Phonetic cues are weighted differently when spectral resolution is degraded

Matthew B. Winn, Monita Chatterjee, William J. Idsardi [a)]

Dept of Hearing and Speech Sciences, University of Maryland, College Park, MD 20742

[a)] Dept of Linguistics, University of Maryland, College Park, MD 20742

Submitted 9/3/2010

ABSTRACT

Two experiments were conducted to test the hypothesis that listeners would modify perceptual strategies for phonetic identification if the spectrum of a speech signal was degraded. The contrasts between tense and lax vowels and between word-final s/z were selected because of the primacy of spectral cues and redundancy of temporal cues used in normal listening situations. In the first experiment, synthesized speech signals with vowels varying in two spectral dimensions (initial formant structure and vowel-inherent spectral change) and in one temporal dimension (vowel duration) were presented with intact spectra or through an 8- or 4-channel noise-excited vocoder, used to simulate a cochlear implant. In the second experiment, similar manipulations were made to vowels and final consonants in a natural utterance, with variables of F1 offset transition, vowel duration, fricative duration and duration of consonant voicing. Identification results showed that as the spectrum was degraded, listeners relied less upon the spectral cues, and compensated with greater reliance on the temporal cues. 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 used by individuals with normal hearing.

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 an increasing research focus on cochlear implants, literature on phonetic cue perception must be expanded to acknowledge the abilities of the continually-growing population of CI users. One avenue to explore this issue is through speech perception tasks using normal-hearing (NH) listeners, where the experimental sounds are spectrally degraded (Shannon et al., 1995). Research on the perception of phonetic cues is typically done in ideal conditions with normal-hearing listeners; it is unclear whether conclusions from these studies can generalize to individuals in degraded conditions. Perhaps the integration of acoustic features is governed by flexible strategies, sympathetic to the constraints encountered by the listener.

Spectral degradation is a challenge faced by CI users because a cochlear implant provides only coarse spectral detail, owing to the limited number of independent spectral processing channels in the device, as well as interactions between the electrodes which carry information from those channels (Chatterjee and Shannon, 1998; Henry et al., 2005). Thus, the subtle fine-grained spectral differences perceptible to those with normal hearing are not reliably distinguishable by those who use electric hearing (Kewley-Port and Zheng, 1998; Loizou and Poroy, 2001). It is thus presumed that phonetic cues driven by spectral contrasts would be most challenging for a listener encountering spectral degradation, such as a CI user.

A. Spectral degradation and electric hearing

Despite severe spectral degradation, NH listeners can still recover a great deal of information from the speech signal. Shannon et al. (1995) observed that only 3-4 channels of

noise-excited spectral bands presented to NH listeners were sufficient for near-perfect recognition of sentences and individual phonemes. Performance by high-achieving CI users has been found to be comparable to NH listeners in these simulations (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). Despite the presence of up to 22 intracochlear electrodes in modern implants, CI users have been shown to demonstrate the use of only 6-8 independent spectral channels (Fishman, 1997; Friesen et al., 2001). NH listeners make use of upwards of 20 channels (Friesen et al., 2001; Shannon and Galvin, 2004), particularly while engaged in more challenging tasks such as listening in background noise or listening to more complex materials.

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 non-spectral cues in speech, such as voice-onset-time, temporal amplitude envelope, or vowel duration.

B. Trading relations in phonetic feature perception

Given the presence of multiple acoustic cues to any particular phonetic contrast, a listener will show consistently greater reliance upon one primary cue over than the other(s) (Repp, 1982). The remaining secondary cues can be manipulated experimentally to reveal the extent to which they can influence a listener's perception of the primary cue and/or the phonetic segment as a whole. 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; Whalen et al., 1993) as well as word-final position (Raphael, 1972). Additional cue interactions are measurable for place of articulation for stops (Repp, 1978) and for fricatives (Repp and Mann, 1981). Trading relations between temporal and spectral cues have also observed for English consonants and vowels (Xu et al., 2005) as well as for Mandarin lexical tones (Xu and Pfingst, 2003; Xu et al., 2002). The relative weighting of temporal and spectral cues in stop consonant identification is sensitive to attention and training (Francis et al., 2000, 2008), suggesting plasticity of phonetic cue use. Though the artificial sounds heard in these experiments may not be faithful representations of everyday speech sounds, they reveal the malleability of phonetic perception in response to varying listening circumstances.

The tendency of listeners to attend to various phonetic features appears to be driven at least in part by stimulus fidelity and relative usefulness of cues which remain intact in the signal. This observation lends itself to the hypothesis that CI users may attend to phonetic features in a manner distinct from that of NH listeners, owing to the degradation of spectral structure. Although CI users' general rate of success in speech recognition has been explored by several researchers, it is not clear how these individuals integrate the various acoustic features which cue the phonetic segments of speech. Two experiments were conducted to test the hypothesis that spectral degradation is a circumstance which promotes a phonetic listening strategy that deviates from the norm. It is hoped that these experiments can additionally lend insight into the experience of CI users.

