Contribution of formant frequency information to vowel perception in steady-state noise by cochlear implant users

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

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

(Received 16 June 2016; revised 12 January 2017; accepted 18 January 2017; published online 23 February 2017)

Cochlear implant (CI) recipients have difficulty understanding speech in noise even at moderate signal-to-noise ratios. Knowing the mechanisms they use to understand speech in noise may facilitate the search for better speech processing algorithms. In the present study, a computational model is used to assess whether CI users' vowel identification in noise can be explained by formant frequency cues (F_1 and F_2). Vowel identification was tested with 12 unilateral CI users in quiet and in noise. Formant cues were measured from vowels in each condition, specific to each subject's speech processor. Noise distorted the location of vowels in the F_2 vs F_1 plane in comparison to quiet. The best fit model to subjects' data in quiet produced model predictions in noise that were within 8% of actual scores on average. Predictions in noise were much better when assuming that subjects used *a priori* knowledge regarding how formant information is degraded in noise (experiment 1). However, the model's best fit to subjects' confusion matrices in noise was worse than in quiet, suggesting that CI users utilize formant cues to identify vowels in noise, but to a different extent than how they identify vowels in quiet (experiment 2).

© 2017 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4976059>]

[TB]

Pages: 1027–1038

I. INTRODUCTION

Cochlear implants (CIs) are prosthetic hearing devices that restore varying degrees of hearing to individuals with severe to profound hearing impairment. CI devices can provide recipients with substantial amounts of open-set word recognition in quiet, albeit with large amounts of individual differences (Mahmoud and Ruckenstein, 2014). Most CI recipients complain about difficulty understanding speech in noise at signal-to-noise ratios (SNRs) that would have little impact on speech understanding in listeners with normal hearing (Gifford and Revit, 2010). Present day devices are equipped with directional microphones and signal processing algorithms that can result in improved speech perception in difficult listening environments (Geibler *et al.*, 2015; Mauger *et al.*, 2014). However, the improvements in speech perception offered by these advances remain poorer than what listeners with normal hearing would achieve under similar listening conditions, especially when signal and noise are co-located in space (Gifford and Revit, 2010). To advance noise reduction algorithms further, it may be helpful to consider the effect of noise on the acoustic cues employed by CI users to understand speech.

The long term goal behind the present study is to develop computational models of speech perception by CI users in order to better understand the mechanisms they employ to understand speech in noise. The modeling framework employed is the Multidimensional Phoneme Identification (MPI) model (Svirsky, 2000, 2002), which predicts a listener's phoneme identification based, in part, on

their ability to discriminate a postulated set of relevant speech cues. The MPI model has been successfully employed in predicting vowel and consonant identification in quiet by CI users (Sagi *et al.*, 2010a, 2010b; Svirsky *et al.*, 2011, 2015) and by normal hearing individuals listening to acoustic simulations of CIs (Won *et al.*, 2016). One advantage of using computational models of speech perception such as the MPI model is the ability to formulate and test hypotheses about which speech cues underlie a listener's phoneme confusion matrix. That is, postulated speech cues that yield a model output that more closely matches a listener's confusion matrix are more likely to be the ones used by that listener than postulated speech cues that yield a poorer match. Another advantage of the MPI model is the capability of predicting speech perception outcomes under different CI device settings (Sagi *et al.*, 2010a, 2010b). A model that is successful at accounting for a CI user's speech perception outcome in noise can potentially be used to predict the effects of various noise-reduction algorithms on that individual's speech perception outcome.

In listeners with normal hearing, the acoustic cues that contribute to vowel identification in quiet include steady-state and dynamic formant frequencies, overall spectral shape, fundamental frequency, and vowel duration (Chistovich and Lublinskaya, 1979; Hillenbrand *et al.*, 1995; Jenkins *et al.*, 1983; Nearey and Assmann, 1986; Peterson and Barney, 1952; Syrdal and Gopal, 1986; Zahorian and Jagharghi, 1993). In the presence of background noise, listeners with normal hearing maintain their reliance on these cues to identify vowels, although at very low SNRs there may be a shift in the relative perceptual weighting assigned to these cues (Ferguson and Kewley-Port, 2002; Parikh and

^{a)}Electronic mail: Elad.Sagi@nyumc.org

Loizou, 2005; Swanepoel *et al.*, 2012). CI listeners use similar speech cues as listeners with normal hearing to identify vowels in quiet, although with different efficiency and perceptual weighting between cues, including steady-state and dynamic formant information as well as vowel duration (Donaldson *et al.*, 2013, 2015; Iverson *et al.*, 2006; Kirk *et al.*, 1992; Sagi *et al.*, 2010b; Winn *et al.*, 2012). The extent to which CI users utilize these cues to identify vowels in noise is not as well understood. In the present study, the MPI model is used to assess the extent to which the first two formant center frequencies, $F1$ and $F2$, can be used to explain CI users' vowel identification in quiet and in noise.

The first question this study addresses is whether $F1$ and $F2$, as represented in the output of a CI, are affected by steady-state background noise. If $F1$ and $F2$ are indeed affected by noise, to what extent do those changes explain the drop in CI subjects' vowel identification scores with decreasing SNR? To address this question, the model was used to predict vowel identification in noise under two different hypotheses about listeners' expectations about how vowels should sound in noise: (a) that expectations are based on the vowel as perceived in quiet, so a single set of exemplars are stored in long term memory, and (b) that listeners adjust their expectations based on *a priori* knowledge about the effect of noise on acoustic cues. The second question this study addresses is whether identification of vowels by CI users in noise can be explained by the same set of speech cues that contribute to their identification of vowels in quiet.

