Phonetic cues are weighted differently when spectral resolution is degraded

Matthew B. Winn, Monita Chatterjee, William J. Idsardi

(Date of JASA submission)

RUNNING TITLE: Phonetic cues / Spectral degradation

Department of Hearing and Speech Sciences
University of Maryland, College Park
0100 Lefrak Hall
College Park, Maryland 20742

4) Department of Linguistics
University of Maryland, College Park
1401 Marie Mount Hall
College Park, Maryland 20742

Corresponding Author and Reprints:
Matthew Winn, University of Maryland Department of Hearing and Speech Sciences
(301) 405-4214 mwinn@hesp.umd.edu
ABSTRACT

An experiment was conducted to test the hypothesis that listeners’ perceptual strategies would change when the spectrum of an input speech signal was degraded. The contrast between tense and lax English vowels (as in the words “heat” and “hit”) was selected because of the known primacy of spectral cues and redundancy of temporal cues used in normal listening situations.

Synthesized speech signals varying in two spectral dimensions (initial formant structure and vowel-inherent spectral change) and in one temporal dimension (vowel duration) were presented to listeners with intact spectra or through an 8- or 4-channel noise-excited vocoder, used to simulate a cochlear implant. Two-alternative forced choice identification results showed that as the spectrum was degraded, listeners relied less upon the spectral cues, and compensated with greater sensitivity to the temporal cue. Preliminary results with cochlear implant users suggest that results from the simulations are generalizable to a target clinical population; those who experience spectral degradation may employ perceptual strategies that differ from those with normal hearing.

PACS numbers: 43.71.Es, 43.71.Ky, 43.71.Gv
I. INTRODUCTION

In view of the remarkable success of the cochlear implant (CI) as a treatment for hearing loss (Hiraumi et al., 2007; Zeng et al., 2008), and in the context of an increasing research focus on cochlear implants, literature on phonetic cue perception must be expanded to incorporate the abilities of this continually growing clinical population. Phonetic perception research has been restricted primarily to listeners with normal hearing; it is unclear whether this knowledge can be generalized to clinical populations for whom the incoming sound is compromised. Though limitations exist, the human auditory system is able to overcome adverse challenges in order to understand speech. What is unknown is how these challenges are overcome.

One kind of challenge encountered by CI users is spectral degradation. The implant provides an electric signal directly to the auditory nerve, but with only a coarse amount of spectral detail; the incoming sound spectrum is limited by the number of independent spectral processing channels, as well as interactions between the electrodes which carry information from those channels (Chatterjee and Shannon, 1998; Henry and Turner, 2003; Henry et al., 2005; Litvak et al., 2007). Thus, the subtle fine-grained spectral differences perceptible to those with acoustic hearing are not reliably distinguishable by those who use electric hearing (Townshend et al., 1987; Nelson et al., 1995; Kewley-Port and Zheng, 1998; Loizou and Poroy, 2001). It is thus predicted that CI users would experience difficulty attending to phonetic contrasts signaled by subtle spectral cues.

A. Overcoming spectral degradation
Despite severe spectral degradation, listeners can still recover a great deal of information from the speech signal. Shannon et al. (1995) observed that 3-4 channels of noise-excited spectral bands presented to normal-hearing (NH) listeners were sufficient for near-perfect recognition of sentences and vowels, with consonant performance nearly as good, approaching 90%. Performance by high-achieving CI users has been found to be comparable to NH listeners in these degraded spectral conditions (Fu et al., 1998), including conditions which simulate the effects of upward spectral shifting as a result of imperfect implant insertion depth (Fu and Shannon, 1999). Although modern implantable devices contain up to 22 different intracochlear electrodes in the implant array, results from experiments using vowel, consonant and sentence recognition tasks suggest that implant users can only take advantage of 6-8 independent spectral channels (Fishman, 1997; Friesen et al., 2001). NH listeners’ performance in quiet tends to plateau once the number of simulated channels reaches 8 (Dorman et al., 1997), or perhaps as high as 12 (Xu et al., 2005), but this plateau is dependent on the type of auditory stimulus. While engaged in more challenging tasks such as listening in background noise or listening to more difficult speech materials, normal-hearing listeners show increased benefit up to 20 channels (Friesen et al., 2001; Shannon and Galvin, 2004).

Not all features of the acoustic input are compromised in electric hearing. Shannon (1992) found that CI users’ temporal sensitivity to amplitude modulations in the electric domain was comparable to or superior to that of NH listeners hearing amplitude-modulated acoustic signals. Thus, although some phonetic cues are obscured by spectral degradation, it is expected that CI users can reliably use timing cues in speech, such as voice-onset-time, temporal amplitude envelope, or vowel duration. Kirk et al. (1992) measured identification accuracy of normal-hearing listeners and CI users listening to vowels embedded in wVw or bVb contexts,
which exploited consonant transitions at slow and fast rates, respectively. CI users were found to make use of static spectral contrasts in vowels, though they were not able to benefit from dynamic spectral changes which played a role in normal-hearing listeners’ responses. Iverson et al. (2006) found that neutralization of dynamic formant movement and vowel duration cues led to significant performance reductions in a 13-alternative vowel identification task performed by CI users and NH participants listening to spectrally-degraded simulations. Specifically, spectral degradation (whether simulated or via electric hearing) precluded accurate perception of dynamic formant information; listeners treated the dynamic cues as if they were steady-state.

