

# Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration

Paul Iverson<sup>a)</sup>

*Department of Phonetics and Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom*

Charlotte A. Smith

*Department of Human Communication Science, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom*

Bronwen G. Evans

*Department of Phonetics and Linguistics, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom*

(Received 21 November 2005; revised 25 September 2006; accepted 26 September 2006)

Previous work has demonstrated that normal-hearing individuals use fine-grained phonetic variation, such as formant movement and duration, when recognizing English vowels. The present study investigated whether these cues are used by adult postlingually deafened cochlear implant users, and normal-hearing individuals listening to noise-vocoder simulations of cochlear implant processing. In Experiment 1, subjects gave forced-choice identification judgments for recordings of vowels that were signal processed to remove formant movement and/or equate vowel duration. In Experiment 2, a goodness-optimization procedure was used to create perceptual vowel space maps (i.e., best exemplars within a vowel quadrilateral) that included F1, F2, formant movement, and duration. The results demonstrated that both cochlear implant users and normal-hearing individuals use formant movement and duration cues when recognizing English vowels. Moreover, both listener groups used these cues to the same extent, suggesting that postlingually deafened cochlear implant users have category representations for vowels that are similar to those of normal-hearing individuals. © 2006 Acoustical Society of America. [DOI: 10.1121/1.2372453]

PACS number(s): 43.71.Es, 43.71.Ky [AJO]

Pages: 3998–4006

## I. INTRODUCTION

Monophthongal English vowels have long been thought to be recognized by their F1 and F2 target frequencies, but it has become clear that finer-grained phonetic variation, such as intrinsic formant movement and duration, is also important for recognition by normal-hearing native English speakers. For example, vowel recognition accuracy in quiet declines by about 15–23 percentage points when vowel formant movement is flattened in synthesized or signal-processed speech (e.g., Assmann and Katz, 2005; Hillenbrand and Nearey, 1999), and vowels can be recognized even when the relatively steady-state portions (i.e., where the formant frequencies meet their targets) have been removed (e.g., Strange, 1989). At least in American English, vowel duration likely has a smaller influence on intelligibility (e.g., lengthening or shortening vowels reduces vowel identification accuracy by about 5 percentage points), although speakers systematically vary vowel duration in their productions (e.g., Hillenbrand *et al.*, 2000). This recent emphasis on fine-grained phonetic variation in vowels parallels work on episodic memory and talker differences, which suggests that such phonetic details are an important contributor to speech

understanding, rather than a nuisance that must be normalized or removed (e.g., Hawkins and Smith, 2001; Johnson, 2005; Nygaard and Pisoni, 1998).

Current evidence suggests that listeners also use fine-grained acoustic variation to recognize vowels under adverse conditions (e.g., noise, hearing impairments, or cochlear implants). For example, Neel (1998) found that normal-hearing and elderly hearing-impaired individuals were affected similarly by manipulations of signal-processed vowels; recognition accuracy declined for both groups when formant movement was removed and duration was equated. Ferguson and Kewley-Port (2002) found that normal-hearing and hearing-impaired individuals use formant movement and duration when listening to natural speech in multi-talker babble, but the errors for specific vowels differed between the two groups, suggesting that the two groups had somewhat different cue weightings. Kirk *et al.* (1992) found that both normal-hearing listeners and cochlear implant users were able to recognize vowels above chance based only on consonantal formant transitions (i.e., edited CVC syllables in which the quasi-steady-state vowel portion was removed), although removing these consonantal formant transitions from natural CVCs (i.e., allowing listeners to hear the quasi-steady-state vowel without the consonants) had no effect on vowel recognition.

The present study examined whether vowel-intrinsic formant movement (i.e., formant movement within the vowel,

<sup>a)</sup>Author to whom correspondence should be addressed.

rather than the consonantal formant transitions examined by Kirk *et al.*, 1992) and duration are used for vowel recognition by postlingually deafened adult cochlear implant users, and normal-hearing individuals listening to cochlear implant simulations. It would be surprising if exactly the same cues were used when recognizing vowels via cochlear implants and normal hearing, because the sensory information provided by acoustic and electric hearing differ substantially. Modern cochlear implants represent the continuous spectrum of speech with a relatively small number of spectral channels, and vowel recognition accuracy seems to be primarily limited by the effective number of spectral channels that are available (e.g., Dorman *et al.*, 1997; Dorman and Loizou, 1998; Fishman *et al.*, 1997; Friesen *et al.*, 2001; Shannon *et al.*, 1995; Xu *et al.*, 2005). Although some modern cochlear implants have as many as 22 electrodes, the neural populations stimulated by different electrodes overlap to a considerable extent, so most implant users effectively have only around 4–7 independent channels (e.g., Friesen *et al.*, 2001). In contrast, normal-hearing individuals are able to utilize about 20 spectral channels. Given that cochlear implant users have poorer resolution for frequency differences, it may be advantageous for them to give more weight to vowel duration than would normal-hearing individuals; temporal resolution via cochlear implants can be as good as with normal hearing (e.g., Busby *et al.*, 1993; Shannon, 1989, 1992; see Shannon, 1993 for a review). However, vowel formant movement may be less informative; in addition to having reduced spectral resolution, some cochlear implant users appear to have difficulty perceiving changes in formant frequencies (e.g., Dorman and Loizou, 1997).

It is particularly plausible that cochlear implant users would learn to rely on different cues than do normal-hearing listeners, because individuals undergo a period of acclimatization after receiving their cochlear implant; vowel recognition accuracy increases by an average of ~35 percentage points over the first 9 months of implant use (e.g., Tyler *et al.*, 1997; Välimaa *et al.*, 2002). The acclimatization process is not well understood, but the improvements in speech perception probably arise from changes in linguistic categorization, not only from changes in lower-level psychophysical processing. For example, Svirsky *et al.* (2004) tracked best exemplar locations in an F1 × F2 vowel space for cochlear implant users following implantation; most individuals had vowels at anomalous locations immediately after implantation, and their best exemplar locations tended to move toward those of normal-hearing individuals as they used their cochlear implant over a 2-year period. This suggests that individuals “re-map” their vowel space in some way after implantation. However, even after this remapping has likely been completed, the position of individual vowels in the F1 × F2 space can differ from those of normal-hearing individuals, and their vowel categories can overlap significantly (e.g., Harnsberger *et al.*, 2001).