II. METHODS

A. Subjects

Twelve post-lingually deafened, unilaterally implanted adult subjects participated in the present study (Table I). Ages at time of testing ranged from 28 to 70 yrs old, with a mean age of 49 yrs. All subjects used Nucleus CI devices (Cochlear Corp., New South Wales, Australia). Nine subjects used Freedom speech processors. Three subjects used ESPrit 3G speech processors. Of these, two subjects were implanted with Nucleus 24 series electrode arrays and one subject was implanted with the Nucleus 22 series electrode array (S6). Subject S6 was programmed with the SPEAK

processing strategy. All other subjects were programmed with the ACE processing strategy. Subjects had at least 11 months experience using their device at time of testing. All subjects were native American English speakers. Prior to testing, informed consent was obtained from all subjects. Recruitment protocols and consent forms were approved by the New York University School of Medicine Institutional Review Board.

B. Stimuli and testing procedures

Subjects were tested for vowel identification in quiet (70 dB SPL), and in the presence of background noise at two SNRs: +10 dB SNR and +5 dB SNR. Stimuli were nine possible vowels in /hVd/ context, i.e., "had," "haved," "head," "heard," "heed," "hid," "hood," "hud," and "who'd" (Cochlear Corporation and the University of Iowa, 1995). They included five tokens of each vowel recorded from the same male talker, who was a native speaker of American English. A testing block for a given listening condition included these 45 stimuli (5 tokens $\times$ 9 vowels) presented in randomized order. Three testing blocks were obtained for each listening condition (i.e., 135 stimuli per listening condition). Testing sessions always began with one testing block in the quiet condition. The order of subsequent testing blocks and listening conditions was blocked across subjects using a Latin-square design. A testing block of one listening condition was always followed by a testing block of a different listening condition.

Background noise was unmodulated, steady-state speech weighted noise obtained from the ICRA noises CD (Track number 1; Dreschler *et al.*, 2001). This type of noise was chosen because it effectively masks the vowel stimuli used in the present study, which vary relatively slowly with time, at relatively low noise levels. Signal and noise were combined digitally. Each vowel stimulus was placed at the center of a 3 s segment of noise. To ensure constancy of both the SNR of each stimulus as well as the noise intensity level from one stimulus presentation to the next, vowel stimuli were digitally normalized to have constant root-mean-square (RMS) power. Noise RMS power was measured over the duration to which the signal would be added, and adjusted to the desired SNR.

TABLE I. Information about demographics, device type, and experience for study subjects including added signal processing options in their everyday map used for testing (ADRO = adaptive dynamic range optimization, AS = autosensitivity).

Subject	Age	Gender	Implant Experience	Implant	Processor	Strategy	Option
S1	53	female	5.8	CI24R (CA)	Freedom	ACE	ADRO
S2	70	male	2.5	Freedom CI24RE (CA)	Freedom	ACE	AS+ADRO
S3	33	male	11.9	CI24M	Freedom	ACE	AS+ADRO
S4	55	male	2.1	Freedom CI24RE (CA)	Freedom	ACE	none
S5	28	female	4	Freedom CI24RE (CA)	Freedom	ACE	ADRO
S6	64	female	15.8	Nucleus 22 series	ESPrIt 3G	SPEAK	none
S7	66	male	0.9	Freedom CI24RE (CA)	Freedom	ACE	ADRO
S8	70	male	7.4	CI24R (CS)?	Freedom	ACE	none
S9	44	female	5.6	CI24R (CA)	ESPrIt 3G	ACE	none
S10	35	female	10.9	CI24M	Freedom	ACE	ADRO
S11	41	female	5.4	CI24R (CA)	ESPrIt 3G	ACE	AS
S12	28	male	8.1	CI24M	Freedom	ACE	ADRO

Testing was carried out in a sound controlled test room (Industrial Acoustics Company). Stimuli and noise were presented from the front over a loudspeaker. The signal presentation level was set to 70 dB-C using vowel stimuli in quiet, and was verified using a sound level meter with the microphone placed at the approximate location where subjects would be seated away from the loudspeaker (about 1-m). Prior to testing, subjects were familiarized with the test stimuli in quiet, which were presented several times in alphabetical order to subjects while they were shown an alphabetical list of the stimuli. During testing, stimuli would be presented and subjects would select the vowel they heard by pointing to a sheet that listed the possible stimuli. Subjects were instructed to select the vowel label that most closely matched the vowel they heard.

C. Speech cue measurements

The application of the MPI model in the present study differs from Sagi *et al.* (2010b) in the set of postulated speech cues. In Sagi *et al.* (2010b), the relevant set of speech cues postulated to underlie vowel identification by CI users was location of mean steady-state formant energy along the implanted array. In addition to steady-state formant information there is some evidence that CI users can also utilize dynamic cues, i.e., vowel duration combined with change in formant center frequency with time throughout the vowel utterance due to co-articulation and vowel-inherent spectral change (VISC) (Donaldson *et al.*, 2013, 2015; Iverson *et al.*, 2006; Kirk *et al.*, 1992; Winn *et al.*, 2012). To allow for the possible contribution of VISC in the present study, locations of mean formant energy along the implanted array were measured for the first and second halves of the vowel duration for each vowel stimulus. The possible contribution of vowel duration to CI users' identification of vowels in quiet and in noise was not explored in the present study.

In Nearey and Assman (1986), VISC is defined as the relatively slowly varying changes in formant frequencies associated with vowels themselves. One unique result of Nearey and Assman (1986) is that they were able to reproduce the rate of vowel identification in normal hearing listeners using very brief sections extracted from "nucleus" and "offglide" portions of the vowel. These were 30 ms segments centered at 24% and 64% of the vowel duration. For the vowel stimuli of the present study, we compared acoustic measures of formant frequencies at these brief 30-ms nucleus and offglide portions with formant measures averaged over the first and second halves of the vowel duration using Praat (Boersma, 2001). The two are highly correlated (R^2 values for $F1$ and $F2$ in the nucleus portion in comparison to $F1$ and $F2$ averaged over the first half were both 0.99; and R^2 values for $F1$ and $F2$ in the offglide portion in comparison to $F1$ and $F2$ averaged over the second half were 0.97 and 0.98, respectively). The main advantage of using a larger time window to track VISC in separate halves of the vowel duration (rather than brief sections extracted from the nucleus and offglide portions) is not in the acoustic domain, but rather in measuring these formant cues from the stimulation pattern that is delivered to a CI.