The authors found that listeners with CIs used these phonetic cues to the same extent as those with normal hearing, suggesting that phonetic representations of vowels are stable despite the participants’ long periods of deafness / experience with electric hearing. Interestingly, the authors found no evidence that decreased spectral resolution led to increased attention to durational cues. In fact, information transfer analysis suggested that as spectral resolution was increasingly degraded, NH listeners recovered \textit{less} duration information, which runs counter to what might be predicted. In contrast, results from a multiple-talker vowel identification task by Chang (2006) revealed that some CI users received a significant amount of phonetic information in the temporal domain, though there was a noteworthy amount of variability. Important information about consonant voicing and manner is contained within the temporal envelope, but this information yields relatively little in terms of vowel identification, which is determined primarily by the number of simulated spectral channels (Xu et al., 2005). The question remains then, as to how high-performing CI users can adapt to spectral degradation in order to successfully identify speech sounds.
B. Trading relations in phonetic feature perception

The particular issue explored in this study is how a listener overcomes spectral degradation in order to accurately perceive speech contrasts signaled by spectral differences. It is for these spectrally-dominated contrasts that CI listeners may be expected to recruit secondary phonetic cues to compensate for the lack of spectral detail. The use of these secondary phonetic cues is not indicated under ideal circumstances of normal hearing and a quiet listening environment. Literature on cue-weighting has revealed that acoustic cues for phonetic features are held in a perceptual hierarchy whereby some exhibit dominance over others. That is, given multiple cues to a particular phonetic contrast, a listener will show consistently greater reliance upon one cue than the other(s) (Repp, 1982). The seemingly-discarded redundant cues, however, can be manipulated experimentally such that the exchange rate of phonetic information can be determined. Specifically, acoustic-phonetic boundaries driven by one cue can be shifted, given great enough change in an alternate cue. For example, trading relations can be observed in the organization of duration, voice pitch, and consonant release burst properties in the perception of stop consonant voicing at word-initial (Lisker, 1975; Summerfield & Haggard, 1977; Haggard et al., 1981; Whalen, 1993) as well as word-final position (Raphael, 1972, 1981). Additional trading relations are measurable for place of articulation for stops (Repp, 1978; Bailey and Summerfield, 1980) and for fricatives (Repp and Mann, 1981).

By systematically varying the number of simulated channels as well as the low-pass filter cutoff of the envelope in those channels, Xu et al. (2005) showed that spectral degradation brings about a corresponding increase in the effects of temporal envelope cues in phonetic identification. Similarly, trading relations have been observed between temporal and spectral
cues in perception of Mandarin lexical tones (Xu and Pfingst, 2003; Xu et al., 2002). NH listeners have been shown to adapt their use of these cues depending on attention and training (Francis et al., 2000, 2008). Though the artificial sounds heard in these experiments may not be faithful representations of everyday speech sounds, the investigation of weighting of perceptual cues is an avenue to understanding the perceptual processes that underlie speech perception.

C. The tense/lax distinction in English

The phonetic contrast explored in the present experiment is that between high-front lax and tense vowels in English. This feature helps to distinguish /I/ and /i/ in word pairs such as “hit/heat, fill/feel, hid/heed and bin/bean.” This contrast is ideal for the evaluation of perceptual weighting by CI users because of the specific acoustic features that define the English vowel space, described below.

1. The role of vowel formants

Since a landmark study by Peterson and Barney (1952), the most commonly examined feature of vowels has been the structure or relationship between frequencies of the first three vocal tract resonances, or formants. Whether in raw acoustic space, auditory space, or part of a normalization scheme, this information alone is regarded by many to be one of the necessary components used to identify vowels. Each vowel exhibits some category overlap with adjacent vowels in acoustic space, consistent with articulatory similarities of the vowels (Peterson and Barney, 1952; Hillenbrand et al., 1995). Still, using only this static formant frequency information, /i/ is identified with near-perfect accuracy of roughly 95%, and /I/ is identified with
roughly 75% accuracy in a 10-choice closed-set vowel identification task (Hillenbrand and Gayvert, 1993), where chance level would be 10%.

2. The role of vowel duration

Vowel duration was suggested by Ainsworth (1972) to be a significant contributor to the identification of English vowels. Decades later, the effect of duration was called into question when pattern classifiers and behavioral results conflicted with the results of Ainsworth. A discriminant analysis used by Hillenbrand et al. (1995) revealed that after organizing vowel categories according to formants, addition of duration information showed an average benefit of just 2.9% for classifier accuracy. Zahorian and Jagharghi (1993) also used an automatic pattern classifier to estimate the effects of durational features; when duration was added to the Peterson and Barney spectral measurements as a factor in this analysis, classification accuracy improved by less than 1%. The effect of vowel duration on listeners’ perceptual judgments was tested by Hillenbrand et al. (2000) in a 12-choice closed set vowel identification task where stimuli were created with either a formant synthesizer or a sinusoidal synthesizer. The latter signals were regarded as more faithful representations of original recordings, and this improvement in stimulus fidelity was accompanied by a decrease in the effect of vowel duration. In this study, duration-based shifts affecting the /i/-/I/ contrast were shown to be especially rare; short /i/ segments rarely were heard as /I/, and long /I/ segments were rarely heard as /i/ (crossovers occurred for only 0.6% of stimuli for both vowels). This negligible effect of duration on vowel identification contrasts with the results of Ainsworth, whose signals (consisting of two steady-state synthesized formants) were relatively low-fidelity in comparison. Hillenbrand et al. (2000) suggests that the differences in spectral resolution in experimental stimuli may bring about
overestimations of the use of duration in speech perception. Further evidence suggests that the nature of stimulus creation and fidelity of spectral cues yields significant effects on the estimation of phonetic cue use for stop sounds (Wardrip-Fruin, 1982, 1985), as well as for vowels (Assman and Katz, 2005). The emergent pattern from this research seems to suggest that as spectral resolution becomes finer, the effect of vowel duration decreases. The coarse spectral resolution offered by cochlear implants may suggest an increased role for this feature.