The present study investigated how formant movement and duration contribute to vowel identification accuracy (Experiment 1) and to the underlying representations of the vowel categories (Experiment 2). Experiment 1 was similar to previous studies that examined the effects of removing

formant movement and duration on vowel identification for normal-hearing listeners (e.g., Assmann and Katz, 2005; Hillenbrand *et al.*, 2000; Hillenbrand and Nearey, 1999); listeners were tested on natural vowels and on signal-processed versions in which the vowel formant movement was removed and duration was equated. Cochlear implant users were tested on these stimuli without additional processing. Normal-hearing listeners were tested on unprocessed versions and on stimuli that had been passed through two, four, and eight-channel noise vocoders simulating a CIS processing strategy (Shannon *et al.*, 1995). Experiment 2 used synthetic stimuli to find locations of best exemplars within a vowel quadrilateral. Previous work has conducted this kind of mapping in a two-dimensional space composed of F1 and F2 target frequencies (e.g., Harnsberger *et al.*, 2001; Johnson *et al.*, 1993). The present study used a multidimensional extension of this method (Iverson and Evans, 2003), to find best exemplars in a five-dimensional space comprising F1 and F2 frequencies at the beginning and end of the vowels, and duration.

## II. EXPERIMENT 1: IDENTIFICATION OF SIGNAL-PROCESSED VOWELS

### A. Method

#### 1. Subjects

The cochlear implant users were 11 postlingually deafened adults with an age range of 50–75 years. All were native speakers of British English. The subjects were not selected based on their implant or processor strategy; there were eight Nucleus, two Clarion, and 1 Med-El users. They were tested 0.6–7.3 years postimplantation.

The normal-hearing subjects were ten native speakers of Standard Southern British English, with an age range of 24–34 years. All reported having no known hearing or learning disabilities.

#### 2. Stimuli and apparatus

The stimuli were recorded from two speakers, male and female, who were native speakers of southern British English. They were recorded saying the carrier sentence *Say /hVd/ again* with 13 words: *heed* (/i:/), *hid* (/ɪ/), *hayed* (/eɪ/), *head* (/ɛ/), *had* (/æ/), *heard* (/ɜ:/), *hud* (/ʌ/), *hod* (/ɒ/), *hard* (/ɑ:/), *hoard* (/ɔ:/), *hood* (/ʊ/), *hoed* (/əʊ/), and *who'd* (/u:/). They were also recorded reading a short passage (Aesop's *The north wind and the sun*). The stimuli were recorded in an anechoic chamber, and downsampled for playback with 11 025 16 bit samples per second.

Three additional versions of the vowels were created that (1) removed all formant movement, (2) equated duration, and (3) removed formant movement and equated duration. The changes to the stimuli were made using Praat (Boersma and Weenink, 2002). Formant movement was removed using LPC analysis and resynthesis. Specifically, LPC analyzed the signal from the start of voicing after the /h/ to the start of the /d/ closure, the signal was inverse filtered to produce an LPC residual, a time slice of the LPC analysis was identified that represented the vowel's target formant frequencies (defined as the point where F1 reached a peak),

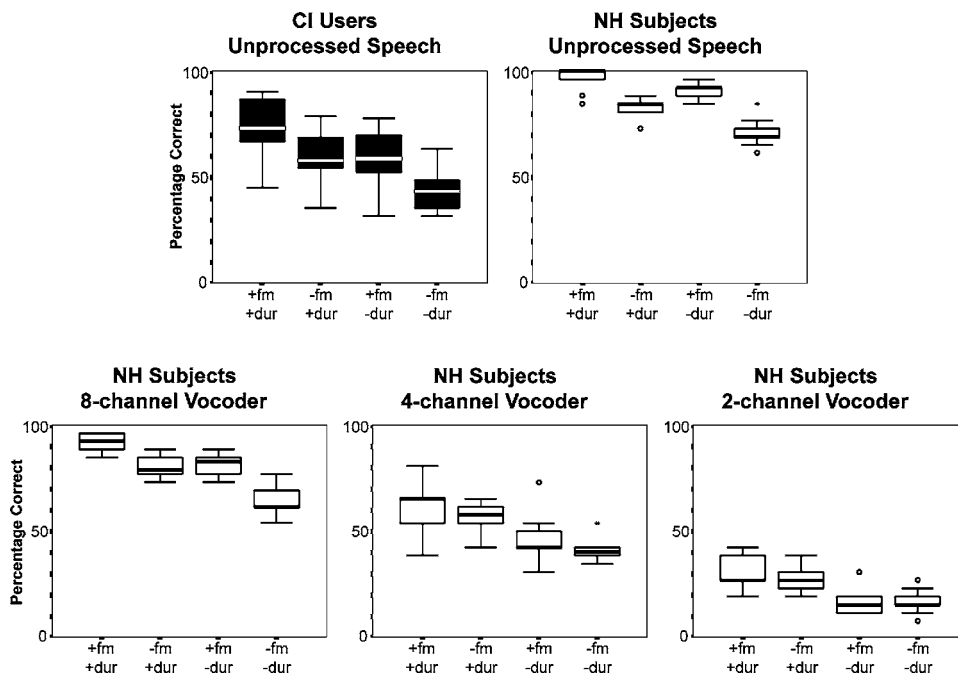


FIG. 1. Boxplots for vowel identification by cochlear implant (CI) users and normal-hearing (NH) subjects, for natural speech (+fm, +dur), vowels with no formant movement but natural duration contrast (–fm, +dur), vowels with natural formant movement and no duration contrast (+fm, –dur), and vowels with no formant movement and no duration contrast (–fm, –dur). The boxes and whiskers display quartile ranges. Circles and asterisks indicate outliers (i.e., more than 1.5 times the inter-quartile range away from the median).

and this single LPC slice was used to filter the entire LPC residual. This process created stimuli that retained the natural F0 of the original stimuli, but had formant frequencies that remained fixed at each vowel’s target values. Duration was equated using PSOLA (Pitch Synchronous Overlap and Add), such that the durations of the /h/, the /d/ closure, and the vowel were set to the mean values for each talker. This general approach to signal processing natural speech was similar to that used by Assmann and Katz (2005), although they used STRAIGHT (Kawahara, 1997; Kawahara *et al.*, 1999) rather than the procedures outlined above.

