained using narrowband tone histories make a poor comparison to those using speech precursors given the discrepancy in constant versus intermittent evidence for acoustic energy in a certain frequency region.

[3] Here and throughout, frequency specificity of filters is assumed. If the preceding acoustic context contains reliable spectral peaks that are not contrastive with the target sound (e.g., positioned at remote frequencies), contrast effects will not be observed (Laing *et al.*, 2012).

[4] The 50% threshold in the inverse logit equation can be calculated by predicting the log odds of 0. Here, 50% thresholds were calculated as −Intercept/Slope for the default high-F_1-filtered psychometric function and −(Intercept + Filter frequency shift)/Slope for the low-F_1-filtered psychometric function.

[5] Alexander and Kluender (2010) reported unexpectedly diminished calibration to a reliable spectral peak common to vowel target and longer-duration preceding acoustic context. They attributed this result to acute sensitivity to acoustic onsets and repetition of acoustic information throughout the trial. Additional research is needed to understand potentially different roles of context duration when narrowband spectral similarities are being de-emphasized (calibration) versus when narrowband spectral differences are being emphasized (contrast).

Ainsworth, W. (**1975**). "Intrinsic and extrinsic factors in vowel judgments," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. Tatham (Academic, London), pp. 103–113.

Alexander, J. M., and Kluender, K. R. (**2010**). "Temporal properties of perceptual calibration to local and broad spectral characteristics of a listening context," J. Acoust. Soc. Am. **128**(6), 3597–3613.

Anderson, L. A., Christianson, G. B., and Linden, J. F. (**2009**). "Stimulus-specific adaptation occurs in the auditory thalamus," J. Neurosci. **29**(22), 7359–7363.

Baer, T., Moore, B. C. J., and Gatehouse, S. (**1993**). "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," J. Rehab. Res. Develop. **30**(1), 49–72.

Barlow, H. B. (**1961**). "Possible principles underlying the transformations of sensory messages," in *Sensory Communication*, edited by W. A. Rosenblith (MIT Press, Cambridge, MA and John Wiley, New York), pp. 53–85.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," J. Mem. Lang. **68**(3), 255–278.

Bates, D. M., Maechler, M., and Bolker, B. (**2012**). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

Boersma, P., and Weenink, D. (**2014**). "Praat: Doing phonetics by computer [Computer program]," Version 5.3.61, retrieved January 1, 2014 from http://www.praat.org/ (Last viewed August 5, 2014).

Broadbent, D. E., and Ladefoged, P. (**1960**). "Vowel judgements and adaptation level," Proc. R. Soc. B **151**, 384–399.

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., and Schwartz, O. (**2007**). "Visual adaptation: Neural, psychological and computational aspects," Vis. Res. **47**(25), 3125–3131.

Creelman, C. D. (**1957**). "Case of the unknown talker," J. Acoust. Soc. Am. **29**, 655.

Dean, I., Harper, N. S., and McAlpine, D. (**2005**). "Neural population coding of sound level adapts to stimulus statistics," Nature Neurosci. **8**(12), 1684–1689.

Dechovitz, D. (**1977**). "Information conveyed by vowels: A confirmation," Haskins Lab Status Report, SR-51/52, pp. 213–219.

Elman, J., and McClelland, J. (**1998**). "Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes," J. Mem. Lang. **27**(2), 143–165.

Ganong, W. F. (**1980**). "Phonetic categorization in auditory word perception," J. Exp. Psychol. **6**(1), 110–125.

Garcia-Lazaro, J. A., Ahmed, B., and Schnupp, J. W. (**2011**). "Emergence of tuning to natural stimulus statistics along the central auditory pathway," PLoS One **6**(8), e22584.

Geisler, W. S. (**2008**). "Visual perception and the statistical properties of natural scenes," Ann. Rev. Psych. **59**, 167–192.

Hartmann, W. M. (**1998**). *Signals, Sound, and Sensation* (AIP Press, Woodbury, NY), 647 pp.

Holt, L. L. (**2005**). "Temporally nonadjacent nonlinguistic sounds affect speech categorization," Psych. Sci. **16**(4), 305–312.

Holt, L. L. (**2006**). "The mean matters: Effects of statistically defined non-speech spectral distributions on speech categorization," J. Acoust. Soc. Am. **120**(5), 2801–2817.

Holt, L. L., and Lotto, A. J. (**2002**). "Behavioral examinations of the level of auditory processing of speech context effects," Hear. Res. **167**, 156–169.

Huang, J., and Holt, L. L. (**2009**). "General perceptual contributions to lexical tone normalization," J. Acoust. Soc. Am. **125**(6), 3983–3994.

Huang, J., and Holt, L. L. (**2012**). "Listening for the norm: Adaptive coding in speech categorization," Front. Psychol. **3**, 10.

Johnson, K. (**1990**). "The role of perceived speaker identity in F0 normalization of vowels," J. Acoust. Soc. Am. **88**(2), 642–654.

Johnson, K., Strand, E. A., and D'Imperio, M. (**1999**). "Auditory–visual integration of talker gender in vowel perception," J. Phonetics **27**(4), 359–384.

Kiefte, M., and Kluender, K. R. (**2008**). "Absorption of reliable spectral characteristics in auditory perception," J. Acoust. Soc. Am. **123**(1), 366–376.

Kingston, J., Kawahara, S., Chambless, D., Key, M., Mash, D., and Watsky, S. (**2014**). "Context effects as auditory contrast," Atten. Percept. Psychophys. **76**, 1437–1464.