3. The role of vowel-inherent spectral change (VISC)

A relatively recent emphasis in vowel literature has been on the dynamic spectral changes that take place over the course of vowel pronunciation. Hillenbrand and Gayvert (1993) evaluated listeners’ use of the spectral information provided by Peterson & Barney (1952) in an experiment where participants heard steady-state vowels synthesized using their measurements of F0, F1, F2, and F3. Identification performance was 72.7%, which, while much higher than the chance level of 10%, is far worse than the recognition reported for the original recordings (94.4%). A falling F0 contour yielded a slight benefit, possibly due to the increased exposure to the spectral envelope resulting from harmonic movement. Generally though, performance in this task suggested that static spectral information is not entirely sufficient to identify English vowels. The vowel opposition relevant to this paper (/i/-/I/) showed a distinct pattern: performance for /i/ was very good (95.4%) but for /I/ was poor (76.8%), with a majority of the incorrectly-identified steady-state /I/ stimuli labeled as /i/ (a pattern which was substantiated by later analysis on the pattern of spectral change seen in /I/). The lax vowel was thus distinguished by a separate feature distinct from steady-state formant frequencies. Hillenbrand et al. (1995) noted that “the formants are moving in such a way as to enhance the contrast between vowels with similar static
positions in formant space” (p. 3106). The aforementioned discriminant analysis showed that F1 and F2 steady-state spectral peaks were not sufficient predictors of vowel identification results, yielding roughly 75% accuracy. Accuracy was improved by roughly 20% when a second pair of F1 and F2 spectral peak measurements were added from a later point in the vowel, thus accounting for spectral change over the course of pronunciation. The lax /I/ vowel was found to exhibit a centralizing pattern of spectral change, where the first formant rises and the second formant drops (by contrast, the tense /i/ vowel remains virtually steady throughout production).

Even limited information about formant change can yield significant improvements in listener identifications, as illustrated by Nearey & Assmann (1986), who presented listeners with two 30ms windowed clips (separated by 10ms silence) of merely the onset and offset a vowel. By listening to only these two concatenated snapshots of the vowel utterance, listeners showed identification error rates close to that for the unmodified vowel recordings, suggesting that brief glimpses at the onset and offset were nearly sufficient for accurate vowel recognition. When the two snapshots consisted of a repetition of the vowel onset, accuracy dropped by roughly 30%.

Hillenbrand and Nearey (1999) directly examined the role of formant change by having listeners identify naturally-spoken vowels, or synthesized vowels with either natural dynamic formant contours or artificially flattened contours which neutralized the spectral change. When the formant contour was flattened, accuracy dropped significantly. Most errors for the flat-contour lax /I/ vowels occurred because they were mis-labeled as tense /i/, highlighting the importance of VISC in the perception of this distinction.

4. The tense/lax distinction in English: summary and hypothesis
The features which contribute to perceptual distinction of /i/ and /I/ appear to include the spectral features of formant structure and dynamic formant movement, and perhaps duration. Studies which have implicated a role for duration in this distinction (Stevens, 1959; Ainsworth, 1972) have used synthetic stimuli which have compromised the integrity of spectral detail, possibly leading listeners to place an abnormally high emphasis on duration to compensate for degraded spectral information. Other studies suggest a negligible role of duration for this contrast (Hillenbrand et al., 1995; 2000; Jenkins, et al., 1983; Zahorian & Jargharghi, 1993). The role of formant structure is unquestioned, as the work of Peterson & Barney (1952) and Hillenbrand et al., (1995) laid the foundation for spectral analysis of vowels, and its relation to articulation.

Experiments by Bohn and Flege (1990) and Bohn (1995) revealed a roughly 98% change in identification scores for vowels at extreme endpoints of a formant structure continuum, while extreme endpoints of a duration continuum yielded only a roughly 9% change in identification. Although the role of formant movement is not entirely clear for all vowel contrasts, the work of Nearey and Assmann (1986), Hillenbrand and Gayvert (1993), Hillenbrand et al. (1995) and Hillenbrand and Nearey (1999) suggests that it is critical to distinguishing the lax and tense vowels which are the focus of the current investigation. Analysis of the methods used in these experiments suggests that as spectral integrity improves, a listener is less likely to rely on duration to determine phonetic identity of vowels in English. Although considerable improvements in speech synthesis and manipulation have improved the quality of signals in perceptual experiments, signal degradation is still commonly encountered for many individuals, including those with cochlear implants. We therefore anticipated that when distinguishing the tense and lax vowels in English, duration would exhibit greater-than-usual influence if participants used cochlear implants or were presented with sounds processed through a CI.
simulator. The question here is whether a real-life auditory constraint can bring a typically-ignored acoustic feature under a new perceptual spotlight. Additionally, it is expected that VISC will play a smaller role in vowel identification when spectral resolution is degraded, as results obtained by Dorman and Loizou (1997) indicated that CI users identified the lax vowel /I/ with an accuracy close to that reported in the literature for NH listeners who were denied access to vowel-inherent spectral change (Hillenbrand and Gayvert, 1993).

Iverson et al. (2006) suggest that the use of phonetic cues can change subtly or considerably, based on the degree of spectral resolution with which the speech signal is delivered to the listener. The current investigation explores the same issues with a finer level of analysis. That is, instead of altering VISC and vowel duration in an all-or-none binary fashion, they were gradually altered in a continuum, so that the perceptual weightings of each feature could be measured as they traded off with each other. This approach can potentially reveal the degree to which VISC and vowel duration (as well as the more obvious feature of formant structure) contribute to listeners’ vowel perception, rather than whether they contribute. Furthermore, the current investigation expands on the work of Iverson et al. by measuring the effects of vowel spectral structure, in addition to duration and VISC. It was hypothesized that the sensitivity to the three aforementioned phonetic features would change as a function of the degree of spectral resolution. Specifically, it was expected that reliance upon the spectral cues would decrease as the spectrum became less clear. Additionally, it was expected that reliance upon the durational cues would increase, since these cues are preserved in the electric or otherwise spectrally-degraded sound.
II. METHODS

A. Participants

Participants were divided into two groups. One group included 13 adult (19-63 years of age) native speakers of American English with normal hearing, defined as having pure-tone thresholds ≤20 dB HL from 250–6000 Hz in both ears (ANSI, 2004). The second group of participants included 4 adult (50-66 years of age) recipients of cochlear implants. CI users were all post-lingually deafened, and all were users of the Cochlear Freedom or N24 devices. See Tables I and II for demographic information and speech processor parameters for each cochlear implant user.