For presentation to normal-hearing subjects, these stimuli were processed by eight, four, and two-channel noise vocoders that were designed to simulate CIS processing (Shannon *et al.*, 1995). Using MATLAB, the stimuli were divided into spectral bands by sixth-order Butterworth filters, amplitude envelopes were calculated by half-wave rectification and low-pass filtering (fourth-order Butterworth, 400 Hz cut-off frequency), a noise carrier was modulated by each envelope, the modulated noise carriers were filtered by the original analysis bands, and the output was summed across bands. Each set of bands spanned a range from 200 to 5000 Hz, and this range was divided into bands based on equal basilar membrane distance (Greenwood, 1990). The filter slopes of the bands crossed at their –3 dB cutoff frequencies.

The stimuli were played to subjects at a comfortable loudness level (adjusted by each listener). Stimuli were delivered over headphones to normal-hearing individuals and in free field (a single speaker placed in a sound-attenuated booth) to cochlear implant patients.

### 3. Procedure

The speech from each talker was presented in separate blocks. To familiarize the subjects with the talker, subjects heard a short passage read by the talker at the start of the block, and they were simultaneously able to view the text of

the passage on a computer screen. In the noise-vocoded conditions, this passage was processed identically to the stimuli that followed. Afterwards, they were presented with *Say hVd again* sentences, and clicked on buttons on a computer screen to indicate which vowel they heard (cochlear implant patients who were unable to use the interface dictated their responses to an assistant). The buttons were labeled both with an hVd word (e.g., *hoed*) and with a familiar word that had the same vowel (e.g., *load*). Subjects did not receive feedback after their response. Before the experiment began, subjects were shown the response interface and completed a few practice trials, until they were satisfied that they understood the task. In each experimental block, subjects heard 52 sentences (13 words  $\times$  four conditions) presented in an order that was randomized for each subject.

Cochlear implant patients heard unprocessed vowels, and completed four blocks for each talker. Normal hearing subjects were presented with unprocessed versions as well as stimuli processed by eight, four, and two-channel vocoders; they completed two blocks (one for each talker) for each of these four conditions.

### B. Results

As displayed in the boxplots of Fig. 1, there were substantial effects of removing formant movement and duration for cochlear implant users. Compared to the natural versions, removing formant movement lowered recognition accuracy by an average of 13.3 percentage points, equating duration lowered accuracy by an average of 14.0 percentage points, and combining the two manipulations lowered accuracy by an average of 29.4 percentage points. These differences were evaluated by transforming the percentages into RAU (Rationalized Arc-sin Units; Studebaker, 1985) and conducting a MANOVA with two within-subject variables coding the differences between conditions (natural vs flat formant movement; natural vs equated duration). As suggested by the

boxplots, the effects of formant movement,  $F(1,10) = 64.7, p < 0.001$ , and duration,  $F(1,10) = 67.1, p < 0.001$ , were both significant; there was no significant interaction of formant movement and duration,  $p > 0.05$ .

For normal-hearing subjects, a similar MANOVA was run with the addition of a four-level within-subjects variable for the listening condition (i.e., unprocessed speech; eight, four, and two-band vocoders). There was a significant effect of duration,  $F(1,9) = 133.0, p < 0.001$ , but no interaction of duration with the other variables,  $p > 0.05$ ; removing duration cues reduced recognition scores by an average of 5.4–15.4 percentage points across the conditions, but there is no clear evidence that subjects put more weight on duration when there was less spectral resolution. There was a significant effect of formant movement,  $F(1,9) = 87.3, p < 0.001$ , and a significant interaction between formant movement and noise-vocoder condition,  $F(1,9) = 33.9, p < 0.001$ . The interaction occurred because the effect of removing formant movement depended on the amount of spectral resolution available; removing formant movement reduced recognition by an average of 13.5 percentage points for unprocessed speech, and 11.2 percentage points for the eight-channel noise vocoder, but had no significant effects in the four- and two-channel conditions. Finally, the main effect of noise-vocoder condition was also significant,  $F(1,9) = 421.1, p < 0.001$ ; vowel recognition became less accurate when the number of channels decreased.

Unsurprisingly, removing formant movement had the largest effect on the recognition of diphthongs (i.e., *hayed* and *hoed*). To test whether more subtle patterns of formant movement also had an effect on vowel recognition, the MANOVA analyses were repeated with *hayed* and *hoed* omitted. For cochlear implant patients, removing formant movement had a smaller effect when diphthongs were not included (reducing correct recognition by 4.4 percentage points), but the effect of formant movement remained significant,  $F(1,10) = 33.5, p < 0.001$ . For normal-hearing subjects, the effect of formant movement remained significant in the eight-channel condition (reducing correct recognition by 5.2 percentage points), but was eliminated in the unprocessed-speech condition when duration cues were also present. This led to a significant interaction between formant movement, vocoder condition, and duration,  $F(3,7) = 6.8, p = 0.018$ .

To further assess the role of formant movement and duration, the percent information transfer (Miller and Nicely, 1955) was calculated for unprocessed vowels, for the features of duration (short vs long) and formant movement (monophthong vs diphthong). For example, the vowels were all classified as long or short, and a  $2 \times 2$  confusion matrix was constructed to tally how often short vowels were identified as a short vowel (e.g., /i/ identified as /i/), short vowels were identified as a long vowel (e.g., /i/ identified as /i:/), long vowels were identified as a long vowel (e.g., /i/ identified as /eI/), and long vowels were identified as a short vowel (e.g., /i:/ identified as /i/). The information transfer statistic ranged from 100% if long vowels were never identified as short vowels (and vice versa) to 0% if the responses for long and short vowels were the same.