In the present study, measurements of mean formant energy locations along the implanted array were inferred from electrodiagrams of vowel stimuli in quiet and in noise specific to each listener's speech processor settings. An electrodiagram represents the assignment of stimulation pulse magnitudes to electrodes as a function of time. In nucleus devices, this pulse magnitude sequence can be captured from the radio-frequency (RF) output of the speech processor. Electrodiagrams were obtained on a subject-specific basis using speech processors donated to our lab by Cochlear Americas. The same model of speech processor programmed with the same processor settings that each subject used for vowel identification testing was used for obtaining the electrodiagrams. An automated MATLAB routine was used to play vowel stimuli in quiet and in noise over a loudspeaker and record the RF output from the speech processor. RF output was captured using the Clinical Programming System (CPS) connected to a PC by an IF5 card (Cochlear Americas, Denver, CO) and controlled with the NICaptureClient commands built into the Nucleus Implant Communicator software package (NIC, provided by Cochlear Americas, Denver, CO). These commands record the pulse sequence represented in the RF output. The speech processor and CPS unit were placed on a tripod 1 m away from the same loudspeaker used for testing and in the same room at a height parallel to the loudspeaker. The speech processor was placed on an artificial pinna pinned to the side of a piece of foam that rested above the CPS unit to ensure that the orientation of the speech processor microphone would be similar to how it would be worn by a listener.

Examples of electrodiagrams thus obtained for the stimuli "hid" and "hood" in quiet and in noise are shown in Fig. 1. The electrode number is displayed on the vertical axis. Nucleus implanted arrays consist of 22 stimulating electrodes ordered in increasing number from the most basal electrode. Electrodes are assigned adjacent frequency bands such that lower frequencies stimulate more apical electrodes (Electrode 22 being the most apical) and higher frequencies stimulate the more basal electrodes (Electrode 1 being the most basal). The specific range of frequencies assigned to electrodes depends on the setting selected in the speech processor. The default total frequency range is 188 to 7938 Hz. In a sense, the electrodiagram is similar to a rough grain spectrogram.

In Fig. 1, vertical lines indicate boundaries for the vowel duration window, i.e., the points in time marking the beginning and end of the vowel utterance. Vowel duration windows were determined from acoustic stimuli in quiet using Praat (Boersma, 2001), i.e., not from electrodiagram recordings. Within Praat, onset of $F0$ was used to determine the time marking the beginning of the vowel, and the time point at which acoustic energy above 800 Hz dropped abruptly was used to determine the end of the vowel duration. For acoustic stimuli in noise, $F0$ onset was used to determine the beginning of the vowel duration window and the vowel durations obtained in quiet were used to infer the end of the vowel duration window in noise. Correct alignment of vowel duration windows within electrodiagrams was verified by visual inspection.

FIG. 1. Examples of $F1$ and $F2$ estimates (circles and squares, respectively) from electrograms of the vowels in “hid” and “hood” in quiet and noise conditions. Electrograms were obtained specific to each subject’s speech processor settings. Formants were averaged over the first and second halves of the vowel duration window (vertical lines).

Locations of mean formant energies along the implanted array were inferred from electrograms as follows. From each half of the vowel duration window, the number of times each electrode was stimulated, weighted by pulse magnitude, was compiled into a histogram. These histograms were partitioned into regions of formant energies, one for the first formant ($F1$) and one for the second formant ($F2$). The mean value within each formant region represented the location of mean energy for the formant within that region. For these calculations, electrode numbers were treated as continuous variables. That is, formant energy is typically distributed over a range of frequencies that spans two or more electrodes. Hence, it is possible for the mean of this energy distribution to be located at a position along the implanted array that lies between adjacent electrodes. Our use of the electrode number as a continuous variable is an approximate way of inferring location along the implanted array without knowing exact interelectrode distances.

The frequency ranges selected to partition histograms into formant regions were based on measures of $F1$ and $F2$ for the acoustic vowel stimuli using Praat, obtained over the first and second halves of the vowel window. In choosing frequency range values, we wanted to avoid making electrogram measurements of formant energy locations based off of narrow frequency ranges determined for each vowel *a priori*. On the other hand, we wanted to avoid including in our measurements contributions from more than one formant. Hence, frequency range values were chosen to be

broad enough to include the range of possible formant frequency values across as many vowels as possible for a given formant, but without including frequency values of other formants. The region of formant energy for $F1$ included all electrodes that encoded frequencies below 711 Hz. The region of formant energy for $F2$ varied for the different vowel tokens. For the vowel in “heard,” the $F2$ region included all electrodes that encode frequencies between 711 and 1575 Hz. For the vowel in “heed” and for the first half of the vowel in “hid,” the $F2$ region included all electrodes that encode frequencies between 711 and 2550 Hz. For the remaining vowels, including the second half of the vowel in “hid,” the $F2$ region included all electrodes that encode frequencies between 711 and 2200 Hz. Examples of speech cue measurements thus obtained are shown in Fig. 1 for the vowels in “hid” and “hood” in quiet and in noise. Electrode locations of $F1$ and $F2$ energies are indicated by circles and squares, respectively.

D. MPI model

1. Model framework

The implementation of the MPI model for vowel identification in the present study closely follows Sagi *et al.* (2010b). The postulated set of speech cues (Sec. II C) are assumed to vary along independent dimensions so that phonemes are represented by these speech cues as points in a Euclidean multidimensional space. The MPI model consists

of two primary components, an internal noise component and a decision component. In the internal noise component, it is assumed that internal noise (both sensory and memory) cause the same phoneme to produce percepts that are distributed about that phoneme's location within the multidimensional space. It is assumed that percepts are normally distributed about phoneme locations and that the standard deviations of this distribution along each dimension are proportional to the listener's discrimination, or just-noticeable-difference (JND) for the speech cues that define the multidimensional space. These JND-like parameters are input parameters to the MPI model.