II. EXPERIMENT 1: THE TENSE/LAX VOWEL DISTINCTION

A. Review of acoustic features and hypothesis

The first experiment explored the high-front lax / tense vowel contrast in English, which distinguishes /I/ and /i/ in word pairs such as hit/heat, fill/feel, hid/heed and bin/bean. The features which contribute to this distinction include the spectral features of formant structure and dynamic formant movement, and possibly vowel duration. Steady-state spectral structure appears to be very informative for tense /i/, yielding 90% correct responses even when other cues are neutralized; lax /I/ is identified with 75% accuracy in these conditions (Hillenbrand and Gayvert, 1993). Although the role of dynamic 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, 2000) and Hillenbrand and Nearey (1999) suggests that time-varying spectral information is critical to distinguishing the lax and tense vowels which are the focus of the current investigation. Ainsworth (1972) showed that duration can modulate vowel identification in a task using two formants to synthesize a vowel sound. Additionally, Bohn and Flege (1990) and Bohn (1995) revealed a small effect of duration for the i/I contrast when synthesized formant structure was held constant (neutralizing dynamic spectral cues). However, these findings have been challenged by subsequent studies which preserved relatively richer spectral detail, including time-varying information (Hillenbrand et al., 1995; 2000; Zahorian and Jagharghi, 1993). In particular, Hillenbrand et al. (2000) reported that duration-based misidentifications of the i/I contrast occurred for less than 1% of stimuli. Hillenbrand et al. (2000), Nittrouer (2004) and Assman and Katz (2005) suggest that acoustic cue weighting for vowels is affected by signal fidelity, to the extent that commonly-used formant synthesizers (presumably such as the one used by Ainsworth) are likely to underestimate the role of time-varying spectral cues, and overestimate the role of temporal cues. Though this is an unfavorable

limitation of Ainsworth (1972), it is an important consideration when predicting the performance of listeners in spectrally-degraded conditions. 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.

The goal of the current study is to explore whether the auditory constraint of spectral degradation will compel listeners to alter their perceptual strategies for integrating acoustic features in phonetic identification. Some prior work indicates that this may not only be useful, but necessary. Kirk et al. (1992) found that CI users were able to make use of static spectral cues in vowels, but did not take advantage of a slow/fast formant transition contrast used by NH listeners. This would imply that the dynamic formant cue for lax vowels may be compromised in degraded conditions. Chang (2006) showed that some CI users made considerable use of temporal cues, though there was a noteworthy amount of variability. Xu et al. (2005) revealed that the use of temporal cues traded off with the use of spectral cues in consonant and vowel recognition tasks in degraded conditions. However, Iverson et al. (2006) found that spectral degradation did not promote durational cue use in a 13-alternative vowel identification task which specifically addressed durational and dynamic spectral cues.

There are some limitations in the analyses used in previous experiments. Xu et al. and Iverson et al. utilized information transfer analysis to attempt to assess each of the vowel height, advancement and durational factors independently. The acoustic properties of English vowels limit the validity of this analysis because vowel pairs identified as varying by duration (such as tense/lax i/I) also vary by spectral composition. Previous literature suggests that even for spectrally-similar vowel pairs, spectral differences are primary to the durational differences;

assessment of "duration" in the aforementioned informational transfer analyses is more reasonably interpreted as assessment of covarying dynamic spectral features. Hence, the use of duration cues may have not been objectively measured in these experiments. 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 question thus remains unclear as to how listeners can overcome spectral degradation to successfully distinguish spectrally-driven contrasts.

The current experiment seeks to clarify this issue by manipulating acoustic features gradually, so that the perceptual weightings of each cue can be measured as they trade off with each other for a particular phonetic contrast. This approach was chosen to 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. It was hypothesized that VISC would 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 VISC (through signal manipulation) (Hillenbrand and Gayvert, 1993). Additionally, it was expected that reliance upon the durational cues would increase, since they are preserved in the electric or otherwise spectrally-degraded sound; this pattern would be consistent with results of previous studies using spectrally sparse signals (Ainsworth, 1972; Raphael, 1972). The current experiment seeks to address limitations of previous studies in an attempt to reveal possible alternative cue-weighting strategies for a phonetic contrast predicted to be difficult for listeners challenged by spectral degradation.

B. Methods

1. Participants

Participants 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). A second group of participants included 4 adult (50-66 years of age) recipients of cochlear implants. This small group of CI users merely provided an initial sampling of performance by this population in the task; data obtained with these listeners were not included in statistical analyses or comparisons. CI users were all post-lingually deafened, and all were users of the Cochlear Freedom or N24 devices. See Table I for demographic information and speech processor parameters for each CI user. All participants were screened for fluency in languages for which vowel duration is a phonemic feature (Finnish, Hungarian, Arabic, Vietnamese, etc), to ensure that no participant entered with *a priori* bias towards durational feature sensitivity.

Table I. Relevant demographic information about the CI participants in this study. All used the ACE processing strategy.

CI user ID#	Etiology / of hearing loss	Approx. duration of hearing loss	Age at testing	Age at implant-ation	M / F	Device	Pulse rate	Stim. mode
1	Unknown	Unknown	66	63	F	Freedom	900	MP1+2
2	Genetic	10 years	66	63	F	Freedom	1800	MP1+2
3	Unknown	22 years	64	57	M	Nucleus 24	900	MP1+2
4	Unknown	Unknown	50	40	M	Nucleus 24	720	MP1+2

