JARO (2017)
DOI: 10.1007/s10162-017-0615-y
© 2017 Association for Research in Otolaryngology

JARO
Journal of the Association for Research in Otolaryngology

Research Article

Acoustic Context Alters Vowel Categorization in Perception of Noise-Vocoded Speech

CHRISTIAN E. STILP[1]

[1]*University of Louisville, 317 Life Sciences Building, Louisville, KY 40292, USA*

Received: 15 July 2016; Accepted: 30 January 2017

ABSTRACT

Normal-hearing listeners' speech perception is widely influenced by spectral contrast effects (SCEs), where perception of a given sound is biased away from stable spectral properties of preceding sounds. Despite this influence, it is not clear how these contrast effects affect speech perception for cochlear implant (CI) users whose spectral resolution is notoriously poor. This knowledge is important for understanding how CIs might better encode key spectral properties of the listening environment. Here, SCEs were measured in normal-hearing listeners using noise-vocoded speech to simulate poor spectral resolution. Listeners heard a noise-vocoded sentence where low-F_1 (100–400 Hz) or high-F_1 (550–850 Hz) frequency regions were amplified to encourage "eh" (/ɛ/) or "ih" (/ɪ/) responses to the following target vowel, respectively. This was done by filtering with +20 dB (experiment 1a) or +5 dB gain (experiment 1b) or filtering using 100 % of the difference between spectral envelopes of /ɛ/ and /ɪ/ endpoint vowels (experiment 2a) or only 25 % of this difference (experiment 2b). SCEs influenced identification of noise-vocoded vowels in each experiment at every level of spectral resolution. In every case but one, SCE magnitudes exceeded those reported for full-spectrum speech, particularly when spectral peaks in the preceding sentence were large (+20 dB gain, 100 % of the spectral envelope difference). Even when spectral resolution was insufficient for accurate vowel recognition, SCEs were still evident. Results are suggestive of SCEs influencing CI users' speech perception as well, encouraging further investigation of CI users' sensitivity to acoustic context.

Keywords: Speech perception, Cochlear implants, Context effects, Spectral contrast effects, Normalization

INTRODUCTION

Many cochlear implant (CI) users experience difficulty with speech perception. Several nonexclusive reasons for this difficulty include prolonged duration of deafness, low neural survival, shallow insertion of the electrode array, poor signal quality delivered by the implant, current spread, and limited experience with an intact speech signal before onset of deafness (cf., Blamey et al. 2012). Difficulty understanding speech is often exacerbated in the presence of background noise, even for many CI users who exhibit little difficulty understanding speech in quiet. These difficulties motivate more research on understanding and improving speech perception for CI users. While much of this research primarily focuses on how CIs transmit acoustic properties of speech sounds, far less attention has been paid to how acoustic context (i.e., recent acoustic history spanning several seconds) also influences speech perception for CI users.

A growing literature reveals that acoustic context plays an important role in speech perception for normal-hearing listeners. Specifically, when acoustic frequencies are reliable across time (i.e., relatively stable or recurring in the acoustic spectrum), the auditory system becomes highly sensitive to changes from these frequencies. This results in spectral contrast effects (SCEs), where perception of a given

Correspondence to: Christian E. Stilp · University of Louisville · 317 Life Sciences Building, Louisville, KY 40292, USA. Telephone: (502) 852-0820; email: christian.stilp@louisville.edu

Published online: 09 March 2017

sound is biased away from reliable spectral properties in preceding sounds. For example, listeners are more likely to perceive /ɪ/ (lower first formant [F_1] frequency) when the preceding acoustic context features reliable spectral energy at higher F_1 frequencies; conversely, perception is biased toward /ɛ/ (higher F_1) when the context features reliable energy at lower F_1 frequencies (Watkins 1991; Sjerps et al. 2011; Stilp et al. 2015). SCEs have been reported for a wide range of speech stimuli (Johnson 1990; Watkins 1991; Watkins and Makin 1996a, 1996b; Holt 2006; Mitterer 2006; Sjerps et al. 2011; Stilp et al. 2015; Assgari and Stilp 2015; Stilp and Assgari 2017). SCEs have also been reported in categorization of musical instruments (Stilp et al. 2010), suggesting that this phenomenon is not limited to speech perception but likely reflects a general operating characteristic of the auditory system.[1]

Spectral contrast effects are an example of *extrinsic* cues to speech sound identity, where acoustic properties of neighboring sounds influence perception of the target speech sound. This is complemented by *intrinsic* cues to speech sound identity, which are acoustic properties of the target sound itself. While intrinsic and extrinsic cues are both important for speech sound recognition by normal-hearing listeners (Ainsworth 1975; Nearey 1989), the vast majority of speech perception research with CI users has focused on intrinsic cues. For example, a large number of investigations measured CI users' vowel perception by presenting isolated syllables, in many cases, recordings of talkers saying /hVd/ (e.g., from Hillenbrand et al. 1995). However, extrinsic cues are noteworthy because they contribute to speech perception via both peripheral and central auditory processing. Extrinsic cues produced SCEs for normal-hearing listeners in dichotic stimulus presentations (preceding acoustic context and target sound are presented to opposite ears; Watkins 1991; Holt and Lotto 2002). This reveals that SCEs are not limited to peripheral processing (i.e., a healthy cochlea) but recruit central auditory processing as well. CI users still have an intact central auditory system, so these listeners should still be sensitive to extrinsic cues to speech sound identity. Exploring the influence of degraded extrinsic cues on speech categorization can shed new light on deficits in speech perception for this listener population.

Yet, the few investigations of extrinsic cues to speech sound recognition for CI users provided conflicting reports. Aravamudhan and Lotto (2004, 2005) reported that unlike normal-hearing listeners (Lotto and Kluender 1998), CI users were not more likely to identify a consonant as /d/ (higher F_3 onset frequency) following /r/ (lower F_3 offset frequency) nor were they more likely to identify that same consonant as /g/ (lower F_3 onset frequency) following /l/ (higher F_3 offset frequency). Aravamudhan and Lotto concluded that CIs do not provide sufficient spectral resolution to produce SCEs. On the other hand, Winn et al. (2013) reported positive evidence of SCEs in CI users' speech perception. When presented with fricative-vowel syllables, CI users and normal-hearing listeners were both more likely to identify the initial consonant as higher-frequency /s/ when it was followed by /u/ with lower F_2 and F_3, and both groups labeled the consonant as lower-frequency /ʃ/ more often when it was followed by /i/ with higher F_2 and F_3. Even though these effects were larger for normal-hearing listeners than CI users, they still supported significant shifts in speech sound categorization due to SCEs.

Many differences exist across these studies of short-term SCEs (where spectral differences between immediately adjacent sounds are perceptually magnified), including factors relating to the implant (electrode placement, implant signal processing strategy), the listener (listening experience with the implant, neural survival), and the stimuli. Two particular stimulus differences warrant closer consideration. The first key difference is that of stimulus timescale. In studies by Aravamudhan and Lotto (2004, 2005), listeners had only tens of milliseconds to detect formant transitions, even less for detecting their offset and onset frequencies. In Winn et al. (2013), the fricatives and vowels each exceeded 100 ms, which may have given listeners more opportunities to sample key spectral properties across time. In long-term SCEs, which are the focus of the present investigation, perception of a target sound is altered by earlier sounds whose spectral properties are relatively stable across several seconds. While CI users might struggle with fine frequency comparisons on short timescales (Donaldson et al. 2013, 2015), having more opportunities to perceive reliable spectral properties throughout a longer acoustic context (such

[1] The focus of this discussion is spectral contrast effects that are produced by stable spectral properties in a preceding listening context. The research cited used speech and nonspeech contexts whose durations were comparable to that of a sentence (one or more seconds). Importantly, the reliable spectral properties were relatively stable or recurring throughout the context and need not occur at its offset in order to produce the contrast effect. A parallel research program investigated SCEs that were produced by spectral characteristics at the offset of a shorter-duration context (e.g., a single syllable or tone). These short-term SCEs have been reported for speech contexts and speech targets (Lotto and Kluender 1998), nonspeech contexts and speech targets (Lotto and Kluender 1998), and speech contexts and nonspeech targets (Stephens and Holt 2003). In these studies, the predictions and patterns of results are consistent with those of longer-term SCEs discussed in this report; the principle difference is the timecourse of the preceding acoustic context. In addition, short-term SCEs have been observed in the nonhuman animals' responses to speech (Lotto et al. 1997), further supporting the proposal that SCEs are a general operating characteristic of the auditory system.

as a sentence) should increase the likelihood of producing an SCE. The second key difference across stimuli is spectral bandwidth. CI users might have lacked the spectral resolution required to accurately perceive changes between narrowband formant transitions (such that no SCEs were observed; Aravamudhan and Lotto 2004, 2005) but did have sufficient spectral resolution to perceive changes in spectral envelopes across fricatives and vowels (producing SCEs; Winn et al. 2013). If CI users lack sufficient spectral resolution to resolve stable spectral peaks in the acoustic context, their spectral resolution might still be sufficient for stable spectral envelopes to produce SCEs in speech categorization.

Even if SCEs influence speech perception for CI users, equally compelling arguments can be made that these effects would be either smaller or larger than those exhibited by normal-hearing listeners. Several lines of research suggest that SCEs would be smaller for CI users than normal-hearing listeners, as they were according to Winn et al. (2013). CI users require greater differences between spectral peaks and valleys to recognize vowels than normal-hearing listeners do (Loizou and Poroy 2001). When reliable spectral peaks in an acoustic context are smaller, diminished SCEs were observed for normal-hearing listeners (Stilp et al. 2015). Thus, a given acoustic context featuring a reliable spectral peak is likely to have lower local spectral contrast following CI processing, which is expected to decrease the size of the resulting SCEs. Additionally, spectrally degraded acoustic contexts produced smaller SCEs than full-spectrum contexts for normal-hearing listeners (signal-correlated noise versus a sentence; Watkins 1991). By extension, acoustic contexts that are degraded by CIs might be expected to produce smaller SCEs than those observed in acoustic hearing. Smaller-than-normal SCEs impair speech sound categorization, particularly for perceptually ambiguous sounds (i.e., hypoarticulated tokens near the center of a phonetic continuum and/or at the boundary between two phonemic categories). Under normal circumstances, these sounds are disambiguated by the preceding acoustic context, but when SCEs are too small, these sounds remain ambiguous and are thus poorly identified (Fig. 1b).