Kohn, A. (**2007**). "Visual adaptation: Physiology, mechanisms, and functional benefits," J. Neurophys. **97**(5), 3155–3164.

Ladefoged, P. (**1989**). "A note on 'Information conveyed by vowels,'" J. Acoust. Soc. Am. **85**(5), 2223–2224.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**(1), 98–104.

Laing, E. J., Liu, R., Lotto, A. J., and Holt, L. L. (**2012**). "Tuned with a tune: Talker normalization via general auditory processes," Front. Psychol. **3**, 203.

Leek, M. R., Dorman, M. F., and Summerfield, Q. (**1987**). "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **81**(1), 148–154.

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

Lotto, A. J., and Kluender, K. R. (**1998**). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," Percept. Psychophys. **60**(4), 602–619.

Malmierca, M. S., Cristaudo, S., Perez-Gonzalez, D., and Covey, E. (**2009**). "Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat," J. Neurosci. **29**(17), 5483–5493.

McClelland, J. L., Mirman, D., and Holt, L. L. (**2006**). "Are there interactive processes in speech perception?," Trends Cogn. Sci. **10**(8), 363–369.

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

Mitterer, H. (**2006**). "Is vowel normalization independent of lexical processing?," Phonetica **63**(4), 209–229.

Mullennix, J., Pisoni, D. B., and Martin, C. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**(1), 365–378.

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**(5), 2088–2113.

Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cog. Psychol. **47**(2), 204–238.

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**(3), 355–376.

Pérez-González, D., Malmierca, M. S., and Covey, E. (**2005**). "Novelty detector neurons in the mammalian auditory midbrain," Eur. J. Neurosci. **22**(11), 2879–2885.

R Core Team (**2014**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (Last viewed August 5, 2014).

Remez, R. E., Rubin, P. E., Nygaard, L. C., and Howell, W. A. (**1987**). "Perceptual normalization of vowels produced by sinusoidal voices," J. Exp. Psych.: Human Percept. Perf. **13**(1), 40–61.

Samuel, A. G. (**2001**). "Knowing a word affects the fundamental perception of the sounds within it," Psych. Sci. **12**(4), 348–351.

Schwartz, O., and Simoncelli, E. P. (**2001**). "Natural signal statistics and sensory gain control," Natural Neurosci. **4**(8), 819–825.

Simoncelli, E. P. (**2003**). "Vision and the statistics of the visual environment," Curr. Op. Neurobiol. **13**(2), 144–149.

Sjerps, M. J., McQueen, J. M., and Mitterer, H. (**2012**). "Extrinsic normalization for vocal tracts depends on the signal, not on attention," Proc. Interspeech **2012**, 394–397.

Sjerps, M. J., McQueen, J. M., and Mitterer, H. (**2013**). "Evidence for precategorical extrinsic vowel normalization," Att. Percept. Psychophys. **75**(3), 576–587.

Sjerps, M. J., Mitterer, H., and McQueen, J. M. (**2011**). "Constraints on the processes responsible for the extrinsic normalization of vowels," Att. Percept. Psychophys. **73**(4), 1195–1215.

Sjerps, M. J., and Smiljanic, R. (**2013**). "Compensation for vocal tract characteristics across native and non-native languages," J. Phonetics **41**(3–4), 145–155.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (**2012**). "Predictive top-down integration of prior knowledge during speech perception," J. Neurosci. **32**(25), 8443–8453.

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

Stilp, C. E., and Anderson, P. W. (**2014**). "Modest, reliable spectral peaks in preceding sounds influence vowel perception," J. Acoust. Soc. Am. **136**(5), EL383–EL389.

Stilp, C. E., and Assgari, A. A. (**2015**). "Does the auditory system efficiently code all languages or just American English?," Assoc. Research Otolaryn. Abstracts 38(A).

Stilp, C. E., and Lewicki, M. S. (**2014**). "Statistical structure of speech sound classes is congruent with cochlear nucleus response properties," Proc. Meet. Acoust. **20**, 050001.

Ulanovsky, N., Las, L., and Nelken, I. (**2003**). "Processing of low-probability sounds by cortical neurons," Nature Neurosci. **6**(4), 391–398.

Voss, R. F., and Clarke, J. (**1975**). "1-over-f-noise in music and speech," Nature **258**(5533), 317–318.

Wainwright, M. J. (**1999**). “Visual adaptation as optimal information transmission,” Vis. Res. **39**(23), 3960–3974.

Watkins, A. J. (**1991**). “Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion,” J. Acoust. Soc. Am. **90**(6), 2942–2955.

Watkins, A. J., and Makin, S. J. (**1994**). “Perceptual compensation for speaker differences and for spectral-envelope distortion,” J. Acoust. Soc. Am. **96**(3), 1263–1282.

Watkins, A. J., and Makin, S. J. (**1996a**). “Effects of spectral contrast on perceptual compensation for spectral-envelope distortion,” J. Acoust. Soc. Am. **99**(6), 3749–3757.

Watkins, A. J., and Makin, S. J. (**1996b**). “Some effects of filtered contexts on the perception of vowels and fricatives,” J. Acoust. Soc. Am. **99**(1), 588–594.

Winn, M., Rhone, A., Chatterjee, W., and Idsardi, W. (**2013**). “The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants,” Front. Psychol. **4**, 824.

Winn, M. B., and Litovsky, R. Y. (**2015**). “Using speech sounds to test functional spectral resolution in listeners with cochlear implants,” J. Acoust. Soc. Am. **137**, 1430–1442.