The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing[a)]

Matthew B. Winn[b)] and Monita Chatterjee
Department of Hearing and Speech Sciences, University of Maryland, College Park, 0100 Lefrak Hall, College Park, Maryland 20742

William J. Idsardi
Department of Linguistics, University of Maryland, College Park, 1401 Marie Mount Hall, College Park, Maryland 20742

(Received 7 September 2010; revised 10 October 2011; accepted 5 December 2011)

Although some cochlear implant (CI) listeners can show good word recognition accuracy, it is not clear how they perceive and use the various acoustic cues that contribute to phonetic perceptions. In this study, the use of acoustic cues was assessed for normal-hearing (NH) listeners in optimal and spectrally degraded conditions, and also for CI listeners. Two experiments tested the tense/lax vowel contrast (varying in formant structure, vowel-inherent spectral change, and vowel duration) and the word-final fricative voicing contrast (varying in F1 transition, vowel duration, consonant duration, and consonant voicing). Identification results were modeled using mixed-effects logistic regression. These experiments suggested that under spectrally-degraded conditions, NH listeners decrease their use of formant cues and increase their use of durational cues. Compared to NH listeners, CI listeners showed decreased use of spectral cues like formant structure and formant change and consonant voicing, and showed greater use of durational cues (especially for the fricative contrast). The results suggest that although NH and CI listeners may show similar accuracy on basic tests of word, phoneme or feature recognition, they may be using different perceptual strategies in the process. © *2012 Acoustical Society of America*. [DOI: 10.1121/1.3672705]

PACS number(s): 43.71.Es, 43.71.Ky, 43.71.Gv, 43.66.Ts [MSS] Pages: 1465–1479

I. INTRODUCTION

In view of the remarkable success of the cochlear implant (CI) as a prosthetic device (Zeng *et al*., 2008), and in the context of continually growing body of research on cochlear implants, literature on phonetic cue perception must be expanded to acknowledge the abilities of individuals fitted with these devices. It is well known that a major obstacle to accurate speech understanding with electric hearing (the use of a CI) is the poor spectral resolution offered by these devices, owing to the limited number of independent spectral processing channels (Fishman *et al*., 1997; Friesen *et al*., 2001), interactions between the electrodes which carry information from those channels (Chatterjee and Shannon, 1998) as well as a distorted tonotopic map (Fu and Shannon, 1999). Thus, the subtle fine-grained spectral differences perceptible to those with normal hearing are not reliably distinguished by those who use CIs (Kewley-Port and Zheng, 1998; Loizou and Poroy, 2001; Henry *et al*., 2005). In view of some of these studies, it is presumed that phonetic cues driven by spectral contrasts would be most challenging for CI listeners. Although numerous studies have explored word, phoneme and feature recognition in various kinds of degraded conditions, few have explored the use of acoustic cues that contribute to these perceptions.

Not all sound components are compromised in electric hearing; temporal processing can be as good or better than that of normal-hearing (NH) listeners, as evidenced by temporal modulation transfer functions Shannon (1992) and gap detection tasks (Shannon, 1989). Thus, although some phonetic cues are obscured by spectral degradation, it is expected that CI listeners should be able to use nonspectral cues in speech, which might be carried by the temporal amplitude envelope or segment duration. Fittingly, experiments have revealed a large number of errors on place-of-articulation perception (which relies primarily upon spectral cues in the signal, such as spectral peak frequencies and formant transitions), while the manner-of-articulation and voicing features are rarely misperceived, because they can be transmitted via temporal cues, which are well-maintained in electric hearing (Dorman *et al*., 1991). Similar results of poor place perception and excellent voicing perception have been shown continually for NH listeners listening to simulations of cochlear implants (i.e., Shannon *et al*., 1995).

A. Trading relations in phonetic feature perception

Phonetic contrasts are signaled by various acoustic dimensions in the temporal and spectral domains. Those dimensions that are used perceptually to identify speech sounds are called "phonetic cues"; they are acoustic cues

[a)]This paper includes some material that appeared previously in M. Winn's doctoral dissertation. Portions of this work were presented in "Phonetic cues are weighted differently when spectral resolution is degraded," the American Auditory Society Annual Meeting, Scottsdale, AZ, March 2010 and in "Modulation of phonetic cue-weighting in adverse listening conditions," 34th Mid-Winter meeting of the Association for Research in Otolaryngology, February 2011.

[b)]Author to whom correspondence should be addressed. Electronic mail: mwinn83@gmail.com

that contribute to phonetic categorization. For example, the first formant (F1) of a vowel sound corresponds to the height of that vowel; as the vowel height decreases, F1 increases. Hence, F1 serves as a phonetic cue for contrastive vowel height. There are multiple co-occurring phonetic cues for any particular contrast, which creates a high amount of redundancy in the signal. A classic example is the contrast between voiced and voiceless stops in word-medial position, which has been claimed to contain at least 16 different acoustic cues (Lisker, 1978). A wealth of literature has revealed that changes in one acoustic dimension can be compensated by conflicting changes in another dimension (for multiple examples, see Repp, 1982). For example, trading relations can be observed in the integration of cues for syllable-initial stop consonant voicing; changes in voice-onset-time that signal voicing can be somewhat offset by changes in the pitch domain that signal voicelessness (Whalen *et al*., 1993). As these and other cues covary in natural speech, the listener must integrate them in a way that yields reliable and accurate identification of the incoming information. It has been shown that the use of acoustic cues for phonetic contrasts is affected by the developmental age (Nittrouer, 2004, 2005) as well as language background (Morrison, 2005) of a listener. Perhaps it is also affected by spectral resolution in a way that is useful for understanding the experience of CI listeners relative to NH listeners.

Perception of acoustic dimensions such as duration, formant frequencies, or the time-varying amplitude envelope all depend on the fidelity of the stimulus. Trading relations between temporal and spectral signal fidelity have been observed for the perception of English consonants and vowels (Xu *et al*., 2005) as well as for Mandarin lexical tones (Xu and Pfingst, 2003). In those studies, as the degree of spectral resolution was decreased, the level of temporal resolution played a larger role in listeners' perceptual accuracy. These experiments were carried out using noise-band vocoding (NBV) (to be described in detail later) based on that used by Shannon *et al*. (1995), which is commonly used to simulate electric hearing. The current study takes a similar approach to ask a different question—beyond showing correct and incorrect performance on word and phoneme recognition tasks, what can we learn about the avenue that listeners take to achieve this performance? There is reason to believe that listeners will adapt to an altered stimulus input by changing the relative importance of signal components (Francis *et al*., 2000, 2008, among many examples). Perhaps cochlear implant listeners and normal-hearing listeners in degraded conditions can adopt new strategies that would suit the challenges and residual abilities available to them.

Readers will recognize the central issue in this paper as one of cue-trading/cue-weighting. There are several models of cue-weighting present in the literature, and the current study was not designed to explicitly test or challenge any of them. Of particular interest, however, are those accounts which specifically acknowledge the reliability with which the signal is represented. The cue weighting-by-reliability model of Toscano and McMurray (2010) suggests that the weighting of acoustic cues in phonetic perception can be predicted by their distributional properties in the input; the basic theme of this research permeates other work as well, such as that of Holt and Idemaru (2011). Specifically, a cue is more reliable (and hence should be more heavily weighted) if the contrastive level means are far apart and have low variance. In the current study, it could be argued that spectral degradation (whether simulated or via electric hearing) would diminish the reliability of spectral cues like vowel formants as well as formant transitions, since there is no clear placement of these peaks in the degraded spectrum. The temporal dimensions (duration, time-varying amplitude envelope), however, should remain relatively unchanged.

In summary, the current experiments were conducted to explore whether spectral degradation affects listeners' use of various acoustic cues in phonetic identification. It was hypothesized that if spectral resolution were poor, listeners would be less affected by phonetic cues in the spectral domain, and more affected by phonetic cues in the temporal domain. This hypothesis would be supported by two kinds of results: (1) normal-hearing listeners using phonetic cues differently when spectral resolution is artificially degraded and (2) cochlear implant listeners using phonetic cues in a way that is different from normal-hearing listeners. The hypothesis was tested using two different phonetic contrasts, described below.

II. EXPERIMENT 1: THE LAX/TENSE VOWEL DISTINCTION

A. Review of acoustic cues

The first experiment explored the high-front lax/tense vowel contrast (/I/ and /i/) in English, which distinguishes word pairs such as hit/heat, fill/feel, hid/heed, and bin/bean. The cues that contribute to this distinction include the spectral dimensions of formant structure and vowel-inherent spectral change (VISC), as well as vowel duration. Formant structure has long been known to correspond to vowel categorization, albeit with a considerable amount of overlap between categories (Hillenbrand *et al*., 1995). Still, this cue is extremely powerful; using only steady-state formants synthesized from measurement taken at one timepoint in a vowel, human listeners identify vowels with roughly 75% accuracy (Hillenbrand and Gayvert, 1993). Automatic pattern classifiers show similar performance with just one sample of formant structure (i.e., a spectral snapshot) (Hillenbrand *et al*., 1995).

VISC refers to the "relatively slowly varying changes in formant frequencies associated with vowels themselves, even in the absence of consonantal context" (Nearey and Assmann, 1986). Throughout production of the lax vowel /I/, F1 increases and F2 decreases; the tense vowel /i/ is relatively steady-state by comparison, with only a negligible amount of formant movement, if any (Hillenbrand *et al*., 1995). VISC plays a role in vowel classification, as indicated by at least four kinds of data: (1) measurement of dynamic formant values from production data (Nearey and Assmann, 1986; Hillenbrand *et al*., 1995), (2) results of pattern classifiers show better performance when spectral change is included as a factor (Zahorian and Jagharghi, 1993; Hillenbrand *et al*., 1995), (3) listeners reliably identify vowels with only snapshots of the onset and offset (with silent or masked center portions) (Jenkins *et al*., 1983; Parker and Diehl,

1984; Nearey and Assmann, 1986), and (4) human listeners show improved identification results when vowels include natural patterns of spectral change; there is generally a 23%–26 % decline in accuracy for vowels whose formant structure lacks spectral change (Hillenbrand and Nearey, 1999; Assmann and Katz, 2005). When VISC is neutralized, there is a significant decline in /I/ recognition, while the vowel /i/ is identified virtually perfectly (Assmann and Katz, 2005), consistent with the acoustics of these vowels.