On the other hand, SCEs might be larger for CI users than normal-hearing listeners. CI electrodes (and the vocoder channels used to simulate them) have broader bandwidths than auditory filters in healthy hearing. These broader bandwidths would increase the bandwidths of reliable spectral peaks, which have been shown to increase the sizes of ensuing SCEs (Stilp et al. 2015). This would be consistent with the findings of Stilp and Alexander (2016), who attributed larger SCEs for listeners with sensorineural hearing loss to broadened auditory filters. Additionally, as many acoustic cues to speech are degraded or eliminated in vocoder processing, normal-hearing listeners displayed increased reliance on cues that remained compared to their usage of those cues in intact speech (Winn et al. 2012; Winn and Litovsky 2015; Moberly et al. 2014, 2016; Kong et al. 2016). For example, Winn and Litovsky (2015) showed that normal-hearing listeners' weighting of formant transitions decreased and weighting of spectral tilt changes increased when categorizing voiced stop consonants that were noise - vocoded with simulated current spread compared to their cue weighting in full-spectrum speech. Stable spectral properties do not require fine frequency resolution to produce SCEs (broadened spectral peaks: Stilp et al. 2015; spectral envelopes: Watkins 1991; Stilp et al. 2010, 2015; Sjerps et al. 2011). As long as these spectral properties are presented above detection thresholds, they should be perceived in spectrally degraded speech. If listeners increase their reliance on these spectral properties compared to their usage when perceiving full-spectrum speech, then larger SCEs will be observed. Larger-than-normal SCEs would impair speech sound categorization by shifting phonemic categories too far apart from one another. This makes perceptually unambiguous sounds (i.e., normally articulated or even hyperarticulated tokens found near each end of a phonetic continuum) more ambiguous and thus misidentified (Fig. 1c).

Whether SCEs are smaller or larger for CI users than normal-hearing listeners, this comparison is complicated by the considerable individual differences among CI users, particularly in how stable spectral properties in speech are coded by electrode locations that may differ across listeners. The present experiments used noise vocoding as an acoustic simulation of CI processing in order to maintain control over how stable spectral properties were coded. These experiments addressed two main questions. First and foremost, experiments marked the first test of whether long-term SCEs can be produced in spectrally degraded speech (as would be transmitted by a CI). The slopes of listeners' response functions were expected to be shallower than those observed for full-spectrum speech owing to spectral degradation and/or increased task difficulty, particularly as fewer spectral channels were presented. Instead, the changes in the intercepts of these functions were of primary interest, as these indicated shifts in the listeners' responses due to the spectral context. Second, results illuminated whether SCEs should be smaller or larger for CI users compared to normal-hearing listeners. Both possibilities are detrimental to speech perception (Fig. 1), but identifying this outcome may reveal new areas in which the CI users' speech recognition can be examined and

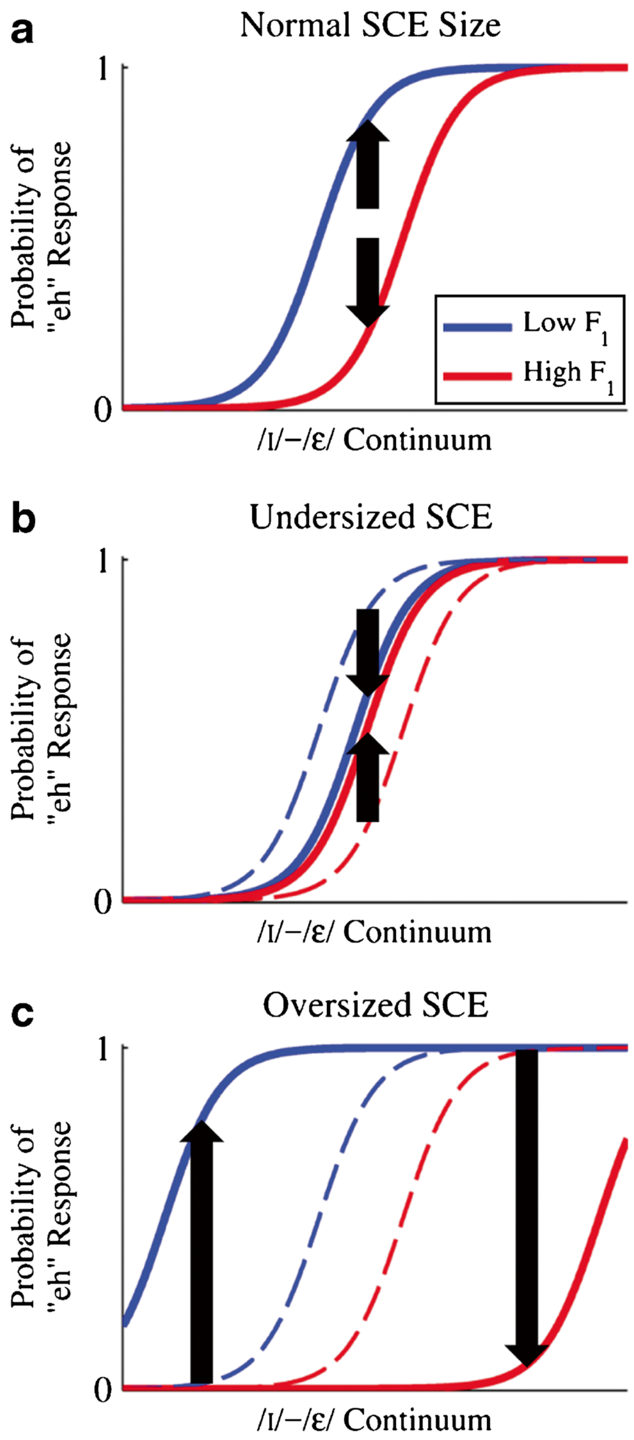

FIG. 1. Hypothetical data illustrating different degrees of the influence of spectral context (i.e., SCEs) on phoneme categorization. In each panel, the probability of responding “eh” is a function of hearing different tokens from a vowel continuum that acoustically varies from /ɪ/ to /ɛ/. Responses vary depending on whether the preceding acoustic context had an emphasis in the low-F_1 frequency region (*left curves*, more “eh” responses) or in the high-F_1 frequency region (*right curves*, fewer “eh” responses). **a** Responses when the SCE is of appropriate magnitude (i.e., as measured in normal-hearing listeners). *Arrows* indicate where a mid-continuum vowel (which is ordinarily perceptually ambiguous) is disambiguated by preceding spectral context (making “eh” responses far more or far less likely). **b** Responses when the SCE is undersized. Here, preceding spectral context exerts too little influence on the target sounds, producing an SCE that is much smaller (*solid lines*) than the “appropriate” size (*dashed curves*, replotted from **a**). *Arrows* indicate where the perceptually ambiguous mid-continuum vowel fails to be disambiguated by the preceding spectral context, increasing perceptual confusions. **c** Responses when the SCE is oversized. Here, preceding spectral context exerts too much influence on the target sounds, producing an SCE that is much larger (*solid lines*) than the appropriate size (*dashed curves*, replotted from **a**). Perceptually ambiguous mid-continuum stimuli are fully disambiguated by the spectral context, but *arrows* indicate where previously unambiguous stimuli toward each end of the continuum are now categorized less consistently, increasing perceptual confusions.

potentially improved. In each experiment, listeners heard a sentence featuring reliable spectral properties (spectral peak in a prescribed frequency region, overall spectral shape) followed by a target vowel that was categorized as /ɛ/ (as in "bet") or /ɪ/ (as in "bit"). By measuring the extent to which the listeners' responses were biased away from reliable spectral properties in the preceding sentence, SCEs characterized the influence of extrinsic cues on noise-vocoded vowel identification.

EXPERIMENT 1: VOWEL IDENTIFICATION FOLLOWING ACOUSTIC CONTEXT WITH A RELIABLE SPECTRAL PEAK

Participants

Thirty-two listeners enrolled in undergraduate courses at the University of Louisville participated in exchange for course credit (18 in experiment 1a; 14 in experiment 1b). No listener participated in multiple experiments. All listeners were native English speakers with self-reported normal hearing.

Stimuli

Acoustic Context. The acoustic context was the sentence "Please say what vowel this is" spoken by the author (2174-ms duration). Frequency regions in the sentence were made reliable through amplification, creating "low-F_1-amplified" and "high-F_1-amplified" versions of the context. The sentence was processed by a 300-Hz-wide finite impulse response filter near F_1 in the target vowels /ɪ/ (100–400 Hz) or /ɛ/ (550–850 Hz). The level of filter gain in the passband (with zero gain at other frequencies) varied across experiments. In experiment 1a, filter gain was set at +20 dB, which has been successful in producing relatively large SCEs in perception of full-spectrum speech (Stilp et al. 2015; Assgari and Stilp 2015; Stilp and Alexander 2016) and thus was likely to (at least partially) overcome signal degradation from noise vocoding and produce SCEs. In experiment 1b, filter gain was set at +5 dB, which also reliably produced SCEs in perception of full-spectrum materials but at smaller magnitudes than the +20 dB peaks. This provided a more sensitive test of whether reliable spectral peaks in the precursor sentence and ensuing SCEs would be distorted by noise vocoding. Filters were created using the fir2 function in MATLAB (MathWorks, Inc., Natick, MA) with 1200 coefficients. The sentence and filtering were the same as in previous studies of SCEs (Stilp et al. 2015; Assgari and Stilp 2015; Stilp and Alexander 2016).

Vowels. Target vowels were the same /ɪ/-to-/ɛ/ continuum as tested in previous studies of SCEs by Stilp and colleagues (Stilp et al. 2015; Assgari and Stilp 2015; Stilp and Alexander 2016). For a detailed description of the generation procedures, see Winn and Litovsky (2015). Briefly, tokens of /ɪ/ and /ɛ/ were recorded by the author. Formant contours were extracted from each token in Praat (Boersma and Weenink 2014). Formant frequencies varied across time throughout the duration of the vowel in order to maintain perceived naturalness. In the /ɪ/ endpoint, F_1 linearly increased across time from 400 to 430 Hz while F_2 linearly decreased from 2000 to 1800 Hz. In the /ɛ/ endpoint, F_1 linearly decreased across time from 580 to 550 Hz while F_2 linearly decreased from 1800 to 1700 Hz. These F_1 and F_2 trajectories were linearly interpolated to create ten sets of formant tracks corresponding to ten representative vowels on the /ɪ/-to-/ɛ/ continuum which were later used as canonical vowel targets. For example, across the vowel continuum, the F_1 trajectory within each vowel target progressed from increasing in frequency (/ɪ/ endpoint) to a relatively flat trajectory (near the middle of the continuum) to decreasing in frequency (/ɛ/ endpoint). The ten-step vowel continuum was generated by applying each of the ten formant tracks to a single voice source (extracted from the /ɪ/ endpoint using LPC inverse filtering). Energy above 2500 Hz for all vowels was set to the energy high-pass filtered from the original /ɪ/ token. Fundamental frequency was set to 100 Hz throughout the vowel. Each vowel token was 246 ms in duration. Each vowel and precursor sentence was set to equal root mean square (RMS) amplitude. Trial sequences were then created by concatenating one vowel to a precursor sentence with a 50-ms silent interstimulus interval.