Table I here

Table II here

B. Stimuli

1. Speech synthesis

A 7x7x5 continuum of words in hVt form was synthesized using HLSYN (Sensimetrics Corp., 2003; Hanson et al., 1997; Hanson and Stevens, 2002). The nucleus of the synthesized words resembled a high-front American English vowel (tense /i/, lax /I/, or an ambiguous vowel intermediate to these). The vowel sound varied between tense-like and lax-like according to three feature continua. Vowel durations were adapted from characteristic durations of /i/ and /I/ (before voiceless stop sounds) reported by House (1961). Following pilot testing, the 100-130ms
range reported by House was expanded to 91-136ms to further disambiguate continuum endpoints. Intermediate durations were linearly interpolated in 7 steps (see Table I). Vowels also varied in initial formant structure according to a 7-step continuum. Parameter levels were based off those reported in the online database corresponding to the results of Hillenbrand et al. (1995). The Bark frequency scale (Zwicker and Terhardt, 1980; Syrdal and Gopal, 1986) was used to determine spectral parameter levels in order to better model the frequency map of the human auditory system. Using this scale, continuum endpoints were expanded by 0.12 of one auditory filter in order to ensure the presence of clearly-identifiable tense and lax segments, in view of potentially conflicting secondary features. Levels in Bark frequency were converted back to conventional frequency for the forthcoming results graphs to facilitate ease of interpretation. See Table II for a detailed breakdown of the parameter levels for formant structure. A third dimension of stimulus construction varied by the amount and direction of vowel-inherent spectral change (VISC), as defined by the amount of change undergone by each of the first three formants as vowels progress through the 20%, 50% and 80% timepoints of their total durations. The second and fourth items in this continuum resembled VISC for lax and tense vowels, respectively, measured by Hillenbrand et al. (1995). For the purposes of this contrast, vowel laxing for the second item involved increasing F1 and decreasing F2 & F3 (all by separate amounts indicated by Hillenbrand et al.). The fourth token had no VISC, as was reflective of tense /i/ vowels measured in the same study. The third item featured half the VISC seen in the lax vowel, interpolated along the Bark frequency scale. The first and fifth items were expanded beyond the range measured by Hillenbrand et al., and featured the same amount of VISC used between the intermediate three items, but in equal and opposite directions. See Table III for a detailed breakdown of this parameter. All synthesized words began and ended with 50ms of
silence. Word-initial [h] featured a 10ms onset ramp and 50ms of voiceless frication noise with steady formant structure that matched that for the initial steady-state portion of the vowel. Vowel pitch began at 120Hz, rose linearly to reach 125Hz at the 33% mark of the vowel, and declined smoothly to a 100 Hz by vowel offset. Word-final [-t] transition was modeled after that used by Bohn & Flege (1990), though formant transitions were shortened due to the duration differences between stimuli in their experiment compared to the current one. F1, F2, F3 and F4 transitions all began at the 80% timepoint in the vowel (to ensure that the entire 20% - 80% vowel formant trajectory could be realized). Offset values for formants 1, 2, 3 and 4 for the [t] transition were 300, 2000, 2900, and 3500 Hz, respectively, as in the experiment by Bohn & Flege. The formant transition was followed by a 65ms silent period, followed by a 30ms diffuse high-frequency burst as airflow faded to zero. Voiceless formants within this burst terminated at 400, 1600, 2600 and 3500 Hz.

Table III here

Table IV here

Table V here

2. Noise-band vocoding

Noise-band vocoding was accomplished using online signal processing within the iCAST software (version 5.04.02), developed by Qian-Jie Fu (2006). Stimuli were bandpass filtered
into 4 or 8 frequency bands using sixth-order Butterworth bandpass filters (24 dB/octave). The corner frequencies and bandwidths of the frequency bands varied with the number of bands. Specific values were determined using the logarithmic equation provided by Greenwood (1990), assuming a 35 mm cochlear length. The temporal envelope was extracted from each frequency band using half-wave rectification and low-pass filtering, with a cutoff frequency of 200 Hz (24 dB/octave) which is sufficient for good speech understanding (Shannon et al., 1995). The fine spectral content was replaced with a given number of noise carrier bands (4 or 8), created by bandpass filtering white noise with sixth-order (24 dB/octave) Butterworth filters. The extracted temporal envelope was then used to modulate the corresponding spectral noise bands, resulting in decreased spectral information and preservation of temporal envelope cues. The frequency content of each stimulus was carefully selected to stay within the NH limits (upper limit of 6000 Hz) measured for each participant. 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. The frequency band allocation was determined based on Greenwood’s logarithmic function (Greenwood, 1990) for all experimental conditions. No spectral shifting was used in the noise-band vocoding scheme.

C. 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 located at 0° azimuth. Each token was presented once, and listeners subsequently clicked on one of two word choices (“heat” or “hit”) to indicate their perception. Stimuli were presented in 122 or 123-token
blocks organized by degree of spectral resolution (unprocessed, 8-channel or 4-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.

D. Analysis

Listeners’ responses were plotted as % of “tense” perceptions as a function of stimulus parameters described in the Methods. Listeners’ sensitivity to vowel features was measured using slopes of these response curves corresponding to the three feature continua. Using a standard function in the SigmaPlot software (SYSTAT, 2004), response functions were fit to a 3-parameter sigmoidal curve with the following function:

\[
y = \frac{a}{1 + e^{-\frac{(x-x_0)^2}{b}}}
\]

where \(y\) = the % of responses which were identified as a tense vowel, \(x\) is the parameter level normalized within the range of values used in the experiment (i.e. a minimum duration of 85 ms reflected as “0” while a maximum duration of 145 ms reflected as “1”), \(x_0\) is the interpolated parameter level which produces a 50% crossover in identification, \(a = 1\) (the maximum possible likelihood of a response), and \(1/b\) determines the slope of the response curve. This slope is used to assess a listener’s sensitivity to the particular feature represented by the curve, as it reflects the change in perceptual judgments as a function of a change in stimulus parameter level. A shallow slope indicates that changes in the parameter level yield smaller perceptual changes, while steep slopes indicate that the same change in stimulus parameter level yields a greater likelihood of
perceptual change. In the remainder of this paper, we will not refer to the units of the slope; it should be understood, however, that the slope for a response function with frequency on the abscissa has different units than the slope of a function with duration on the abscissa. Response curves were found to consistently fit well to the sigmoidal function described above.

