A mathematical model of medial consonant identification by cochlear implant users

Mario A. Svirsky^{a)} and Elad Sagi

Department of Otolaryngology, New York University School of Medicine, New York, New York 10016

Ted A. Meyer

Department of Otolaryngology – Head and Neck Surgery, Medical University of South Carolina, Charleston, South Carolina 29425

Adam R. Kaiser and Su Wooi Teoh

DeVault Otologic Research Laboratory, Department of Otolaryngology, Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, Indiana 46202

(Received 22 December 2009; revised 8 November 2010; accepted 21 November 2010)

The multidimensional phoneme identification model is applied to consonant confusion matrices obtained from 28 postlingually deafened cochlear implant users. This model predicts consonant matrices based on these subjects' ability to discriminate a set of postulated spectral, temporal, and amplitude speech cues as presented to them by their device. The model produced confusion matrices that matched many aspects of individual subjects' consonant matrices, including information transfer for the voicing, manner, and place features, despite individual differences in age at implantation, implant experience, device and stimulation strategy used, as well as overall consonant identification level. The model was able to match the general pattern of errors between consonants, but not the full complexity of all consonant errors made by each individual. The present study represents an important first step in developing a model that can be used to test specific hypotheses about the mechanisms cochlear implant users employ to understand speech.

© 2011 Acoustical Society of America. [DOI: 10.1121/1.3531806]

PACS number(s): 43.71.An, 43.66.Ts, 43.71.Es, 43.71.Ky [MSS]

Pages: 2191–2200

I. INTRODUCTION

One important long term goal of our research is to help understand the mechanisms that underlie speech perception by cochlear implant (CI) users. We try to get closer to this goal in the present study by developing mathematical models of consonant identification that test hypotheses about the mechanism CI users employ to encode and process the input auditory signal to extract information relevant for identifying different speech sounds. These models are based on the CI users' discrimination along specified perceptual dimensions and aim to predict not only overall percent correct scores but also detailed patterns of consonant confusions in CI users spanning a wide range of perceptual performance, implant device, stimulation strategy, age at testing, and age at implantation.

As an acoustic event, most consonants can be defined as a controlled constriction (partial or complete) of the vocal tract during the process of speaking. The constriction typically occurs by way of the lips, the tongue blade, and/or the tongue body, and the airflow can be continuous or periodic. Although the acoustic signal emitting from the vocal tract is analog in nature, consonants are perceived as discrete linguistic segments. These segments can be defined, in part, as combinations of binary distinctive features that describe the behavior of the articulators in terms of the manner in which they produce the constriction, the place within the oral cavity where the constriction is produced, and whether the airflow is

continuous (i.e., voiceless), or periodic (i.e., voiced). Within the acoustic signal one can find quasi-discrete correlates, i.e. speech cues, to these linguistic features near the boundaries which mark the discontinuities in the short-time spectrum of the acoustic signal that result from the constriction, and subsequent release, of the vocal tract airway (Stevens, 2000).

The relationship between consonant identity and their underlying speech cues differ between listeners with normal hearing and users of CIs. For example, CI users are capable of distinguishing between voiced and voiceless consonants (Teoh *et al.*, 2003) even though present-day CIs are not designed to specifically deliver to the listener F0 information, which is a relatively strong acoustic cue for the voicing feature in listeners with normal hearing (Stevens, 2000). Furthermore, whether due to the degraded signal provided by the CI or the user's inability to discriminate or process all the information that is actually transmitted by the implant, the number of speech cues available to CI users is smaller than the number of speech cues available to listeners with normal hearing. This limitation allows for the possibility to predict consonant identification in CI users with a model that incorporates a smaller subset of speech cues than would be required to explain the same phenomenon in listeners with normal hearing.

In the present study we attempted to fit consonant confusion data by CI users based on three types of consonantal speech cues: (1) mean center frequencies of the first three formant energies F1, F2, and F3; (2) proportion of acoustic energy below 800 Hz; and (3) the duration of an intervening silent gap between the initial closure and subsequent release of the consonantal constriction. These cues are related to the

^{a)}Author to whom correspondence should be addressed. Electronic mail: mario.svirsky@nyumc.org

various distinctive features. For example, the center frequencies of formant energies reflect the vocal tract configuration during the consonant constriction and can serve as indicators for place of articulation. The proportion of acoustic energy below 800 Hz reveals whether a consonant includes frication, and/or whether it was voiced (voiced consonants have more low-frequency energy than their voiceless counterparts). The presence or absence of a brief silent gap indicates whether the consonantal constriction was complete or partial, and the different silent gap durations can be used to differentiate among the obstruent consonants. We examined these cues at the level of the electrical stimulation pattern output of the CI, as opposed to the acoustic waveform received by the CI. This way, we are examining the speech information transmitted by the CI after it has been processed by the speech processor, which is different from the information that is available in the acoustic waveform.

The model employed in the present study is an application of the multidimensional phoneme identification (MPI) model (Svirsky, 2000, 2002; Svirsky *et al.*, 2001; Sagi *et al.*, 2010a, 2010b) which is a general framework that predicts phoneme identification based on listeners' ability to discriminate speech cues. These speech cues are combined to form a multidimensional perceptual space so that a phoneme's location within the space is defined by that phoneme's set of speech cue values. Two important potential causes of confusions among phonemes are the proximity of the target phoneme to other phonemes within the space, and the listener's just-noticeable-difference (JND) for each acoustic cue that defines phoneme location. That is, the larger the JND, the greater the likelihood of confusing the target phoneme with phonemes that are further away within the space. In the present study, this model is applied to consonant identification data obtained from a group of CI subjects. The JNDs that serve as subject-specific inputs to the model were allowed to vary to determine the model's best fit to the observed data, and to assess the extent to which the hypothesized perceptual dimensions can explain consonant identification in this population.