Noise Vocoding. All trial sequences were noise - vocoded from 100–5000 Hz in MATLAB. Vocoding down to this low-frequency edge can cause errors due to filter instability, but it was essential to maintain the integrity of the low-F_1 reliable spectral peak (100–400 Hz) in order to compare the present results to those obtained with full-spectrum stimuli. Therefore, stimuli were first spectrally rotated about 8000 Hz in Praat. The low-frequency edge of 100 Hz in the original signal was transposed to 8000 Hz in the rotated signal, and the high-frequency edge of 5000 Hz in the original signal was transposed to 3000 Hz in the rotated signal. Corner frequencies for 6, 12, and 24 vocoder channels from 100–5000 Hz were computed using Greenwood's (1990) formula (Table 1). These corner frequencies were then subtracted from 8100 Hz so that signal frequencies and channel corner frequencies were both inverted but properly aligned as in typical vocoding. The spectrally rotated signal was then noise - vocoded in MATLAB using fourth-order Butterworth filters for channel analysis and synthesis (24 dB/octave rolloff)

TABLE 1
Channel numbers and upper cutoff frequencies in the noise vocoder

Channel number (6)	*Channel number (12)*	*Channel number (24)*	*Upper cutoff (Hz)*
1	1	1	*135*
		2	*175*
	2	3	*219*
		4	*270*
2	3	5	*327*
		6	*392*
	4	7	465
		8	549
3	5	9	*643*
		10	*749*
	6	11	*869*
		12	1006
4	7	13	1160
		14	1334
	8	15	1532
		16	1755
5	9	17	2008
		18	2294
	10	19	2618
		20	2984
6	11	21	3399
		22	3868
	12	23	4399
		24	5000

Sentences were vocoded with 6 (first column), 12 (second column), or 24 channels (third column). Total signal bandwidth spanned 100–5000 Hz. Italicized cutoff frequencies indicate low-F_1 (100–400 Hz) or high-F_1 (550–850 Hz) regions which were amplified in the precursor sentence to create reliable spectral peaks

and second-order Butterworth filter for amplitude envelope extraction (12 dB/octave rolloff, low-pass cutoff at 400 Hz). Finally, the vocoded signal was spectrally rotated again about 8000 Hz in Praat to return all frequencies to their original positions. Following noise vocoding, stimuli were set to a fixed RMS amplitude. Sample trial sequences are presented in Figure 2.

Table 1 illustrates how intrinsic (spectral properties within the target vowel) and extrinsic (reliable spectral properties in the preceding acoustic context) cues to vowel perception can be dissociated. With regard to intrinsic cues to vowel identity, listeners are expected to have difficulty differentiating target vowels with only six vocoder channels, as F_1 information for almost all vowels in the continuum falls in the same (second) vocoder channel. Conversely, for the extrinsic cues to vowel identity, reliable spectral peaks in the precursor sentences can be differentiated even with only six vocoder channels. This is sufficient to make F_1 in the target vowel appear to be at a higher or lower frequency by comparison, producing shifts in vowel categorization (SCEs). In the 24-channel case, vowel intelligibility should be greatly aided by improved signal quality and F_1 information being transmitted across distinct vocoder channels (seventh and eighth channels in particular). Extrinsic cues (low-F_1 and high-F_1 regions) remain highly differentiable in the 24-channel case, raising the question of whether the extent of influence from the acoustic context would remain equal or diminish compared to the six-channel case.

Experimental Setup and Procedure

Listeners participated individually in single-wall sound-isolating booths (Acoustic Systems, Inc., Austin, TX). Following the acquisition of informed consent, listeners were given instructions and told to respond whether the target vowel at the end of each trial sounded more like "ih (as in 'bit')" or "eh (as in 'bet')". A custom MATLAB script led the participants through the experiment. Stimuli were D/A converted by RME HDSPe AIO sound cards (Audio AG, Haimhausen, Germany) on personal computers and passed through a programmable attenuator (TDT PA4, Tucker-Davis Technologies, Alachua, FL) and headphone buffer (TDT HB6). Trial sequences were upsampled to 44,100 Hz and presented diotically at an average of 70 dB SPL via circumaural headphones (Beyerdynamic DT-150, Beyerdynamic Inc. USA, Farmingdale, NY). Listeners responded by clicking the mouse to indicate which vowel they heard on each trial. The number of spectral channels in vocoded stimuli (6, 12, 24) was blocked and tested in random orders across listeners. Each block consisted of 200 trials (10 target vowels × 2 filter conditions [low-F_1 amplified, high-F_1 amplified] × 10 repetitions) tested in quasi-random order, such that each combination of target vowel and precursor filtering was tested once every 20 trials. Each block lasted approximately 12 min, between which listeners took short breaks.

Statistical Analysis

The first step in data analysis was identifying and removing outlier participants. Exclusion criteria from past studies (e.g., at least 80 % correct on vowel continuum endpoints; Assgari and Stilp 2015) were not appropriate because of the increased difficulty of categorizing spectrally degraded stimuli. Additionally, high variability in speech intelligibility for CI users is widely documented, but this cannot be accurately modeled by analyzing the data from only higher-performing normal-hearing listeners. Instead, the exclusionary criterion was defined by overreliance on a single response category. If a listener gave the same response to 24-channel vocoded stimuli (which was the most intelligible speech listeners heard) at least

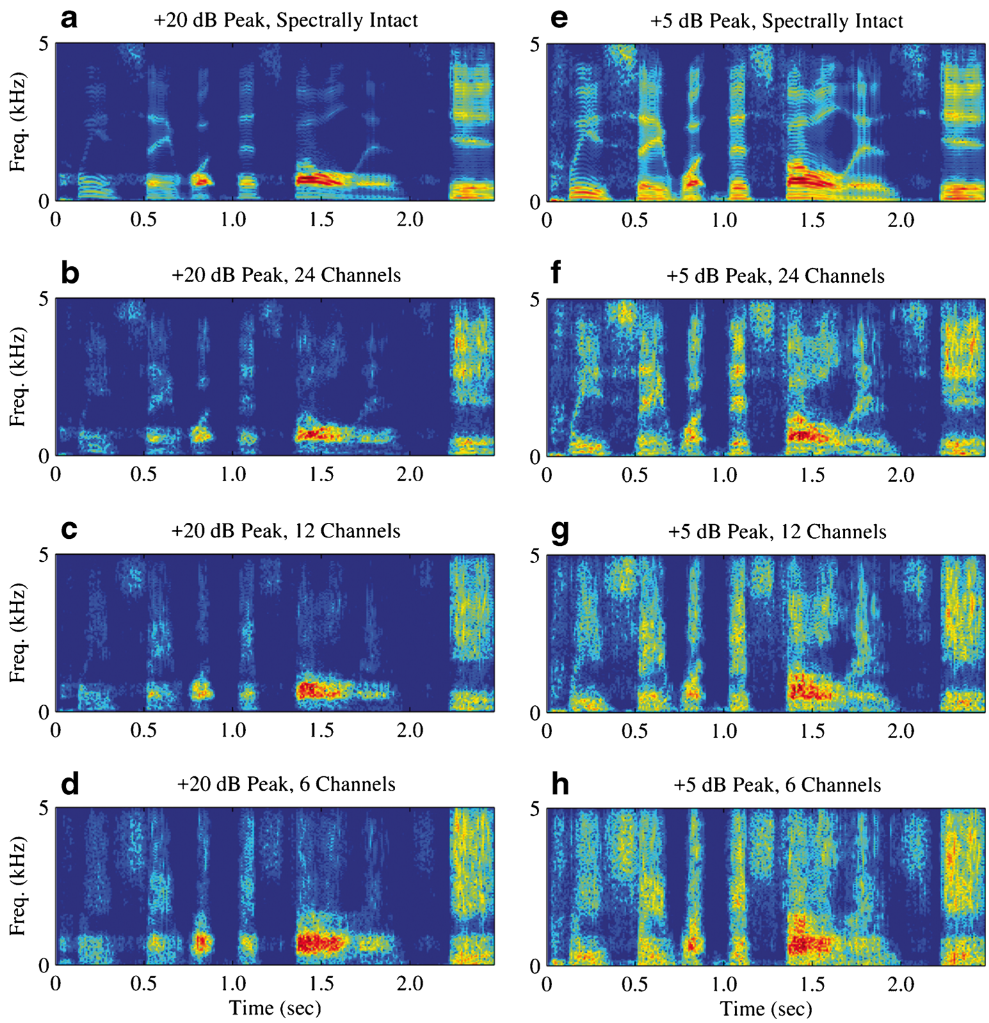

FIG. 2. Spectrograms of sample trials in experiment 1. The *left column* illustrates trials from experiment 1a, where the precursor sentence was processed by a +20 dB filter in the high-F_1 region (550–850 Hz). The *right column* illustrates trials from experiment 1b, where the precursor was processed by the same high-F_1 filter but only +5 dB gain. The *top row* shows a sample trial with full-spectrum stimuli (from Stilp et al. 2015); *successive rows* show those stimuli vocoded with 24, 12, and 6 spectral channels, respectively. The target vowel is the /ɪ/ endpoint.

80 % of the time, all data from that listener were removed from analyses. This resulted in the removal of one listener's data from experiment 1a and none from experiment 1b, resulting in analyses of 17 and 14 complete data sets, respectively.