The duration of tense vowels tend to be longer than that of lax vowels by roughly 33%–80%, depending on the particular contrast and context (House, 1961; Hillenbrand *et al*., 1995). However, the role of duration in vowel perception has not always been clear; it appears to be driven at least in part by the fidelity of the stimulus. Ainsworth (1972) showed that duration can modulate identification of vowels synthesized with two steady-state formants. Bohn and Flege (1990) and Bohn (1995) revealed a small effect of duration for the i/I contrast when using three steady-state formants. However, these results are challenged by other studies that preserved relatively richer spectral detail, including time-varying spectral information (Hillenbrand *et al*., 1995, 2000; Zahorian and Jagharghi, 1993). Using modified natural speech, Hillenbrand *et al*. (2000) reported that duration-based misidentifications of the I/i contrast were especially rare (with an error rate of less than 1%). An emergent theme from Hillenbrand *et al*. (2000), Nittrouer (2004), and Assmann and Katz (2005) is that the use of acoustic cues in vowels is affected by signal fidelity, to the extent that commonly used formant synthesizers are likely to underestimate the role of time-varying spectral cues, and to overestimate the role of durational cues. That is, listeners use phonetic cues differently depending on the quality with which the sound is presented.

Although considerable improvements in speech synthesis and manipulation have improved the quality of signals in perceptual experiments, signal degradation is inescapable for individuals with cochlear implants. Iverson *et al*. (2006) remarked, "It would be surprising if exactly the same cues were used when recognizing vowels via cochlear implants and normal hearing, because the sensory information provided by acoustic and electric hearing differ substantially." Despite the aforementioned trend observed in spectral and temporal signal fidelity, Iverson *et al*. (2006) did not find evidence to suggest that duration was more heavily used by CI listeners or NH listeners in degraded conditions. In fact, as spectral resolution was degraded from 8 to 4 to 2 channels (each representing progressively worse resolution, to be explained further in Sec. II B), NH listeners showed *less* recovery of duration information in the signal. This counterintuitive result may have arisen because of the methods by which duration cue use was assessed. The experimenters used information transfer analysis (ITA) (Miller and Nicely, 1955) to track phonetic features that were recovered or mistaken in the identification tasks. Although these features are commonly thought to correspond regularly to acoustic dimensions (i.e., vowel height as variation in F1 frequency, vowel advancement as variation in F2 frequency, lax/tense as duration), ITA by itself does not reveal the mechanisms (cues) by which the features are recovered. This is particularly important for the duration cue; most dialects of English do not contain vowel pairs that contrast exclusively by duration. Thus, for any long or short vowel in English (as coded in ITA), there are accompanying covarying spectral cues. If a listener relies on these spectral cues (as would be predicted on the basis of aforementioned work), then it is not surprising that "duration" information transmission declined as spectral resolution decreased. In the ITA sort of analysis, "duration" could be merely a different name for spectral information, unless the latter has been specifically controlled. The question remains then, as to whether changes in vowel duration play a greater role in vowel identification when spectral resolution is degraded.

Despite of the limitations of the ITA-based analysis, the work by Iverson *et al*. (2006) is to be commended for laying the groundwork for studying the role of varying acoustic cues with varying degrees of temporal and spectral resolution. This approach has been only sparingly applied to the problem of speech perception by CI listeners (Dorman *et al*., 1991, is a rare example), and it is the aim of the present paper to explore it further using two contrasts that have been shown to involve both spectral and temporal cues. Many previous experiments (Hillenbrand and Nearey, 1999; Hillenbrand *et al*., 2000; Iverson *et al*., 2006) have assessed the role of multiple cues by retaining them or neutralizing them in a dichotic fashion. The current experiment seeks to expand upon this work by manipulating acoustic cues gradually and orthogonally, so as to assess their effects in a more fine-grained way that is unfeasible in experiments that test for many vowels and consonants concurrently.

Some prior work indicates that listeners with hearing impairment do exhibit altered use of acoustic cues in speech perception. In a place-of articulation identification task, Dorman *et al*. (1991) showed that, compared to NH listeners, CI listeners were affected more heavily by the spectral tilt of a stop consonant; NH listeners relied instead on formant transitions. Kirk *et al*. (1992) found that CI listeners were able to make use of static formant cues in vowels, but did not take advantage of the formant transition contrasts used by NH listeners. This would suggest that the dynamic formant cue for lax vowels may be compromised in degraded conditions. Accordingly, Dorman and Loizou (1997) indicated that CI listeners identified the lax vowel /I/ with accuracy similar to that of NH listeners in conditions where VISC is neutralized (Hillenbrand and Gayvert, 1993). We therefore expected the perception of speech sounds by CI listeners to fall in line with predictions informed by the aforementioned work that implicates signal degradation as an influential force on the use of durational cues. We thus predicted that as spectral resolution became poorer, use of formant cues would decline, the use of VISC cues would decline (if at all present), and the use of temporal cues would increase.

B. Methods

1. Participants

Participants included 15 adult (14 between the ages of 19–26; mean age 22.7 years, and one 63 year-old) listeners

TABLE I. Relevant demographic information about the CI participants in this study. All used the ACE processing strategy and the MP1+2 stimulation mode except for C30, who used the CIS strategy.

ID No.	Gender	Etiology of HL	Duration of HL	Age at testing	Age at impl.	Device	Pulse rate
C1	F	Unknown	Unknown	66	63	Freedom	900
C2	F	Genetic	10 years	66	63	Freedom	1800
C3	M	Unknown	22 years	64	57	N 24	900
C4	M	Labyrinthitis	11 years	50	40	N 24	720
C5	M	Unknown	Unknown	56	54	Med-El	1515
C6	F	Measles	59 years	71	66	Freedom	1800
C7	F	Unknown	4 years	73	69	Freedom	2400

with normal hearing, defined as having pure-tone thresholds $\leq$20 dB HL from 250–8000 Hz in both ears (ANSI, 2004). A second group of participants included seven adult (age 50–73; mean age 63.5 years) recipients of cochlear implants. CI listeners were all post-lingually deafened. Six were users of the Cochlear Freedom or N24 devices; one used the Med-El device. See Table I for demographic information and speech processor parameters for each CI user. All participants were native speakers of American English and were screened for fluency in languages for which vowel duration is a phonemic feature (i.e., Finnish, Hungarian, Arabic, Vietnamese, etc.), to ensure that no participant entered with *a priori* bias towards durational feature sensitivity. Normal-hearing participants 01 (the first author) and 02 were highly familiar with the stimuli, having been involved in pilot testing and the construction of the materials. It should be noted that the age difference between the normal-hearing and cochlear implant listener group is substantial, and can influence auditory processing in a way that is relevant to this study. Specifically, auditory temporal processing is known to be deficient in older listeners (Gordon-Salant and Fitzgibbons, 1999). The current study explores whether auditory cues in the temporal domain can overcome those that are compromised in the spectral domain. These listeners may or may not experience deficiencies in the temporal domain that could complicate this matter. Aside from this, there also exists variability in the durations and etiologies of deafness among the impaired listener group (as is the case in virtually all studies that use CI listeners). For these reasons, direct statistical comparisons between the normal-hearing listeners and cochlear implant listeners are limited in their utility and thus omitted from this paper.

2. Stimuli

a. Speech synthesis. Words were synthesized to resemble "hit" and "heat." The vowels in these words varied by formant structure (in seven steps, with the first four formants all simultaneously varying), vowel-inherent spectral change (in five steps, with the first three formants all varying dynamically) and vowel duration (in seven steps). See Table II for a detailed breakdown of the levels for each parameter. This $7 \times 7 \times 5$ continuum of words was synthesized using HLSYN (Hanson *et al.*, 1997; Hanson and Stevens, 2002). Formant structure was based off values reported in the online database of Hillenbrand *et al.* (1995); it was expanded beyond the average values in their corresponding publication to represent a realistic natural range of production. Formant continuum steps were interpolated using the Bark frequency scale (Zwicker and Terhardt, 1980) to reflect the nonlinear frequency spacing in the human auditory system. Levels in Bark frequency were converted to Hz in this article to facilitate ease of interpretation. A second dimension of stimulus construction varied by the amount and direction of vowel-inherent spectral change (VISC). Although there are various ways of modeling this cue (Morrison and Nearey, 2007), it is represented here in terms of the difference in the F1, F2, or F3 frequency (in Hz) from the 20% to the 80% timepoints in the vowel. All three formants were changed in accordance with data from Hillenbrand *et al.* (1995), except the fourth formant, which was kept constant. The penultimate items in this VISC continuum were modeled after typical lax and tense vowels, and the continuum endpoints were expanded along this parameter, again to account for productions outside the means reported by Hillenbrand

TABLE II. Acoustic parameter levels defining the continua of formants, vowel-inherent spectral change, and vowel duration. Each parameter was varied orthogonally.