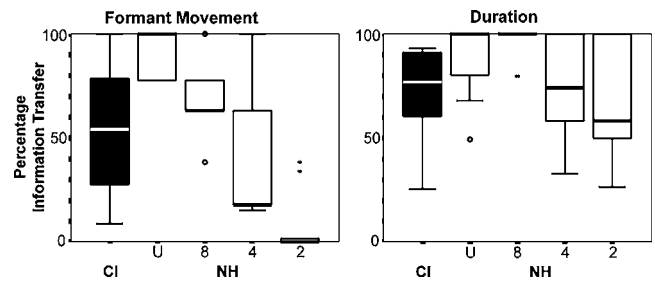


FIG. 2. Boxplots of percent information transfer for formant movement and duration in each of the conditions.

For formant movement, the percentage of information transfer for normal-hearing individuals declined as a function of the number of channels (see Fig. 2); listeners received almost no formant movement information when there were only two channels, but the majority of listeners received 100% of the formant movement information in unprocessed speech. Information transfer for cochlear implant users had a range similar to normal-hearing listeners in the four-channel condition, and a median level of information transfer similar to normal-hearing listeners in the eight-channel condition. This fits with previous findings that cochlear implant users are able to utilize 4–7 channels of spectral information (Friesen *et al.*, 2001). It thus appears that cochlear implant users use formant movement to about the same extent as normal-hearing individuals listening to cochlear implant simulations.

For duration, normal-hearing individuals received less information when the number of channels was low. This likely occurred because the duration feature covaries with spectral differences between vowels in English (e.g., *heed* and *hid* differ both in formant frequencies and duration), and spectral differences would have been clearer in the unprocessed and eight-channel conditions. The range of information transfer scores for cochlear implant users was similar to that of normal-hearing listeners in the four-channel condition, which suggests that both groups of listeners used duration to similar degrees. Given that the perception of duration should not be dependent on spectral resolution, it seems as if all listeners should have been able to achieve 100% information transfer for duration. The upper quartile of normal-hearing subjects achieved 100% information transfer even under the two-channel conditions, but most normal-hearing individuals and all cochlear implant users were below this optimum level. It thus seems as if the vowel recognition accuracy of cochlear implant users could be further improved if they learned to make more effective use of duration.

### III. EXPERIMENT 2: VOWEL-SPACE MAPPING WITH FORMANT MOVEMENT AND DURATION

Despite the fact that the cue use of long-term cochlear implant users has the potential to change over time, the results of Experiment 1 suggested that cochlear implant users use formant movement and duration cues to the same extent as do normal-hearing listeners. The present experiment assessed their cue weightings in more detail by examining what combinations of formant frequencies, formant movement, and duration produce best exemplars (i.e., prototypes)

of vowel categories. Although most phonetic categorization research has focused on how different acoustic cues alter the locations of category boundaries (e.g., Hoffman, 1958), examinations of best exemplars can also reveal the structure of phonetic categories and the relative use of different acoustic cues (e.g., Allen and Miller, 2001; Evans and Iverson, 2004; Iverson and Kuhl, 1996). In the present case, mapping boundaries is difficult because they exist as four-dimensional surfaces within the five-dimensional stimulus space; locating best exemplars is easier computationally because they can be represented as a single point within the space.

Previous work (Harnsberger *et al.*, 2001; Svirsky *et al.*, 2004) has mapped best exemplars of vowels using a graphical computer interface in which cochlear implant users interactively clicked on stimuli in a  $F1 \times F2$  grid, until they found vowels that they thought matched words printed on the computer screen (e.g., *heed*). Such an approach is not feasible for higher-dimensional stimulus sets, because the additional dimensions increase the number of possible stimuli (e.g., there were 100,700 stimuli in the present experiment). We have developed a goodness optimization method to search spaces like this more efficiently (Evans and Iverson, 2004; Iverson and Evans, 2003). On each trial, subjects see a word printed on the computer screen, hear a synthesized vowel, and rate how closely the word that they hear matches the printed target. After each rating, a computational algorithm analyzes the goodness ratings and changes the acoustic parameters on the next trial to iteratively converge on the best exemplar location. Using this technique, we are able to locate best exemplars within this large stimulus space after 35 trials per vowel.

## A. Method

### 1. Subjects

The subjects were the same as in Experiment 1.

### 2. Stimuli and apparatus

The stimuli consisted of hVd syllables embedded in recordings of the carrier sentence *Say\_again*. The carrier sentence was produced by the male speaker in Experiment 1. Initial and final words, plus the burst of the /d/, were edited from a natural recording. The hVd syllables were created using a cascade-parallel synthesizer (Klatt and Klatt, 1990) to match the vocal timbre, pitch, and higher formant frequencies of the talker. Each syllable had static formant frequencies during the /h/ that matched the onset of the vowel. The F1 and F2 formant frequencies changed linearly from the onset to the offset of the vowel, and F1 fell at the end of the vowel to simulate the /d/ closure. The durations of /h/ and the /d/ closure were fixed, and the duration of the vowel was allowed to vary from 148 to 403 ms. F1 frequency was restricted so that it had a lower limit of 5  $ERB_N$  (Glasberg and Moore, 1990) and an upper limit of 15  $ERB_N$ . F2 frequency was restricted so that it had a lower limit of 10  $ERB_N$ , was always at least 1  $ERB_N$  higher than F1, and had an upper limit defined by the equation  $F2 = 26 - (F1 - 5)/2$ . The stimuli were synthesized in advance with a 1  $ERB_N$  spacing of the vowel space, and with seven levels of log-spaced duration

values, for a total of 100 700 individual stimuli. The  $ERB_N$ , and log-duration transforms were chosen so that the stimuli would be spaced roughly equally with regard to perception. This spacing allowed us to efficiently distribute the stimuli, although the goodness optimization procedure does not require this equal perceptual spacing.

In addition to these unprocessed stimuli, a second set of stimuli was created by passing them through an eight-channel noise vocoder, following the procedures detailed in Experiment 1.