2. Stimuli

a. Speech synthesis. A 7x7x5 continuum of words in hVt form was synthesized using HLSYN (Hanson et al., 1997; Hanson and Stevens, 2002). The nucleus of the synthesized words resembled a high-front American English vowel which varied between tense-like /i/ and lax-like /I/. Vowel durations were adapted from characteristic durations of /i/ and /I/ (before voiceless stop sounds) reported by House (1961), and linearly interpolated in 7 steps (see Table II). Vowels also varied in initial formant structure according to a 7-step continuum which was based off values reported in the online database of Hillenbrand et al. (1995). The Bark frequency scale (Zwicker and Terhardt, 1980; Syrdal and Gopal, 1986) was used to interpolate spectral parameter levels in order to produce a balanced and symmetrical continuum with respect to the human auditory system. Levels in Bark frequency were converted to Hz in this article 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 penultimate items in this continuum resembled VISC for lax and tense vowels, respectively, measured by Hillenbrand et al. (1995). See Table III for a detailed breakdown of this parameter. Ranges for all three parameters were extended slightly beyond the typical natural production range at continuum endpoints, to ensure perceptibility in the presence conflicting co-occurring features. All stimuli began and ended with 50ms of silence. Word-initial [h] featured 50ms of steady voiceless formant structure which matched the initial steady-state portion of the vowel. Vowel pitch began at 120Hz, rose to 125Hz at the 33% mark of the vowel, and declined smoothly to 100 Hz by vowel offset. Word-final [-t] transitions contained

F1, F2, F3 and F4 offset targets at 300, 2000, 2900, and 3500 Hz, respectively, as used by Bohn & Flege (1990). These transitions all began at the 80% timepoint in the vowel (to ensure that the entire 20% - 80% vowel formant trajectory could be realized). The formant transition was followed by a 65ms of silent stop closure, 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 II. Acoustic parameter levels defining the vowel duration and initial formant structure continua, which were independently manipulated in the experiment.

Duration (ms)	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)
85	446	1993	2657	3599
100	418	2078	2717	3618
108	403	2122	2747	3628
115	389	2167	2778	3637
122	375	2213	2809	3647
130	362	2260	2841	3657
145	335	2357	2905	3677

TABLE III. 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)

	Formant change (Hz) from 20% to 50% timepoints				Formant change (Hz) from 20% to 80% timepoints			
Token	F1	F2	F3	F4	F1	F2	F3	F4
1	38	-137	-63	0	49	-287	-33	0
2	25	-91	-42	0	33	-191	-22	0
3	13	-46	-21	0	16	-96	-11	0
4	0	0	0	0	0	0	0	0
5	-13	46	21	0	-16	96	11	0

b. Noise-band vocoding. Noise-band vocoding (Shannon et al., 1995) was accomplished using online signal processing within the iCAST software (version 5.04.02; Fu, 2006). Stimuli were bandpass filtered into 4 or 8 frequency bands using sixth-order Butterworth bandpass filters (24 dB/octave). Specific values were determined assuming a 35 mm cochlear length (Greenwood, 1990). The cutoff frequency of the envelope lowpass filter was 200 Hz, which is sufficient for good speech understanding (Shannon et al., 1995). 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.

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 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 in each condition of spectral resolution.

C. Analysis

Listeners' responses were plotted as proportion of "tense" perceptions as a function of stimulus parameter values described earlier. For analysis, each parameter range was normalized so that the most "lax" level was 0.0, and the most "tense" level was 1.0. Listeners' sensitivity to vowel features was measured using slopes of these response curves corresponding to the three feature continua. Response functions were found to be well described by a 3-parameter sigmoidal curve of the following form:

$$y = \frac{a}{1 + e^{\left(\frac{x - x_0}{b}\right)}} \quad (1)$$

For equation (1), y is the proportion 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 (100%:

the maximum possible likelihood of a response), and 1/b determines the slope of the response curve. 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.

D. RESULTS

1. Initial formant structure

Figure 1 shows identification response functions along parameter levels of initial formant structure, collapsed across all levels of the other two parameters; similar analyses were carried out for all features used in this and the subsequent experiment. Figure 2 shows the slopes corresponding to these 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 (spectral resolution [three levels: unprocessed, 8 channel, 4 channel]) 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 simulation conditions heard by the normal-hearing listeners. The small number of CI users is insufficient for cross-group comparison for this and the forthcoming analyses.

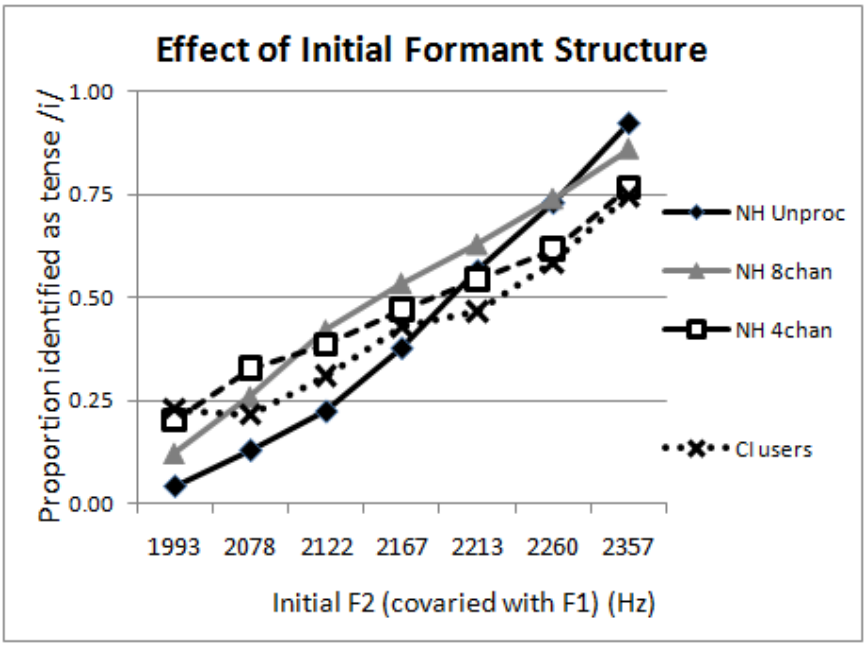

FIG. 1. Responses from 13 normal-hearing listeners and four cochlear implant users in Experiment I, showing proportion of vowels identified as tense as formant structure was altered from prototypically lax to tense. Similar stimulus-response curves were collected for all other variables used in each experiment.