II. METHODS

A. CI subjects

Twenty-eight postlingually deafened adult CI users participated in the present study. Subjects provided informed consent and were compensated for their time. Table I lists the demographics for these subjects including age at implantation, experience with their device at time of testing, type of implant device and stimulation strategy, and number of stimulation channels. Subjects' age at implantation ranged from 16 to 76 yr with a mean age at implantation of 50 yr. With the exception of subjects N14 and N19, CI users had at least 1 yr of experience with their device before participating in this study. Six subjects (C1 through C6) were implanted with the Clarion 1.2 device (Advanced Bionics Corp, Valencia, CA) and used the continuous interleaved sampling (CIS) stimulation strategy. The remaining subjects (N1 through N22) were implanted with either the Nucleus-22 or Nucleus-24 devices (Cochlear Corp, NSW, Australia) and used either the spectral-peak (SPEAK) or the advanced combination encoder (ACE) processing strategies.

TABLE I. Demographics of CI subjects tested for this study, six users of the Advanced Bionics device (C) and 22 users of the Nucleus device (N). Age at implantation and experience with implant before testing on 24-consonant identification task are stated in years. Speech processing strategies are CIS, SPEAK, and ACE.

Subject	Implanted age (yr)	Experience (yr)	Implanted device	Proc. strategy	No. channels
C1	67	2.0	Clarion 1.2	CIS	8
C2	32	4.1	Clarion 1.2	CIS	8
C3	23	5.0	Clarion 1.2	CIS	8
C4	53	5.6	Clarion 1.2	CIS	5
C5	39	3.5	Clarion 1.2	CIS	6
C6	43	4.2	Clarion 1.2	CIS	8
N1	31	5.0	NucCI22M	SPEAK	18
N2	71	3.1	NucCI22M	SPEAK	16
N3	60	11.5	NucCI22M	SPEAK	7
N4	71	4.9	NucCI22M	SPEAK	14
N5	67	5.4	NucCI22M	SPEAK	19
N6	45	4.1	NucCI22M	SPEAK	20
N7	48	8.3	NucCI22M	SPEAK	16
N8	16	3.5	NucCI22M	SPEAK	18
N9	66	4.1	NucCI22M	SPEAK	18
N10	49	1.0	NucCI24M	ACE	20
N11	42	2.1	NucCI24M	SPEAK	16
N12	45	2.1	NucCI24M	SPEAK	20
N13	59	1.3	NucCI24M	SPEAK	6
N14	76	0.5	NucCI24M	ACE	19
N15	65	2.2	NucCI24M	SPEAK	20
N16	54	1.0	NucCI24M	SPEAK	20
N17	48	1.0	NucCI24M	SPEAK	20
N18	46	2.5	NucCI24M	SPEAK	20
N19	25	0.9	NucCI24M	ACE	22
N20	46	1.0	NucCI24M	ACE	14
N21	37	1.0	NucCI24M	SPEAK	20
N22	67	2.0	NucCI24M	SPEAK	20

B. Consonant identification

CI subjects were given a closed set consonant identification task comprising 24 consonants in an intervocalic /a/-consonant-/a/ context (e.g., the consonant /k/ would be presented as /aka/). The consonants tested were /b, d, tʃ, f, g, h, dʒ, k, l, m, n, ŋ, p, r, s, ʃ, t, ð, θ, v, j, w, z, and ʒ/. Ten of the 28 subjects were also tested using 16-consonant confusion matrices but results of the mathematical modeling conducted with the 16-consonant and the 24-consonant data sets turned out to be very similar, so in the interest of brevity the present report focuses on the 24-consonant data only. Stimuli were obtained from the Iowa Audiovisual Speech Perception Laser Video Disc (Tyler *et al.*, 1987) and consisted of three utterances (i.e., tokens) of each consonant recorded from one female speaker. Stimuli were presented in an auditory only, quiet condition at 70 dB C-weighted sound pressure level to subjects seated approximately 1 m from an Acoustic Research loudspeaker. Before the experiment, subjects were familiarized with the stimuli. During the experiment, subjects were presented with a stimulus and required to select one of the possible consonants as a response. If they were unsure, subjects were prompted to guess, and stimuli were repeated if necessary (for example, in the rare cases when a

listener was distracted during a presentation). Feedback was provided. The list of consonant tokens was presented three times to each listener (i.e., 3 presentations $\times$ 3 tokens per consonant = 9 responses per consonant), each time with a different randomization of presentation order. CI subjects' responses were compiled into confusion matrices. These matrices were the data to be fit by the MPI model, i.e., these matrices were used to determine the model output that best fit the subjects' consonant data.

C. MPI model

Three steps were employed to implement the MPI model. First, perceptual dimensions and their associated speech cues were postulated, providing a framework for the model's mathematical formulation. Second, measurement of speech cues were obtained for all consonant tokens from the electrical stimulation pattern produced by a given CI device's speech processor. In effect, these measurements are the electrical counterparts of their corresponding acoustical speech cues (i.e., "electrical speech cues") which provide information about consonant features. Third, Monte Carlo simulation was used to determine the model's input JNDs that produced output confusion matrices that best fit each CI user's consonant matrix. This implementation closely follows [Sagi et al. \(2010b\)](#) wherein a MPI model of vowel identification by CI users was presented.

1. Step 1: Model development

Five perceptual dimensions were postulated corresponding to the following speech cues (described in more detail in step 2): locations of first, second, and third mean formant energies along the implanted electrode array (F1, F2, and F3); proportion of charge above threshold in stimulating electrodes that encode frequencies below 800 Hz (A); and duration of the silent gap interval between the consonantal constriction and release (G). Eight combinations of perceptual dimensions were assessed. These were F1F2F3AG, F1F2F3A, F1F2F3G, F2F3AG, F1F3AG, F1F2AG, F1F2F3, and AG. For simplicity, it was assumed that perceptual dimensions are orthogonal and that distances in the multidimensional perceptual space are Euclidean.

The MPI model consists of two components, the internal noise model and the decision model. In the internal noise model, because of imperfect representations of a given stimulus due to sensory and memory limitations ([Durlach and Braida, 1969](#)), successive presentations of the same consonant token will result in slightly different percepts. These percepts are modeled as slightly different locations that vary about the location of the consonant token in the multidimensional perceptual space. The distribution of these percept locations relative to the consonant token location is modeled with a multidimensional Gaussian distribution [cf. Eq. (1) in [Sagi et al., \(2010b\)](#), for a three-dimensional example]. Along a given dimension, the mean of this distribution is equal to the value of the speech cue that defines the consonant token's location, and the standard deviation is equal to the observer's JND for that speech cue.