Results were analyzed using generalized linear (logistic) mixed effect models in R (R Development Core Team 2016) using the lme4 package (Bates et al. 2014). This offered an advantage over traditional logistic regression, as the random effect structure allowed estimation of variance attributable to the listener sample to be partitioned from variance attributable to the fixed effects. This was desirable given that large individual differences were anticipated for perception of noise-vocoded speech, particularly when few spectral channels were available. Separate models with identical fixed and random effects were conducted for each experiment. The model architecture was patterned after the model utilized by Stilp et al. (2015) who tested the same stimuli but without noise vocoding. Fixed effects in the model included vowel target (coded as a continuous variable from 1 to 10 then mean centered), filter F_1 frequency (categorical variable with two levels: low F_1, high F_1, with high F_1 set as the default level), spectral resolution (SR; continuous variable with values of 6, 12, and 24 spectral channels then mean centered), and all possible interactions therein in the effort to explain as much of the variability as possible. Random slopes were included for each main fixed effect, but random slopes were not included for interactions between main fixed effects due to model convergence errors. A random intercept of listener was also included in the model. The final model had the following form:

$$\begin{aligned}\text{Response} \sim\ & \text{vowel} + \text{filter} + \text{SR} + \\ & \text{vowel} \times \text{filter} + \text{vowel} \times \text{SR} + \text{filter} \times \text{SR} + \\ & \text{vowel} \times \text{filter} \times \text{SR} + \\ & (1 + \text{vowel} + \text{filter} + \text{SR} \mid \text{listener})\end{aligned}$$

RESULTS

Model estimates for experiments 1a and 1b are presented in Table 2. Statistical significance was evaluated using the criterion of α = .05. Both experiments exhibited a significant effect of vowel (experiment 1a: $Z = 6.282$, $P < 0.001$; experiment 1b: $Z = 4.182$, $P < 0.001$), revealing an increase in the log odds of responding "eh" following a one-step increase along the vowel continuum toward the /ɛ/ endpoint (vowel number 10). Both experiments exhibited a significant effect of filter frequency (experiment 1a: $Z = 6.199$, $P < 0.001$; experiment 1b: $Z = 3.698$, $P < 0.001$). Positive coefficients for this effect revealed an increase in the log odds of responding "eh" when moving this spectral peak from the high-F_1 frequency region to the low-F_1 region (i.e., more positive intercept for the psychometric function). This confirms the presence of SCEs in perception of noise-vocoded stimuli. The coefficient for the filter effect is markedly larger for experiment 1a (1.664) than experiment 1b (0.732), replicating larger contrast effects being produced by larger F_1 peaks in the precursor sentence as reported for full-spectrum versions of these stimuli (Stilp et al. 2015; Assgari and Stilp 2015; Stilp and Alexander 2016). Both experiments exhibited significant negative effects of spectral resolution (experiment 1a: $Z = -2.693$, $P = 0.007$; experiment 1b: $Z = -2.392$, $P = 0.017$), indicating decreases in the log odds of responding "eh" when the number of spectral channels increased. This is particularly evident in the initial bias toward responding "eh" (Fig. 3a) becoming less strong as signal quality improved (Fig. 3c). Finally, the only significant interaction was between vowel and spectral resolution [statistically significant in experiment 1a ($Z = 10.755$, $P < 0.001$); trend toward statistical significance in experiment 1b ($Z = 1.941$, $P = 0.052$)]. This interaction indicates steeper psychometric function slopes as spectral resolution increased, consistent with decreased difficulty categorizing speech sounds as signal quality improved.

Similar to the selection of an exclusionary criterion, some ways of measuring SCEs were inappropriate for experiments with noise-vocoded stimuli. Previous investigations with full-spectrum versions of these stimuli fit logistic regressions to each listener's responses following low-F_1-amplified or high-F_1-amplified precursor sentences, measured the distance between regression function midpoints to obtain the listener's SCE, then averaged SCEs across all listeners (Assgari and Stilp 2015; Stilp and Alexander 2016). This approach is limited when the several listeners' regression functions lack sigmoidal shapes and/or well-defined midpoints, as in the present data. Instead, SCEs were operationalized as the overall change in the listeners' mean probabilities of responding "eh" across the entire vowel continuum following low-F_1-amplified precursors versus high-F_1-amplified precursors. This metric indicates how much more likely listeners were to respond "eh" following low-F_1-amplified precursors (which is spectrally contrastive with an "eh," or high-F_1, response) than high-F_1-amplified precursors (which is not spectrally contrastive with an "eh" response). This operationalization is similar to previous investigations of SCEs where the main effect of acoustic context spectrum was tested in a repeated measures ANOVA where the dependent measure was a single response category (Holt 2005; 2006).

SCEs were calculated at each level of spectral resolution in each experiment. In experiment 1a (+20 dB filter gain), SCEs were response shifts of 26.76 (6 channels; Fig. 3a), 25.06 (12 channels; Fig. 3b), and 29.18 % (24 channels; Fig. 3c). All of these SCE magnitudes were considerably larger than those observed using full-spectrum versions of the same stimuli (14.15 %; Fig. 3d; Stilp et al. 2015). In experiment 1b (+5 dB filter gain), SCEs were response shifts of 9.93 (6 channels; Fig. 3e), 9.86 (12 channels; Fig. 3f), and 11.86 % (24 channels; Fig. 3g), again all of which were larger than SCEs for full-spectrum stimuli (6.82 %; Fig. 3h; Stilp et al. 2015).

DISCUSSION

Identification of noise-vocoded vowels was clearly influenced by SCEs in experiment 1. Noise vocoding degraded overall signal quality, but spectral peaks in the preceding sentence still altered identification of the target vowel. It is noteworthy that even with the relatively high spectral resolution of 24 vocoder channels, responses did not fully approach those observed for full-spectrum speech (comparing Fig. 3c to d; comparing Fig. 3g to h). While words and sentences are largely intelligible when vocoding with 24 channels, here, listeners categorized vowels differing primarily in F_1, which was transmitted by only two vocoder channels (channel 7 with center frequency of 465 Hz, channel 8 with center frequency of 549 Hz). This retained an element of task difficulty

TABLE 2

Results of generalized linear mixed effects models for experiments 1a (left) and 1b (right)

	Experiment 1a				*Experiment 1b*			
	Estimate	*SE*	*Z*	*P*	*Estimate*	*SE*	*Z*	*P*
Intercept	0.255	0.248	1.028	0.304	0.363	0.132	2.743	0.006
Vowel (V)	0.260	0.041	6.282	<0.001	0.391	0.094	4.182	<0.001
Filter (F)	1.664	0.268	6.199	<0.001	0.732	0.198	3.698	<0.001
SR	−0.052	0.019	−2.693	0.007	−0.023	0.010	-2.392	0.017
V by F	−0.028	0.019	−1.438	0.151	−0.029	0.020	−1.488	0.137
V by SR	0.019	0.002	10.755	<0.001	0.004	0.002	1.941	0.052
F by SR	0.010	0.007	1.350	0.177	0.008	0.007	1.085	0.278
V by F by SR	−0.003	0.003	−1.022	0.307	−0.001	0.003	−0.198	0.843

SR spectral resolution (number of vocoder channels), *SE* standard error of the mean

that was evident in the different psychometric function slopes across 24-channel and full-spectrum results.

There was an unexpected bias toward "eh" responses to the vowel targets, especially at lower spectral resolutions. Inspection of the vowel spectra suggested that this was likely due to spectral smearing in the vocoding process. High-energy harmonics immediately above the F_1 peak in the vowel were smeared, producing a spectral shoulder. This shoulder extended from 549–1006 (third channel in the 6-channel condition) or from 549–749 Hz (fifth channel in the 12-channel condition), but no such shoulder was observed in the 24-channel conditions. At frequencies immediately above this spectral shoulder, acoustic energy was very low due to the spectral valley

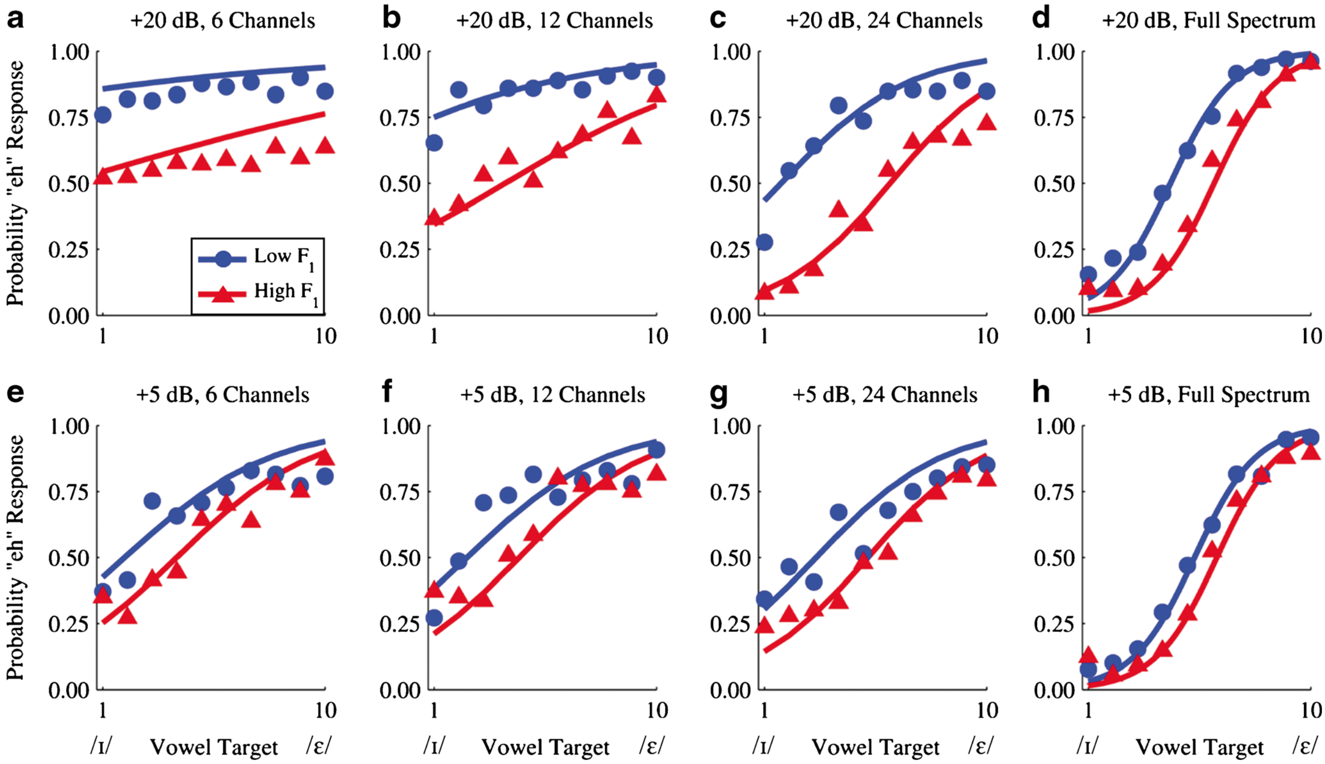

FIG. 3. Listener performance and model predictions for experiment 1 and related results for full-spectrum stimuli from Stilp et al. (2015). *Circles* represent mean probabilities of responding "eh" following the low-F_1-amplified precursor sentence; *triangles* represent mean probabilities of responding "eh" following the high-F_1-amplified precursor sentence. *Solid lines* depict predicted responses from the mixed effects models. The *top row* (**a**–**d**) depicts results for precursor sentences with +20 dB reliable spectral peaks; the *bottom row* (**e**–**h**) depicts results for precursor sentences with +5 dB reliable spectral peaks. **a**–**c** Results from experiment 1a for the 6-, 12-, and 24-channel conditions, respectively. **d** Listener performance and model predictions for full-spectrum stimuli with +20 dB peaks from Stilp et al. (2015). **e**–**g** Results from experiment 1b for the 6-, 12-, and 24-channel conditions, respectively. **h** Listener performance and model predictions for full-spectrum stimuli with +5 dB peaks from Stilp et al. (2015).

between F_1 and F_2. Given this large difference in energy across neighboring frequency regions, the spectral shoulder was likely perceived as a high-F_1 spectral peak, resulting in the bias toward responding "eh." Even with this bias toward a single response category, SCEs (i.e., relative shifts in responses) were still observed at all spectral resolutions.