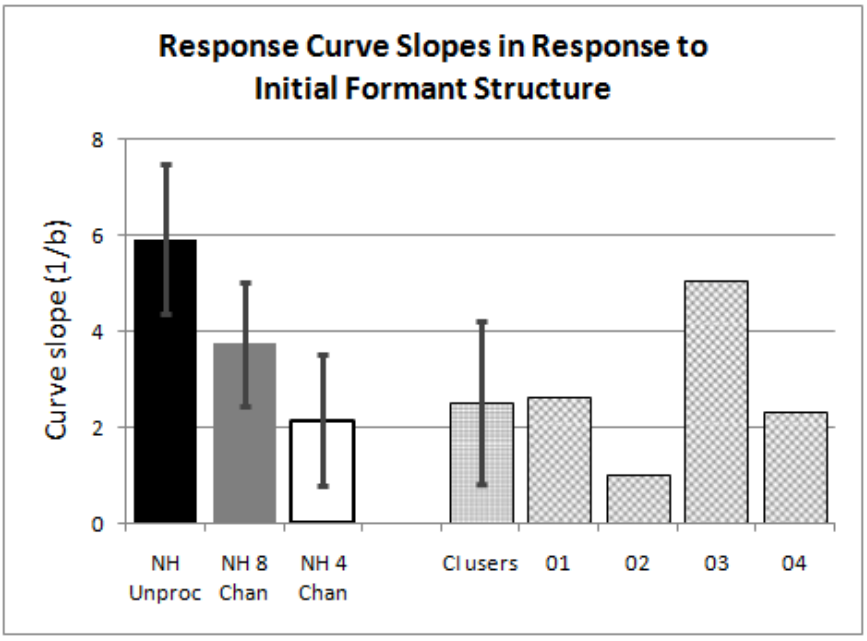

FIG. 2. Group mean and individual CI user curve slopes in response to the continuum of initial formant structure for Experiment I. Error bars indicate one standard deviation.

2. Vowel-inherent spectral change (VISC)

Identification response functions along parameter levels of VISC were collapsed across all levels of the other two parameters. Figure 3 shows the slopes corresponding to these 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 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, $p < 0.01$. 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.

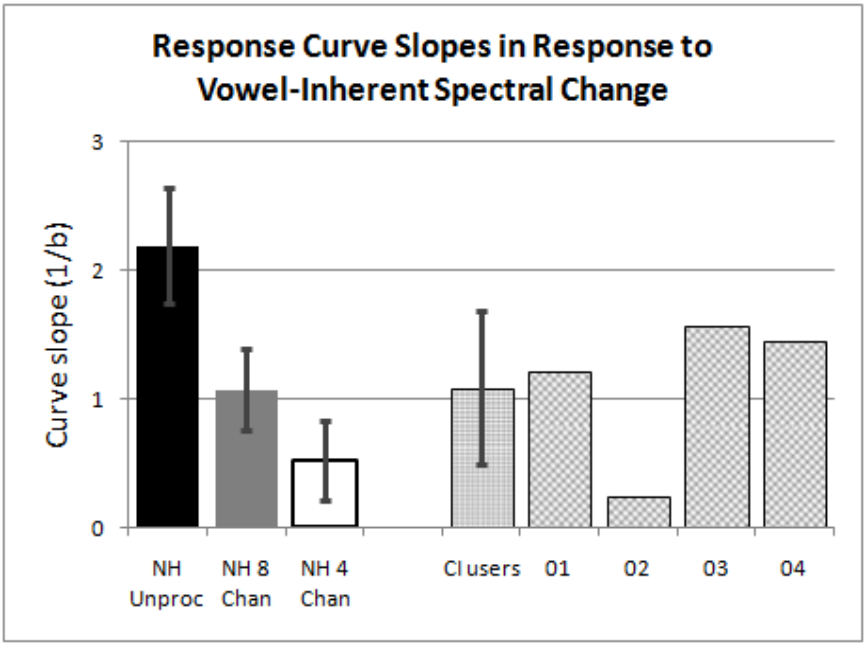

FIG. 3. Group mean and individual CI user curve slopes in response to the continuum of vowel-inherent spectral change for Experiment I. Error bars indicate one standard deviation.

3. Vowel duration

Identification response functions along parameter levels vowel duration were collapsed across all levels of the other two parameters. Figure 4 shows the slopes corresponding to these 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 effects of initial formant structure and VISC. Results show a significant main effect of degree of spectral resolution [$F(1.10, 20.34) = 7.22$, $p < 0.01$]; 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 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.