In the decision component of the model, percepts are categorized by selecting the closest response center, i.e., the average location, or centroid, of where the listener expects to find a phoneme within the multidimensional space. If the listener's expectations are unbiased, as may be the case with normal hearing listeners or CI listeners who have completely adapted to their CI, then response centers are assumed equal to stimulus centers, i.e., the actual average location of phonemes within the multidimensional space. This process of categorization of percepts by selecting the closest response center is inherently related to the boundaries that define each vowel category. Boundaries for vowel categories within the MPI model framework are defined as the locations within the multi-dimensional space that are equidistant from any two adjacent response centers. For example, if a percept happens to be equidistant between two or more adjacent response centers, then it falls on the boundary between these vowel categories and will be categorized with equal probability among the vowel categories. By conducting many iterations of the model components, i.e., generating a percept and classifying it based on the location of the closest response center, and repeating this procedure for all phonemes of interest, one can construct a confusion matrix.

The MPI model can be implemented in two steps. The first step is to define the set of speech cues postulated to underlie subjects' performance, and to measure these cues for each stimulus used for testing. Four perceptual dimensions were used in the present study. Namely, the locations of mean $F1$ and $F2$ formant energies along the implanted array averaged over the first half of the vowel duration, as well as $F1$ and $F2$ averaged over the second half of the vowel duration (Sec. IIC). When measured for a given vowel stimulus, these four values serve as coordinates that specify that vowel's location within the four-dimensional $F2$ vs $F1$ perceptual space. The second step is to establish the model construct, including assumptions about input parameters (e.g., JNDs) and response centers depending on the hypotheses to be tested.

With our modeling implementation, we sought to answer two questions. First, to what extent can the drop in CI subjects' vowel scores with decreasing SNR be explained by the effects of noise on estimates of formant positions within the $F2$ vs $F1$ space (cf. Fig. 2)? Furthermore, are model predictions better when we assume that listeners use the same response centers in quiet and in noise, or when we assume that listeners use *a priori* knowledge about how response centers change in noise? These questions are

addressed in experiment 1. Second, is the model as capable of describing CI users' confusion matrices in the noise conditions as it is in explaining their confusion matrices in the quiet condition? This question is addressed in experiment 2.

a. Experiment 1. To assess the relation between distortion of formant estimates in noise to decreased vowel identification in noise by CI users, model JND parameters were fixed between quiet and noise conditions, and optimized to fit subjects' vowel confusion matrices in the quiet condition. The optimized JND parameters are defined as those that produced a best-fit between a model matrix and a subject's observed matrix. The underlying assumption here is that CI subjects' discrimination for place of stimulation is unaffected by noise. By fixing the JND parameters to those of the quiet condition, the difference in model output between the quiet and noise conditions is due primarily to differences in formant estimates between quiet and noise, as well as on our choice of response centers. Response centers, i.e., the vowel exemplars that serve as targets for how percepts are categorized within the decision model, were chosen in one of two ways. In one case, response centers were set equal to the values of stimulus centers (i.e., the average formant estimates for each vowel obtained from electrode output) depending on the listening condition. That is, in the +10 dB SNR condition, response centers for each vowel were set equal to stimulus centers obtained in the +10 dB SNR condition; and in the +5 dB SNR condition, response centers were set equal to stimulus centers obtained in the +5 dB SNR condition. The assumption here is that listeners adjust their expectations of how the vowels should sound as they are affected by the presence of noise. In the second case, response centers were set equal to stimulus centers in the quiet condition irrespective of the listening condition. In this case, listeners' expectations of how vowels should sound are fixed to how those sounds are perceived in quiet.

It is important to reiterate that the predictions in noise were made based entirely on the JND parameters optimized to fit subjects' response matrices obtained in the quiet condition. There were zero degrees of freedom for the predictions in noise. That is, no free parameters were adjusted to obtain predicted scores in noise. Essentially, the primary factor that affected subjects' predicted scores in noise in this comparison was the effect noise had on estimates of speech cues.

b. Experiment 2. The goal of this experiment is to assess the performance of the MPI model in the noise conditions in comparison to its performance in the quiet condition. It has already been established by the present investigators that vowel recognition in quiet by CI users is well described by listeners' abilities to discriminate locations of mean formant energy along the implanted array (Sagi *et al.*, 2010b). Are these speech cues equally sufficient to account for CI users' vowel recognition in noise? To answer this question, JND parameters were optimized to fit subjects' confusion matrices, separately for each listening condition (i.e., the same number of degrees of freedom were used for each condition). Response centers were set equal to the stimulus centers used in each condition.

FIG. 2. Effect of noise conditions on average formant measures of vowels used for testing. Formants in Hz obtained from acoustic vowels using Praat (left panel), and in electrode units (eu) from electrograms of vowels from one representative subject's CI processor (right panel).

2. Parameter estimation

Optimized JNDs were obtained as follows. Using speech cue measurements from subjects' electrograms, model predicted confusion matrices were generated while systematically varying the JND input parameters. Two degrees of freedom were used, one for the JND parameters along the $F1$ dimensions, and one for the JND parameters along the $F2$ dimensions. The values for JND parameters ranged from 0.02 electrode units to 4 electrode units in steps of 0.02 (i.e., $200 \times 200 = 40\,000$ possible combinations for JND values of $F1$ and $F2$). One electrode unit refers to the distance between two adjacent electrodes along the length of the electrode array. For a given set of specified JND parameters, each model predicted confusion matrix was generated using 2500 repetitions per vowel. Goodness-of-fit between model predicted confusion matrices and subjects' observed confusion matrices was assessed by calculating the RMS difference between predicted and observed matrices. To calculate RMS difference, matrix rows were first expressed as percentages so that each row summed to 100%. Then

pair-wise elements of predicted and observed matrices were subtracted, squared, and summed, dividing the total by the number of elements and taking the square root. Hence, the RMS difference reflects the standard deviation, in percent, between elements of predicted and observed matrices. The optimized JND parameters are defined as those that produced a model matrix with minimum RMS difference from a subject's observed matrix.