		Step number						
		1	2	3	4	5	6	7
Formants	F1	446	418	403	389	375	362	335
(Hz)	F2	1993	2078	2122	2167	2213	2260	2357
	F3	2657	2717	2747	2778	2809	2841	2905
	F4	3599	3618	3628	3637	3647	3657	3677
VISC	F1	49	33	16	0	−16		
(change	F2	−287	−191	−96	0	96		
in Hz)	F3	−33	−22	−11	0	11		
	F4	0	0	0	0	0		
Duration (ms)		85	100	108	115	122	130	145

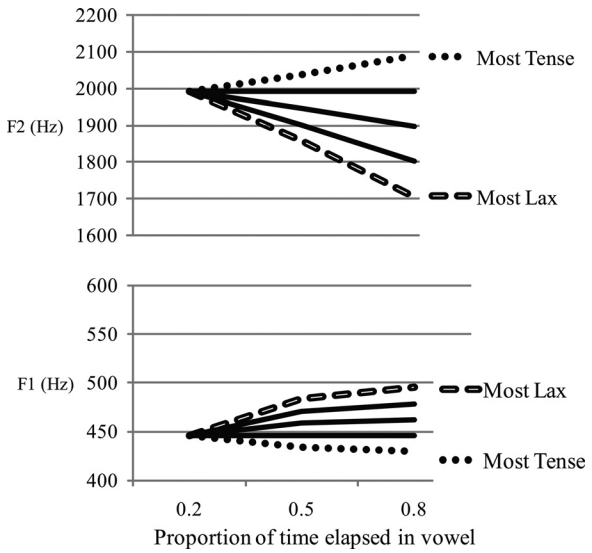

FIG. 1. Stylized representation of different levels of VISC applied to the same formant structure.

et al. See Table II for a detailed breakdown of this parameter, and Fig. 1 for a schematic illustration of its effects on formant structure. Vowel durations were modeled from characteristic durations of /i/ and /I/ (before voiceless stop sounds) reported by House (1961), and linearly interpolated (see Table II). Word-initial [h] was 60 ms of steady voiceless/aperiodic formant structure that matched that at the onset of the vowel; necessarily, the initial consonant was also varied as a result of the formant continuum. Word-final [−t] transitions targets for F1, F2, F3, and F4 were 300, 2000, 2900, and 3500 Hz, respectively, as used by Bohn and Flege (1990). These transitions all began at the 80% timepoint in the vowel (although this decision resulted in slightly different transition speeds depending on overall duration, it was necessary to ensure that the entire 20 %–80 % VISC trajectory could be realized). The formant transition was followed by a 65 ms of silent stop closure, followed by a 65 ms diffuse high-frequency (t) burst. Vowel pitch began at 120 Hz, rose to 125 Hz at the 33% timepoint of the vowel, and fell to 100 Hz by vowel offset.

b. Spectral degradation: Noise-band vocoding. Spectral resolution was degraded using noise-band vocoding (NBV), which has become a common way to simulate a cochlear implant (see Shannon *et al*., 1995). This was accomplished using online signal processing within the ICAST stimulus delivery software (version 5.04.02; Fu, 2006). Stimuli were bandpass filtered into four or eight frequency bands using sixth-order Butterworth filters (24 dB/octave). This number of bands was chosen to best approximate the performance of CI listeners (Friesen *et al*., 2001). The temporal envelope in each band was extracted by half-wave rectification and low-pass filtering with a 200-Hz cutoff frequency, which is sufficient for good speech understanding (Shannon *et al*., 1995). The envelope of each band was used to modulate the corresponding bandpass-filtered noise. Specific band frequency cutoff values were determined assuming a 35 mm cochlear length (Greenwood, 1990) and are listed in Table III below. The lowest frequency of all analysis bands (141 Hz, 31 mm from the base, approximately) was selected to approximate those commonly used in modern CI speech processors. The highest frequency used (6000 Hz, approximately 9 mm from the base) was selected to be within the normal limits of hearing for all listeners, and to correspond with the upper limits of the frequency output of HLSYN. No spectral energy above this frequency was available to listeners in the unprocessed condition. Spectrograms of the word "hit" in the unprocessed (regularly synthesized), eight-channel NBV and four-channel NBV versions are illustrated in Fig. 2. The images show that specific formant frequency bands are no longer easily recoverable; the spectral fine structure is replaced by coarse/blurred sampling. Formant differences that remain unresolved within the same spectral channel are coded by the relative level of the noise band carrying that channel, as well as the time-varying amplitude (i.e., beating) owing to the interaction of multiple frequencies added together.

TABLE III. Specification of analysis and carrier filter bands for the noise-band vocoding scheme for experiment 1.

	Channel number							
4-channel	1		2		3		4	
8-channel	1	2	3	4	5	6	7	8
High-pass (Hz)	141	275	471	759	1181	1801	2710	4044
Low-pass (Hz)	275	471	759	1181	1801	2710	4044	6000

3. Procedure

All speech recognition testing was conducted in a double-walled sound-treated booth. Stimuli were presented at 65 dBA in the free field through a single loudspeaker. Each token was presented once, and listeners subsequently used a computer mouse to select one of two word choices ("heat" or "hit") to indicate their perception. Stimuli were presented in blocks organized by degree of spectral resolution (unprocessed, eight-channel or four-channel). Ordering of blocks was randomized, and presentation of tokens within each block was randomized. In this self-paced task, the 245 stimuli were each heard 5 times in each condition of spectral resolution.

4. Analysis

Categorical responses were fit using logistic regression, in accordance with recent trends in perceptual analysis (Morrison and Kondaurova, 2009). Listeners' binary responses (tense or lax) were fit using a generalized linear (logistic) mixed-effects model (GLMM). This was done in the R software interface (R Development Core Team, 2010), using the lme4 package (Bates and Maechler, 2010). A random effect of participant was used, and the fixed-effects were the stimulus factors described above. The binomial family call function was used because the possibility of a "tense" response could not logically exceed 100% or fall below 0%. This resulted in the use of the logit link function, and an assumption that variance increased with the mean according to the binomial distribution. Parameter levels were centered around

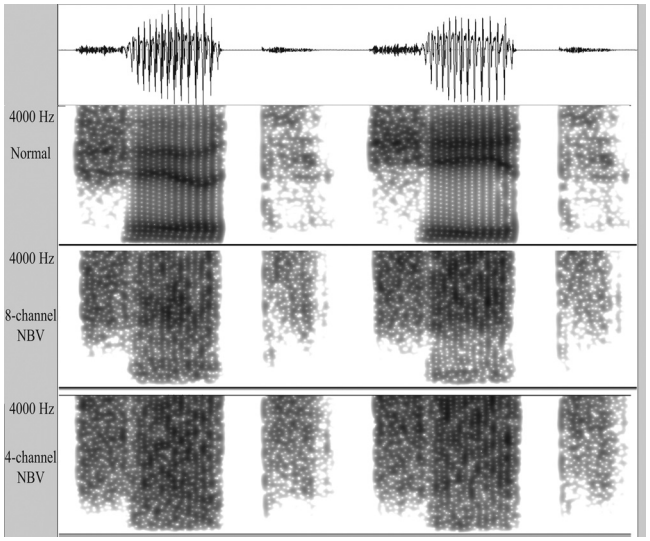

FIG. 2. Spectrograms illustrating synthesized words "hit" (left) and "heat" (right) in the normal/unprocessed condition (top), eight-channel noise-band vocoder (middle) and four-channel noise-band vocoder (bottom) conditions.

0, since the R GLM call function sets "0" as the default level while estimating other parameters. Thus, since the median duration was 115 ms, a stimulus with duration of 85 ms was coded as −30, and one with duration of 122 ms was coded as 8 ms. All factors and interactions were added via a forward-selection hill-climbing process. The model began with the intercept and factors (e.g., the inclusion of duration as a response predictor) competed one at a time; that which yielded the highest significance was kept. Subsequent factors (or factor interactions) were retained in the model if they significantly improved the model without unnecessarily overfitting. The ranking metric was the Akaike information criterion (AIC) (Akaike and Hirotugu, 1974), as it has become a popular method for evaluating mixed effects models (Vaida and Blanchard, 2005; Fang, 2011). This criterion measures relative goodness of fit of competing models by balancing accuracy and complexity of the model. This method contrasts with backward-elimination models that would be judged according to the Wald statistic. The goal of each model was similar to that used by Peng *et al*. (2009); it tested whether the coefficient of the resulting estimating equation for an acoustic cue was different from 0 and, crucially, whether the coefficient was different across conditions of spectral resolution.

Previous literature suggested that 4 or 8 is a suitable number of channels in a noise-band vocoder as a simulation of a cochlear implant. Both of these were tested in this experiment, not for a regression of cue usage against spectral degradation, but instead to find the best proxy value to simulate electric hearing for the problem at hand. Inspection of the psychometric functions of the NH listeners and CI listeners revealed that the eight-channel simulation was the best model of electric hearing, in accordance with previous assessment of better-performing CI listeners (Dorman and Loizou, 1998; Friesen *et al*., 2001). Furthermore, the amount of variability in the four-channel condition made it difficult to draw firm conclusions about how listeners perceived the signals. A small number of listeners demonstrated non-monotonic effects of spectral degradation on the use of the phonetic cues (i.e., they showed greater use of formant cues in four-channel compared to eight-channel conditions, but sometimes reported hearing neither the /i/ nor the /I/ vowel), suggesting that reducing the number channels below 8 did not necessarily change the resolution in a meaningful way *vis a vis* this experimental task. In the four channel case, the reduced spectral degradation was likely accompanied by increased availability of temporal envelope cues in voiced portions (because of increased numbers of harmonics falling into the broader filters), which may have been accessed/utilized differentially by different participants, depending on the precision of their temporal resolution. Some were able to capitalize on this, while some were not. Although (variations in) this ability is an interesting consideration in the use of noise-band vocoded signals, it is outside the scope of this investigation. Subsequent analysis of the data discarded the four-channel condition, yielding two sets of data models: (1) normal hearing listeners in both listening conditions (unprocessed and degraded using an eight-channel NBV) and (2) cochlear implant listeners hearing the unprocessed stimuli.

C. Results

Identification functions along the three parameter continua are shown in Figs. 3–5. The following models were found to describe the data optimally.