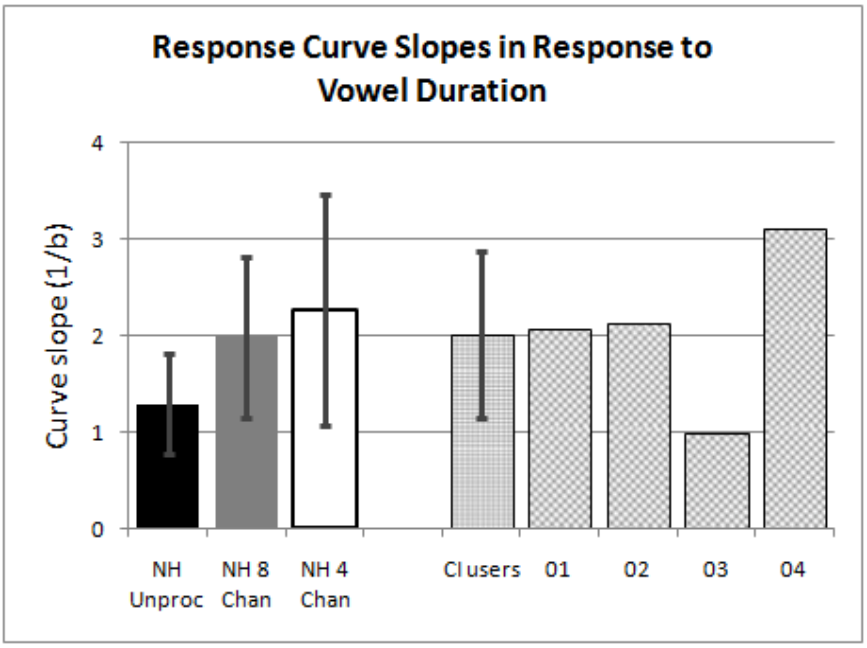

FIG. 4. Group mean and individual CI user curve slopes in response to the continuum of vowel duration for Experiment I. Error bars indicate one standard deviation.

III. EXPERIMENT 2: THE POSTVOCALIC S/Z CONTRAST

A. The role of vowel duration and F1 transition at vowel offset

A second experiment was conducted to supplement the results of the first experiment, with particular focus on dynamic spectral cues and vowel duration. In this task, listeners chose between two words (loss and laws) which contrasted phonologically by final consonant voicing. Several investigators have examined the role of spectral and temporal phonetic features in this distinction, and have focused primarily on stop consonants. Influential experiments by Chen (1970) and Raphael (1972) have suggested vowel duration to be the operative acoustic feature in this distinction. As for the aforementioned study by Ainsworth (1972), the limited spectral integrity of Raphael's stimuli (three steady-state synthesized formants) may have yielded artificially-high reliance upon duration. Furthermore, between ("critically") long or short vowel

durations in that experiment, the synthetic dichotic "voicing" difference (implemented as a falling F1 transition as well as durational manipulations of the final consonant) exhibited an appreciable but rather neglected effect on listeners' judgments of several cognate contrasts, including that for s/z. Since the publication of the study by Raphael (1972) several authors have argued against the dominance of vowel duration in the final consonant voicing distinction. For example, Hillenbrand et al. (1984) noted that shortening the duration of vowels before voiced stops does not significantly alter listeners' perceptions. When vowel offset transitions were deleted, listeners exhibited an appreciable bias towards voiceless consonant perception, implicating the role of dynamic spectral features at vowel offset (rather than vowel duration) in this distinction. Similar findings were found by Revoile (1982) and Wardrip-Fruin (1982), and summarized in a review by Walsh and Parker (1984), who noted that vowel length exhibits an effect only for "artificial or abnormal circumstances." Revoile (1982) reported that vowel duration was used as a consonant voicing cue only by individuals with hearing impairment. Vowel duration was observed by Nittrouer (2004) to exert an appreciable effect on perceptual judgments in an experiment that used synthetic speech, but this effect was strongly overpowered by F1 offset transition when natural speech tokens were used. Thus, just as for previous experiments with vowels, the effect of duration on perceptual judgments appears to be modulated by spectral fidelity of the signal.

The current experiment was designed to measure listeners' use of spectral cues (F1 offset transition in a vowel) and temporal cues (vowel duration) in the perception of speech sounds which were spectrally intact or spectrally degraded. The voicing contrast for fricatives was used because the word-final stop consonant voicing contrast has been explored thoroughly (see Walsh and Parker, 1984 and Nittrouer, 2008 for reviews of such work). Furthermore, the cue of stop

closure voicing varies dichotomously, which is signaled less by spectral differences than the mere presence or absence of acoustic energy.

B. Methods

1. Participants

Participants for Experiment 2 were comprised of 10 adult (18-26 years of age) listeners with normal-hearing (from 250 – 8000 Hz) and 4 cochlear implant users whose demographics were the same as those for Experiment 1 (see section II B. 2.). 4 of the NH listeners and all 4 CI users also participated in Experiment 1.