In the decision model (the second component of the MPI model), the percepts generated by the internal noise

model are categorized by selecting the "response center" that is closest to the percept in the perceptual space. The response center represents the listener's internal template or exemplar of how a given phoneme should sound and is interpreted as the average location of where the listener expects a given phoneme to lie in the perceptual space. With an ideal experienced listener, it is reasonable to hypothesize that response centers match the average locations of the stimuli in the perceptual space. In other words, the listener's expectations match the actual physical characteristics of the stimuli. If response centers differ from the physical characteristics of the stimuli, then the listener's responses are said to be biased. In the present study, subjects are treated as ideal experienced listeners such that their response centers are equal to the average locations of consonants in the perceptual space, as measured over the tokens used for each consonant. Having defined the locations of response centers, one can generate a percept and then classify this percept by selecting the consonant category associated with the closest response center. That is, for each percept we calculate the Euclidean distance between that percept and each response center (normalized along each dimension in terms of the number of JNDs) and select the closest response center as the listener's response [cf. Eq. (2) and associated text in [Sagi et al. \(2010b\)](#), for a three-dimensional example].

In summary, the MPI model generates a noisy percept from the coordinates that define a given consonant stimulus (i.e., internal noise model) and then selects the response center closest to the percept (i.e., decision model). Using a Monte Carlo algorithm, this model of the listener's response is repeated hundreds or thousands of times for each consonant token, and the output can be tabulated in a model confusion matrix. Consonant confusions in this predicted matrix are caused by both the relative overlap of consonant tokens in the perceptual space as well as the listener's JND across each perceptual dimension.

2. Step 2: Measurement of electrical speech cues

Electrical speech cue measurements were obtained from the output of two CI speech processors; the Nucleus-24 Sprint body-worn processor (Cochlear Corp.) and the Clarion 1.2 device (Advanced Bionics Corp.). Details of the hardware and software setup used to obtain these outputs can be found in [Sagi et al. \(2010b\)](#). From the output recorded from both devices, information about which electrode was stimulated at a given point in time, the charge amplitude (i.e., charge per phase), and pulse duration were compiled and stored for offline analysis.

Electrical speech cue measurements were obtained from the consonantal portion of each stimulus, which was determined by visual inspection of the temporal envelope of stimulation pulses summed across electrodes, from the point at which this envelope decreased from the preceding vowel, up until this envelope increased into the following vowel. That is, the consonant duration included the transitions from the preceding vowel and into the following vowel.

Locations of mean formant energies along the implanted array were calculated based on histograms of the number of

FIG. 1. Electrodegrams of the consonants in /ama/ (top) and /adʒa/ (i.e., j as in “jaw”) (bottom) obtained with the Nucleus device. Higher electrode numbers refer to more apical or low-frequency encoding electrodes. Charge magnitude is depicted as a gray-scale from 0% (light) to 100% (dark) of dynamic range. Rectangle represents consonantal portion used to compile histograms on the right of electrodegrams, which represent a weighted count of the number of times each electrode was stimulated. F1, F2, and F3 are locations of mean formant energies. Right-most vertical bar indicates proportion of charge above threshold in electrodes encoding frequencies below 800 Hz. Silent gap duration indicated by vertical dashed lines in electrodegram.

times each electrode was stimulated over the duration of the consonantal utterance, weighted by the amount of electrical charge above threshold for each current pulse (measured in terms of the percentage of the dynamic range). For example, in Fig. 1 electrodegrams of the consonants /ama/ and /adʒa/ (i.e., j as in “jaw”) as processed by the Nucleus-24 processor are shown. The rectangle that encloses portions of each electrodegram represents the consonant duration. To the right of each electrodegram is a histogram (horizontal bar graph) of the weighted number of times each electrode was stimulated over this duration. Each histogram was subdivided into regions of formant energies (one for each formant) and the mean location for each formant was calculated from its respective portion of the histogram. Because the consonant stimuli used in the present study were utterances from a female speaker, the boundaries that subdivided the histogram into regions of formant energies were derived from the formant center-frequency data of Peterson and Barney (1952) for female speakers. Based on these values, the F1–F2 frequency boundary was 1000 Hz, and the F2–F3 frequency boundary was 2250 Hz. The histogram bin belonging to the electrode that encoded a given frequency boundary was subdivided using linear interpolation. For example, in Fig. 1, the frequency band assigned to electrode 16 is 950–1150 Hz. When treated linearly, the 1000 Hz frequency boundary divides this band so that 25% of the energy in the histogram bin belonging to electrode 16 was assigned to F1, and 75% to F2. Although the implanted electrodes in a CI are located at discrete positions along the cochlea, formant energy can be distributed over more than one electrode so that the location of mean formant energy can vary continuously along the length of the electrode

array. In Fig. 1, estimates of mean formant energy locations are indicated by arrows. Using the known inter-electrode distance for each device (0.75 mm for the Nucleus-24 and 2 mm for the Clarion 1.2), locations of mean formant energy were expressed in units of millimeters along the length of the implanted array. It is important to note that mean formant location measurements are sensitive to the CI device’s frequency table settings. The setup described above (which used standard frequency tables) could be used to obtain measurements for many of the subjects. For the remaining subjects who had slightly different frequency tables, linear interpolation was used to transform the formant location measurements programmed with standard frequency table settings into a set of measurements specific to each CI subject’s frequency table (cf. Sagi et al., 2010b, for more details).

The proportion of charge in electrodes encoding frequencies below 800 Hz was determined over the duration of the consonantal utterance (enclosed in rectangles in Fig. 1) for each consonant token as follows. The amount of charge in each pulse was expressed in terms of charge above threshold (i.e., charge at threshold was subtracted from the charge value recorded for each pulse). The amount of charge above threshold in stimulation pulses that occurred only in electrodes encoding frequencies below 800 Hz was summed and divided by the total amount of charge above threshold summed over all electrodes. When evaluating the numerator, charge values of stimulation pulses that came from the electrode that included the 800 Hz boundary were multiplied by a fraction, depending on where this boundary occurred within the frequency band assigned to that electrode using linear interpolation. In Fig. 1, this calculation is depicted by the vertical bar

graphs to the right of the horizontal histograms. The horizontal histogram represents the total charge (above threshold) delivered to each electrode over the consonantal portion of the electrodiagram to the left (enclosed in rectangles), and the vertical bar to the right depicts the proportion of charge that falls below the horizontal line indicating the 800 Hz boundary in the horizontal histograms. In the case of /ama/, nearly half of the charge is delivered to the proportion of the electrodes below the 800 Hz boundary. In the case of /ad3a/, the proportion of charge delivered to electrodes encoding frequencies below 800 Hz is less than 20%.