(1) Perception by NH listeners in different conditions:

$$\begin{aligned}\text{Tense} \sim\ & Formant + Duration + VISC + SR \\ & + Formant : SR + VISC : SR + Duration : SR \\ & + (1|Participant).\end{aligned}$$

(2) Perception by CI listeners:

$$\text{Tense} \sim Formant + VISC + Duration + (1|Participant).$$

For these two models, the interaction between two factors A and B is indicated by A:B. Independent factors are indicated by "+." "SR" refers to spectral resolution (normal or degraded/NBV), and (1|Participant) is a random effect of participant.

For both models, all three main cues were significant (all $p < 0.001$), and interactions between each cue and spectral resolution was also significant for the normal-hearing listeners (all $p < 0.001$). The parameter estimates all went in the predicted direction, and are listed in Table IV. Results suggest that when spectral resolution was degraded, normal-hearing listeners' responses were affected less by formants, less by VISC, and more by duration, compared to when spectral resolution was intact. The CI simulations were predictive of the CI listeners' results (smaller effect of formants and VISC, greater effect of duration), although direct statistical comparison between the NH and CI groups was not conducted (to be discussed further in the summary and discussion). Surprisingly, there were no significant interactions between cues. Typically, one would expect the effects of VISC and duration to be strongest in an ambiguous range of formant values; raw data suggested this, but the

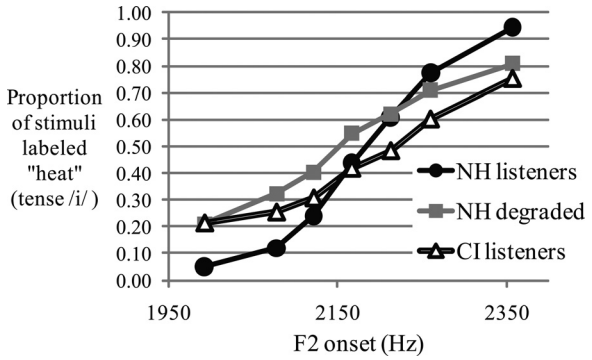

FIG. 3. Group mean response functions from 15 normal-hearing listeners and seven cochlear implant listeners along the continuum of vowel formant structure. Although these results are plotted by F2, the other formants were covarying (see Table II).

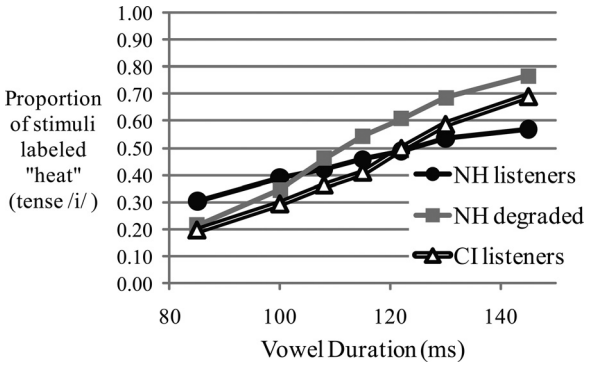

FIG. 5. Group mean response functions from 15 normal-hearing listeners and seven cochlear implant listeners along the continuum of vowel duration.

interaction did not reach significance in the model. Although error bars were omitted from the group psychometric functions (Figs. 3–5), variability in the use of acoustic cues is presented in Table IV.

Although direct statistical comparison is not valid for the groups in this study, the CI listener data is encouraging, as it falls along the same general trend as the NH listeners in the simulated conditions. The individual variability is apparently not limited to one group or the other; just as NH listeners have variations in listening strategies, so do the CI listeners, and both groups fall within similar ranges.

D. Conclusions

In this experiment, listeners were presented with words whose vowels varied along three acoustic dimensions. Normal-hearing listeners heard these words with clear unprocessed spectral resolution and also through eight- and four-channel noise-band vocoding schemes; the eight-channel condition was a better match to the CI listeners' performance. Cochlear implant listeners heard only the unprocessed words.

In conditions that are thought to simulate the use of a cochlear implant, normal-hearing listeners showed decreased use of spectral cues (formant structure and vowel-inherent spectral change), and showed increase use of vowel duration when identifying tense and lax vowels. Results from CI listeners suggested that they may be affected less by formant and VISC cues, and may be affected more by duration cues compared to NH listeners. Although this experiment tests merely one phonetic contrast, it appears to suggest that the NBV simulations hold some predictive value in determining the use of phonetic cues by CI uers.

In view of previous studies using synthesized speech, it is possible, despite the high quality of the speech synthesized by HLSYN, that the role of duration for NH listeners in the unprocessed condition was overestimated. Previous work suggests that duration is largely neglected by NH listeners for this vowel contrast when natural speech quality is preserved (Hillenbrand *et al.*, 2000). Thus, the differences in the use of duration by NH in diferent conditions (and possibly the differences in the use of duration by NH listeners and CI listeners) may be larger than what these data suggest. Another important consideration is the relatively advanced age of the CI user group, which will be discussed later in the summary and discussion section.

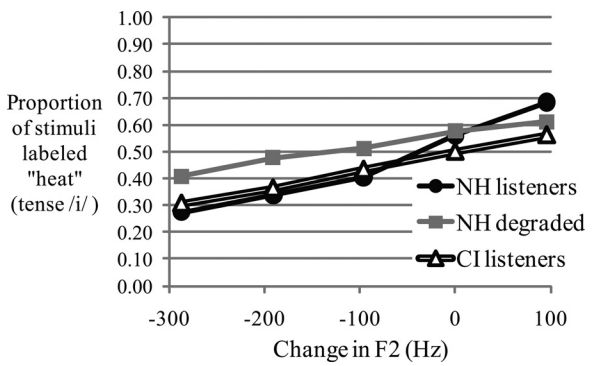

FIG. 4. Group mean response functions from 15 normal-hearing listeners and seven cochlear implant listeners along the continuum of vowel-inherent spectral change. Although these results are plotted by change in F2, the other formants were covarying (see Table II).

III. EXPERIMENT 2: THE WORD-FINAL S/Z CONTRAST

A. Review of acoustic cues

A second phonetic contrast was explored to supplement the first experiment. The second experiment explored the final consonant voicing contrast, which distinguishes /s/ and /z/ in word pairs such as bus-buzz, grace-graze, and loss-laws. The cues that contribute to this distinction include (but are not limited to) the offset frequency/transition of the first formant of the preceding vowel, the duration of the preceding vowel, the duration of the consonant, and the amount of voicing (low-frequency energy/amplitude modulation) within that consonant. Vowel duration has received the most consideration in the literature; vowels are longer before voiced sounds than before voiceless ones (House and Fairbanks, 1953; House, 1961). Chen (1970) and Raphael (1972) suggested that this duration difference is an essential

TABLE IV. Intercepts and parameter estimates for the optimal logistic models for experiment 1. The top portion reflects the group model; raw data could be reconstructed using these variables in an inverse logit equation. Rows in the lower portion reflect parameter estimates from individual listeners within each group.

	Formants			VISC			Duration		
	NH	NBV	CI	NH	NBV	CI	NH	NBV	CI
Group Est.	0.026	0.011	0.010	0.011	0.004	0.004	0.046	0.061	0.052
Int.	−1.368	−0.22	−0.773	−1.368	−0.22	−0.773	−1.368	−0.22	−0.773
Indiv. ests.									
01	0.034	0.014	0.008	0.016	0.007	0.003	0.074	0.091	0.047
02	0.024	0.020	0.015	0.013	0.007	0.007	0.099	0.096	0.040
03	0.028	0.019	0.003	0.012	0.006	0.001	0.041	0.042	0.036
04	0.027	0.015	0.009	0.010	0.005	0.006	0.033	0.045	0.067
05	0.034	0.016	0.015	0.019	0.006	0.006	0.039	0.031	0.056
06	0.025	0.014	0.011	0.012	0.004	0.004	0.038	0.038	0.045
07	0.027	0.013	0.007	0.015	0.004	0.004	0.058	0.049	0.073
08	0.037	0.009		0.012	0.003		0.057	0.067	
09	0.024	0.003		0.008	0.001		0.052	0.049	
10	0.016	0.006		0.010	0.002		0.045	0.040	
11	0.019	0.006		0.013	0.002		0.029	0.050	
12	0.023	0.010		0.006	0.002		0.023	0.069	
13	0.024	0.008		0.003	0.001		0.031	0.097	
14	0.025	0.012		0.007	0.002		0.035	0.085	
15	0.018	0.005		0.004	0.002		0.027	0.065	

perceptual cue for this distinction. However, just as for the aforementioned study by Ainsworth (1972), the limited spectral integrity of Raphael's stimuli (three steady-state synthesized formants) may have caused an overestimation of the effect of vowel duration. Furthermore, stimuli in Raphael's study that contained vowels of intermediate duration were contrasted reliably by the presence or absence of a vowel-offset F1 transition (F1T). When the F1T appeared at the end of the vowel, listeners tended to hear the following consonant as voiced.

Warren and Marslen-Wilson (1989) also suggested vowel duration to be an essential cue for consonant voicing. Their experiment used a gating paradigm, whereby a signal is truncated before completion; listeners attempted to identify the complete word. This method is problematic for this contrast, however, because it confounds the cues of vowel duration and F1T. When a signal is truncated before the F1T, the duration is shortened *and* the F1T is removed; the contributions of each cue are not recoverable in this paradigm. When truncation points fell before the region of the F1T, perception of voicing dramatically declined, but perhaps because of the absence of F1T rather than because of the shortened vocalic duration. Virtually no effect of vowel duration is observed when vowel portions are deleted from the middle (Revoile, 1982) or beginning (Wardrip-Fruin, 1982) of the segment; only when portions were deleted from the offset (area of F1T) does the perception change from voiced to voiceless (Hogan and Rozsypal, 1980; Wardrip-Fruin, 1982; Hillenbrand *et al.*, 1984; Warren and Marslen-Wilson, 1989). Hillenbrand *et al.* (1984) noted that compressing the duration of vowels before voiced stops does not significantly alter listeners' perceptions. Similar findings were reported by Wardrip-Fruin (1982), who showed that a falling F1T signaled voicing across the whole range of vowel durations tested, while syllables without this transition yielded no more than 60% voiced responses even at the longest vowel duration. Summers (1988) suggested that F1T differences are not limited to vowel offset; F1 is lower before voiced consonants at earlier-occurring times in the vowel as well. The importance of F1 is also underscored by the results of Hogan and Rozsypal (1980), who observed that excising the vowel offset had a smaller effect on high vowels; for these segments, the F1 is already low and therefore a less-useful cue since there is no room for transition.