When assessing the generalizability of these results to CI users' speech perception, one point merits further consideration. Normal-hearing listeners' ability to resolve spectral peaks exceeds that of CI users, even in spectrally degraded materials. As such, normal-hearing listeners might be more sensitive to reliable spectral peaks in the acoustic context than CI users are. CI users relied less on spectrally narrow cues (formants, formant transitions) than normal-hearing listeners did to recognize speech and relied more on other, secondary cues (Winn et al. 2012; Winn and Litovsky 2015; Moberly et al. 2014). According to Winn and Litovsky (2015), CI users weighted formant transitions less and weighted changes in spectral envelope more than normal-hearing listeners who heard noise-vocoded stimuli with steep channel cutoffs. Reliable spectral envelopes have produced SCEs in perception of full-spectrum speech (Watkins 1991; Watkins and Makin 1996a, 1996b; Sjerps et al. 2011; Stilp et al. 2015). Demonstrating SCEs in perception of noise-vocoded speech following reliable spectral envelopes instead of narrowband peaks would further promote generalizability of results to CI users.

Experiment 2 introduced filtering that was distributed across the entire frequency spectrum. Experiment 2 utilized spectral envelope difference filters (Watkins 1991), where stimuli spectra were altered by a complex frequency response corresponding to the difference between two vowel spectra (here, /ɪ/ and /ɛ/). Stimuli were processed by filters that reflected 100 % of the total difference between these spectral envelopes (experiment 2a) or only 25 % of this difference (experiment 2b). As in experiment 1, this manipulation of spectral shape reliability provided two tests of whether reliable spectral shape information would overcome signal degradation from noise vocoding and still produce SCEs.

EXPERIMENT 2: VOWEL IDENTIFICATION FOLLOWING ACOUSTIC CONTEXT WITH A RELIABLE SPECTRAL ENVELOPE

Participants

Thirty-one listeners enrolled in undergraduate courses at the University of Louisville participated in experiment 2 in exchange for course credit (15 in experiment 2a; 16 in experiment 2b). No listener participated in multiple experiments or in experiment 1. All listeners were native English speakers with self-reported normal hearing.

Stimuli

Acoustic Context. The acoustic context was the same sentence as presented in experiment 1. Frequency regions in the sentence were made reliable through spectral envelope difference filtering (Watkins 1991). Spectral envelope difference filtering was conducted before noise vocoding to facilitate comparison of results to those obtained using full-spectrum stimuli (Stilp et al. 2015). Following Stilp et al. (2010, 2015), spectral envelopes for each endpoint of the vowel continuum were derived using 512-point Fourier transforms smoothed by a 512-point Hamming window with a 512-point overlap. Spectral envelopes were equated for peak power then subtracted from one another in both directions (/ɪ/ minus /ɛ/, /ɛ/ minus /ɪ/). A 500-point finite impulse response was obtained for each spectral envelope difference using inverse Fourier transform. Impulse responses were either maintained at 100 % of the original spectral difference (experiment 2a) or scaled down to 25 % of the spectral difference (experiment 2b; Fig. 4). Scaling was calculated using linear amplitude values. The sentence was then filtered using each impulse response, producing precursor sequences with an overall low-F_1 emphasis (/ɪ/ minus /ɛ/) or high-F_1 emphasis (/ɛ/ minus /ɪ/).

Vowels. The same target vowels from experiment 1 were presented in experiment 2. As in experiment 1, each vowel and precursor sentence was set to equal RMS amplitude. Trials were created by concatenating one vowel to a precursor sentence with a 50-ms silent interstimulus interval. Sample trials are depicted in Figure 5.

Noise Vocoding. All trial sequences were noise-vocoded in the same manner as detailed in experiment 1.

Experimental Setup and Procedure

The setup and procedure for experiment 2 matched that of experiment 1.

RESULTS

The same exclusionary criterion (>80 % responses to 24-channel stimuli coming from a single vowel category) was applied to the results of experiment 2. Data from three listeners in experiment 2a and three listeners in experiment 2b were removed, resulting in analyses of 12 and 13 complete data sets, respectively.

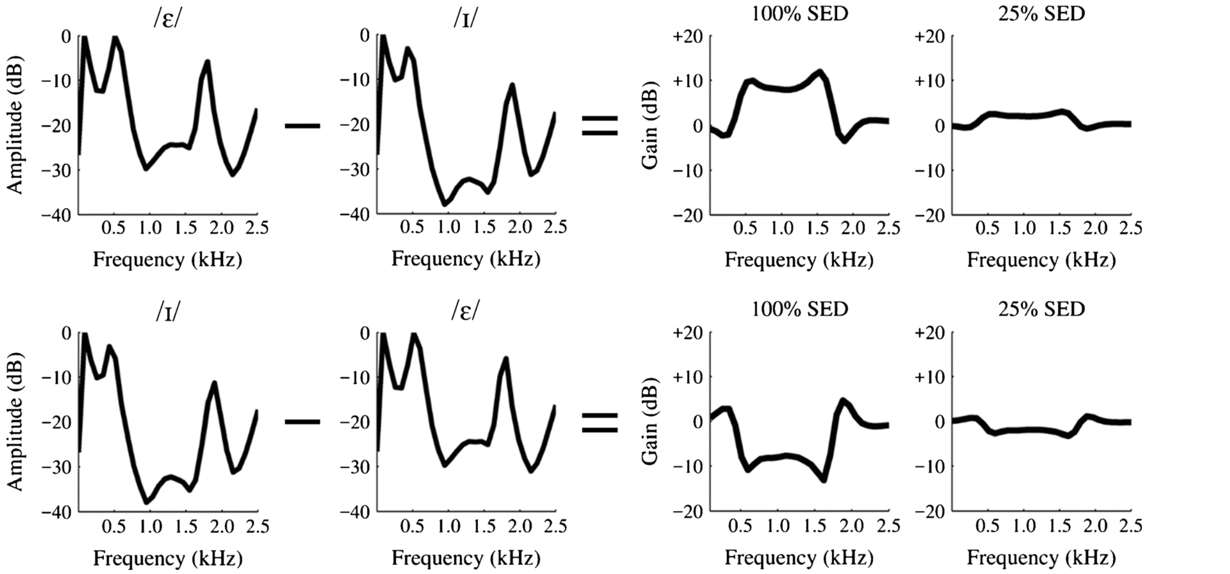

FIG. 4. Generation of spectral envelope difference filters. The *top row* depicts generation of the high-F_1-emphasis difference filter, where the spectrum of /ɪ/ is subtracted from the spectrum of /ɛ/. The *bottom row* depicts generation of the low-F_1-emphasis difference filter, where the spectrum of /ɛ/ is subtracted from the spectrum of /ɪ/. The *third column* depicts filter profiles for 100 % of the spectral envelope difference (SED; experiment 2a), and the *fourth column* depicts filter profiles for 25 % of the SED (experiment 2b).

Responses were analyzed using the same mixed effects model structure reported in experiment 1. Model estimates for experiments 2a and 2b are presented in Table 3.

Both experiments displayed significant effects of vowel (experiment 2a: $Z = 5.505$, $P < 0.001$; experiment 2b: $Z = 6.930$, $P < 0.001$), again indicating that listeners changed their responses depending on the target vowel presented. Both experiments exhibited significant effects of spectral envelope difference filter (i.e., SCEs; experiment 2a: $Z = 6.326$, $P < 0.001$; experiment 2a: $Z = 4.555$, $P < 0.001$), with much larger effects for filters that added larger spectral peaks to the precursor sentence (100 % of the spectral envelope difference in experiment 2a, estimate = 1.585) compared to more modest filtering (only 25 % of the spectral envelope difference in experiment 2b, estimate = 0.612), replicating the pattern reported by Stilp et al. (2015) using full-spectrum speech. Spectral resolution did not significantly influence the listeners' responses in experiment 2a ($Z = -0.071$, $P = 0.943$) but it did in experiment 2b ($Z = -2.723$, $P = 0.006$). The significant intercept in experiment 2b revealed an overall bias toward “eh” responses, and the negative coefficient on the SR factor indicates that this bias decreased as the number of spectral channels increased. The interaction between vowel and filter frequency was statistically significant in experiment 2b ($Z = -2.484$, $P = 0.013$), indicating psychometric function slopes were shallower for low-F_1-amplified difference filters than high-F_1-amplified difference filters. Significant interactions between vowel and spectral resolution in both experiments (experiment 2a: $Z = 6.679$, $P < 0.001$; experiment 2b: $Z = 3.721$, $P < 0.001$) indicated that psychometric function slopes became steeper as more spectral channels were used in vocoding, similar to experiment 1.

The interaction between SCE and spectral resolution was statistically significant in experiment 2a ($Z = -2.805$, $P = 0.005$). SCEs were calculated at each level of spectral resolution following the same procedure as detailed in experiment 1. SCEs were smaller for 24-channel speech (response shift of 20.00 %, Fig. 6c) than at lower spectral resolutions (6 channels: response shift of 26.92 %, Fig. 6a; 12 channels: response shift of 28.92 %, Fig. 6b). All of these effects exceeded SCEs for full-spectrum versions of these stimuli (9.08 %, Fig. 6h; Stilp et al. 2015). The interaction between SCE and spectral resolution was not significant in experiment 2b ($Z = 1.028$, $P = 0.304$). SCEs were unexpectedly smaller at 12 channels than other spectral resolutions (6 channels: response shift of 9.85 %, Fig. 6e; 12 channels: response shift of 4.69 %, Fig. 6f; 24 channels: response shift of 11.69 %, Fig. 6g). Given that SCEs in the 12-channel condition of experiment 2a were not smaller than at other spectral resolutions, the reason for this result is unclear. Six- and 24-channel vocoded stimuli filtered by 25 % of the spectral envelope difference produced larger SCEs than the same processing for full-spectrum stimuli (5.09 %; Stilp et al. 2015).