2. Stimuli

Stimuli for the second experiment were constructed using natural recordings of the words “loss” and “laws.” A single /l/ segment of intermediate duration was chosen as the onset of all stimuli in the experiment, to neutralize it as a cue for final voicing (Hawkins and Nguyen, 2004). The vowel was segmented from a recording of “laws,” and thus contained a "voiced" F1 offset transition from roughly 615 Hz to 450 Hz, which is in the range of transitions in natural speech observed 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 615 Hz F1 offset) and expanding the duration to the original value using the pitch synchronous overlap-add (PSOLA) feature in Praat (Boersma and Weenink, 2010). Rather than using recordings from “loss” and “laws” separately, this manipulation was preferable, in order to maintain stable phonation quality, as well as equivalent onset and steady-state vowel formant values, which have been shown to affect voicing judgments (Summers, 1988). This earlier-occurring acoustic information is not within the scope of this investigation, which is focused instead on the use of

dynamic (rather than steady-state) spectral information. A decaying amplitude envelope was applied to the final 60 ms of all vowels, as in Flege (1985); it resembled a contour intermediate to those seen in the natural productions (where pre-voiceless vowels terminated at virtually zero amplitude, while pre-voiced vowels terminated at roughly 10 dB less than peak volume. This consistent artificial amplitude contour was used to neutralize offset amplitude decay as a cue for voicing (Hillenbrand et al., 1984), since amplitude perception is not within the scope of this investigation. Vowel durations were manipulated using PSOLA to create a 7-step continuum between 175ms and 325 ms, based on values from natural production reported by House (1961) and Stevens (1992), and used by Flege (1985) in perceptual experiments. The resulting 14 vowels (7 pre-voiced and 7 pre-voiceless) were manipulated using PSOLA to contain the same falling pitch contour (which started at 96 Hz and ended at 83 Hz), because pitch has been shown to affect judgment of final fricative voicing (Derr and Massaro, 1980, Gruenenfelder and Pisoni, 1980), but was not within the scope of this investigation. 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 resemble those used by Soli (1982), Flege (1985) and Summers (1988). 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. One-third of the fricatives were kept entirely voiceless, leaving a 3-level roving factor of voicing varying between 0 - 50 ms, which resembles that 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 periodic fricative noise was blended with the last pitch period of the vowel (each at 50% volume) to produce a smooth

transition between segments. The resulting stimulus set consisted of 126 items which varied in four dimensions: presence/absence of vowel-offset falling F1 transition (2 levels), vowel duration (7 levels), duration of fricative (3 levels), and duration of voicing within that fricative (3 levels). The former two factors are presented in the analysis, as they pertained to the focus of temporal vs. spectral cue-trading. Fricative duration was not predicted to affect listeners' judgments (Soli, 1982), and voicing within the fricative is frequently absent altogether in natural speech (Haggard, 1978); these two roving values served to provide natural variability in the stimulus set, but were not considered in the analysis.

3. Noise-band vocoding

Noise-band vocoding was accomplished using a procedure similar to that used in Experiment 1 (described in section II B. 2. b.), except that the upper-limit of the analysis and filter bands was changed from 6 kHz to 7 kHz, to ensure that a substantial amount of frication noise was represented within the spectrally-degraded output.

4. Procedure

The procedure for Experiment 2 was the same as that for Experiment 1 (described in section II B. 3.), with minor modifications to account for the different stimulus set. Visual word choices were "Loss" and "Laws," and the 126-member stimulus set was presented in alternating blocks of unprocessed and 8-channel noise-band vocoder conditions. Each block was heard 5 times in each condition of spectral resolution.

C. Analysis

Listeners' use of the dynamic spectral feature (F1 transition at vowel offset) was measured as the difference in percentage of stimuli judged as "laws" (voiced offset) in the presence of the F1

transition compared to those in the absence of the transition; a higher difference score indicates greater influence of the dynamic spectral cue. Listeners' use of the temporal feature (vowel duration) was measured using the same method as for Experiment 1 (described in section II C.), with higher curve slope values indicating greater use of the temporal cue.

D. Results

1. F1 transition at vowel offset

Figure 5 shows the difference in percentage of stimuli judged as voiced when the F1 offset was falling, compared to that when the F1 offset was flat. For all three conditions (NH unprocessed, NH degraded, CI user), more stimuli were judged as voiced when the F1 offset was falling. The average NH difference score for unprocessed stimuli was 52% (S.D. = 10%) and for the 8-channel simulations was 19% (S.D. = 10%). For NH listeners, a paired-samples t-test revealed the effect of spectral degradation to be significant $t(9) = 15.89$, $p < 0.001$; use of the dynamic spectral cue decreased when the spectrum was degraded. CI users' difference score was 15% (S.D. = 10%), which resembled that for the NH listeners in the 8-channel condition. As for Experiment I, comparison between NH and CI groups was not conducted due to the limited CI group sample size.