The duration of silent gap interval was obtained by examining the record of stimulation pulses for brief pauses in stimulation that occurred in all but the most apical channel. This channel (the one encoding the lowest frequency band) was ignored because there may be some energy in this low-frequency range even during complete vocal tract closure, e.g., during voiced obstruents. The silent gap duration was measured from the time stamp of the last stimulation pulse (plus its pulse duration) that occurred before the brief pause in stimulation until the time stamp of the first stimulation pulse that occurred after this pause. In Fig. 1, this measurement is depicted within the rectangle that isolates the consonantal portion of /ad3a/. The duration for which stimulation was absent for all but the lowest frequency encoding electrode (indicated by the vertical dashed lines) is 39 ms. In contrast, no silent gap occurred in /ama/.

3. Step 3: Determination of input JNDs

Predicted matrices were generated for the eight perceptual dimension combinations outlined in step 1, based on the electrical speech cue measurements described in step 2 and a set of input JNDs. Predicted matrices were generated for different combinations of JND parameter values to determine how well the model could fit each CI subject's consonant confusion matrix. Two measures, described below, were used to compare predicted and CI subject matrices in order to determine an optimal set of JNDs that produced best-fit predicted matrices. For the perceptual dimension combinations involving mean location of formant energy along the implanted array (F1, F2, and F3), JND was varied with one degree of freedom regardless of the number of formant dimensions included in the combination. That is, as a simplifying assumption, JND (in units of millimeter along the electrode array) was assumed equal for each formant dimension. Note that this assumption is a reasonable first order approximation because of the units that were employed (distance along the cochlea). Additional degrees of freedom were used to vary JND for the other two perceptual dimensions, i.e., proportion of charge in electrodes encoding frequencies below 800 Hz (A) and duration of silent gap (G). Hence, between one and three degrees of freedom were used to vary JND across all perceptual dimension combinations.

For the F1, F2, and F3 dimensions, JND was varied from 0.03 to 4.00 mm in steps of 0.01 mm. The lower bound of 0.03 mm represents a reasonable estimate of the lowest JND for place of stimulation in the cochlea achievable with present-day CI devices (Kwon and van den Honert, 2006; Firszt *et al.*,

2007). For the A dimension, JND was varied from 0.05 to 0.5 in steps of 0.05. For the G dimension, JND was varied from 1 to 30 ms for consonant tokens with non-zero silent gaps. For consonant tokens that contained no silent gap (i.e., 0 ms) a fixed JND of 5 ms was assumed, reflecting the perfect or near perfect identification of gapped vs non-gapped stimuli by CI users observed in Sagi *et al.* (2009). The selection of JND = 5 ms used in the present study for the non-gapped consonants was derived from the observation that on the acoustic cue task in Sagi *et al.* (2009), all subjects scored a d' of 3 or greater between the 0 and 15 ms silent gap stimuli. Hence, in this case, defining the JND as occurring at $d' = 1$ gives JND = 5 ms for the non-gapped stimuli. In contrast, all stimuli with silent gaps in the present study had silent gap durations larger than 30 ms, and so the JND for these stimuli was varied to allow for different levels of discrimination across subjects (in more technical terms, the cumulative- d' for gap duration was treated as bilinear with a fixed slope for gap durations less than or equal to 15 ms, and a variable slope for durations greater than 15 ms).

With JNDs specified, multiple iterations of the MPI simulation algorithm would read in the set of electrical cue measurements and populate a consonant confusion matrix. Each predicted matrix consisted of 1500 entries per consonant. This number of iterations allowed us to complete the simulations within a reasonable amount of time, while at the same time ensuring that the cell entries in the resulting model matrix would not differ by more than 1% (when matrix rows are expressed as percentages) in comparison to a model matrix produced with 150 000 entries per consonant.

When observed and predicted matrices were compared, matrix elements were represented in percentages so that each row summed to 100%. Two measures were used to determine which JND values resulted in the best prediction of a listener's observed consonant confusion matrix. One measure was the root-mean-square (rms) difference between the observed and predicted matrices, i.e., the square root of the sum-square difference between corresponding elements of each matrix divided by the number of matrix elements. With this measure, the optimal JNDs that produced the best-fit model matrix were those that minimized the rms difference between observed and predicted matrices. In addition to using rms, goodness-of-fit between best-fit predicted and observed matrices was assessed by squaring the element-wise correlation (i.e., the coefficient of determination, R^2) between the two matrices. In order to compare modeling results across perceptual dimensions, a repeated-measures analysis of variance (RMANOVA) was applied to rms values.

The other measure assessed how the MPI model predicted the pattern of confusions among pairs of consonants. Matrices were analyzed to determine which pairs of consonants were confused above a certain threshold percentage of the time. An example of this process for a threshold of 10% is presented in Table II, where a comparison is made between an observed 16-consonant matrix obtained from one subject and a predicted 16-consonant matrix obtained with our model. No distinction was made as to the direction of the confusion within a pair, e.g., the percentage of "aba"/"apa" confusions was averaged with the percentage of "apa"/"aba" confusions, and the average was compared with the predefined threshold percentage.

TABLE II. An example of matching of error patterns between observed (O) and predicted (P) 16-consonant matrices. Percentage of consonant-pair confusions above 10% in either O or P presented in top and bottom panels. Resulting 2×2 comparison matrix (inset bottom panel) counts number of true positives (bold), false positives (italics), false negatives (regular text), and true negatives (omitted from top and bottom panels) between O and P at 10% threshold.