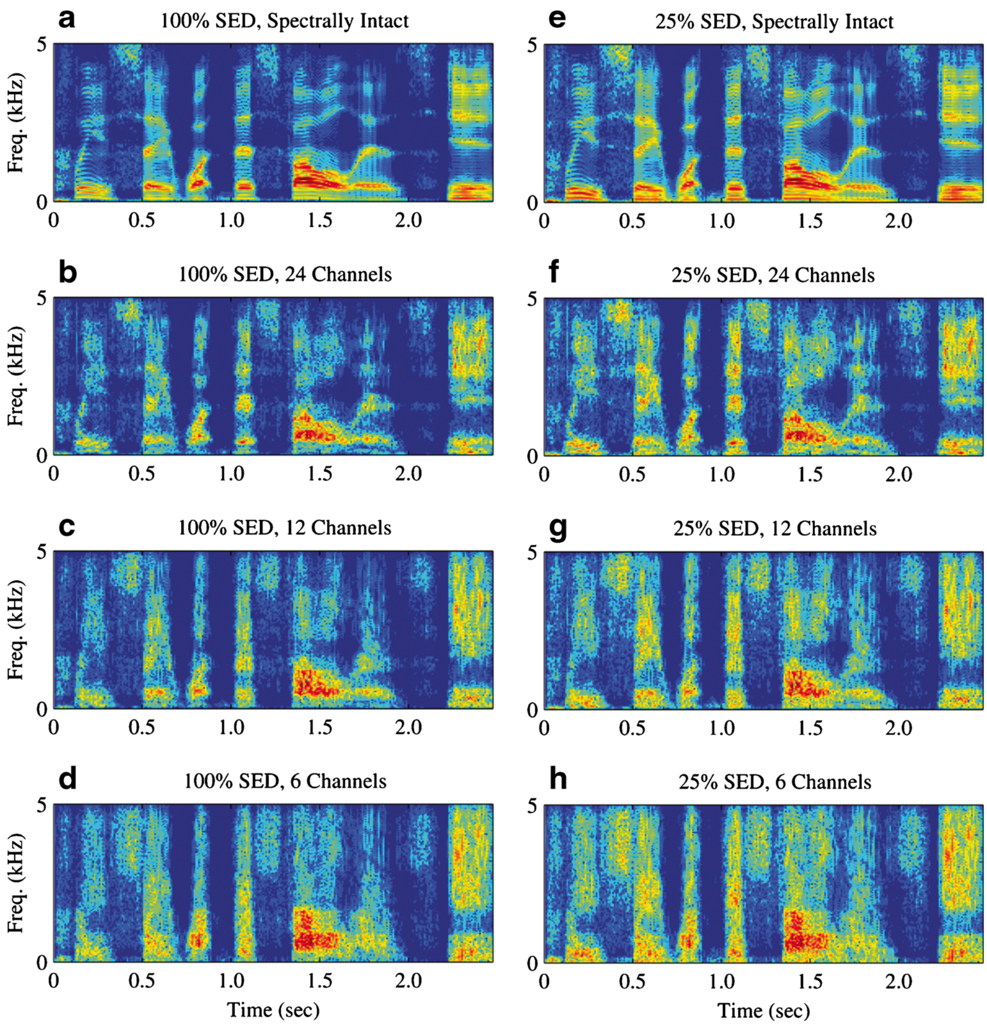

FIG. 5. Spectrograms of sample trials in experiment 2. The *left column* illustrates trials from experiment 2a, where the precursor sentence was processed by a filter reflecting 100 % of the spectral envelope difference (SED) of /ɛ/ minus /ɪ/ (high-F_1 emphasis). The *right column* illustrates trials from experiment 2b, where the precursor was processed by a filter reflecting only 25 % of the SED of /ɛ/ minus /ɪ/. The *top row* shows a sample trial with full-spectrum stimuli (from Stilp et al. 2015); *successive rows* show those stimuli vocoded with 24, 12, and 6 spectral channels, respectively. The target vowel is the /ɪ/ endpoint.

Finally, the three-way interaction between vowel, filter, and spectral resolution approached statistical significance for experiment 2a ($Z = -1.943$, $P = 0.052$) and was significant for experiment 2b ($Z = -2.092$,

TABLE 3

Results of generalized linear mixed effects models for experiments 2a (left) and 2b (right)

	Experiment 2a				*Experiment 2b*			
	Estimate	*SE*	*Z*	*P*	*Estimate*	*SE*	*Z*	*P*
Intercept	0.156	0.172	0.908	0.364	0.576	0.127	4.537	<0.001
Vowel (V)	0.351	0.064	5.505	<0.001	0.501	0.072	6.930	<0.001
Filter (F)	1.585	0.251	6.326	<0.001	0.612	0.134	4.555	<0.001
SR	−0.001	0.016	−0.071	0.943	−0.055	0.020	−2.723	0.006
V by F	−0.012	0.023	−0.537	0.591	−0.055	0.022	−2.484	0.013
V by SR	0.014	0.002	6.679	<0.001	0.008	0.002	3.721	<0.001
F by SR	−0.024	0.008	−2.805	0.005	0.008	0.008	1.028	0.304
V by F by SR	−0.006	0.003	−1.943	0.052	−0.006	0.003	−2.092	0.036

SR spectral resolution (number of vocoder channels), *SE* standard error of the mean

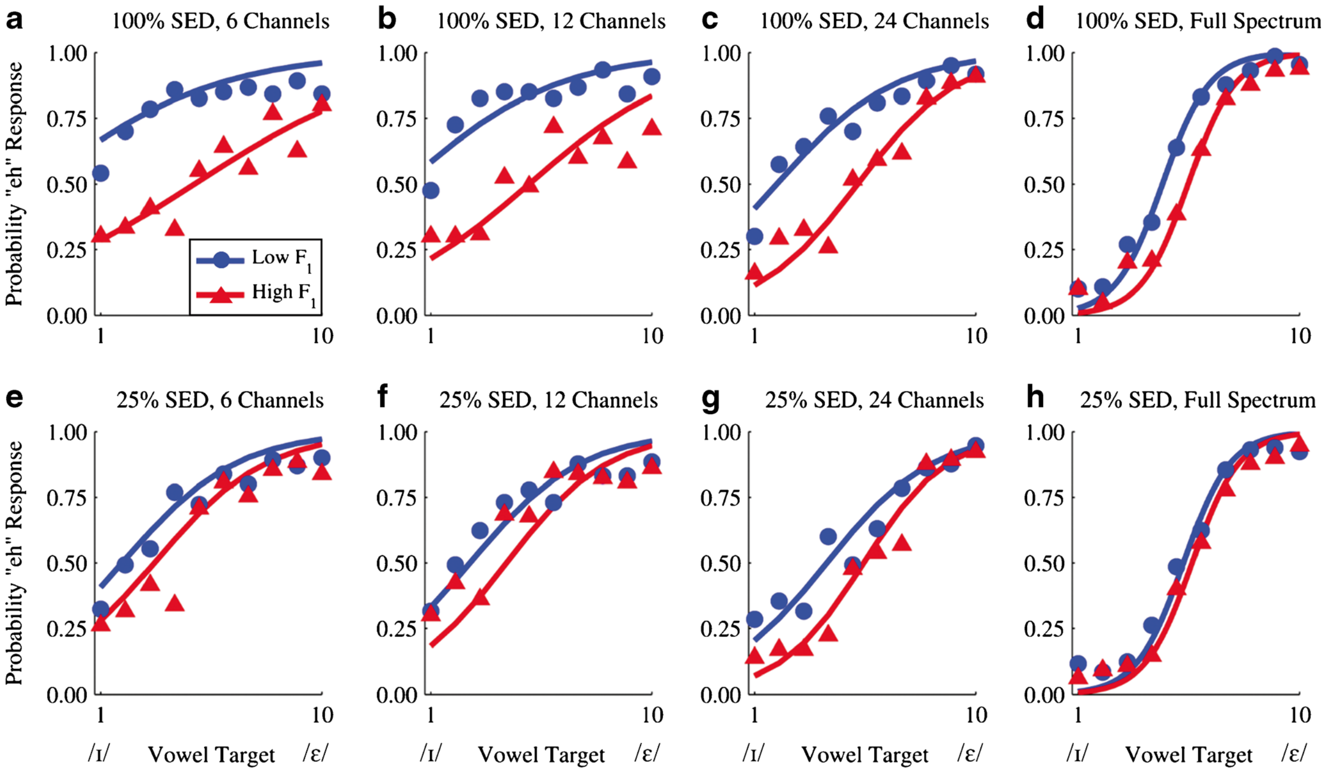

FIG. 6. Listener performance and model predictions for experiment 2 and related results for full-spectrum stimuli by Stilp et al. (2015). *Circles* represent mean probabilities of responding "eh" following the low-F_1-amplified precursor sentence (/ɪ/ minus /ɛ/); *triangles* represent mean probabilities of responding "eh" following the high-F_1-amplified precursor sentence (/ɛ/ minus /ɪ/). *Solid lines* depict predicted responses from the mixed effects models. The *top row* (**a**–**d**) depicts results for precursor sentences filtered by 100 % of the spectral envelope difference (SED) between vowel continuum endpoints; the *bottom row* (**e**–**h**) depicts results for precursor sentences filtered by 25 % of the SED. **a**–**c** Results from experiment 2a for the 6-, 12-, and 24-channel conditions, respectively. **d** Listener performance and model predictions for full-spectrum stimuli filtered by 100 % of the SED from Stilp et al. (2015). **e**–**g** Results from experiment 2b for the 6-, 12-, and 24-channel conditions, respectively. **h** Listener performance and model predictions for full-spectrum stimuli filtered by 25 % of the SED from Stilp et al. (2015).

$P = 0.036$). In experiment 2b, logistic regression slopes for low-F_1-amplified precursors changed minimally across spectral resolutions (slopes = 0.43, 0.44, and 0.46 for 6-, 12-, and 24-channel stimuli, respectively), but slopes for high-F_1-amplified precursors grew steeper with more spectral channels (slopes = 0.44, 0.48, and 0.58 for 6-, 12-, and 24-channel stimuli, respectively). The trend toward significance in experiment 2a followed this same general pattern but was less clear due to all regression slopes steepening as a function of spectral resolution, as in experiment 1a (low-F_1-amplified slopes = 0.28, 0.32, and 0.42; high-F_1-amplified slopes = 0.24, 0.32, and 0.49 for 6-, 12-, and 24-channel stimuli, respectively).