3. Performance metrics

In experiment 1, model predicted percent correct scores were compared with subjects' observed percent correct scores using a measure of mean absolute difference (MAD, in units of percent correct), i.e., the absolute difference between observed and predicted scores averaged across subjects. The lower the MAD the better the model prediction. In addition, predicted and observed percent correct scores were compared using binomial 95% confidence intervals based on the number of identifications used to calculate subjects' observed scores (i.e., 135 per subject). If model predicted

TABLE II. Subjects' observed vowel percent scores in quiet and noise conditions (sd = standard deviation).

Subject	Observed Scores (%)		
	Quiet	+10 dB SNR	+5 dB SNR
S1	78	54	50
S2	87	69	63
S3	91	74	68
S4	100	89	91
S5	91	70	63
S6	97	72	65
S7	97	87	75
S8	89	75	63
S9	81	57	48
S10	90	85	77
S11	53	37	30
S12	85	60	51
Mean (%)	87	69	62
sd (%)	12	15	16

scores fall outside the confidence interval for observed scores, then the two scores are significantly different.

In experiment 2, goodness-of-fit, in the form of minimized RMS values (explained in Sec. IID2), was compared between quiet and noise conditions using a one-way repeated measures analysis of variance (ANOVA). This comparison differs from that of experiment 1 in that the best possible fit of the model to subjects' entire confusion matrices was sought in each listening condition (i.e., JND optimized separately for each condition). In experiment 1, differences between model predictions and observed data were limited to percent correct scores only. Presumably, any discrepancies in goodness-of-fit between quiet and noise conditions could be due to the set of postulated cues. That is, a poorer best-fit of the model in a particular condition implies that the postulated acoustic cues are relatively poorer at describing how phonemes in that condition are identified.

III. RESULTS

A. Experiment 1: MPI model predictions for vowel scores in noise

Subjects' observed vowel percent scores in quiet and noise conditions are presented in Table II. Scores were significantly poorer in the noise conditions than in the quiet condition, and scores in the +5 dB SNR condition were significantly poorer than scores in the +10 dB SNR condition ($p < 0.001$, Holm-Sidak paired comparisons following a significant one-way repeated measures ANOVA). The first question we sought to address was to what extent the drop in CI subjects' vowel scores with decreasing SNR was explained by the effects of noise on estimates of formant locations. The change in formant estimates with increasing background noise is depicted in Fig. 2 in which average first and second formant frequencies measured acoustically (left panel) and measured from electrograms (right panel) are plotted for the vowel stimuli used in the present study. Acoustic measures of formant values were obtained using

Praat's linear predictive coding algorithm with default values. Electric location values of formants along the implanted array (in electrode units, or eu) were obtained from electrograms of vowels using one representative subject's speech processor settings. In both cases, i.e., acoustic and electric measures, formants are averaged over the entire duration of the vowel window for simplicity. Plots of $F2$ vs $F1$ in quiet (which were quite consistent for acoustic measures with the classic Peterson and Barney, 1952, values), +10 dB SNR, and +5 dB SNR are arranged in the top, middle, and bottom panels of Fig. 2, respectively. In the presence of background noise, formant values tend to shift, becoming more clustered, and the vowel quadrilateral typical for English vowels becomes distorted.

The formant values used for modeling in the present study, which were obtained from electrograms, track the acoustic values relatively closely, albeit with some differences. Overall, the two sets of measures are highly correlated. For example, when Praat-derived formants (in Hz) were converted to electrode units using linear interpolation and the frequency allocation table for subject S1, and then compared with formant location measurements obtained from this subject's electrograms, correlation coefficients of 0.81 and 0.95 were obtained for $F1$ and $F2$ values in the first half of the vowel window, and correlation coefficients of 0.77 and 0.86 were obtained for $F1$ and $F2$ values in the second half of the vowel window (quiet and noise conditions combined in all cases, $p < 0.001$). Hence, the distorting effect of noise on formant estimates that one might expect from the degraded spectral representation offered by a CI can be shown to occur in the acoustic signal as well.

Having demonstrated that background noise distorts formant estimates, the MPI model was employed to determine the extent to which this distortion can account for the drop in subjects' vowel percent correct scores with increasing background noise. Modeling predictions of subjects' observed vowel percent correct scores in quiet and noise listening conditions were generated using subject-specific formant estimates obtained from electrograms in quiet and noise conditions (Sec. IIC), and using subjects' best-fit JND parameters obtained in the quiet condition. These predictions are presented in Figs. 3 and 4. In Fig. 3, response centers were set equal to the stimulus centers used in each listening condition, i.e., it was assumed listeners could adjust their expectations of how vowels should sound in the presence of increasing background noise. Model predictions were assessed in two ways. MADs between model and observed percent correct scores across subjects were reported for each condition. The lower the MAD the better the prediction. Furthermore, 95% binomial confidence intervals (gray curves) were reported for the observed scores. Any model produced scores that fall within the upper and lower confidence interval suggest predictions that were not significantly different than corresponding observed scores. Not surprisingly, a very good match between model and observed percent correct scores were obtained in quiet (MAD = 3%, only 1 out of 10 subjects' model produced scores was significantly different than their observed score). This result occurred because model produced scores in quiet were

FIG. 3. Model predictions of subjects' vowel scores in noise (middle and right panels) based on formant measures in noise and best-fit JNDs obtained from subjects' vowel matrices in quiet (fit between model and observed scores in quiet in left panel). Response centers assumed equal to stimulus centers in each condition.

derived, in part, from subjects' observed scores in quiet. That is, model produced scores in quiet were generated using the best-fit JND parameters that minimized RMS difference between model produced and observed confusion matrices in quiet. However, model produced scores were highly predictive of observed scores in noise even though the predicted scores were obtained using best-fit JND parameters obtained in quiet (for +10 dB SNR, MAD = 8% and 6 out of 12 subjects' model produced scores were not significantly different than their observed score; for +5 dB SNR, MAD = 7% and 7 out of 12 subjects' model produced scores were not significantly different than their observed score). That is, knowing only the best-fit JND parameters in quiet and estimates of speech cue measurements in noise, and assuming listeners could adjust their expectations of how the vowels should sound in noise, the model provided good predictions of subjects' observed scores in noise with zero degrees of freedom.