### 3. Procedure

On each trial, subjects heard one sentence and rated on a continuous scale whether the hVd was close to being a good exemplar of a word that was displayed on the computer screen. Their ratings were given by mouse clicking on a continuous bar presented on a computer screen. They gave ratings for 12 words: *heed*, *hid*, *hayed*, *head*, *had*, *hard*, *hod*, *hoard*, *heard*, *hoed*, *hood*, and *who'd*. To familiarize subjects with the speaker and task, they first heard the speaker read *The North Wind and The Sun* (as in Experiment 1), and gave a set of practice ratings for the word *hud*.

The goodness optimization procedure involved searching along individual vectors through the stimulus space (i.e., one-dimensional straight-line paths), and finding the best exemplar on each vector. There were a total of seven search vectors and five trials per vector for each vowel. The vectors were chosen so that Vector 1 would allow most subjects to find a close approximation of their best exemplar (the search path passed through formant frequencies measured from natural productions), Vectors 2–6 orthogonally varied the five acoustic dimensions over a wide range, and Vector 7 fine tuned the position of the best exemplar. Specifically, Vector 1 was a straight-line path that passed through two points: (1) the F1 and F2 formant frequencies at the beginning and ending of the natural productions of the target word, and (2) a neutral stimulus in the middle of the vowel space (F1 = 500 Hz and F2 = 1500 Hz, at both the onset and offset); duration was not varied along Vector 1. Vector 2 varied duration, keeping formant frequencies fixed. Vector 3 varied the onset F1 and F2 formant frequencies (i.e., duration and offset formant frequencies were fixed) along the same basic path as the first vector (i.e., through a straight-line path including a neutral vowel and the onset formant frequencies of the natural production). Vector 4 was orthogonal to Vector 3 in the F1/F2 onset space. Vectors 5 and 6 were analogous to Vectors 3 and 4, except that the offset F1 and F2 frequencies were varied. Vector 7 varied all dimensions, passing through the best values found thus far on all dimensions and the neutral vowel.

The end points of all vectors were constrained by the boundaries of the vowel space. For example, Vector 1 for *heed* crossed diagonally across the vowel space, starting from the high-front boundary of the space (i.e., low F1 and high F2), passing through the middle of the space, and ending at the low-back boundary of the space (i.e., high F1 and low F2).

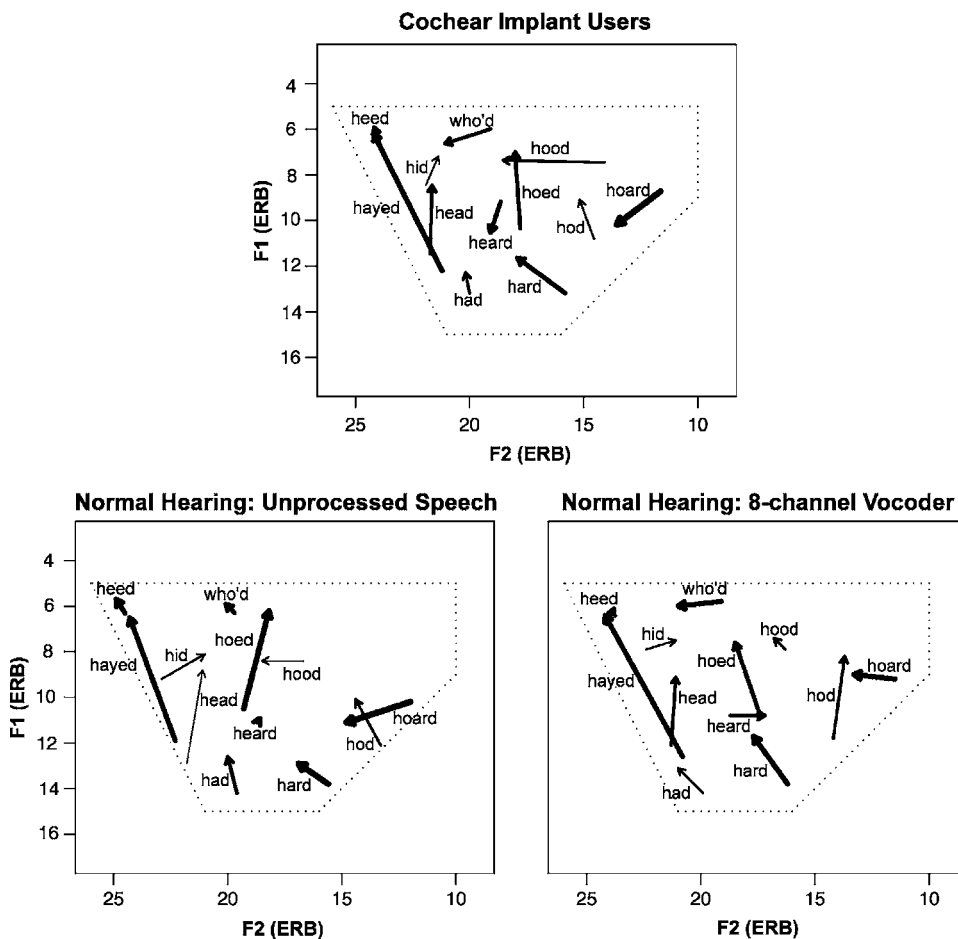


FIG. 3. Best exemplars of vowels with formant movement and duration, for cochlear implant users and normal-hearing individuals. Vowels are plotted as an arrow from the F1 and F2 frequencies at the start of the vowel to the F1 and F2 frequencies at the end. The thickness of the line indicates the preferred duration, with thicker lines for longer vowels.