Observed (O)															
p															
d															
t	14														
g		<i>4</i>													
k	33		20												
v	<i>30</i>														
f	<i>14</i>														
z				<i>24</i>		10									
s						40									
ð					24	<i>14</i>	10								
ʃ															
m							<i>0</i>								
n															
l					<i>0</i>	<i>0</i>	<i>0</i>	17	34						
i									<i>0</i>						
	b	p	d	t	q	k	v	f	z	s	ð	ʃ	m	n	l

Predicated (P)															
p															
d															
t	19														
g		<i>23</i>													
k	15		22												
v	<i>8</i>														
f	<i>5</i>														
z				<i>5</i>		11									
s						22									
ð					12	<i>1</i>	12								
ʃ															
m							<i>10</i>								
n															
l					<i>10</i>	<i>11</i>	<i>14</i>	11	11						
i									<i>26</i>						
	b	p	d	t	q	k	v	f	z	s	ð	ʃ	m	n	l

2×2 Comparison matrix			
		$O \geq 10\%$	$O < 10\%$
$P \geq 10\%$	9	6	
$P < 10\%$	<i>4</i>	<i>101</i>	

From the percentage of consonant-pair confusions in both the observed and predicted matrices, a 2×2 comparison matrix was constructed (e.g., Table II, inset bottom panel) with the four cells indicating: (1) the number of consonant pairs that the model correctly predicted would be confused more often than the threshold percentage (true positives); (2) the number of consonant pairs that the model incorrectly predicted would be confused more frequently than the threshold percentage (false positives); (3) the number of consonant pairs that were confused more often than the threshold percentage and that the model failed to predict (false negatives); and (4) the number of consonant pairs that the model correctly predicted would not be confused more often than the threshold percentage (true negatives). Separate analyses were conducted for thresholds of 3%, 5%, and 10%. Different thresholds were necessary to

assess errors made by subjects with very different performance levels. For example, some subjects had good performance and confused very few consonant pairs more than 10% of the time so a lower threshold was required.

Comparison matrices were obtained for all subjects' observed consonant matrices and for MPI predicted matrices at all combinations of perceptual dimensions and for all JNDs tested. The following criteria were applied sequentially to determine the set of JNDs that produced the best-fit model matrix: (1) 2×2 comparison matrices were selected so that the maximum value between the number of false negatives and false positives was minimized; (2) if more than one matrix fulfilled the first criteria, then the matrix for which the metric $2 \times (\text{true positives}) - (\text{false negatives}) - (\text{false positives})$ was maximized was selected; and (3) if more than one matrix fulfilled the second criteria, then the matrix that minimized rms was selected. The predicted matrix in Table II is an example of a best-fit model matrix obtained with the above criteria.

Additionally, a best-fit 2×2 comparison matrix was labeled "satisfactory" if both true positives and true negatives were greater than or equal to false positives and false negatives. The satisfactory label means that several things happened simultaneously. Namely, that the MPI model was able to predict at least half of the consonant pairs confused by a CI subject, and to do so with a limited number of false positives (more specifically, at least half of the consonant pairs that the model predicted would be confused were indeed confused). It also means that the MPI model was able to predict at least half of the consonant pairs that would not be confused by the subject, while making a limited number of false negatives (at least half of the consonant pairs that the model predicted would not be confused were indeed not confused). In Table II, the 2×2 comparison matrix qualifies as satisfactory.

In summary, two methods were used to optimize the model's fit to the data. The first method minimized the rms difference between observed and predicted matrices and the second method optimized the pattern of consonant pairs that were or were not confused. Predicted matrices obtained with the second method were further analyzed to see how many would meet our definition of a satisfactory criterion.

D. Information transfer analysis

To assess the MPI model's ability to predict subjects' consonant feature identification, information transfer (IT) analysis (Miller and Nicely, 1955) was used to compare subjects' 24-consonant observed and best-fit predicted matrices (for the F1F2F3AG combination at lowest rms) in terms of the features voicing, manner, and place.

III. RESULTS

A. Fit between observed and MPI predicted matrices

In terms of the minimized rms, the fit of the MPI model to CI subjects' confusion matrices is summarized in Table III. Percent correct scores varied from 5% to 78% on the 24-consonant task with a mean of 45%. These scores are, on average, lower than consonant scores reported for other CI users of the devices listed in Table I, but the variability of

TABLE III. Minimum rms difference between CI users' observed and predicted 24-consonant confusion matrices. Only lowest rms values across perceptual dimensions (in bold) and values within 1% of this minimum were reported. Also reported are observed consonant percent correct (c24%), rms difference between observed matrices and a purely random matrix (Rand.), mean rms and average goodness-of-fit (R^2) across subjects.

CI user	c24%	rms: 24-Consonant								
		F1F2 F3AG	F1F2 AG	F1F3 AG	F2F3 AG	F1F2 F3G	F1F2 F3A	F1F2 F3	AG	Rand.
C1	57.7	10.3	9.7	—	10.2	—	—	—	—	15.2
C2	77.6	9.2	8.8	—	9.4	9.3	—	—	—	18.1
C3	47.9	11.6	11.7	11.5	11.6	11.9	—	—	12.4	15.2
C4	8.5	8.5	8.7	8.4	8.7	8.7	—	—	—	7.8
C5	40.0	10.0	9.9	10.5	10.1	10.1	10.6	10.7	—	12.8
C6	30.6	10.8	10.7	11.1	10.9	10.8	11.4	11.5	—	12.1
N1	56.5	10.8	11.0	10.7	11.0	10.9	11.7	—	—	15.5
N2	38.6	13.4	13.3	13.8	13.4	13.7	14.3	14.1	—	16.3
N3	4.9	9.6	9.6	9.4	—	9.6	9.4	—	—	8.7
N4	30.6	11.7	11.6	11.9	11.8	11.7	11.9	11.8	—	12.9
N5	41.9	12.2	12.0	12.3	12.5	12.3	12.5	12.4	—	14.6
N6	67.6	10.3	10.8	10.1	10.4	10.4	—	—	—	17.0
N7	43.1	10.6	10.0	10.6	10.5	—	—	—	—	13.0
N8	23.7	8.6	8.5	9.0	8.7	9.0	9.3	9.2	9.4	10.1
N9	12.5	8.8	8.8	8.9	8.9	9.2	—	—	8.9	8.4
N10	77.3	7.8	—	8.0	7.8	7.9	—	—	—	17.0
N11	62.8	11.4	12.1	11.5	11.3	11.9	—	—	—	16.6
N12	66.2	10.1	10.3	10.5	10.2	10.4	—	—	—	16.7
N13	20.0	9.7	9.6	9.7	9.7	10.3	—	—	10.1	10.5
N14	24.5	7.0	7.0	7.2	7.2	7.1	7.9	8.0	7.9	8.7
N15	58.2	11.2	11.4	11.3	11.6	11.2	—	—	—	16.3
N16	60.3	8.6	8.9	8.5	8.7	8.8	—	—	—	15.0
N17	69.0	9.9	10.4	10.1	10.0	9.9	—	—	—	16.9
N18	47.6	11.1	11.2	10.9	11.3	11.2	—	—	—	14.3
N19	67.6	9.7	10.5	9.8	9.7	9.8	—	—	—	16.6
N20	52.1	10.1	10.3	9.8	10.1	10.4	—	—	—	15.0
N21	48.6	11.1	11.4	11.3	11.3	11.2	—	—	—	15.6
N22	27.4	11.2	11.7	11.3	11.5	11.2	11.6	12.1	—	12.7
mean	45.1	10.2	10.3	10.4	10.3	10.4	11.2	11.7	12.3	13.9
	R^2	0.42	0.41	0.40	0.41	0.41	0.33	0.29	0.23	0