When response centers were set equal to stimulus centers in quiet irrespective of the listening condition, i.e., it was assumed that listeners' expectations of how vowels should sound are fixed to how they are perceived in quiet, the model produced scores that were poorly predictive of subjects' observed scores in noise. This result is demonstrated in Fig. 4 where for both noise conditions MAD is greater than 20%, and 11 out of 12 subjects' model produced scores were significantly different than observed scores. The only difference in model predictions between Figs. 3 and 4 are in the selection of response centers.

FIG. 4. Model predictions of subjects' vowel scores in noise based on formant measures in noise and best-fit JNDs obtained from subjects' vowel matrices in quiet, but with response centers assumed equal to stimulus centers in quiet (RCQ).

B. Experiment 2: MPI model performance in quiet vs noise

The second question we sought to address was whether the model is as capable of describing CI users' confusion matrices in the noise conditions as it is in explaining their confusion matrices in the quiet condition. Goodness-of-fit of the MPI model to subjects' vowel confusion data is presented in Figs. 5 and 6. In Fig. 5, observed and best-fit predicted confusion matrices in quiet and noise conditions are presented for one subject (S5). Observed and predicted response frequencies for each vowel (along rows of each matrix) are presented side-by-side as bar graphs. The horizontal dashed line along each row represented a response frequency of 50%. The modeling data for this particular subject were selected because it is most representative of the average best-fit model across subjects (for subject S5, RMS was 8.3%, 12.3%, and 15.2% for quiet, +10 dB SNR, and +5 dB SNR conditions, respectively). In Fig. 6, minimized RMS difference between model predicted and subjects' observed confusion matrices, averaged across subjects, is shown for the quiet and noise conditions (error bars indicate standard error). The average RMS values were 7.8%, 11.4%, and 12.9% for quiet, +10 dB SNR, and +5 dB SNR conditions, respectively. A repeated measures ANOVA revealed a significant effect of listening condition on minimized RMS difference values ($p < 0.001$). *Post hoc* comparisons (Holm-Sidak) show significant differences between quiet and both noise conditions ($p < 0.05$), but not between the two noise conditions ($p = 0.15$). In contrast to the data in Fig. 3, the RMS values in Fig. 6 were obtained using the best-fit JNDs for each condition (i.e., 2 degrees of freedom for each condition). To help interpret the goodness of fit represented by the RMS values reported in Fig. 6 (gray bars), the solid horizontal lines in Fig. 6 indicate the upper 95% confidence interval in RMS difference due to sampling error. That is, if one were to conduct two vowel identification experiments in the same individual under the same conditions using 135 identifications, compute the RMS difference between these two matrices, and then repeat the process many times over, the solid horizontal lines would indicate the upper 95% confidence interval in the sampling distribution of RMS differences produced by this process. Hence, a best-fit model matrix that produces an RMS difference below the horizontal lines in Fig. 6 would be statistically indistinguishable

FIG. 5. Example of best-fit model predicted matrices obtained for subject S5. Bar graphs along each row represent response frequencies for each vowel for observed (black filled bars) and predicted (gray filled bars) matrices. Horizontal dashed gray lines represent response frequencies of 50%.

from the observed matrix. Based on this gauge of goodness of fit, it is clear that the model implementation of the present study produced a better description of subjects' observed confusion matrices in quiet than in noise. These results suggest that without any additional modeling assumptions, the postulated speech cues selected for modeling were relatively better for predicting vowel confusions in quiet than in noise.

FIG. 6. Minimized RMS difference between best-fit model and subjects' observed vowel confusion matrices in quiet and noise conditions. Two degrees of freedom for JND parameters were used to fit model in each condition. Response centers were assumed equal to stimulus centers in each condition.

IV. DISCUSSION

A. Experiment 1: MPI model predictions for vowel scores in noise

In the present study, the MPI model was fairly successful in providing predictions of CI users' vowel scores in steady-state background noise based on their vowel confusion matrices in quiet. Three factors drove this result. One factor was the changes in formant location measurements with increasing background noise. The other two factors were the modeling assumption that subjects' ability to discriminate mean formant locations along the array (i.e., JND parameters determined from subjects' vowel confusion matrices in quiet) remained fixed when predicting subjects' vowel scores in noise, and the modeling assumption that response centers were assumed to be equal to stimulus centers in noise.

1. Distortion of formant information with noise

The size of the distortion in formant estimates depicted in Fig. 2 is much larger than those obtained by Parikh and Loizou (2005). For example, in their study, the MAD between formant estimates in quiet and those in +5 dB SNR steady-state speech weighted noise was only 16 Hz for $F1$ and 23 Hz for $F2$. In contrast, in the present study, the MAD between formant estimates in quiet and those in +5 dB SNR was nearly 100 Hz for $F1$ and 200 Hz for $F2$. However, Parikh and Loizou (2005) used a very strict criterion on whether to include a formant estimate within the average measure. That is, in every 20 ms vowel segment in noise, any formant estimate that did not give a clear peak or was not "vowel-like" in either $F1$ or $F2$ was rejected. In their measurements, this criteria amounted to rejection of nearly 25% and 40% of formant estimates in the +10 dB and +5 dB SNR speech-shaped noise conditions, respectively. In our measures, all $F1$ and $F2$ estimates provided by Praat were included in the average measure.