The best exemplars were found for each vector over five trials. On the first two trials, subjects heard the most extreme stimuli that it was possible to synthesize along the vector (e.g., in the case of *heed*, they heard extreme high-front and low-back vowels, with the order of these two trials randomized). The selection of stimuli on the remaining trials was based on the subjects' judgments, using formulas that were designed to find stimuli along the path that would be perceived as better exemplars. On the third trial, subjects heard a stimulus that was selected by a weighted average of the first two stimuli, according to the equation

$$c = a * \frac{f(b)}{f(a)+f(b)} + b * \frac{f(a)}{f(a)+f(b)}, \quad (1)$$

where  $a$  and  $b$  are the positions on the search path for the first two trials,  $f(a)$  and  $f(b)$  are the goodness ratings for the stimuli on those trials (the goodness responses of close to far away were scaled from 0 to 1), and  $c$  is the new path position selected for the third trial. On the fourth and fifth trials, the stimuli were selected by finding the minimum of a parabola that was defined by the equation

$$\min = \frac{b - 0.5 * \{ [b-a]^2 * [f(b)-f(c)] - [b-c]^2 * [f(b)-f(a)] \}}{[b-a] * [f(b)-f(c)] - [b-c] * [f(b)-f(a)]}, \quad (2)$$

where  $b$  is the path position of the best stimulus found thus far;  $a$  and  $c$  are the most recently tested positions on either side of  $b$ ; and  $f(a)$ ,  $f(b)$ , and  $f(c)$  are the goodness ratings for those stimuli. At the completion of the fifth trial, subjects were allowed to repeat the search if it had produced a poor

exemplar. If the best exemplar was correct, the parameters of the best stimulus found thus far were passed onto the next stage of the search algorithm (i.e., to search along the next vector).

## B. Results

As displayed in Fig. 3, there were few overall differences between the vowel spaces of cochlear implant users and normal-hearing individuals. On average, both groups of listeners preferred vowels with similar formant frequencies, similar amounts of formant movement, and similar duration contrasts. The results of cochlear implant and normal-hearing individuals (listening to unprocessed speech) were compared using MANOVA analyses with subject type as a between-subject variable and word as a within-subject variable; the analyses were conducted separately for F1 and F2 target frequencies (average of the onset and offset values), F1 and F2 formant movement (offset minus onset), and duration. There was a significant effect of subject type for F1 target frequencies,  $F(1,16)=15.1, p=0.001$ ; the average F1 formant frequencies were slightly lower for cochlear implant users (9.2 ERB<sub>N</sub>) than for normal-hearing listeners (9.8 ERB<sub>N</sub>). However, there were no other significant effects of subject for any of the other acoustic dimensions,  $p > 0.05$ , demonstrating that there were few overall differences between the best exemplars of normal-hearing and cochlear implant-using individuals. Unsurprisingly, there were significant dif-

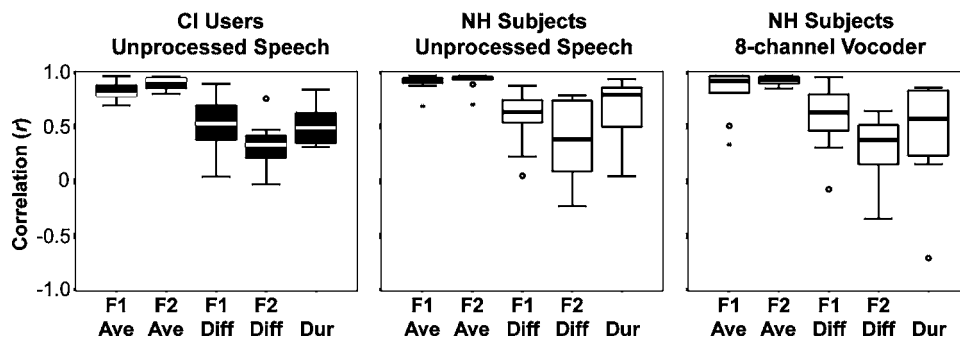


FIG. 4. Boxplots of correlations between the best exemplars of individual subjects and the average best exemplars for normal-hearing individuals listening to unprocessed speech. The pattern of results was similar for cochlear implant (CI) users, normal-hearing (NH) subjects listening to unprocessed speech, and normal-hearing subjects listening to eight-channel vocoders.

ferences between the words in terms of F1 target frequencies,  $F(11,6)=269.4, p < 0.001$ , F2 target frequencies,  $F(11,6)=42.6, p < 0.001$ , F1 formant movement,  $F(11,6)=40.5, p < 0.001$ , F2 formant movement,  $F(11,6)=4.7, p = 0.034$ , and duration,  $F(11,6)=10.8, p = 0.004$ , but there were no interactions between word and subject type,  $p > 0.05$ . The words thus differed significantly for both groups of listeners on all acoustic dimensions.

The best exemplars of normal-hearing individuals for unprocessed and noise-vocoded speech were compared with repeated-measure ANOVAS, with word and condition (unprocessed vs vocoded) coded as within-subject variables. The results were very similar to the comparison between cochlear implant users and normal-hearing individuals above. There was a significant effect of condition for F1 target frequencies,  $F(1,9)=12.2, p = 0.007$ ; the average F1 formant frequencies were slightly lower in the noise-vocoder condition (9.5 ERB<sub>N</sub>) than for unprocessed speech (9.8 ERB<sub>N</sub>). There were no other significant effects of condition for any of the other acoustic measurements,  $p > 0.05$ . There were significant differences between the words in terms of F1 target frequencies,  $F(4.4, 39.7)=69.4, p < 0.001$ , F2 target frequencies,  $F(4.3, 38.7)=111.9, p < 0.001$ , F1 formant movement,  $F(3.0, 27.2)=10.2, p < 0.001$ , F2 formant movement,  $F(4.3, 38.6)=4.4, p = 0.004$ , and duration,  $F(2.8, 24.8)=10.5, p < 0.001$ , but there were no interactions between word and condition,  $p > 0.05$ .

In order to examine individual differences, the average best exemplar for normal-hearing subjects listening to unprocessed speech was calculated for each vowel, and Pearson correlations were calculated between these averages and the data for individual subjects along each acoustic dimension (i.e., F1 and F2 target frequencies, F1 and F2 formant movement, and duration). This measure thus quantified how closely individual subjects approximated standard British English vowels on each of these dimensions. When this analysis was conducted for normal-hearing individuals listening to unprocessed speech, each individual was compared to an average that did not include themselves.