the scores is consistent for this population of CI users (Fu, 2002; Donaldson and Kreft, 2006). In addition to the eight perceptual dimension combinations tested in the present study (columns 3 through 10), Table III also includes the rms difference between subjects' observed matrices and a purely random matrix (column 11) for purposes of comparison. To make the table easier to read, only the lowest of the minimized rms values achieved across perceptual dimension combinations (highlighted in bold) and those within 1% of the lowest value are reported in Table III. On first inspection, it is clear that columns 3 through 7 are populated with data and that the data in columns 8 through 10 are relatively sparse. That is, those combinations of perceptual dimensions which included formants (F1, F2, and/or F3) and silent gap duration (G) as perceptual dimensions yielded best-fit rms values that were the lowest across perceptual dimension combinations, or within 1% of the lowest values. The lowest rms values achieved across perceptual dimensions ranged from 7% to 13%. The large majority of these lowest rms values occurred for the F1F2F3AG and F1F2AG perceptual dimension combinations. For subjects C4, N3, and N9, the lowest rms values occurred in the random condition. These subjects performed poorest on the consonant identification

task, from 5% to 15% correct, so the MPI model could not provide a better description of these subjects' consonant matrices than a purely random matrix. In terms of the model's goodness-of-fit, i.e. R^2 , averaged across subjects (last row in Table III), the best-fit condition accounted for over 40% of the observed variation in subjects' matrices, in contrast to the random condition which yields $R^2 = 0$.

The observation that combinations that included formants and silent gap duration as perceptual dimensions provided better descriptions of subjects' observed matrices was confirmed by a RMANOVA on ranks applied to rms values. For this test, perceptual dimension combinations as well as the random matrix comparison were considered as different treatment groups applied to the same CI subjects. The RMANOVA was significant ($p < 0.001$). *Post-hoc* analysis (Student–Newman–Keuls method) yielded the following significant group rms differences ($p < 0.01$): F1F2F3AG < F1F2AG $\equiv$ F1F3AG $\equiv$ F2F3AG $\equiv$ F1F2F3G < F1F2F3A < F1F2F3 < AG < random rms ($\equiv$ sign indicates no significant difference). To summarize, the F1F2F3AG and/or F1F2AG models produced the best fits as a group, and the models that included formants and silent gap durations as perceptual dimensions produced better fits than models that did not include both of these dimensions.

TABLE IV. Number of satisfactory 2×2 comparison matrices between observed and predicted 24-consonant matrices, at thresholds of 3%, 5%, and 10% for each perceptual dimension.

Subject	No. satisfactory 2×2 comparison matrices: 24-consonant							
	F1F2 F3AG	F1F2 AG	F1F3 AG	F2F3 AG	F1F2 F3G	F1F2 F3A	F1F2 F3	AG
C1	2	1	1	1	0	0	0	0
C2	0	0	0	0	0	0	0	0
C3	2	2	1	2	0	0	0	0
C4	1	1	1	1	1	0	0	1
C5	0	2	0	0	0	0	0	0
C6	0	0	0	0	0	0	0	0
N1	0	0	0	0	0	0	0	0
N2	0	0	0	0	0	0	0	0
N3	0	0	0	0	0	0	0	0
N4	1	0	1	1	1	1	0	0
N5	0	1	1	0	0	0	0	0
N6	1	0	1	0	0	1	0	0
N7	3	3	0	2	0	0	0	0
N8	0	1	0	0	0	0	0	0
N9	1	1	1	0	1	0	1	1
N10	0	0	0	0	0	0	0	0
N11	0	0	0	0	0	0	0	0
N12	3	1	1	0	3	1	0	0
N13	2	2	2	2	0	0	0	1
N14	3	3	2	3	2	2	1	1
N15	3	2	1	0	2	2	0	0
N16	0	1	0	0	0	0	0	0
N17	2	0	0	0	2	0	0	0
N18	0	0	0	0	0	1	0	0
N19	0	0	0	0	0	0	0	0
N20	0	0	0	0	0	0	0	0
N21	0	1	0	0	0	0	0	0
N22	0	0	1	0	0	0	0	0
Total	24	22	14	12	12	8	2	4

In terms of matching of error patterns, the fit of the MPI model to CI subjects' consonant confusion matrices is summarized in Table IV for the eight perceptual dimension combinations tested in the present study (columns 2 through 9). More specifically, Table IV lists the number of satisfactory 2×2 comparison matrices obtained for each subject at each perceptual dimension (maximum of 3 representing satisfactory comparison matrices obtained at thresholds of 3%, 5%, and 10%). The 2×2 comparison matrix, which is a contingency table of the model's ability to predict the consonant-pair confusions made by subjects above a specified threshold percentage, was termed satisfactory if the number of true positives and true negatives were both greater than (or equal to) the number of false positives and false negatives. As observed in Table IV, at least one satisfactory matrix was obtained for 19 of the 28 subjects.