III. RESULTS

A. Initial formant structure

Figure 1 shows the response functions along parameter levels of initial formant structure collapsed across all levels of the other two parameters. The x-axis on this graph reflects F2 change, which covaried with F1 and F3 change. Various functions reflect responses for NH listeners in the unprocessed, 8-channel and 4-channel simulations, as well as those for CI users. Figure 2 shows the average slopes fit to these response curves. Average curve slope for the unprocessed stimuli was 5.91 (S.D.=1.57), for 8-channel simulation was 3.73 (S.D.=1.30) and for 4-channel simulations was 2.14 (S.D.=1.35). A repeated-measures analysis of variance (ANOVA) with one within-subject factor (number of channels [three levels: unprocessed, 8, 4]) was used to analyze vowel identification curve slope data. Results show a significant main effect of degree of spectral resolution [F(2, 24) = 38.79, p < 0.01]; as spectral resolution is degraded, the curve slopes corresponding to the initial formant structure become shallower. Tukey HSD post-hoc testing revealed significant differences between slopes for the unprocessed and 8-channel conditions, p < 0.01, between the unprocessed and 4-channel condition, p < 0.01, and between the 8-channel and 4-channel conditions, p = 0.02. The mean curve slope for the group of four CI users was 2.53 (S.D. = 1.69), which approximated the group mean for the 4-channel
The small number of CI users is insufficient for cross-group comparison for this and the forthcoming analyses.

FIGURE 1 HERE

FIGURE 2 HERE

B. Vowel-inherent spectral change (VISC)

Figure 3 shows the response functions along parameter levelss of VISC collapsed across all levels of the other two parameters. The x-axis on this graph reflects F2 change, which covaried with F1 and F3 change. Various functions reflect responses for NH listeners in the unprocessed, 8-channel and 4-channel simulations, as well as those for CI users. Figure 4 shows the average slopes fit to these response curves. Average slope for the unprocessed stimuli was 2.19 (S.D.=0.45), for 8-channel simulation was 1.07 (S.D.=0.32) and for 4-channel simulations was 0.52 (S.D.=0.31). A repeated-measures ANOVA was carried out, as for the main effect of initial formant structure. Results show a significant main effect of degree of spectral resolution [F(2, 24) = 107.02, p < 0.01]; as spectral resolution is degraded, the curve slopes corresponding to VISC become shallower. Tukey HSD post-hoc testing revealed significant differences between slopes for the unprocessed and 8-channel conditions, p < 0.01, between the unprocessed and 4-channel condition, p < 0.01, and between the 8-channel and 4-channel conditions, p <
The mean curve slope for the group of CI users was 1.08, which approximated the group mean for the 8-channel simulation conditions heard by the normal-hearing listeners.

Figure 3 Here

Figure 4 Here

C. Vowel duration

Figure 5 shows the response functions along parameter levels of vowel duration collapsed across all levels of the other two parameters. Various functions reflect responses for NH listeners in the unprocessed, 8-channel and 4-channel simulations, as well as those for CI users. Figure 6 shows the average slopes fit to these response curves. Average slope for the unprocessed stimuli was 1.29 (S.D.=0.52), for 8-channel simulation was 1.98 (S.D.=0.84) and for 4-channel simulations was 2.26 (S.D.=1.20). A repeated-measures ANOVA was carried out, as for the main effects of initial formant structure and VISC. Results show a significant main effect of degree of spectral resolution [F(1.10, 20.342) = 7.22, p <0.01]; as spectral resolution is degraded, the curve slopes corresponding to vowel duration become steeper. Tukey HSD post-hoc testing revealed significant differences between slopes for the unprocessed and 8-channel conditions, p < 0.01, between the unprocessed and 4-channel condition, p = 0.02. However, no significant differences were found between the unprocessed and 8-channel conditions or between
the 8-channel and 4-channel conditions. The mean curve slope for the group of CI users was 2.01, which approximated the group mean for the 8-channel simulation conditions heard by the normal-hearing listeners.

FIGURE 5 HERE

FIGURE 6 HERE

D. Interactions

1. Formant structure and vowel-inherent spectral change

When the initial formant pattern of a vowel is unambiguous, a listener is expected to rely on that cue and give relatively little attention to secondary cues such as spectral change or vowel duration. When that formant pattern lies ambiguously between two vowels, a listener may be expected to recruit those secondary features in order to identify the vowel. To assess the use of secondary cues, the vowel formant continuum was divided into three categories (lax, tense, and ambiguous) based on the values identified by Hillenbrand et al. (1995). The ultimate and penultimate endpoints were labeled as lax or tense, and the remaining three parameter levels in the center were regarded as ambiguous. Table III illustrates the divisions along the continuum of formant structure.