DISCUSSION

Spectral context effects span several decades in the speech perception literature, but reports of spectral context effects in the CI literature, whether with CI users or acoustic simulations of CIs, are scant (see "INTRODUCTION"). The present results revealed that stable properties of the acoustic context, such as a spectral peak or spectral shape that is relatively reliable across time, modified speech categorization in acoustic simulations of CI processing. This extends a host of earlier work reporting SCEs in perception of full-spectrum speech (Watkins 1991; Watkins and Makin 1996a, 1996b; Holt 2006; Sjerps et al. 2011; Stilp et al. 2015; Assgari and Stilp 2015; Stilp and Assgari 2017) or acoustically degraded speech (Watkins 1991; Sjerps et al. 2011). Importantly, these results demonstrate the separability of spectral context effects from speech intelligibility. This issue had not come up in previous investigations of SCEs, because precursor sounds and/or target speech sounds were highly intelligible due to being spectrally intact and presented in quiet. In experiment 1a, listeners exhibited great difficulty distinguishing vowels with only six channels of spectral resolution (see shallow psychometric function slopes in Fig. 3a). However, the long-term average spectrum of the preceding sentence still exerted substantial influence on the listeners' responses. While acoustic cues for a given speech sound may be difficult to detect or use for accurate speech recognition (i.e., intrinsic cues), acoustic properties of surrounding sounds can still exert considerable influence on identification of the target sound (i.e., extrinsic cues; Ladefoged and Broadbent 1957; Ainsworth 1975; Nearey 1989).

There are several reasons to expect that CI users' speech perception would be influenced by SCEs. First, healthy hearing is not a prerequisite for spectral context effects to influence speech perception. Stilp and Alexander (2016) reported that listeners with sensorineural hearing loss not only exhibited SCEs in vowel perception but exhibited significantly larger effects than those observed for normal-hearing listeners presented with the same stimuli. They attributed these enlarged SCEs to broadened auditory filters in sensorineural hearing loss. Broadened auditory filters would broaden the bandwidths of reliable spectral peaks, which has been shown to increase the magnitudes of SCEs (Stilp et al. 2015). Current CI processing strategies electrically stimulate more neurons than would be engaged by acoustic stimulation, which would also broaden the bandwidths of reliable spectral peaks. This increases the likelihood of observing SCEs in the CI users' speech perception, if not also larger SCEs than those reported for normal-hearing listeners.

Second, recent reports by Wang et al. (2012; 2015; 2016) demonstrated CI users' sensitivity to preceding acoustic context in speech and nonspeech tasks. Most germane to the present report, formant peaks in a target vowel sound were more detectable (and vowel recognition more accurate) when preceded by a precursor stimulus with spectral notches in the frequency regions corresponding to formant peaks (Wang et al. 2012; see also Goupell and Mostardi 2012). This resulted in enhancement effects, where differences between precursor and target spectra were perceptually enhanced (Viemeister 1980; Viemeister and Bacon 1982; Summerfield et al. 1984). Enhancement effects have been proposed to be related to SCEs, as both are perceptual magnifications of spectral changes in the signal (Kluender et al. 2003). Additionally, CI users incorporated acoustic context when making loudness judgments (Wang et al. 2015, 2016). Loudness comparisons between two stimuli were modified by the addition of a precursor stimulus before the target sound, its frequency content, and the temporal interval between precursor and target, among other stimulus properties. CI users displayed qualitatively similar context effects to normal-hearing listeners, demonstrating CI users' sensitivity to preceding acoustic context.

Third, the enhancement of spectral changes (as in auditory enhancement effects and SCEs) is not exclusive to peripheral (cochlear) processing but occurs in the central auditory system as well. Physiological correlates of psychophysical enhancement have been reported at the auditory nerve (Palmer et al. 1995), the cochlear nucleus (Scutt and Palmer 1998), and the inferior colliculus (Nelson and Young 2010). Dichotic stimulus presentation, where the preceding acoustic context and target sound are presented to opposite ears, still produced SCEs (Watkins 1991; Holt and Lotto 2002) and enhancement effects (Erviti et al. 2011; Carcagno et al. 2012). While peripheral neural encoding differs widely across acoustic and electrical hearing, the rest of the auditory system is similarly predicated on enhancing changes in the acoustic input. This suggests that spectral changes are emphasized in the CI users' auditory perception as well.

The present results suggest that SCEs in CI users' speech perception would be larger than those observed for NH listeners (Fig. 7). This is in the opposite direction from that of Winn et al. (2013) and a host of other investigations where smaller perceptual effects for CI users are attributed to the impoverished signal coming from the CI. It is important to note that overly large SCEs can be detrimental to accurate speech perception. Specifically, large SCEs can result in the miscategorization of previously unambiguous speech sounds (Fig. 1c). This is particularly evident when the preceding acoustic context possesses large reliable spectral peaks, as in experiments 1a (Fig. 3a–c) and 2a (Fig. 6a–c). Even with 24 spectral channels, the /ɪ/ endpoint of the vowel continuum (stimulus number 1) was categorized as /ɛ/ on roughly 30 % of the trials presenting the low-F_1-amplified precursor, and all other members of the continuum were categorized as /ɛ/ the majority of the time. These errors were exacerbated at lower spectral resolutions to the point where no vowel was categorized as /ɪ/ on the majority of trials when the preceding acoustic context had low-F_1 emphasis (Figs. 3a and 6a). Speech sound miscategorization owing to overly large SCEs was also reported for listeners with sensorineural hearing loss (Stilp and Alexander 2016), indicating that this difficulty cannot be solely attributed to noise vocoding. If CI users also exhibit oversized SCEs, digital signal processing approaches would be required to modify the influence of preceding acoustic context in order to mitigate such detrimental effects on speech recognition.

The present experiments used noise vocoding to model perceptual consequences of spectral degradation in CI processing. However, differences in intensity resolution across normal-hearing listeners and CI

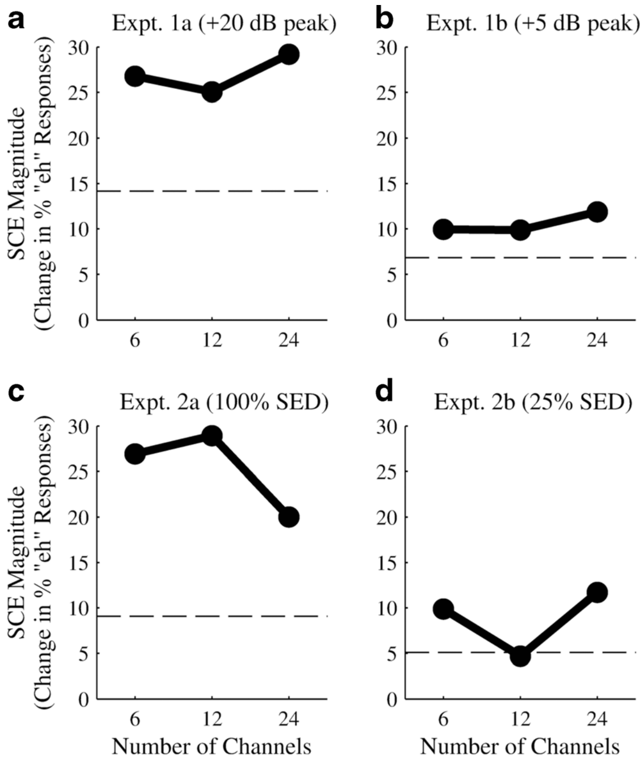

FIG. 7. Magnitudes of spectral contrast effects (SCEs, measured as the change in mean percentage of "eh" responses following low-F_1 versus high-F_1 precursor sentences) as a function of the number of spectral channels in noise vocoding. *Dashed lines* indicate SCE magnitudes reported for full-spectrum versions of the present stimuli by Stilp et al. (2015). In 11 of the 12 cases, larger SCEs were observed in perception of noise-vocoded stimuli than full-spectrum stimuli, especially when reliable spectral properties were more pronounced (**a**, **c**). *SED* spectral envelope difference.

users must also be considered. For CI users, intensity resolution is often a trade-off between superior intensity discrimination (Shannon 1983; Zeng 2004; but see Rogers et al. 2006) and vastly inferior dynamic ranges compared to normal-hearing listeners (Zeng and Shannon 1994, 1999; Nelson et al. 1996). Other factors relating to the stimulus, stimulus presentation, CI, and listener can influence intensity resolution, but dynamic range is of particular interest. Reduced dynamic range impairs speech recognition for CI users (Loizou et al. 2000; Zeng and Galvin 1999), and it might also limit the influence of extrinsic cues on speech sound recognition. For normal-hearing listeners, larger stable spectral peaks in the preceding acoustic context produced larger SCEs in both full-spectrum and noise-vocoded materials (Stilp et al. 2015; Assgari and Stilp 2015; Stilp and Alexander 2016; Stilp and Assgari 2017). For CI users, this relationship might be lessened if all stable spectral peaks above a certain magnitude (i.e., a given electrode's dynamic range) were peak clipped and produced the same-sized SCEs. This is an important question for future research, not just only whether CI users demonstrate SCEs but also the nature of their sensitivity to stable spectral properties of the listening environment.

CONCLUSION

Preceding acoustic context significantly influenced perception of speech that was spectrally degraded via noise vocoding. This influence of spectral context was observed even when stable spectral properties in preceding sounds were very modest (+5 dB spectral peak in experiment 1b; filtering by only 25 % of the total spectral envelope difference in experiment 2b). This generality of SCEs bears considerable importance for speech perception by CI users. The present results suggest that stable spectral properties, whether band limited or broadband and whether dramatic or modest, would likely alter their speech categorization. Additionally, in all conditions but one, SCEs were larger than those observed for full-spectrum versions of the same stimuli. This raises important questions about not just whether preceding acoustic context influences CI users' speech perception but the extent of this influence as well. This knowledge could then inform digital signal processing in CIs to more faithfully encode reliable acoustic properties of the listening environment and ensure appropriate perceptual magnification of changes from these properties.

ACKNOWLEDGEMENTS

We thank Matthew Winn and three anonymous reviewers for helpful comments on an earlier draft of this manuscript. We also thank Kara Hendrix, Almina Klanco, Orlando Madriz, Asim Mohiuddin, Emily Nations, and Caitlyn Stromatt for assistance with data collection.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of Interest The authors declare that they have no conflict of interest.