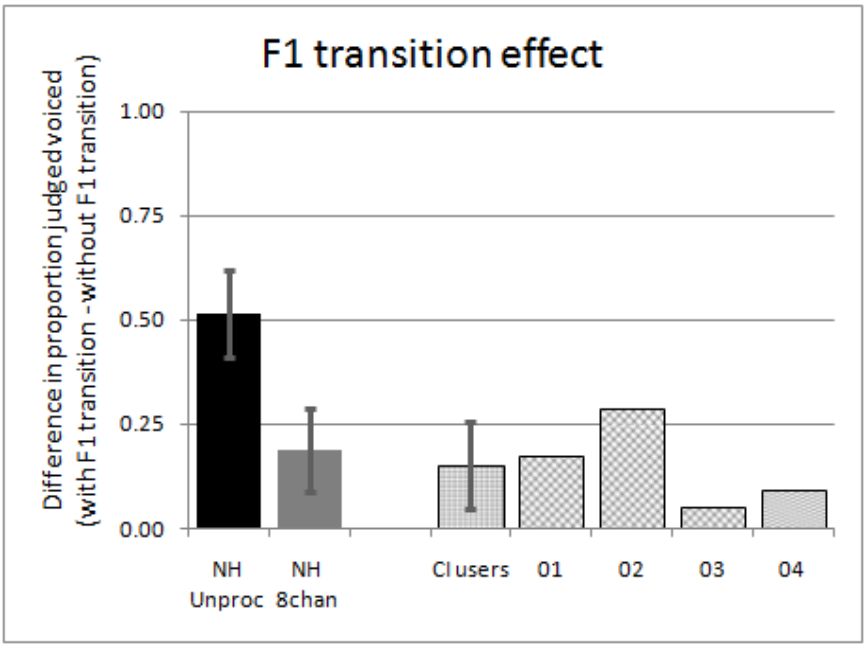

FIG. 5. Group mean and individual CI user effects on "voiced" responses as function of vowel offset F1 transition. Y-axis indicates proportion voiced responses with falling F1 transition minus proportion voiced responses with flat F1 offset. Error bars indicate one standard deviation.

2. Vowel duration

Identification response functions along parameter levels of vowel duration were collapsed across all levels of the other parameters. Figure 6 shows the slopes corresponding to these curves. Average slope for the unprocessed stimuli was 1.83 (S.D. = 1.08), and for the 8-channel simulation was 2.56 (S.D.=2.28). A paired-samples t-test revealed the effect of spectral resolution to be significant $t(9) = 2.75$, $p < 0.02$; use of the temporal cue increased as the spectrum was degraded. The average slope of the CI users' response curves was 4.11 (S.D. = 1.81), which was greater than that for the NH listeners in simulated conditions; the small number of CI users precludes the application of comparative statistics.

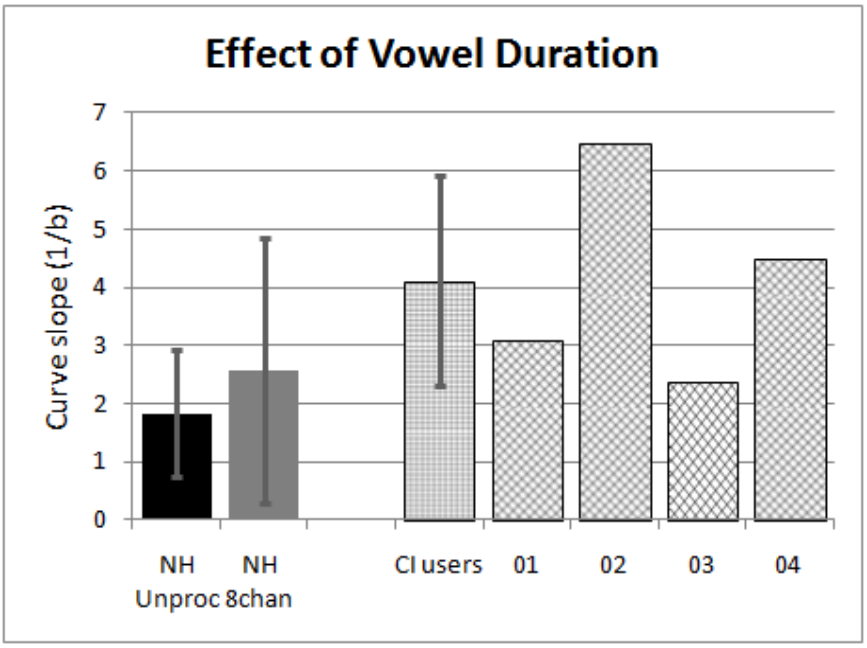

FIG. 6. Group mean and individual CI user curve slopes in response to the continuum of vowel duration for Experiment II. Error bars indicate one standard deviation.

IV. DISCUSSION

The results shown in Figures 2 through 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 static and dynamic formant structure played a smaller role in NH listeners' phonetic identifications. Conversely, vowel duration played a larger role. Although the differing units of measurement for formant frequencies, formant change and vowel duration do not allow for direct comparison to assess relative weighting (vis a vis the "exchange rate" of spectral versus temporal cues), the range of parameter levels used in the experiment reflect the best available data on natural production and thus reflect ostensibly similar acoustic space between the phonetic categories involved. A normalized scale was used to assess the slope of all feature continua, to allow for approximate comparisons.

In Experiment I, there was a large amount of variability in the 4-channel simulation, suggesting heterogeneity in listeners' abilities to attend to spectral and temporal cues in this condition. Interestingly, some 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. Other listeners in this condition reported attempting to attend to unforeseen and unmanipulated sound features in this degraded condition, including voice quality, stop release, and spurious vowel qualities (neither /i/ nor /I/). Others were able to complete the task seemingly without much interference from the spectral degradation. There was no apparent pattern underlying these differences.

Results obtained with CI users showed considerable variability. 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 (seemingly relying solely on vowel duration). Two other listeners showed intermediate patterns, which mirrored the NH listeners' patterns in the 8-channel condition. The results obtained with the CI listeners fall along the general pattern seen 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. 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 strategies which differ from those used by NH listeners attending to spectrally-intact stimuli. In agreement with previous literature, the CI listeners in the current experiments showed an average response pattern that matched well with the normal-hearing listeners in the 4- or 8-channel simulated conditions.