A meta-analysis by Walsh and Parker (1984) suggests that vowel length exhibits an effect only for "artificial or abnormal circumstances." For example, Revoile (1982) showed that vowel duration was used as a voicing cue by individuals with hearing impairment, but not those with normal hearing. Wardrip-Fruin (1985) observed vowel duration effects for words presented in low-pass filtered noise, but not in quiet (Wardrip-Fruin, 1982). In experiments by Nittrouer (2004, 2005), vowel duration served as a voicing cue for synthetic speech, but this effect was strongly reduced and overpowered by the F1T cue when natural speech tokens were used. Thus, just as for previous experiments with vowels, the effect of duration on perceptual judgments appears to be driven at least partly by spectral fidelity of the signal.

Not surprisingly, there are acoustic cues that correspond to the voicing contrast within the fricative consonant itself. Voiceless fricatives are longer than voiced ones (Denes, 1955; Haggard, 1978), further increasing the vowel:consonant duration ratio (VCR) for voiced fricatives. VCR and duration of voicing within the fricative noise were shown by Hogan and Rozsypal (1980) to be reliable cues for perception of voicing

in sounds in an experiment where extension of vowel duration by itself did not force a change in voicing perception. Voicing during the consonant is not thought to be essential for perception of the voicing feature, since voiced fricatives are routinely devoiced in natural speech (Klatt, 1976; Haggard, 1978). Listeners reliably perceive voicing despite this apparent omission (Hogan and Rozsypal, 1980).

There are even more cues to the s/z contrast than are discussed here, but the aforementioned cues have been given the most consideration in the literature, and are thought to play a crucial role in perception of this contrast. The second experiment in this paper was designed to assess the use of these acoustic cues in listening conditions similar to those used in experiment 1. It was hypothesized that when spectral resolution was degraded, the F1 transition cue would be used less, and the durational cues (vowel and consonant duration, or a ratio of the vowel and consonant durations) would be used more. It was not clear whether the voicing duration cue would be used more or less, since it is implemented in both the spectral domain (via low-frequency energy) and the temporal domain (as temporal amplitude modulations of varying duration).

B. Methods

1. Participants

Participants for experiment 2 were comprised of 11 adult (ages 18–37; average 28.9 years) listeners with normal-hearing, defined as having pure-tone thresholds $\leq$20 dB HL from 250–8000 Hz in both ears (ANSI, 2004) and seven cochlear implant listeners whose demographics were the same as those for experiment 1 (see Table I). Four of the NH listeners and all seven CI listeners also participated in experiment 1. Normal-hearing participants 01 (the first author) and 02 were highly familiar with the stimuli, having been involved in pilot testing and the construction of the materials.

2. Stimuli

a. Natural speech manipulation. Stimuli for the second experiment were constructed using modified natural recordings of the words "loss" and "laws." The stimulus set consisted of 126 items that varied in four dimensions: presence/absence of vowel-offset falling F1 transition (two levels), vowel duration (seven levels), duration of fricative (three levels), and duration of voicing within that fricative (three levels). See Table V for a detailed breakdown of the levels for each parameter. A single /l/ segment of was chosen as the onset of all stimuli in the experiment, to neutralize it as a cue for final voicing (see Hawkins and Nguyen, 2004). The low-back vowel in "laws" was chosen because the F1 transition cue present in low vowels has been hypothesized to be compromised or absent in high vowels (Summers, 1988). The vowel was segmented from a recording of "laws," and thus contained a "voiced" F1 offset transition from roughly 635 Hz at vowel steady-state to 450 Hz at vowel offset, which is in the range of transitions observed in natural speech by Hillenbrand *et al.* (1984). A "voiceless" offset transition was created by deleting the final five pitch periods of the vowel in "laws," (maintaining a flat 635 Hz F1 offset) and expanding the duration to the original value using the pitch synchronous overlap-add (PSOLA) function in the PRAAT software (Boersma and Weenink, 2010). Rather than using recordings from "loss" and "laws" separately, this manipulation was preferable, in order to maintain consistent volume, phonation quality and other cues that may have inadvertently signaled the feature in question. In other words, it permitted the attribution of influence directly to the F1 offset level, since earlier portions of the vowel were consistent across different levels of this parameter. A uniform decaying amplitude envelope was applied to the final 60 ms of all vowels, as in Flege (1985); it resembled a contour intermediate to those observed in the natural productions, and was used to neutralize offset amplitude decay as a cue for voicing (see Hillenbrand *et al.*, 1984). Vowel durations were manipulated using PSOLA to create a seven-step continuum between 175 and 325 ms, based on values from natural production reported by House (1961) and Stevens (1992), and used by Flege (1985) in perceptual experiments. All vowels were manipulated using PSOLA to contain the same falling pitch contour (which started at 96 Hz and ended at 83 Hz), to neutralize pitch as a cue for final fricative voicing (see Derr and Massaro, 1980; Gruenenfelder and Pisoni, 1980). 250 ms of frication noise were extracted from a natural /s/ segment. An amplitude contour was applied to the fricative offset to create a 50 ms rise time and 30 ms decay-time. Two other durations (100 and 175 ms) of frication noise were created by applying the offset envelope at correspondingly earlier times. The resulting values ranging from 100–250 ms frication duration resembled those used by Soli (1982) and Flege and Hillenbrand (1985). Voicing was added to these fricatives by replacing 30 or 50 ms onset portions with equivalently long onset portions of a naturally produced voiced /z/ segment. These three levels of voicing thus varied in the range of 0–50 ms, which resembles the range used in perceptual experiments by Stevens (1992). These fricatives were appended to all 14 of the aforementioned vowel segments with onset /l/. For fricatives with onset voicing, the first pitch period of voiced fricative noise was blended with the last pitch period of the vowel (each at 50% volume) to produce a smooth transition between segments. Although the stimuli were not designed explicitly to vary the vowel-consonant duration ratio, this ratio naturally changed as a function of each independently varied duration factor.

TABLE V. Acoustic parameter levels defining the four factors in experiment 2.

First formant offset (Hz)	450	615					
Vowel duration (ms)	175	200	225	250	275	300	325
Consonant duration (ms)	100	175	250				
Voicing duration (ms)	0	30	50				

b. Spectral degradation: Noise-band vocoding. Noise-band vocoding was accomplished using the same procedure described for experiment 1 (described earlier in Sec. II B 2 b), except that the upper-limit of the analysis and filter

TABLE VI. Specification of analysis and carrier filter bands for the noise-band vocoding scheme for experiment 2.

Channel:	1	2	3	4	5	6	7	8
High-pass (Hz)	141	283	495	812	1285	1994	3052	4634
Low-pass (Hz)	283	495	812	1285	1994	3052	4634	7000

bands was changed from 6 to 7 kHz, to ensure that a substantial amount of frication noise was represented within the spectrally degraded output. Analysis/carrier band cutoff frequencies for experiment 2 are displayed in Table VI.

3. Procedure

The procedure for experiment 2 was the same as that for experiment 1 (described earlier in Sec. II B 3), with minor modifications to account for the different stimulus set. Visual word choices were "loss" and "laws," and the 126-item stimulus set was presented in alternating blocks of unprocessed and eight-channel noise-band vocoder conditions. In view of the results of the first experiment, no four-channel NBV condition was used for experiment II. The 126 stimulus items were heard five times in both conditions of spectral resolution. Cochlear implant listeners only heard the natural (unprocessed) items five times each.

4. Analysis

Listeners' binary responses (voiced or voiceless) were fit using a generalized linear (logistic) mixed-effects model (GLMM), using the same procedure as in experiment 1 (see Sec. II C 4). This experiment produced two sets of data: (1) NH listeners in both conditions of spectral resolution and (2) CI listeners listening to the modified natural sounds with intact spectral resolution.

C. Results

Identification functions along the four parameter continua are shown in Figs. 6–9. Although vowel:consonant duration ratio was not explicitly planned in stimulus construction, it was easily calculated and included as a separate factor in the model (since this factor was not fully crossed with the others, listeners responses were not plotted

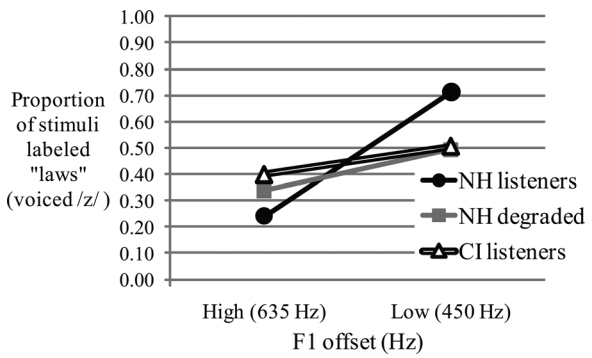

FIG. 6. Group mean response functions from 11 normal-hearing listeners and seven cochlear implant listeners for both levels of the F1 transition offset.

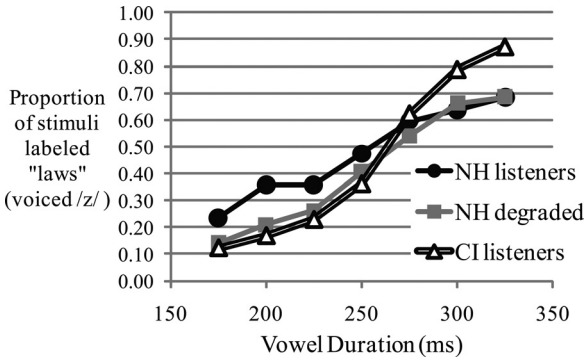

FIG. 7. Group mean response functions from 11 normal-hearing listeners and seven cochlear implant listeners along the continuum of vowel duration.

for this cue). The following models were found to describe the data optimally.