Considering the large amounts of distortion in formant estimates depicted in Fig. 2 at what would be considered moderate amounts of noise, one could ask how vowel

identification by listeners with normal hearing is little affected at these noise levels. One possibility is that formant discrimination is better in normal hearing listeners than in CI users (Fitzgerald *et al.*, 2007; Sagi *et al.*, 2010b) and so contraction and distortion of the vowel quadrilateral (or triangle, depending on language) may not cause appreciable errors until larger amounts of noise are encountered. Another possibility may be in how formants are encoded in the temporal patterns of auditory nerve responses in an intact auditory periphery. Young and Sachs (1979) demonstrated how formant spectral peaks are well preserved in synchronous phase locking activity across a population of auditory nerve fibers. Such a neural representation of temporal fine structure, possibly with the additional contribution of the olivocochlear bundle (Winslow and Sachs, 1987), may be less affected by stochastic background noise than the more place-based neural encoding of the spectral envelope delivered by a CI. Along this line of reasoning, perhaps the relatively modest shift in formant estimates due to the effects of noise reported by Parikh and Loizou (2005) after rejecting non-vowel like samples is more consistent with the representation of formants in auditory nerves of normal hearing listeners. In contrast, the measures in Fig. 2 of the present study may be more representative of the formant information received by CI users. Last, another possibility as to why vowel identification by normal hearing listeners would not be substantially affected at moderate noise levels is that they have access to a broad range of acoustic cues and are able to shift weighting of cues when presented with distortion of the acoustic signal (Winn *et al.*, 2012).

2. Modeling assumptions

In experiment 1, CI users' vowel scores in noise were predicted assuming JND parameters that did not change with listening condition (i.e., they were fixed to the best-fit JND parameter values obtained using subjects' confusion matrices in quiet). Is this a reasonable assumption, given that formant frequency discrimination is degraded in noise (e.g., Liu and Kewley-Port, 2004)? It is possible to argue that this assumption is reasonable because when a vowel or a formant is masked by noise the resulting stimulus has different physical characteristics than the original clean signal. The fact that formant frequency discrimination "degrades" in noise is at least partly related to the fact that the stimulus itself (with noise added) is more difficult to discriminate than the original clean signal. However, degradation of the acoustic stimulus does not necessarily mean that the underlying ability of a listener to discriminate place of stimulation in the cochlea has been affected. By optimizing the JND parameters for model fits to each subject's data in the quiet condition, we are controlling for individual factors that affect CI users' discrimination for place of stimulation in the cochlea such as electrode placement, cochlear size, neural survival, etc. (Bierer, 2010; Long *et al.*, 2014). From a modeling perspective it is more parsimonious to minimize the number of degrees of freedom, and assuming an underlying constant JND for place of stimulation allows us to do so. Providing model predictions by allowing JND parameters to vary

freely with different noise conditions provides, at best, a *post hoc* description of the data. That is, one would not know the necessary increase in JND with noise without first obtaining the data in noise. On the other hand, using a fixed JND parameter within the context of stimulus and response centers that change with noise, as in the present study, can provide a fairly good *a priori* prediction of the data. The accuracy of our predictions in noise based exclusively from subjects' data in quiet suggests that degradation of vowel identification in noise is closely tied to the fact that the stimuli themselves are degraded. In other words, $F1$ and $F2$ as estimated from the electrical stimulation patterns are strongly affected by background noise (see Fig. 2).

In our modeling approach, optimal predictions were obtained when stimulus and response centers were assumed equal within each listening condition. The stimulus centers were the average values of mean formant energy locations along the implanted array for each vowel, and the response centers were listeners' expectations of which formant values were best exemplars for each vowel. As the stimulus centers shifted between quiet and noise listening conditions (as in Fig. 2), the response centers that produced the best-fit models to subjects' data did also. This would mean that listeners' expectations of how vowels should sound were dependent on the level of background noise. Presumably, CI listeners were able to adjust their expectations based on the level and nature of the background noise. The extent of shift in vowels displayed in Fig. 2 seems large, but the amount of shift experienced by a CI user becomes less dramatic when considered in terms of electrode locations. For example, the average shift between formant locations of vowels in the quiet condition and vowels in +5 dB SNR amounted to less than one electrode unit with a standard deviation of ± 1 electrode unit. Nevertheless, the difference in selection of response centers produced a dramatic difference in model predictions (Fig. 3 vs Fig. 4). Although this may be a computational nuance of the framework that underlies the model, it may also point to flexibility in how CI users develop central representations of speech sounds. This may be true particularly for vowel perception which, in listeners with normal hearing, has been shown to be relatively more continuous than categorical, i.e., such that two vowels that receive the same category label are nevertheless discriminable (Feldman *et al.*, 2009).

B. Experiment 2: MPI model performance in quiet vs noise

Although the modeling approach in the present study proved successful in predicting vowel percent correct scores in CI users based on their performance in quiet, the ability of the model to predict vowel confusions became poorer in the noise conditions. As shown in Fig. 6, a minimized RMS difference was significantly higher in the noise conditions than in the quiet conditions. This result could simply be because it is more difficult to model random errors encountered in the noisy conditions than the more patterned confusions that occur in the quiet condition. Alternatively, it is possible that the speech cues employed in the present study, i.e., formant energy locations along the implanted array, may not be as predictive of CI

users' vowel confusions in noise as they are in quiet. Normal hearing listeners can utilize a large number of acoustic cues when identifying speech sounds, and the particular cues they employ in quiet listening conditions may not be the same as the cues they employ in noisy listening conditions (Ferguson and Kewley-Port, 2002; Parikh and Loizou, 2005; Swanepoel *et al.*, 2012). Although CI users are more limited in the number and type of speech cues they can employ to understand speech, the speech cues they use to understand speech in noise may differ than those used for quiet.