Table III here
Within each vowel category, listeners heard all levels of the VISC parameter. The levels which reflect those seen in natural speech production were items 2 and 4 from Table II. The difference in vowel identification scores in response to these two levels reflects the effect of VISC within the vowel category described above. A small VISC-driven difference score indicates that the natural range of VISC did not exhibit a large effect on listeners’ judgments, while a large score indicates that this formant change had relatively more influence. The VISC-driven difference scores for each condition of spectral resolution are plotted in Figure 7 as a function of the initial formant pattern. A repeated-measures ANOVA revealed a significant difference between the three categories of formant structure for unprocessed stimuli \(F(2, 24) = 33.66, p < 0.01\) and for 4-channel simulations \(F(2, 24) = 7.18, p < 0.01\) but not for the 8-channel simulations. Tukey HSD post-hoc testing for the unprocessed conditions revealed significant differences between VISC effects between lax and ambiguous vowels, between tense and ambiguous vowels, and between lax and tense vowels, all \(p < 0.01\). For the four-channel simulation, Tukey HSD post-hoc testing revealed significant differences in the VISC effect between lax and ambiguous vowels, \(p = 0.01\), and between tense and ambiguous vowels, \(p = 0.05\), but not between tense and lax vowels. The results suggest that spectral change exhibits a greater effect on listeners’ phonetic identifications when the formant structure is not clearly lax or tense. This effect is decreased as spectral resolution becomes poorer.

FIGURE 7 HERE

2. Formant structure and vowel duration
Just as for the effect of VISC, duration was expected to yield negligible effects for classically lax or tense formant structures, and larger effects for spectrally-ambiguous formant structures. The durations that reflect those seen in natural speech production (identified by House, 1961) were items 2 and 6 from Table I. The difference in vowel identification scores in response to these two levels reflects the effect of vowel duration within the aforementioned vowel categories (lax, ambiguous and tense). The duration-driven difference scores for each condition of spectral resolution are plotted in Figure 8 as a function of the initial formant pattern. A repeated-measures ANOVA revealed a significant difference between the three categories of formant structure for unprocessed stimuli [F(2, 24) = 9.29, p < 0.01] but not for 8-channel or 4-channel simulations. For unprocessed conditions, Tukey HSD post-hoc testing revealed significant differences between duration effects between lax and ambiguous vowels, p < 0.01, and between tense and ambiguous vowels, p = 0.04. No significant difference in duration effect was found between lax and tense vowels. The results suggest that vowel duration exhibits a greater effect on listeners’ phonetic identifications when the formant structure is not clearly lax or tense.

FIGURE 8 HERE

E. Correlation and cue-trading

Listeners showed a trend in their responses to formant structure as it related to their responses to vowel duration. Figure 9 illustrates a scatterplot of individual curve slopes for initial formant structure matched with corresponding individual curve slopes for duration,
representing all three conditions of spectral resolution. As curve slope increased for formant structure, it correspondingly decreased for vowel duration. Pearson product-moment correlations were -.52, -.40, and -.33 for the unprocessed, 8-channel and 4-channel simulations, respectively; none of these reached significance at the level of 0.05. For CI users, the correlation, r(4) = -0.70, did not reach significance, perhaps because of the low number of data points.

FIGURE 9 HERE

IV. DISCUSSION

The results shown in Figures 2, 4 and 6 demonstrate a distinct trend in terms of phonetic cue use as a function of spectral degradation. As the spectrum was degraded from the unprocessed signal to a coarse 8- or 4-channel CI simulation, the spectral features of formant structure and VISC played a smaller role in NH listeners’ identification. Conversely, vowel duration played a larger role. Although the units of measurement for formant frequencies, formant change and vowel duration do not allow for direct comparison to assess relative weighting, the range of parameter levels used in the experiment reflect the best available data on natural vowel production and thus reflect the same acoustic space between the phonetic categories involved. Hence, rates of perceptual change were calculated from “completely lax” to “completely tense,” with appropriate interpolations available for all of the three parameters.
Figures 7 and 8 reveal the conditional use of secondary vowel features. Within the middle of the continuum of initial formant structure, listeners were denied access to the informative primary cue for vowel identification. Within this range, they consequently exhibited patterns of heavier reliance upon secondary cues such as VISC as well as vowel duration. As the spectrum was degraded, this interaction grew weaker, perhaps because the spectral distinctions between plainly lax, tense and ambiguous vowels were compromised. Accordingly, the group of cochlear implant users did not show significant interactions for either secondary vowel feature.

There was a large amount of variability in the 4-channel condition, suggesting heterogeneity in listeners’ abilities to attend to spectral cues in this condition. Correspondingly, there was variability for the use of temporal cues in this condition as well. This relationship is illustrated by the scatterplot in Figure 9, which suggests a trading relation between the use of spectral and temporal features across all conditions of spectral resolution. In this identification task, losses in one perceptual domain were complimented by gains in another domain.