(1) Perception by NH listeners in different conditions:

$$\begin{aligned}\text{Voiced} \sim\ & VCRatio + F1T + VDuration + Voicing \\ & + SR + F1T:SR + VCRatio:VDuration \\ & + Voicing:SR + CDuration + VCRatio:Voicing \\ & + VDuration:SR + (1|Participant).\end{aligned}$$

(2) Perception by CI listeners:

$$\begin{aligned}\text{Voiced} \sim\ & VDuration + CDuration + Voicing \\ & + F1T + VDuration:F1T + F1T:Voicing \\ & + VDuration:Cduration + CDuration:F1T \\ & + VDuration:CDuration:Voicing \\ & + (1|Participant).\end{aligned}$$

For these two models, the interaction between two factors A and B is indicated by A:B. Independent factors are indicated by "+." "VCRatio" refers to the ratio of vowel duration to consonant duration. "SR" refers to spectral resolution (normal or degraded/NBV), and (1|Participant) is a random effect of participant. Predictors are listed in the order in which they were added to the model (this was determined by the AIC metric). Parameter estimates for the groups and for each participant are listed in Tables VII and VIII.

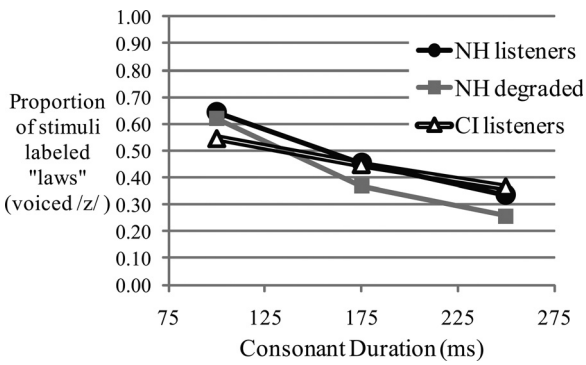

FIG. 8. Group mean response functions from 11 normal-hearing listeners and seven cochlear implant listeners along the continuum of consonant duration.

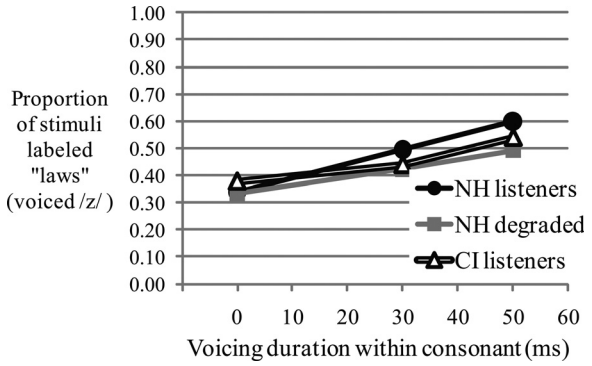

FIG. 9. Group mean response functions from 11 normal-hearing listeners and seven cochlear implant listeners along the continuum of consonant voicing duration.

For the NH listener model, all five main factors (including VCRatio) were significant (all $p < 0.001$), although consonant duration was by far the least powerful main factor in the model, according to the AIC metric. Spectral resolution significantly interacted with F1 transition ($p < 0.001$; F1 transition was a weaker cue in the degraded condition), and with voicing duration ($p < 0.001$; voicing duration was a weaker cue in the degraded condition), but not with VCRatio nor consonant duration. The interaction between spectral resolution and vowel duration did not reach statistical significance ($p = 0.11$; vowel duration was a slightly stronger cue in the degraded condition), but its inclusion improved the model according to the AIC metric. The effect of VCRatio changed slightly in the expected direction in the degraded condition, but this interaction did not reach significance ($p = 0.60$), and was not included in the model. Although the raw data suggested that the interaction between consonant duration and spectral resolution would go in the expected direction, the model did not confirm this; it did not reach significance ($p = 0.42$), and was not included in the model. There were significant interactions between VCRatio and vowel duration ($p < 0.001$), and between VCRatio and voicing duration ($p = 0.005$), indicating a complex interdependence of multiple cues for this contrast.

CI listeners were able to use the F1 transition cue ($p < 0.001$), but apparently not to the same extent as NH listeners (the parameter estimate was lower for the CI group). CI listeners showed use of the vowel duration cue ($p < 0.001$) that appears to be greater than that by NH listeners (the parameter estimate was higher for the CI group). The effect of consonant duration was significant ($p < 0.001$) and appears to be similar to that observed in the NH group. F1 transition significantly interacted with vowel duration ($p < 0.001$), with voicing duration ($p < 0.001$) and with consonant duration ($p = 0.017$). Although the effect of VCRatio did not reach significance, vowel duration significantly interacted with consonant duration ($p = 0.019$). There was a three-way interaction between vowel duration, consonant duration and voicing duration that did not reach significance ($p = 0.15$), but its inclusions produced a significant improvement in the model, according to the AIC metric. Just as for NH listeners, the listeners with hearing impairment showed complex inter-dependence of cues for this contrast. A large amount of variability was seen in the CI listener group for all cues, especially for voicing duration and VCRatio, where several individuals' parameter estimates actually went in the reverse direction.

D. Conclusions

In Experiment 2, listeners were presented with words that varied along four acoustic dimensions. In conditions that are thought to roughly simulate the use of a cochlear implant, normal-hearing listeners maintained use of all four cues, but showed decreased use of the F1 transition and consonant voicing cues. Reliance upon vowel duration did not change significantly when the resolution was degraded. The

TABLE VII. Intercepts and parameter estimates for the optimal logistic models for experiment 2. The top portion reflects the group model; raw data could be reconstructed using these variables in an inverse logit equation. Rows in the lower portion reflect parameter estimates from individual listeners within each group. NH and NBV refer to normal-hearing listeners in the unprocessed and degraded (eight-channel noise-band vocoded) conditions, respectively. CI refers to CI listeners. NH and NBV intercepts for VCRatio and Consonant duration were derived from a separate model where they could be individually computed with an interaction with spectral degradation.

	V:C ratio			Vowel duration			F1 transition		
Group	NH	NBV	CI	NH	NBV	CI	NH	NBV	CI
Est.	**0.783**	**0.910**	**−0.036**	**0.015**	**0.015**	**0.032**	**0.017**	**0.005**	**0.003**
Int.	0.190	−0.137	−0.307	0.213	−0.163	0.295	0.213	−0.163	0.295
Indiv. ests.									
01	−1.36	1.30	−0.48	0.053	0.066	0.030	0.030	0.005	0.011
02	1.47	1.45	−0.18	0.018	0.022	0.075	0.028	0.010	0.030
03	2.32	1.08	−0.33	0.025	0.025	0.027	0.022	0.007	−0.001
04	0.35	1.82	−0.10	0.057	0.058	0.104	0.020	0.003	0.011
05	0.95	1.09	1.15	0.015	0.018	0.079	0.019	0.006	0.016
06	2.71	4.11	−0.30	0.046	0.034	0.101	0.027	0.008	0.026
07	1.30	2.30	0.17	0.032	0.030	0.014	0.027	0.015	0.007
08	1.31	0.10		0.024	0.023		0.014	0.005	
09	−0.14	1.28		0.040	0.061		0.018	0.005	
10	0.46	−0.22		0.021	0.017		0.007	0.001	
11	0.56	1.19		0.018	0.025		0.012	0.004	

TABLE VIII. Intercepts and parameter estimates for the optimal logistic models for experiment 2 (continued). The top portion reflects the group model; raw data could be reconstructed using these variables in an inverse logit equation. Rows in the lower portion reflect parameter estimates from individual listeners within each group. NH and NBV refer to normal-hearing listeners in the unprocessed and degraded (eight-channel noise-band vocoded) conditions, respectively. CI refers to CI listeners. NH and NBV intercepts for VCRatio and Consonant duration were derived from a separate model where they could be individually computed with an interaction with spectral degradation.

	Voicing duration			Consonant duration		
Group	NH	NBV	CI	NH	NBV	CI
Est.	**0.037**	**0.018**	**0.020**	**−0.008**	**−0.006**	**−0.008**
Int.	0.213	−0.163	0.295	0.190	−0.137	−0.295
Indiv. ests.						
01	0.128	0.063	−0.083	−0.055	−0.022	−0.028
02	0.055	0.037	−0.144	−0.007	−0.008	0.019
03	0.081	0.045	−0.049	0.005	−0.003	−0.025
04	0.045	0.014	−0.011	−0.008	0.008	0.044
05	0.063	0.044	0.102	−0.007	−0.002	−0.061
06	0.054	0.047	0.045	−0.014	−0.001	0.027
07	0.072	0.030	0.003	−0.020	−0.004	−0.038
08	0.059	0.045		0.000	−0.017	
09	0.032	0.015		−0.023	−0.004	
10	0.010	0.007		−0.003	−0.010	
11	0.027	0.006		−0.005	0.009	

effect of vowel duration for NH listeners was larger than what was expected based on previous literature (perhaps because early occurring spectral information in the vowel was neutralized).

Statistical comparisons were not made between NH and CI listeners, but a rough qualitative assessment of the data suggests that CI listeners made less use of the F1 transition and consonant voicing cues, and made more use of the vowel duration cue. These results are in agreement with experiment 1, namely, that listeners alter their use of phonetic cues when spectral resolution is degraded, and that CI listeners may use phonetic cues differently than NH listeners. It should be noted, however, that the use of the F1 transition cue is probably dependent on vowel environment. The F1 cue would be less useful for consonants following the /i/ or /u/ vowels; the F1 value in these segments is already low, so any F1 movement would be subtle, if at all present (Hogan and Rozsypal, 1980; Hillenbrand *et al.*, 1984). It is thus possible that durational cues might already be more dominant in these contexts, and therefore not demand significantly different perceptual strategies by CI listeners or NH listeners in degraded conditions.