Experiment II reinforced the conclusions of Experiment I by testing a different phonetic contrast. While dynamic spectral cues heavily influenced NH listeners' identification responses

when spectra were intact, this effect was reduced when the spectrum was degraded. Conversely, vowel duration played a larger role. Both of these features were extrinsic to the phonologically contrastive segment; listeners were able to use (and adjust weighting) of cues in a vowel to determine the identity of a following consonant.

V. CONCLUSIONS

The current study attempted to highlight the potential differences in phonetic processing strategies that arise when speech sounds are spectrally degraded. As speech signals were degraded, NH listeners showed decreased use of the spectral cues of formant structure, VISC and formant transitions, and increased use of vowel duration. Results from a limited number of CI users suggest that individuals who use these prostheses may re-organize the weighting of acoustic cues as they identify tense and lax vowels, or final consonant voicing. 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. Compared to NH listeners listening to the same stimuli, CI users appeared to use spectral cues to a lesser extent, and appeared to use durational cues to a greater extent.

The current analysis 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. That is, even if a CI 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 segment, the contrasts explored in this study may represent just a fraction of those for which CI users employ alternative perceptual strategies. Thus, caution should be used when comparing results of NH listeners and CI users in the same tasks; similar performance may not verify similar perception.

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 software used for the experiment. This research was supported by NIH grant no. R01 DC004786 to MC. MBW was partially supported by NIH grant no. T32 DC000046-17 (PI: Arthur N. Popper).

REFERENCES

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

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

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

Boersma, Paul & Weenink, David (2010). Praat: doing phonetics by computer [Computer program]. Version 5.1.23, retrieved 1 January 2010 from http://www.praat.org/

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

Bohn, O.-S. & 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., & Shannon, R. V. (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.

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., & 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. F., Loizou, P. C., & Rainey, D. (1997). “Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs.” J. Acoust. Soc. Am. 102, 2403–2411.

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., & Nusbaum, H. C. (2000). “Effects of training on attention to acoustic cues.” Percept. Psychophys. 62, 1668–1680.

Francis, Kaganovich, & Driscoll-Huber (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., & 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., & Shannon, R. V. (1999). "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing." J. Acoust. Soc. Am.105, 1889–1900.

Fu, Q.-J., Shannon, R. V., & Wang, X. (1998). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing." J. Acoust. Soc. Am. 104, 3586–3596.

Fu, Q.-J. (2006). TigerSpeech technology: Innovative speech software, version 1.05.02, available from http://www.tigerspeech.com/tst_tigercis.html (Last viewed May 16, 2010).

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 speech 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/ inEnglish. J. Phonetics, 32, 199–231.

Henry, B. A., Turner, C. W., & 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., 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.

Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K., (1995). "Acoustic characteristics of American English vowels." J. Acoust. Soc. Am. 97, 3099-3111.

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. J. Acoust. Soc. Am. 108, 3013–3022.

Hillenbrand, J. M., & 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., & Nearey, T. M. (1999). "Identification of resynthesized /hVd/ utterances: Effects of formant contour." J. Acoust. Soc. Am. 105, 3509–3523.

House, A. (1961). "On vowel duration in English." J. Acoust. Soc. Am. 33, 1174-1178.

Iverson, P., Smith, C., & Evans, B. (2006). "Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration." J. Acoust. Soc. Am. 120, 3998-4006.

Kewley-Port, D. & Zheng, Y. (1998). "Modeling formant frequency discrimination for isolated vowels." J. Acoust. Soc. Am. 103, 1654-1666.

Kirk, K. I., Tye-Murray, N., & Hurtig, R. R. (1992). "The use of static and dynamic vowel cues by multichannel cochlear implant users." J. Acoust. Soc. Am. 91, 3487–3498.

Lisker, L. (1975). “Is it VOT of a first-formant transition detector?” J. Acoust. Soc. Am. 57, 1547-1551.

Loizou, P. & Poroy, O. (2001). “Minimum spectral contrast needed for vowel identification by normal hearing and cochlear implant listeners.” J. Acoust. Soc. Am. 110, 1619-1627.

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. and Lowenstein, J. (2008). “Spectral structure across the syllable specifies final-stop voicing for adults and children alike.” J. Acoust. Soc. Am. 123, 377-385.

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. (1978). “Perceptual integration and differentiation of spectral cues for intervocalic stop consonants.” Percept. Psychophys. 24, 471-485.

Repp, B. & Mann, V. (1981). “Perceptual assessment of fricative-stop coarticulation.” J. Acoust. Soc. Am. 69, 1154-1163.

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.

Shannon, R. V. (1992). "Temporal modulation transfer functions in patients with cochlear implants." J. Acoust. Soc. Am. 91, 2156–2164.

Shannon, R., Fu, Q-J., and Galvin, J. (2004). "The number of spectral channels required for speech recognition depends on the difficulty of the listening situation." Acta Otolaryngol Suppl. 552, 50-54

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.

Syrdal, A. K. & Gopal, H. S. (1986). A perceptual model of vowel recognition based on auditory representation of American English vowels. J. Acoust. Soc. Am. 79, 1086-1100.

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.

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 (2003). "Relative importance of temporal envelope and fine structure in lexical-tone perception." J. Acoust. Soc. Am. 114, 3024–3027

Xu, L., Tsai, Y., and Pfingst, B. E. (2002). ''Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses,'' J. Acoust. Soc. Am. 112, 247–258.

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. in Biomed. Eng., 1, 115-142.

Zwicker, E., & Terhardt, E. (1980). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency." J. Acoust. Soc. Am. 68, 1523-1525.