At the bottom of Table IV is a sum of the total number of satisfactory matrices obtained as a function of the different perceptual dimensions. The largest numbers of satisfactory matrices were obtained for the F1F2F3AG and the F1F2AG perceptual dimension combinations. Furthermore, models that included formants and silent gap duration as perceptual dimensions tended to produce more satisfactory matrices than models

TABLE V. Feature categories assigned to the 24 consonants used in present study. Voicing: 1, voiced; 2, voiceless; Manner: 1, stops; 2, fricatives and affricates; 3, nasals; 4, liquids and glides; Place: 1, front; 2, middle; 3, back.

C24	Voicing	Manner	Place
b	1	1	1
tʃ	2	2	2
d	1	1	2
dʒ	1	2	2
f	2	2	1
g	1	1	3
h	2	2	3
k	2	1	3
l	1	4	2
m	1	3	1
n	1	3	2
ŋ	1	3	3
p	2	1	1
r	1	4	3
s	2	2	2
ʃ	2	2	3
t	2	1	2
θ	2	2	2
ð	1	2	2
v	1	2	1
w	1	4	1
y	1	4	3
z	1	2	2
ʒ	1	2	3

that did not include both of these dimensions. Hence, in terms of comparing combinations of perceptual dimensions, results obtained using matching of predicted and observed error patterns were largely consistent with results obtained by minimizing rms differences between predicted and observed matrices.

B. IT analysis

In IT analysis (Miller and Nicely, 1955), phoneme confusion matrices are reduced to feature matrices by assigning feature categories to each consonant and then tabulating the correct and incorrect responses (i.e., within or between feature categories). The feature categories used for the 24 consonants of the present study in terms of voicing, manner and place are listed in Table V. Percent IT for observed and best-fit MPI predicted matrices (the latter being matrices obtained for the F1F2F3AG combination at lowest rms) are presented as scatter plots in Fig. 2. The correlation in IT estimates between observed and predicted matrices was fairly high, i.e. $R=0.850$ for voicing, $R=0.900$ for manner, and $R=0.926$ for place ($p < 0.0001$ in all cases). Comparing the regression line through the scatter plots in Fig. 2 with the diagonal line of slope equal to 1, it appears that IT estimates from predicted matrices tended to underestimate IT estimates from observed matrices. However, this result may have been due to the relatively small number of stimulus presentations in the observed matrices (on the order of 200) in comparison to the predicted matrices (on the order of 36 000). That is, as demonstrated by Sagi and Svirsky (2008), there exists an overestimation bias in the IT measure when applied to matrices with relatively small numbers of

FIG. 2. Percent IT estimates of best-fit predicted matrices (for F1F2F3AG combination in terms of rms) plotted against IT estimates of subjects' observed matrices in terms of the features voicing, manner, and place. Line through data only represents regression line. Diagonal line extended to axes represents regression line of slope 1.

stimulus presentations in comparison to matrices with large numbers of stimulus presentations.

IV. DISCUSSION

In the Introduction it was stated that the purpose of constructing mathematical models such as the one in the present study is to understand the mechanisms that underlie speech perception by CI users. The MPI model helps us get closer to this goal by providing a platform with which to test specific hypotheses about the interrelation between speech cues transmitted by a CI, a listener's ability to discriminate these cues, and how these cues are combined to arrive at a confusion matrix. When the best fit between a MPI predicted matrix and a subject's observed matrix is poor or deemed inadequate, one can reject the hypothesized speech cues and/or modeling assumptions as descriptors of underlying mechanisms. That is, the hypothesized speech cues and modeling assumptions are validated by the extent to which predicted matrices account for observed matrices, and when required one can revise ones choice of speech cues and/or modeling assumptions.

TABLE VI. Correlation statistics (R - and p -values) among electrical speech cue measurements across formant (F1, F2, and F3) amplitude ratio (A) and silent gap duration (G) perceptual dimensions. The large R values for the A dimension (in bold) suggests redundancy in how consonant tokens are represented by this dimension in comparison to other dimensions.

		F2	F3	A	G
F1	R	-0.24	0.267	0.369	-0.317
	p	0.042	0.023	0.0014	0.0067
F2	R	—	0.119	0.641	-0.276
	p	—	0.318	<0.0001	0.019
F3	R	—	—	0.551	-0.093
	p	—	—	<0.0001	0.436
A	R	—	—	—	-0.352
	p	—	—	—	0.0024

In the present study three types of speech cues were proposed, location of mean formant energy along the implanted array for the first three formants (F1, F2, and F3), proportion of charge in electrodes encoding frequencies below 800 Hz (A), and duration of silent gap (G). Best-fit model matrices were obtained for different combinations of perceptual dimensions comprising these speech cues. When comparing best-fit observed and predicted matrices across perceptual dimension combinations it was found that models that included the formant and silent gap dimensions outperformed models that did not include either. That is, it appears that the A dimension did not provide as much explanatory power as the other two dimensions. In Table VI, correlations between electrical speech cue measurements across F1, F2, F3, A, and G are reported. As one can observe, moderate correlations occurred between the A dimensions and two of the formant dimensions (F2 and F3) suggesting redundancy in how consonant tokens are represented by this dimension in comparison to other dimensions and may account for this dimension's lack of explanatory power. This is not to say that the A dimension provided nothing (the F1F2F3AG combination did tend to provide the best description among those compared), but that it may be necessary to replace this dimension in future applications of the model.

Clearly, even the most successful combination of dimensions in the present study does not explain every aspect of every listener's confusion matrix. The model does not predict very well the confusions made by listeners whose consonant identification scores are random or barely above random, as demonstrated in Table III for the three subjects for whom best-fit rms values were no better than a purely random matrix. Similarly, the model implemented cannot predict the complete pattern of consonant confusions in several subjects as demonstrated in Table IV for those subjects for whom the satisfactory criteria did not apply for any perceptual dimension combination. For these subjects, it is clear that the proposed perceptual dimensions and/or modeling assumptions require revision.

Nevertheless, it is encouraging that a relatively simple model with only three degrees of freedom can explain so many aspects of consonant identification by a diverse group of CI users. In the present study, it was proposed that a substantial source of the performance variability observed in CI

users is attributable to differences in these users' ability to discriminate the basic perceptual cues that are important for speech recognition. The model presented was constructed off of this premise utilizing three types of cues as represented by the output of a CI. Although these cues are not sufficient to explain all aspects of consonant confusions by CI users, they accounted for over 40%, on average, of the variability in consonant identification among the CI users tested in the present study (see last row in Table III) as well as 70%–80% of the variability when these subjects' matrices were analyzed for voicing, manner, and place (see Fig. 2).