IV. SUMMARY AND DISCUSSION

In these experiments, listeners categorized speech tokens that varied in multiple dimensions. The influence of each of those dimensions was modulated by the degree of spectral resolution with which the signal was delivered, or by whether the listener used a cochlear implant. We offer the following general conclusions.

(1) As spectral resolution is degraded, spectral cues (such as formant structure, vowel-inherent spectral change, and a vowel-offset formant transition) played a smaller role, and some temporal cues played a larger role in normal-hearing listeners' phonetic identifications.
(2) Cochlear implant listeners appeared to show less use of spectral cues, and greater use of temporal cues for phonetic identification, compared to normal-hearing listeners. This effect was more pronounced for the final consonant voicing contrast than for the lax/tense vowel contrast.
(3) There was a high amount of variability in the individual data; some normal-hearing listeners showed different use of cues in degraded conditions while others did not. Similarly, some cochlear implant listeners showed patterns similar to the normal-hearing group, while others showed distinctively different patterns. It is not yet known whether either of these patterns can be associated with more general success in speech perception.
(4) Under conditions of normal redundancy of acoustic cues, a normal-hearing listener and a CI user can thus potentially achieve the same performance on a speech recognition task (word recognition, phoneme recognition, confusion matrix/information transfer analysis), but through the use of different acoustic cues.

More generally, this work accords with previous literature that indicates greater use of vowel duration by impaired listeners (Revoile, 1982), and adds a new layer to work comparing the use of cues in natural and synthesized signals (Assmann and Katz, 2005; Nittrouer, 2004, 2005). The variability in the data is problematic for drawing general conclusions, but it might potentially be a fruitful avenue of exploration. A small number of CI listeners in this study appeared to rely heavily on the same cues used by NH listeners, while the others were relatively more influenced by other cues. While auditory prostheses and amplification devices are designed generally to transmit the acoustic cues used by normal-hearing listeners, not all listeners use the cues in the same way. Thus, effort might be wasted in

delivering information to listeners who could subsequently discard it. It is not known whether successful CI listeners are those that are able to extract and decode spectral cues despite device limitations, or if they are diverting attention/resources away from those cues in favor of those that remain intact in the temporal domain.

It should be noted that there are various limitations in the generalization of the CI simulations to real CI listeners. Among these are (1) noise-band vocoding is only a crude approximation of the experience of electric hearing, (2) the two-alternative forced-choice task is an atypical listening scenario, lacking top-down influences such as contextual clues and visual information, which could resolve perceptual ambiguity, (3) the NH participants listening to the simulations were generally much younger than the CI listeners, and (4) the simulated conditions are essentially simulating an initial activation rather than an everyday experience; most NH listeners in this experiment had no prior experience with noise-band vocoding, whereas the CI listeners had all been wearing their devices for multiple years. It is thus possible that the degraded conditions simulated the novelty of a cochlear implant but not the eventual everyday performance.

The issue of age differences between the NH and CI groups introduces some complications in the analysis of the current data. It has been shown numerous times that older listeners show deficiencies in auditory temporal processing in basic psychophysical tasks (Gordon-Salant and Fitzgibbons, 1993, 1999), and tasks involving perception of temporal phonetic cues (Gordon-Salant *et al.*, 2006). They therefore might be less able to capitalize on the duration cue available in this study and in natural speech. Furthermore, older listeners have been shown to experience more difficulty with spectrally degraded speech in general (Schvartz *et al.*, 2008). If one presumes that psychophysical capabilities/deficiencies influence behavior in this identification task, the trend of the CI listeners in this study is opposite to that which might be predicted by their age; they showed increased use of durational cues compared to the young NH listeners. However, it is evident that capability is not entirely predictive of cue usage; the CI listeners in this study did not use the fricative voicing cue even though this population has been shown to exhibit very fine sensitiviy to temporal modulations. Perhaps younger CI listeners, with hypothetical advantages in temporal processing, would show more reliable use of the vowel duration and/or voicing cues than the older listeners in this study. The one older NH listener (n04 in Experiment 1) does not provide sufficient basis for age-matched group comparison, but it is reassuring that this listener's data were not markedly different from the NH group mean (Table IV). Young postlingually deafened CI listeners are generally more scarce in the population though, and were not available at the time of this experiment; the question of the role of aging in the use of phonetic cues in electric hearing invites future work.

It could be argued that the difference in cue-weighting or cue usage makes no difference in the "bottom line" of word recognition. After all, if a listener correctly perceives a word, he/she might not care about the method by which it was done. However, it is not clear whether all perceptual cue weighting strategies are equally reliable, efficient or taxing for the listener. The data in this paper cannot speak to any potential differences in processing speed, efficiency or listening effort, but it should be noted that if normal-hearing listeners tend to rely on a particular cue for a contrast, there is probably a reason for that tendency (it may be explained by acoustic reliability; see Holt and Lotto, 2006; Toscano and McMurray, 2010). Future work might address this issue by exploring neurological responses to multidimensional speech stimuli (see Pakarinen *et al.*, 2007, 2009), or by more sophisticated measurements.

The concept of trading relations between spectral and temporal information is not a new one. Although the hypotheses supported in this paper are not particularly novel or unexpected, they have been largely neglected in previous literature on listeners with hearing impairment. The reader is encouraged to finely distinguish phonetic feature recovery from phonetic cue use; measuring the recovery of "lax/tense," "voicing" or other features by a listener with hearing impairment does not imply that it was because of the same perceptual cue used by normal-hearing listeners. In view of the multiple acoustic cues available for any particular phonetic segment, the contrasts explored in this study may represent just a fraction of those for which CI listeners could employ alternative perceptual strategies. Thus, caution should be used when comparing results of NH listeners and CI listeners in the same tasks; similar performance may not verify similar perception or perceptual processes.

ACKNOWLEDGMENTS

The authors would like to thank the participants for their time and willingness to contribute to this study, as well as Rochelle Newman, Shu-Chen Peng, Nelson Lu, Ewan Dunbar, and Shannon Barrios for their helpful comments and expertise. We also express appreciation to Mitchell Sommers for his patience and helpful suggestions on this manuscript and to two anonymous reviewers for helpful comments on an earlier version of this manuscript. We are grateful to Qian-Jie Fu for the software used for the experiment. This research was supported by NIH Grant No. R01 DC004786 to M.C. M.B.W. was supported by NIH Grant No. T32 DC000046-17 (PI: Arthur N. Popper).

Ainsworth, W. A. (**1972**). "Duration as a cue in the recognition of synthetic vowels," J. Acoust. Soc. Am. **51**, 648–651.

Akaike, H. (**1974**). "A new look at the statistical model identification," IEEE Trans. Autom. Control **19**(6), 716–723.

ANSI (**2004**). ANSI S3.6-2004, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

Assmann, P. F., and Katz, W. F. (**2005**). "Synthesis fidelity and time-varying spectral change in vowels," J. Acoust. Soc. Am. **117**, 886–895.

Bates, D., and Maechler, M. (**2010**). "lme4: Linear mixed-effects models using S4 classes," R package version 0.999375-37, http://CRAN.R-project.org/package=lme4 (Last viewed January 9, 2011).

Boersma, Paul, and Weenink, David (**2010**). "Praat: doing phonetics by computer (Version 5.1.23) [Computer program]," http://www.praat.org/ (Last viewed January 1, 2010).

Bohn, O.-S. (**1995**). "Cross-language speech perception in adults; First language transfer doesn't tell it all," in *Speech Perception and Linguistic Experience; Issues in Cross-Language Research*, edited by W. Strange (New York Press, Baltimore, MD), pp. 279–304.