As displayed in Fig. 4, F1 and F2 target frequencies for individuals consistently approximated the normal averages, but there were poorer correlations and more variability for formant movement and duration. This pattern was generally the same for cochlear implant users and normal-hearing subjects. The individual variability in formant movement and duration likely reflects the fact that these are secondary cues (i.e., not as critical for identification as target F1 and F2

frequencies) rather than being indicative of perceptual difficulties. Independent-samples *t* tests were used to compare the correlations for normal-hearing individuals and cochlear implant users listening to unprocessed speech. The differences were significant for F1 target frequencies,  $t(18.9)=2.49, p = 0.022$ ; the preferred F1 frequencies of normal-hearing individuals (mean  $r=0.90$ ) matched the normal-hearing averages better than did those of cochlear implant users (mean  $r=0.82$ ). There were no significant differences between normal-hearing individuals and cochlear implant users for the other acoustic measures,  $p > 0.05$ . Paired *t* tests were used to compare the correlations for normal-hearing listeners for unprocessed and noise-vocoded speech; there were no significant differences,  $p > 0.05$ .

Pearson correlations were used to assess the relationship between these correlations and vowel identification accuracy for cochlear implant users in Experiment 1. For F1 target frequencies, there were significant correlations for unprocessed vowels,  $r=0.68, p = 0.021$ , vowels with formant movement removed,  $r=0.69, p = 0.020$ , vowels with duration equated,  $r=0.71, p = 0.014$ , and vowels with both formant movement removed and duration equated,  $r=0.72, p = 0.012$ . This demonstrates that subjects had higher vowel recognition scores when their best exemplars more closely matched the F1 values of normal-hearing individuals. There were no significant correlations for the other acoustic measures,  $p > 0.05$ .

#### IV. GENERAL DISCUSSION

The results demonstrate that formant movement and duration are important cues for vowel recognition via cochlear implants and noise-vocoder simulations. In Experiment 1, removing both formant movement and duration contrast reduced recognition accuracy for cochlear implant users by an average of 29.4 percentage points. The results were similar for normal-hearing individuals recognizing noise-vocoded vowels, although the effect of formant movement was diminished with reduced numbers of channels. In Experiment 2, both normal-hearing individuals and cochlear implant users preferred vowels with formant movement and duration contrast, although their preferences for these secondary cues were less consistent than their preferences for target F1 and F2 frequencies.

Despite the fact that cochlear implant users undergo a substantial period of relearning following implantation, there was no evidence that they used formant movement and du-

ration any differently than did normal-hearing individuals who were listening to cochlear implant simulations for the first time. The only significant difference was in the use of F1 target frequencies; the F1 frequencies chosen by cochlear implant users in Experiment 2 were less closely correlated with the normal-hearing averages, and individual differences in these correlations were related to vowel recognition accuracy. It is unknown what caused the variability in best F1 frequencies; listeners may have had anomalous F1 frequency targets in their underlying category representations (e.g., Svirsky *et al.*, 2004) or they may have had impaired spectral resolution in this frequency range.

It was particularly surprising that cochlear implant users and normal-hearing individuals had similar best exemplars in Experiment 2, considering that Harnsberger *et al.* (2001) found that cochlear implant users were often more variable in their best exemplar locations than were normal-hearing individuals, and the present cochlear implant users made many errors when recognizing natural vowels in Experiment 1 (averaging 74% correct). In contrast to Harnsberger *et al.*, which used isolated vowels, our method played vowels embedded in natural sentences, and subjects listened to a short story read by the talker before starting the experiment. Subjects were thus able to make their goodness judgments with reference to a particular talker, and this may have made their responses more reliable. Several subjects in our experiment reported that they had difficulty hearing any difference between the vowels in the practice of Experiment 2 (even stimuli at opposite ends of the vowel space), but they were encouraged to repeatedly play the stimuli and make responses based on any small differences that they could hear. The task demands were thus lower than in Experiment 1 (i.e., where individuals heard stimuli only once), which may have improved accuracy and consistency.

Why did normal-hearing individuals and cochlear implant users make such similar use of formant movement and duration? All of the cochlear implant users were postlingually deafened, so presumably they had had phoneme representations, at some point in their history, that matched those of normal-hearing individuals. It is possible that acclimatization to a cochlear implant involves a change in perceptual processing that does not alter these phoneme representations. For example, shallow insertions of an electrode array into the cochlea can cause a mismatch between the electrode sites and the frequency bands analyzed by the cochlear implant processor, such that the stimulation patterns that a cochlear implant user receives are shifted to higher frequencies relative to the stimulation patterns that they experienced when they had normal hearing; learning to adjust to these spectral shifts is thought to be a major component of acclimatizing to a cochlear implant (e.g., Fu *et al.*, 2005; Rosen *et al.*, 1999). It is plausible that such an adjustment involves a change in how the sensory information from the cochlear implant is mapped onto existing phoneme representations, without involving changes to the phoneme representations themselves. Training may be required (e.g., Fu *et al.*, 2005) in order for cochlear implant users to alter their category representations to place more weight on cues such as duration.

A complicating factor is that the same underlying category representations may be involved in production. That is, the best exemplars in Experiment 2 may reflect hyperarticulated target values that speakers aim to achieve when speaking clearly (Johnson *et al.*, 1993). If a cochlear implant user were to modify their use of secondary cues in order to identify vowels more accurately, then this would presumably have consequences for production. For example, an individual who learns to place greater weight on duration when distinguishing /ɪ/ and /i/ may begin to produce these vowels with more duration contrast and less spectral contrast, and thus produce speech that is less intelligible to normal-hearing speakers. Production considerations may thus tend to promote cochlear implant users to rely on the same acoustic cues as normal-hearing individuals.

## ACKNOWLEDGMENT

We thank Tim Green for comments on this manuscript.