REFERENCES

Ainsworth W (1975) Intrinsic and extrinsic factors in vowel judgments. In: Fant G, Tatham M (eds) Auditory analysis and perception of speech. Academic Press, London, pp 103–113

Aravamudhan R, Lotto AJ (2004) Perceptual overshoot in listeners with cochlear implants. J Acoust Soc Am 116:2523. doi:10.1121/1.4785070

Aravamudhan R, Lotto AJ (2005) Phonetic context effects in adult listeners with cochlear implants. J Acoust Soc Am 118:1962. doi:10.1121/1.4781551

Assgari AA, Stilp CE (2015) Talker information influences spectral contrast effects in speech categorization. J Acoust Soc Am 138(5):3023–3032. doi:10.1121/1.4934559

Bates DM, Maechler M, Bolker B, Walker S (2014) lme4: linear mixed-effects models using Eigen and S4. R package version 1:1–7 http://cran.r-project.org/package=lme4

Blamey P, Artieres F, Başkent D, Bergeron F, Beynon A, Burke E et al (2012) Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: an update with 2251 patients. Audiol Neurotol 18:36–47. doi:10.1159/000343189

Boersma P, 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 July 7, 2016).

Carcagno S, Semal C, Demany L (2012) Auditory enhancement of increments in spectral amplitude stems from more than one source. J Assoc Res Otolaryngol 13(5):693–702. doi:10.1007/s10162-012-0339-y

Donaldson GS, Rogers CL, Cardenas ES, Russell BA, Hanna NH (2013) Vowel identification by cochlear implant users: contributions of static and dynamic spectral cues. J Acoust Soc Am 134(4):3021–3028. doi:10.1121/1.4820894

Donaldson GS, Rogers CL, Johnson LB, Oh SH (2015) Vowel identification by cochlear implant users: contributions of duration cues and dynamic spectral cues. J Acoust Soc Am 138(1):65–73. doi:10.1121/1.4922173

Erviti M, Semal C, Demany L (2011) Enhancing a tone by shifting its frequency or intensity. J Acoust Soc Am 129(6):3837–3845. doi:10.1121/1.3589257

Goupell MJ, Mostardi MJ (2012) Evidence of the enhancement effect in electrical stimulation via electrode matching. J Acoust Soc Am 131(2):1007–1010. doi:10.1121/1.3672650

Greenwood DD (1990) A cochlear frequency-position function for several species—29 years later. J Acoust Soc Am 87(6):2592–2605. doi:10.1121/1.399052

Hillenbrand J, Getty LA, Clark MJ, Wheeler K (1995) Acoustic characteristics of American English vowels. J Acoust Soc Am 97(5):3099–3111. doi:10.1121/1.411872

Holt LL (2005) Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psych Sci 16(4):305–312. doi:10.1111/j.0956-7976.2005.01532.x

Holt LL (2006) The mean matters: effects of statistically defined nonspeech spectral distributions on speech categorization. J Acoust Soc Am 120(5):2801–2817. doi:10.1121/1.2354071

Holt LL, Lotto AJ (2002) Behavioral examinations of the level of auditory processing of speech context effects. Hear Res 167(1):156–169. doi:10.1016/S0378-5955(02)00383-0

Johnson K (1990) The role of perceived speaker identity in F0 normalization of vowels. J Acoust Soc Am 88(2):642–654. doi:10.1121/1.399767

Kluender KR, Coady JA, Kiefte M (2003) Sensitivity to change in perception of speech. Sp Comm 41(1):59–69. doi:10.1016/S0167-6393(02)00093-6

Kong YY, Winn MB, Poellmann K, Donaldson GS (2016) Discriminability and perceptual saliency of temporal and spectral cues for final fricative consonant voicing in simulated cochlear-implant and bimodal hearing. Trends Hear. doi:10.1177/2331216516652145

Ladefoged P, Broadbent DE (1957) Information conveyed by vowels. J Acoust Soc Am 29(1):98–104. doi:10.1121/1.1908694

Loizou PC, Poroy O (2001) Minimum spectral contrast needed for vowel identification by normal-hearing and cochlear implant listeners. J Acoust Soc Am 110(3):1619–1627. doi:10.1121/1.1388004

Loizou PC, Dorman M, Fitzke J (2000) The effect of reduced dynamic range on speech understanding: implications for patients with cochlear implants. Ear Hear 21(1):25–31

Lotto AJ, Kluender KR (1998) General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. Percept Psychophys 60(4):602–619. doi:10.3758/BF03206049

Lotto AJ, Kluender KR, Holt LL (1997) Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). J Acoust Soc Am 102(2):1134–1140. doi:10.1121/1.419865

Mitterer H (2006) Is vowel normalization independent of lexical processing? Phonetica 63(4):209–229. doi:10.1159/000097306

Moberly AC, Lowenstein JH, Tarr E, Caldwell-Tarr A, Welling DB, Shahin AJ, Nittrouer S (2014) Do adults with cochlear implants rely on different acoustic cues for phoneme perception than adults with normal hearing? J Speech Lang Hear Res 57(2):566–582. doi:10.1044/2014_JSLHR-H-12-0323

Moberly AC, Lowenstein JH, Nittrouer S (2016) Word recognition variability with cochlear implants: the degradation of phonemic sensitivity. Otol Neurotol 37(5):470–477. doi:10.1097/MAO.0000000000001001

Nearey TM (1989) Static, dynamic, and relational properties in vowel perception. J Acoust Soc Am 85(5):2088–2113. doi:10.1121/1.397861

Nelson PC, Young ED (2010) Neural correlates of context-dependent perceptual enhancement in the inferior colliculus. J Neurosci 30(19):6577–6587. doi:10.1523/JNEUROSCI.0277-10.2010

Nelson DA, Schmitz JL, Donaldson GS, Viemeister NF, Javel E (1996) Intensity discrimination as a function of stimulus level with electric stimulation. J Acoust Soc Am 100:2393–2414. doi:10.1121/1.417949

Palmer AR, Summerfield Q, Fantini DA (1995) Responses of auditory-nerve fibers to stimuli producing psychophysical enhancement. J Acoust Soc Am 97(3):1786–1799. doi:10.1121/1.412055

R Development Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna http://www.r-project.org/

Rogers CF, Healy EW, Montgomery AA (2006) Sensitivity to isolated and concurrent intensity and fundamental frequency increments by cochlear implant users under natural listening conditions. J Acoust Soc Am 119:2276–2287. doi:10.1121/1.2167150

Scutt MJ, Palmer AR (1998) Physiological enhancement in cochlear nucleus using single tone precursors. Assoc Res Otolaryngol Abs:381.

Shannon RV (1983) Multichannel electrical stimulation of the auditory nerve in man. I Basic psychophysics Hear Res 11(2):157–189. doi:10.1016/0378-5955(83)90077-1

Sjerps MJ, Mitterer H, McQueen JM (2011) Constraints on the processes responsible for the extrinsic normalization of vowels. Atten Percept Psychophys 73(4):1195–1215. doi:10.3758/s13414-011-0096-8

Stephens JD, Holt LL (2003) Preceding phonetic context affects perception of nonspeech (L). J Acoust Soc Am 114(6):3036–3039. doi:10.1121/1.1627837

Stilp CE, Alexander JM (2016)Spectral contrast effects in vowel categorization by listeners with sensorineural hearing loss. J Acoust Soc Am 139(4):2047. doi: 10.1121/2.0000233

Stilp CE, Assgari AA (2017) Consonant categorization exhibits a graded influence of surrounding spectral context. J Acoust Soc Am 141(2):EL153–EL158. doi:10.1121/1.4974769

Stilp CE, Alexander JM, Kiefte M, Kluender KR (2010) Auditory color constancy: calibration to reliable spectral properties across speech and nonspeech contexts and targets. Atten Percept Psychophys 72(2):470–480. doi:10.3758/APP.72.2.470

Stilp CE, Anderson PW, Winn MB (2015) Predicting contrast effects following reliable spectral properties in speech perception. J Acoust Soc Am 137(6):3466–3476. doi:10.1121/1.4921600

Summerfield Q, Haggard M, Foster J, Gray S (1984) Perceiving vowels from uniform spectra: phonetic exploration of an auditory aftereffect. Percept Psychophys 35(3):203–213. doi:10.3758/BF03205933

Viemeister NF (1980) Psychophysical, physiological, and behavioral studies in hearing. In: Bring GVD, Bilsen FA (eds) Adaptation of masking. University Press, Delft, pp 190–197

Viemeister NF, Bacon SP (1982) Forward masking by enhanced components in harmonic complexes. J Acoust Soc Am 71(6):1502–1507. doi:10.1121/1.387849

Wang N, Kreft H, Oxenham AJ (2012) Vowel enhancement effects in cochlear-implant users. J Acoust Soc Am 131(6):EL421–EL426. doi:10.1121/1.4710838

Wang N, Kreft H, Oxenham AJ (2015) Loudness context effects in normal-hearing listeners and cochlear-implant users. J Assoc Res Otolaryngol 16(4):535–545. doi:10.1007/s10162-015-0523-y

Wang N, Kreft H, Oxenham AJ (2016) Induced loudness reduction and enhancement in acoustic and electric hearing. J Assoc Res Otolaryngol. doi:10.1007/s10162-016-0563-y

Watkins AJ (1991) Central, auditory mechanisms of perceptual compensation for spectral envelope distortion. J Acoust Soc Am 90(6):2942–2955. doi:10.1121/1.401769

Watkins AJ, Makin SJ (1996a) Some effects of filtered contexts on the perception of vowels and fricatives. J Acoust Soc Am 99(1):588–594. doi:10.1121/1.414515

Watkins AJ, Makin SJ (1996b) Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. J Acoust Soc Am 99(6):3749–3757. doi:10.1121/1.414981

Winn MB, Litovsky RY (2015) Using speech sounds to test functional spectral resolution in listeners with cochlear implants. J Acoust Soc Am 137(3):1430–1442. doi:10.1121/1.4908308

Winn MB, Chatterjee M, Idsardi WJ (2012) The use of acoustic cues for phonetic identification: effects of spectral degradation and electric hearing. J Acoust Soc Am 131(2):1465–1479. doi:10.1121/1.3672705

Winn MB, Rhone AE, Chatterjee M, Idsardi WJ (2013) The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. Front Psych 4. doi:10.3389/fpsyg.2013.00824

Zeng FG (2004) Compression and cochlear implants. In: Bacon SP (ed) Compression: from cochlea to cochlear implants. Springer-Verlag, New York, pp 184–220

Zeng FG, Galvin JJ III (1999) Amplitude mapping and phoneme recognition in cochlear implant listeners. Ear Hear 20(1):60–74

Zeng FG, Shannon RV (1994) Loudness-coding mechanisms inferred from electric stimulation of the human auditory system. Science 264(5158):564–565. doi:10.1126/science.8160013

Zeng FG, Shannon RV (1999) Psychophysical laws revealed by electric hearing. Neuroreport 10(9):1931–1935