Bohn, O.-S., and J. E. Flege (**1990**). "Interlingual identification and the role of foreign language experience in L2 vowel perception," Appl. Psycholing. **11**, 303–328.
Chang, Y-P., and Fu, Q-J. (**2006**). "Effects of talker variability on vowel recognition in cochlear implants," J. Speech Lang. Hear. Res. **49**, 1331–1341.
Chatterjee, M., and Shannon, R. (**1998**). "Forward masked excitation patterns in multielectrode cochlear implants," J. Acoust. Soc. Am. **103**, 2565–2572.
Chen, M. (**1970**). "Vowel length variation as a function of the voicing of the consonant environment," Phonetica **22**, 129–159.
Denes, P. (**1955**). "Effect of duration on the perception of voicing," J. Acoust. Soc. Am. **27**, 761–764.
Derr, M. A., and Masaaro, D. M. (**1980**). "The contribution of vowel duration, F0 contour, and fricative duration as cues to the /juz/-jus/ distinction." Percept. Psychophys. **27**, 51–59.
Dorman, M. F., and Loizou, P. C. (**1997**). "Mechanisms of vowel recognition for Ineraid patients fit with continuous interleaved sampling processors," J. Acoust. Soc. Am. **102**, 581–587.
Dorman, M., and Loizou, P. (**1998**). "The identification of consonants and vowel by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels," Ear Hear. **19**, 162–166.
Dorman, M., Dankowski, K., McCandless, G., Parkin, J., and Smith, L. (**1991**). "Vowel and consonant recognition with the aid of a multichannel cochlear implant," Q. J. Exp. Psych. Sect. **43A**, 585–601.
Fang, Y. (**2011**). "Asymptotic equivalence between cross-validations and Akaike Information Criteria in mixed-effects models," J. Data Sci. **9**, 15–21.
Fishman, K., Shannon, R., and Slattery, W. (**1997**). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," J. Speech Lang. Hear. Res. **40**, 1201–1215.
Flege, J., and Hillenbrand, J. (**1985**). "Differential use of temporal cues to the /s/-/z/ contrast by non-native speakers of English," J. Acoust. Soc. Am. **79**, 508–517.
Francis, A. L., Baldwin, K., and Nusbaum, H. C. (**2000**). "Effects of training on attention to acoustic cues," Percept. Psychophys. **62**, 1668–1680.
Francis A., Kaganovich, N., and Driscoll-Huber, C. (**2008**). "Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English," J. Acoust. Soc. Am. **124**, 1234–1251.
Friesen, L., Shannon, R., Başkent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.
Fu, Q-J. (**2006**). "Internet-based computer-assisted speech training (iCAST)" [Computer program], TigerSpeech Technology, version 5.04.02, http://www.tigerspeech.com/tst_icast.html (Last viewed Febuary 8, 2010).
Fu, Q.-J., and Shannon, R. V. (**1999**). "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," J. Acoust. Soc. Am. **105**, 1889–1900.
Gordon-Salant, S., and Fitzgibbons, P. (**1993**). "Temporal factors and speech recognition performance in young and elderly listeners," J. Speech Hear Res. **36**, 1276–1285.
Gordon-Salant, S., and Fitzgibbons, P. (**1999**). "Profile of auditory temporal processing in older listeners," J. Speech Lang. Hear. Res. **42**, 300–311.
Gordon-Salant, S., Yeni-Komshian, G., Fitzgibbons, P., and Barrett, J. (**2006**). "Age-related differences in identification and discrimination of temporal cues in speech segments," J. Acoust. Soc. Am. **119**, 2455–2466.
Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Gruenenfelder, T., and Pisoni, D. (**1980**). "Fundamental frequency as a cue to postvocalic consonantal voicing: Some data from perception and production," Percept. Psychophys. **28**, 514–520.
Haggard, M. (**1978**). "The devoicing of voiced fricatives," J. Phonetics **6**, 95–102.
Hanson, H. M., Stevens, K. N., and Beaudoin, R. E. (**1997**). "New parameters and mapping relations for the HLsyn speech synthesizer," J. Acoust. Soc. Am. **102**, 3163.
Hanson, H. M., and Stevens, K. N. (**2002**). "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn," J. Acoust. Soc. Am **112**, 1158–1182.
Hawkins, S., and Nguyen, N. (**2004**). "Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English," J. Phonetics **32**, 199–231.
Henry, B. A., Turner, C. W., and Behrens, A. (**2005**). "Spectral peak resolution and speech recognition in quiet: Normal hearing, hearing impaired, and cochlear implant listeners," J. Acoust. Soc. Am. **118**, 1111–1121.
Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (**2000**). "Some effects of duration on vowel recognition," J. Acoust. Soc. Am. **108**, 3013–3022.
Hillenbrand, J. M., and Gayvert, R. T. (**1993**). "Identification of steady-state vowels synthesized from the Peterson–Barney measurements," J. Acoust. Soc. Am. **94**, 668–674.
Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**, 3509–3523.
Hillenbrand, J., Ingrisano, D., Smith, B., and Flege, J. (**1984**). "Perception of the voiced-voiceless contrast in syllable-final stops," J. Acoust. Soc. Am. **76**, 18–26.
Hogan, J., and Rozsypal, A. (**1980**). "Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant," J. Acoust. Soc. Am. **67**, 1764–1771.
Holt, L. L., and Idemaru, K. (**2011**). "Generalization of dimension-based statistical learning," *Proceedings of the 17th International Congress on Phonetics Science*, Hong Kong, China, pp. 882–885.
Holt, L., and Lotto, A., (**2006**). "Cue weighting in auditory categorization: Implications for first and second language acquisition," J. Acoust. Soc. Am. **119**, 3059–3071.
House, A. (**1961**). "On vowel duration in English," J. Acoust. Soc. Am. **33**, 1174–1178.
House, A., and Fairbanks, G. (**1953**). "The influence of consonant environment upon the secondary acoustical characteristics of vowels," J. Acoust. Soc. Am. **25**, 105–113.
Iverson, P., Smith, C., and Evans, B. (**2006**). "Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration," J. Acoust. Soc. Am. **120**, 3998–4006.
Jenkins, J., Strange, W., and Edman, T. (**1983**). "Identification of vowels in 'vowelless' syllables," Percept. Psychophys. **34**, 441–450.
Kewley-Port, D., and Zheng, Y. (**1998**). "Modeling formant frequency discrimination for isolated vowels." J. Acoust. Soc. Am. **103**, 1654–1666.
Kirk, K. I., Tye-Murray, N., and Hurtig, R. R. (**1992**). "The use of static and dynamic vowel cues by multichannel cochlear implant users," J. Acoust. Soc. Am. **91**, 3487–3498.
Klatt, D. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.
Lisker, L. (**1978**). "Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction," Haskins Lab. Status Rep. Speech Res. **SR-54**, 127–132.
Loizou, P., and Poroy, O. (**2001**). "Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners," J. Acoust. Soc. Am. **110**, 1619–1627.
Miller, G. A., and Nicely, P. A. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Morrison, G., and Kondaurova, M. (**2009**). "Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis," J. Acoust. Soc. Am. **126**, 2159–2162.
Morrison, G., and Nearey, T. (**2007**). "Testing theories of vowel inherent spectral change." J. Acoust. Soc. Am. **122**, EL15–EL22.
Morrison, G. (**2005**). "An appropriate metric for cue weighting in L2 speech perception: Response to Escudero & Boersma (2004)," Stud. Sec. Lang. Acq. **27**, 597–606.
Nearey, T. M., and Assmann, P. (**1986**). "Modeling the role of vowel inherent spectral change in vowel identification," J. Acoust. Soc. Am. **80**, 1297–1308.
Nittrouer, S. (**2004**). "The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults," J. Acoust. Soc. Am. **115**, 1777–1790.
Nittrouer, S. (**2005**) "Age-related differences in weighting and masking of two cues to word-final stop voicing in noise," J. Acoust. Soc. Am. **118**, 1072–1088.
Pakarinen, S., Takegata, R., Rinne, T., Huotilainen, M., and Näätänen, R. (**2007**). "Measurement of extensive auditory discrimination profiles using the mismatch negativity (MMN) of the auditory event-related potential (ERP)," Clin. Neurophysc. **118**, 177–185.
Pakarinen, S., Lovio, R., Huotilainen, M., Alku, P., Näätänen, R., and Kujala, T. (**2009**). "Fast multi-feature paradigm for recording several mismatch negativities (MMNs) to phonetic and acoustic changes in speech sounds," Biol. Psych. **82**, 219–226.

Parker, E. M., and Diehl, R. L. (**1984**). "Identifying vowels in CVC syllables: Effects of inserting silence and noise," Percept. Psychophys. **36**, 369–380.
Peng, S-C., Lu, N., and Chatterjee, M. (**2009**). "Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners," Audiol. Neurotol. **14**, 327–337.
R Development Core Team (**2010**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (Last viewed January 9, 2011).
Raphael, L. (**1972**). "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acoust. Soc. Am. **51**, 1296–1303.
Repp, B. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," Psychol. Bul. **92**, 81–110.
Revoile, S., Pickett, J., and Holden, L. (**1982**). "Acoustic cues to final stop voicing for impaired- and normal-hearing listeners," J. Acoust. Soc. Am. **72**, 1145–1154.
Schvartz, K., Chatterjee, M., and Gordon-Salant, S. (**2008**). "Recognition of spectrally degraded phonemes by younger, middle-aged, and older normal-hearing listeners," J. Acoust. Soc. Am. **124**, 3972–3988.
Shannon, R. (**1989**). "Detection of gaps in sinusoids and pulse trains by patients with cochlear implants," J. Acoust. Soc. Am. **85**, 2587–2592.
Shannon, R. V. (**1992**). "Temporal modulation transfer functions in patients with cochlear implants," J. Acoust. Soc. Am. **91**, 2156–2164.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Soli, S. (**1982**). "Structure and duration of vowels together specify fricative voicing," J. Acoust. Soc. Am. **72**, 366–378.
Stevens, K., Blumstein, S., Glicksman, L., Burton, M., and Kurowski, K. (**1992**). "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters," J. Acoust. Soc. Am. **91**, 2179–3000.
Summers, W. V. (**1988**). "F1 structure provides information for final consonant voicing," J. Acoust. Soc. Am. **84**, 485–492.
Toscano, J., and McMurray, B. (**2010**). "Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics," Cogn. Sci. **34**, 434–464.
Vaida, F., and Blanchard, S. (**2005**). "Conditional Akaike information for mixed-effects models," Biometrika **92**, 351–370.
Walsh, T., and Parker, F. (**1984**). "A review of the vocalic cues to [+- voice] in post-vocalic stops in English," J. Phonetics **12**, 207–218.
Wardrip-Fruin, C. (**1982**). "On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants," J. Acoust. Soc. Am. **71**, 187–195.
Wardrip-Fruin, C. (**1985**). "The effect of signal degradation on the status of cues to voicing in utterance-final stop consonants," J. Acoust. Soc. Am. **77**, 1907–1912.
Warren, P., and Marslen-Wilson, W. (**1989**). "Cues to lexical choice: Discriminating place and voice," Percept. Psychophys. **43**, 21–30.
Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1993**). "F0 gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **93**, 2152–2159.
Xu., L., and Pfingst, B. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception." J. Acoust. Soc. Am. **114**, 3024–3027.
Xu, L., Thompson, K., and Pfingst, B. (**2005**). "Relative contributions of spectral and temporal cues for phoneme \recognition," J. Acoust. Soc. Am. **117**, 3255–3267.
Zahorian, S. A., and Jagharghi, A. J. (**1993**). "Spectral-shape features versus formants as acoustic correlates for vowels," J. Acoust. Soc. Am. **94**, 1966–1982.
Zeng, F-G., Rebscher, S., Harrison, W., Sun, X., and Feng, H. (**2008**). "Cochlear implants: System design, Integration and Evaluation," IEEE Rev. Biomed. Eng. **1**, 115–142.
Zwicker, E., and Terhardt, E. (**1980**). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am. **68**, 1523–1525.