Indeed, our model's ability to do a relatively good job at predicting subjects' feature category confusions and only a moderate job at predicting subjects' consonant confusions does contribute to our understanding of the mechanisms that underlie consonant identification by CI users. It has been known for some time that CI users rely on the spectral and temporal cues transmitted by their device in order to identify consonants (e.g., Donaldson and Nelson, 2000; Fu, 2002). By postulating perceptual dimensions that included both, the model did well at predicting subjects' feature category confusions. However, to surpass this and provide an accurate account for confusions at the level of individual consonants may require isolating a subject-specific set of spectral and temporal dimensions, increasing the number of dimensions employed, and/or using other methods of integrating these dimensions.

V. CONCLUSIONS

A simple implementation of the MPI model that incorporated a small number of speech cues (as represented electrically by a CI) explained many aspects of consonant identification across a diverse group of CI users. The MPI model produced confusion matrices that matched many aspects of CI users' consonant confusion matrices (including, for many subjects, the specific consonant pairs that were or were not confused), despite differences in age at implantation, implant experience, device and stimulation strategy used (Table I), as well as overall consonant identification level (Table III). These results are promising, because they indicate that the framework that underlies the MPI model, even if implemented in a simple fashion, provides a fair degree of explanatory power and brings us closer to understanding the specific mechanisms that drive an individual CI user's ability to identify the sounds important for understanding speech.

ACKNOWLEDGMENTS

Norbert Dillier from Swiss Federal Institute of Technology Zurich provided us with his sCILab computer program, which we used to record stimulation patterns generated by the Nucleus speech processors. Advanced Bionics Corporation provided an implant-in-a-box so we could monitor stimula-

tion patterns generated by their implant. We are extremely grateful to Heidi Neuburger, who conducted part of the subject testing and helped with data management. This study was supported by National Institutes of Health–National Institute on Deafness and Other Communication Disorders (NIH–NIDCO) grants R01-DC03937 (P.I.: Mario Svirsky) and T32-DC00012 (P.I.: David B. Pisoni) as well as by grants from the Deafness Research Foundation and the National Organization for Hearing Research.

- Donaldson, G. S., and Kreft, H. A. (2006). "Effects of vowel context on the recognition of initial and medial consonants by cochlear implant users," *Ear Hear.* **27**, 658–677.
- Donaldson, G. S., and Nelson, D. A. (2000). "Place-pitch sensitivity and its relation to consonant recognition by cochlear implant listeners using the MPEAK and SPEAK speech processing strategies," *J. Acoust. Soc. Am.* **107**, 1645–1658.
- Durlach, N. I., and Braid, L. D. (1969). "Intensity perception. I. Preliminary theory of intensity resolution," *J. Acoust. Soc. Am.* **46**, 372–383.
- Firszt, J. B., Koch, D. B., Downing, M., and Litvak, L. (2007). "Current steering creates additional pitch percepts in adult cochlear implant recipients," *Otol. Neurotol.* **28**, 629–636.
- Fu, Q.-J. (2002). "Temporal processing and speech recognition in cochlear implant users," *NeuroReport* **13**, 1635–1639.
- Kwon, B. J., and van den Honert, C. (2006). "Dual-electrode pitch discrimination with sequential interleaved stimulation by cochlear implant users," *J. Acoust. Soc. Am.* **120**, EL 1–EL 6.
- Miller, G. A., and Nicely, P. E. (1955). "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.* **27**, 338–352.
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Sagi, E., Fu, Q.-J., Galvin, J. J. III, and Svirsky, M. A. (2010a). "A model of incomplete adaptation to a severely shifted frequency-to-electrode mapping by cochlear implant users," *J. Assoc. Res. Otolaryngol.* **11**, 69–78.
- Sagi, E., Kaiser, A. R., Meyer, T. A., and Svirsky, M. A. (2009). "The effect of temporal gap identification on speech perception by users of cochlear implants," *J. Speech Lang. Hear. Res.* **52**, 385–395.
- Sagi, E., Meyer, T. A., Kaiser, A. R., Teoh, S. W., and Svirsky, M. A. (2010b). "A mathematical model of vowel identification by users of cochlear implants," *J. Acoust. Soc. Am.* **127**, 1069–1083.
- Sagi, E., and Svirsky, M. A. (2008). "Information transfer analysis: A first look at estimation bias," *J. Acoust. Soc. Am.* **123**, 2848–2857.
- Stevens, K. N. (2000). "Diverse acoustic cues at consonantal landmarks," *Phonetica* **57**, 139–151.
- Svirsky, M. A. (2000). "Mathematical modeling of vowel perception by users of analog multichannel cochlear implants: Temporal and channel-amplitude cues," *J. Acoust. Soc. Am.* **107**, 1521–1529.
- Svirsky, M. A. (2002). "The multidimensional phoneme identification (MPI) model: A new quantitative framework to explain the perception of speech sounds by cochlear implant users," in *Etudes et Travaux Vol. 5*, edited by W. Serniclaes (Institut de Phonétique et des Langues Vivantes of the ULB, Brussels), pp. 143–186.
- Svirsky, M. A., Silveira, A., Suarez, H., Neuburger, H., Lai, T. T., and Simmons, P. M. (2001). "Auditory learning and adaptation after cochlear implantation: A preliminary study of discrimination and labeling of vowel sounds by cochlear implant users," *Acta Oto-Laryngol.* **121**, 262–265.
- Teoh, S. W., Neuburger, H. S., and Svirsky, M. A. (2003). "Acoustic and electrical pattern analysis of consonant perceptual cues used by cochlear implant users," *Audiol. Neuro-Otol.* **8**, 269–285.
- Tyler, R. S., Preece, J. P., and Lowder, M. W. (1987). "The Iowa audiovisual speech perception laser videodisc," *Laser Videodisc and Laboratory Report* (Department of Otolaryngology-Head and Neck Surgery, University of Iowa at Iowa City).