Results from CI users reveal a large amount of variability, which is not surprising in the context of previous literature on speech perception by this population (Lee et al., 2005; Lazard et al., 2010; Peterson et al., 2010). One CI user appeared to use the phonetic cues in a way similar to the average NH listener (with heavy emphasis on spectral rather than temporal cues), while another CI user showed the opposite pattern (judging virtually solely on vowel duration). Two other listeners showed intermediate patterns, which mirror the NH listeners’ patterns in the 8-channel condition. It is to be noted that the points on the scatterplot obtained with the CI listeners fall along the general function relating the slopes in the NH listeners, suggesting that listening strategies used by NH listeners in spectrally-degraded simulations are representative of those used by the small number of CI listeners used in this experiment.
It is clear that CI users can perform very well on consonant, vowel, word and sentence identification tasks, and that they can demonstrate good ability to converse in the real world (Hiraumi et al., 2007). The method by which they can attain this success has not been deeply explored beyond assessment of percentage of items correctly or incorrectly identified in a task. Analysis of phoneme identification by Xu et al. (2005) looked more deeply than this, as it accounted for separate effects of temporal and spectral features in overall performance. The current study attempts an even finer analysis by measuring listeners’ sensitivity to these features within a particular contrast that contains multiple cues. The present results suggest that spectral degradation of the signal brings about a change in the relative weighting of spectral versus temporal cues in vowel perception. This finding is in agreement with previous work (Ainsworth, 1972; Hillenbrand et al., 2000), which suggests that even small deviations from the full-spectrum signal yield increased influence of duration on listeners’ identification judgments. It is also in agreement with the findings of Xu et al. (2005), who identified a trading relation between spectral and temporal features in determining accuracy in consonant and vowel recognition. The current results contrast, however, with those of Iverson et al. (2006), who found that duration cues did not interact with spectral resolution in a 13-alternative vowel identification task. The latter two experiments utilized information transfer analysis to assess each of the vowel height, advancement and durational factors independently. For the phonetic contrast investigated by the current experiment, these factors are all shown to covary, so the previous analyses may have overestimated the amount of information transferred due to salience of cooperating covarying cues. No vowel pair in standard American English contrasts solely by duration, and all participants in the current experiment were screened for fluency in languages for which vowel duration is a phonemic feature (Finnish, Hungarian, Arabic, Vietnamese, etc). “Duration” (as
defined by the analyses used by Iverson et al. and Xu et al.) for English vowels can be signaled by spectral cues. Hence, listeners who used spectral cues to distinguish the tense and lax vowels in that study would be treated in the analysis as if they were taking advantage of durational cues. Hence, when spectral resolution was degraded, this spectral distinction was compromised, and duration information was consequently suggested to be compromised. Furthermore, the large number of phonemes tested in traditional vowel identification experiments includes contrasts for which duration would hypothetically play no role. Spectrally distinct vowels (such as /a/, /i/, /u/) would probably not require the use of secondary acoustic cues since even coarse spectral representation would still preserve a salient distinction (Kirk et al., 1992). For a more spectrally similar vowel pair like /i/ vs. /I/ (a distinction which carries a substantial semantic load in English), temporal cues could play a more crucial role which may be difficult to capture in a 12- or 13-choice task; spectrally-similar vowel pairs are outnumbered by spectrally-dissimilar vowel pairs. The informational transfer analysis by Iverson et al. incorporated, but did not specifically target spectrally-similar vowel pairs (such as i-I, u-U, æ-E). Additionally, the dichotic nature of the intact/neutralized parameters in some experiments (Hillenbrand and Nearey, 1999; Hillenbrand et al., 2000; Iverson et al., 2006) may obscure listeners’ sensitivity to each feature; perhaps there are gradations of sensitivity that underlie performance differences in various conditions of spectral degradation.

The effect of duration was not observed in all listeners in the most spectrally-degraded (4-channel) condition. In this condition, the subtle differences between adjacent levels in the formant structure continuum were expected to be obscured due to the wider bandwidths of the noise-excited spectral bands. Still, 3 NH participants appeared to show excellent discrimination of these sounds as indicated by their relatively steeply-sloped (>3.0) response functions along the
formant structure continuum in the 4-channel simulation. Interestingly, some of these listeners reported that they were attempting to use vowel duration in these conditions without any explicit instructions to do so, yet still yielded response curves showing a bias for spectral features. Anecdotal reports from some of these listeners indicated that they did not clearly hear the vowel qualities, and in some cases thought that they were relied heavily on duration even as the results indicated a preference for formant structure. Decomposition of first-formant information using temporal processing may be one explanation for this high level of spectral sensitivity. The limited number of CI users precludes a definitive conclusion regarding the perceptual weighting of these acoustic features by this population, though there appears to be evidence that they use perceptual strategies which differ from those used by NH listeners attending to spectrally-intact stimuli. As has been shown by previous literature, the CI listeners showed an average response pattern that matched well with the normal-hearing listeners in the 4- or 8-channel simulated conditions.

V. CONCLUSIONS

The current study attempts to highlight the potential differences in phonetic processing strategies that arise when speech sounds are spectrally degraded. Within this experiment, this focus was limited to the tense / lax vowel contrast, which has been previously shown to be cued primarily by spectral cues, and to remain unaffected by changes in vowel duration. As the spectrum was increasingly degraded, NH listeners showed decreased use of the spectral cues of formant structure and VISC, and show increased use of the temporal cue of vowel duration. This particular vowel contrast was chosen because it creates a specific kind of challenge for cochlear implant users, who are thought to maintain good temporal resolution while experiencing poor
spectral resolution (that is, they have poor access to the primary cue while maintaining good access to a normally-unused cue). Results from a limited number of CI users suggest that it is likely that individuals who use these prostheses re-organize the weighting of phonetic cues as they identify tense and lax vowels. Furthermore, results from these implant users appear to resemble those of normal-hearing listeners in conditions that are thought to best exemplify the effective spectral resolution of modern cochlear implants; the 4- or 8-channel noise-band vocoder was a reasonable simulation of electric hearing.

The current analysis attempts to shed light on the mechanisms by which a sound might be correctly or incorrectly perceived, and this method highlights a potential limitation in the traditional interpretation of vowel recognition scores achieved by CI users. Specifically, a simple correct/incorrect metric may not reveal the extent to which a listener perceives sounds in a typical fashion. Results show that a vowel can be identified by using more than one acoustic feature; even if a cochlear implant user correctly identifies a vowel in an experimental task, we cannot assume that it was because of the same perceptual strategies employed by normal-hearing listeners. Just as for a traffic detour, arrival at the final destination (correct sound identification) may have been a result of an alternative strategy of potentially higher difficulty, slower progress, or lower reliability (the use of secondary acoustic phonetic cues). Though the aforementioned trading relations between spectral and temporal cues may not be relevant for all phonetic contrasts (and thus may not show prominence in a task involving many phonemes), there exist some spectrally-similar phoneme pairs which are more likely to require some amount of perceptual adjustment. In view of the multiple acoustic cues available for any particular phonetic contrast, the vowel feature explored in this experiment may represent just a fraction of phonetic contrasts for which CI users employ alternative perceptual strategies. Thus, when
tallying percentage of correct responses in a vowel (or consonant) identification task, we may wish to exhibit caution when comparing results from cochlear implant users and normal-hearing listeners in the same tasks.
ACKNOWLEDGEMENTS

We would like to thank the participants for their time and willingness to contribute to this study, as well as Rochelle Newman and Shu-Chen Peng for their helpful comments and expertise. We are grateful to Qian-Jie Fu for the use of the software used for the experiment. This research was supported by National Institutes of Health R01 DC 004786.