- Allen, J. S., and Miller, J. L. (2001). "Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate." *Percept. Psychophys.* **63**, 798–810.
- Assmann, P. F., and Katz, W. F. (2005). "Synthesis fidelity and time-varying spectral change in vowels." *J. Acoust. Soc. Am.* **117**, 886–895.
- Boersma, P., and Weenink, D. (2002). Praat (Computer Software), Amsterdam, The Netherlands.
- Busby, P. A., Tong, Y. C., and Clark, G. M. (1993). "The perception of temporal modulations by cochlear implant patients." *J. Acoust. Soc. Am.* **94**, 124–131.
- Dorman, M. F., and Loizou, P. C. (1997). "Mechanisms of vowel recognition for Ineraid patients fit with continuous interleaved sampling processors." *J. Acoust. Soc. Am.* **102**, 581–587.
- Dorman, M. F., and Loizou, P. C. (1998). "The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine-channels." *Ear Hear.* **19**, 162–166.
- Dorman, M. F., Loizou, P. C., and 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.
- Evans, B. G., and Iverson, P. (2004). "Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences." *J. Acoust. Soc. Am.* **115**, 352–361.
- Ferguson, S. H., and Kewley-Port, D. (2002). "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners." *J. Acoust. Soc. Am.* **112**, 259–211.
- Fishman, K. E., Shannon, R. V., and Slattery, W. H. (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.
- Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (2001). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants." *J. Acoust. Soc. Am.* **110**, 1150–1163.
- Fu, Q. J., Nogaki, G., and Galvin, J. III (2005). "Auditory training with spectrally shifted speech: Implications for cochlear implant patient auditory rehabilitation." *J. Assoc. Res. Otolaryngol.* **6**, 180–189.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data." *Hear. Res.* **47**, 103–138.
- Greenwood, D. D. (1990). "A cochlear frequency-position function for several species—29 years later." *J. Acoust. Soc. Am.* **87**, 2592–2605.
- Harnsberger, J. D., Svirsky, M. A., Kaiser, A. R., Pisoni, D. B., Wright, R., and Meyer, T. A. (2001). "Perceptual "vowel spaces" of cochlear implant users: Implications for the study of auditory adaptation to spectral shift." *J. Acoust. Soc. Am.* **109**, 2135–2145.
- Hawkins, S., and Smith, R. (2001). "Polysp: a polysystemic, phonetically-rich approach to speech understanding." *J. Italian Linguistics – Rivista di Linguistica* **13**, 99–188.

- Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (2000). "Some effects of duration on vowel recognition," *J. Acoust. Soc. Am.* **108**, 3013–3022.
- Hillenbrand, J. M., and Nearey, T. M. (1999). "Identification of resynthesized /hVd/ utterances: Effects of formant contour," *J. Acoust. Soc. Am.* **105**, 3509–3523.
- Hoffman, H. S. (1958). "Study of some cues in the perception of the voiced stop consonants," *J. Acoust. Soc. Am.* **30**, 1035–1041.
- Iverson, P., and Evans, B. G. (2003). "A goodness optimization procedure for investigating phonetic categorization." *Proceedings of the International Congress of Phonetic Sciences*, Barcelona, August, 2003, pp. 2217–2220.
- Iverson, P., and Kuhl, P. K. (1996). "Influences of phonetic identification and category goodness on American listeners' perception of /t/ and /l/," *J. Acoust. Soc. Am.* **99**, 1130–1140.
- Johnson, K. (2005). "Speaker normalization in speech perception." *The Handbook of Speech Perception* (Blackwell, Oxford, England).
- Johnson, K., Flemming, E., and Wright, R. (1993). "The hyperspace effect: Phonetic targets are hyperarticulated," *Language* **69**, 505–528.
- Kawahara, H. (1997). "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," *Proceedings of the ICASSP*, Munich, Germany, April, 1997, pp. 1303–1306.
- Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Commun.* **27**, 187–207.
- Kirk, K. I., Tye-Murray, N., and Hurtig, R. R. (1992). "The use of static and dynamic vowel cues by multichannel cochlear implant users," *J. Acoust. Soc. Am.* **91**, 3487–3498.
- Klatt, D. H., and Klatt, L. C. (1990). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.* **87**, 820–857.
- Miller, G. A., and Nicely, P. (1955). "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.* **27**, 338–352.
- Neel, A. (1998). "Factors influencing vowel identification in elderly hearing-impaired listeners," Doctoral dissertation, Indiana University, 1989.
- Nygaard, L. C., and Pisoni, D. B. (1998). "Talker-specific learning in speech perception," *Percept. Psychophys.* **60**, 355–376.
- Rosen, S., Faulkner, A., and Wilkinson, L. (1999). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," *J. Acoust. Soc. Am.* **106**, 3629–3636.
- Shannon, R. V. (1989). "Detection of gaps in sinusoids and pulse trains by patients with cochlear implants," *J. Acoust. Soc. Am.* **85**, 2587–2592.
- Shannon, R. V. (1992). "Temporal modulation transfer functions in patients with cochlear implants," *J. Acoust. Soc. Am.* **91**, 2156–2164.
- Shannon, R. V. (1993). "Psychophysics," in *Cochlear Implants: Audiological Foundations*, edited by R. S. Tyler (Singular, San Diego), pp. 357–388.
- Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Strange, W. (1989). "Dynamic specification of coarticulated vowels spoken in sentence context," *J. Acoust. Soc. Am.* **85**, 2135–2153.
- Studebaker, G. A. (1985). "A 'rationalized' arcsine transform," *J. Speech Hear. Res.* **28**, 455–462.
- Svirsky, M. A., Silveira, A., Neuburger, H., Teoh, S. W., and Suarez, H. (2004). "Long-term auditory adaptation to a modified peripheral frequency map," *Acta Oto-Laryngol.* **124**, 381–386.
- Tyler, R. S., Parkinson, A. J., Woodworth, G. G., Lowder, M. W., and Gantz, B. J. (1997). "Performance over time of adult patients using the Ineraid or nucleus cochlear implant," *J. Acoust. Soc. Am.* **102**, 508–522.
- Välilä, T. T., Määttä, T. K., Löppönen, H. J., and Sorri, M. J. (2002). "Phoneme recognition and confusions with multichannel cochlear implants: Vowels," *J. Speech Lang. Hear. Res.* **45**, 1039–1054.
- Xu, L., Thompson, C. S., and Pfingst, B. E. (2005). "Relative contributions of spectral and temporal cues for phoneme recognition," *J. Acoust. Soc. Am.* **117**, 3255–3267.