Table I. Relevant demographic information about the CI participants in this study.

<table>
<thead>
<tr>
<th>CI user ID#</th>
<th>Etiology / type of hearing loss</th>
<th>Approx. duration of hearing loss</th>
<th>Age at testing</th>
<th>Age at implantation</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Unknown</td>
<td>Unknown</td>
<td>66</td>
<td>63</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>Genetic</td>
<td>10 years</td>
<td>66</td>
<td>63</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>Unknown</td>
<td>22 years</td>
<td>64</td>
<td>57</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>Unknown</td>
<td>Unknown</td>
<td>50</td>
<td>40</td>
<td>M</td>
</tr>
<tr>
<td>CI user ID#</td>
<td>Device</td>
<td>Processing strategy</td>
<td>Pulse rate</td>
<td>Stimulation mode</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>------------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Freedom</td>
<td>ACE</td>
<td>900</td>
<td>MP1+2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Freedom</td>
<td>ACE</td>
<td>1800</td>
<td>MP1+2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Nucleus 24</td>
<td>ACE</td>
<td>900</td>
<td>MP1+2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Nucleus 24</td>
<td>ACE</td>
<td>720</td>
<td>MP1+2</td>
<td></td>
</tr>
</tbody>
</table>
TABLE III. Acoustic parameter levels defining the vowel duration continuum. Items 2 and 6 reflect durations of naturally-spoken lax and tense vowels, respectively, as measured by House (1961)

<table>
<thead>
<tr>
<th>Token</th>
<th>Duration (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>108</td>
</tr>
<tr>
<td>4</td>
<td>115</td>
</tr>
<tr>
<td>5</td>
<td>122</td>
</tr>
<tr>
<td>6</td>
<td>130</td>
</tr>
<tr>
<td>7</td>
<td>145</td>
</tr>
</tbody>
</table>
Table IV. Acoustic parameter levels defining the initial formant structure continuum

<table>
<thead>
<tr>
<th>Token</th>
<th>F1 (Hz)</th>
<th>F2 (Hz)</th>
<th>F3 (Hz)</th>
<th>F4 (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>446</td>
<td>1993</td>
<td>2657</td>
<td>3599</td>
</tr>
<tr>
<td>2</td>
<td>418</td>
<td>2078</td>
<td>2717</td>
<td>3618</td>
</tr>
<tr>
<td>3</td>
<td>403</td>
<td>2122</td>
<td>2747</td>
<td>3628</td>
</tr>
<tr>
<td>4</td>
<td>389</td>
<td>2167</td>
<td>2778</td>
<td>3637</td>
</tr>
<tr>
<td>5</td>
<td>375</td>
<td>2213</td>
<td>2809</td>
<td>3647</td>
</tr>
<tr>
<td>6</td>
<td>362</td>
<td>2260</td>
<td>2841</td>
<td>3657</td>
</tr>
<tr>
<td>7</td>
<td>335</td>
<td>2357</td>
<td>2905</td>
<td>3677</td>
</tr>
</tbody>
</table>
TABLE V. Acoustic parameter levels defining the continuum of vowel-inherent spectral change.  
Items 2 and 4 reflect levels of spectral change in naturally-spoken lax and vowels, respectively, 
as measured by Hillenbrand et al. (1995)

<table>
<thead>
<tr>
<th>Token</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>-137</td>
<td>-63</td>
<td>0</td>
<td>49</td>
<td>-287</td>
<td>-33</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>-91</td>
<td>-42</td>
<td>0</td>
<td>33</td>
<td>-191</td>
<td>-22</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>-46</td>
<td>-21</td>
<td>0</td>
<td>16</td>
<td>-96</td>
<td>-11</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>-13</td>
<td>46</td>
<td>21</td>
<td>0</td>
<td>-16</td>
<td>96</td>
<td>11</td>
<td>0</td>
</tr>
</tbody>
</table>
TABLE III. Acoustic parameter levels defining the Lax, Ambiguous and Tense categories along the continuum of initial formant structure.

<table>
<thead>
<tr>
<th></th>
<th>Lax</th>
<th>Ambiguous</th>
<th>Tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>446</td>
<td>418</td>
<td>403</td>
<td>389</td>
</tr>
<tr>
<td>1993</td>
<td>2078</td>
<td>2122</td>
<td>2167</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2213</td>
<td>2260</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2357</td>
</tr>
</tbody>
</table>
FIG. 1. Responses from 13 normal-hearing listeners and four cochlear implant users, showing % of vowels identified as tense as formant structure was altered from prototypically lax to tense.
FIG. 2. Group mean and individual CI user curve slopes in response to the continuum of initial formant structure. Error bars indicate one standard deviation.
FIG. 3. Responses from 13 normal-hearing listeners and four cochlear implant users, showing % of vowels identified as tense as vowel-inherent spectral change was altered from prototypically lax to tense.
FIG. 4. Group mean and individual CI user curve slopes in response to the continuum of vowel-inherent spectral change. Error bars indicate one standard deviation.
FIG. 5. Responses from 13 normal-hearing listeners and four cochlear implant users, showing % of vowels identified as tense as vowel duration was altered from prototypically lax to tense.
FIG. 6. Group mean and individual CI user curve slopes in response to the continuum of vowel duration. Error bars indicate one standard deviation.
FIG. 7. VISC-driven difference scores organized by degree of spectral resolution and formant structure. Error bars indicate one standard deviation.
FIG. 8. Duration-driven difference scores organized by degree of spectral resolution and formant structure. Error bars indicate one standard deviation.
FIG. 9. Scatterplot of individual listeners’ curve slopes in response to vowel duration paired with the corresponding curve slope for initial formant pattern.