C. Future directions

Future avenues to explore using the analysis of the present study include the roles of vowel duration and relative weighting between speech cues on CI users' perception of vowels in noise. Just as formants are affected by noise, one could expect that vowel duration may also be affected. With increasing background noise, lower intensity portions of the vowel near the beginning and end will be less audible thereby decreasing the vowel duration as it would be perceived by the CI user. Hence, a longer duration vowel may be more likely confused with a shorter duration vowel in noise than in quiet. In the present study, formants were assumed to be equally weighted. It is possible to incorporate relative weighting between speech cue dimensions, particularly in the decision component of the model. Given that listeners with normal hearing can utilize different perceptual weighting of speech cues when understanding speech in noise in comparison to quiet (Ferguson and Kewley-Port, 2002; Parikh and Loizou, 2005; Swanepoel *et al.*, 2012), it may be worthwhile to explore this ability in CI users. Last, the present study offers an immediate practical application. One can use the model of the present study to predict the effect of different noise-reduction algorithms on CI users' vowel scores. To do this, one would obtain electrodiagrams of vowel stimuli in noise for each noise-reduction algorithm of interest, measure formants from those electrodiagrams, and use the model of the present study to generate predicted percent correct scores. This procedure could be done using a generic set of input JND parameters, or with a set of best-fit JND parameters obtained from a group of CI users using vowel stimuli in quiet.

ACKNOWLEDGMENTS

This research was supported by a grant from the U. S. Department of Education, NIDRR Grant No. H133F090031 (PI: E.S.), by a grant from NIH/NIDCD Grant No. R01-DC03937 (PI: M.A.S.), and by a research contract from Cochlear Americas (PI: J. Thomas Roland, Jr.). We would like to thank Susan Waltzman, Bill Shapiro, and the audiologists at the NYU Cochlear Implant Center. We also thank Mahan Azadpour, David Landsberger, Arlene Neuman, and Ann Todd for their helpful comments.

Bierer, J. A. (2010). "Probing the electrode-neuron interface with focused cochlear implant stimulation," *Trends Amplif.* **14**(2), 84–95.
Boersma, P. (2001). "Praat, a system for doing phonetics by computer," *Glott Int.* **5**, 341–345.

Chistovich, L. A., and Lublinskaya, V. V. (1979). "The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli," *Hear. Res.* **1**, 185–195.
Cochlear Corporation and the University of Iowa (1995). "Cochlear Corporation/the University of Iowa Revised Cochlear Implant Test Battery," Englewood, CO.
Donaldson, G. S., Rogers, C. L., Cardenas, E. S., Russell, B. A., and Hanna, N. H. (2013). "Vowel identification by cochlear implant users: Contributions of static and dynamic spectral cues," *J. Acoust. Soc. Am.* **134**, 3021–3028.
Donaldson, G. S., Rogers, C. L., Johnson, L. B., and Oh, S. H. (2015). "Vowel identification by cochlear implant users: Contributions of duration cues and dynamic spectral cues," *J. Acoust. Soc. Am.* **138**, 65–73.
Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiol.* **40**, 148–157.
Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference," *Psychol. Rev.* **116**(4), 752–782.
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**, 259–271.
Fitzgerald, M. B., Shapiro, W. H., McDonald, P. D., Neuburger, H. S., Ashburn-Reed, S., Immerman, S., Jethanamest, D., Roland, J. T., and Svirsky, M. A. (2007). "The effect of perimodiolar placement on speech perception and frequency discrimination by cochlear implant users," *Acta Oto-Laryngol.* **127**, 378–383.
Geibler, G., Arweiler, I., Hehrmann, P., Lenarz, T., Hamacher, V., and Buchner, A. (2015). "Speech reception threshold benefits in cochlear implant users with an adaptive beamformer in real life situations," *Cochlear Implants Int.* **16**, 69–76.
Gifford, R. H., and Revit, L. J. (2010). "Speech perception for adult cochlear implant recipients in a realistic background noise: Effectiveness of preprocessing strategies and external options for improving speech recognition in noise," *J. Am. Acad. Audiol.* **21**, 441–451.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.* **97**, 3099–3111.
Iverson, P., Smith, C. A., and Evans, B. G. (2006). "Vowel recognition via cochlear implants and noise vocoders: Effects of formant movement and duration," *J. Acoust. Soc. Am.* **120**, 3998–4006.
Jenkins, J. J., Strange, W., and Edman, T. R. (1983). "Identification of vowels in 'vowel-less' syllables," *Percept. Psychophys.* **34**, 441–450.
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**, 3487–3497.
Liu, C., and Kewley-Port, D. (2004). "Formant discrimination in noise for isolated vowels," *J. Acoust. Soc. Am.* **116**, 3119–3129.
Long, C. J., Holden, T. A., McClelland, G. H., Parkinson, W. S., Shelton, C., Kelsall, D. C., and Smith, Z. M. (2014). "Examining the electro-neural interface of cochlear implant users using psychophysics, CT scans, and speech understanding," *J. Assoc. Res. Otolaryngol.* **15**, 293–304.
Mahmoud, A. F., and Ruckenstein, M. J. (2014). "Speech perception performance as a function of age at implantation among postlingually deaf adult cochlear implant recipients," *Otol. Neurotol.* **35**, e286–e291.
Mauger, S. J., Warren, C. D., Knight, M. R., Goorevich, M., and Nel, E. (2014). "Clinical evaluation of the Nucleus® 6 cochlear implant system: Performance improvements with SmartSound iQ," *Int. J. Audiol.* **53**, 564–576.
Nearey, T. M., and Assmann, P. F. (1986). "Modeling the role of inherent spectral change in vowel identification," *J. Acoust. Soc. Am.* **80**, 1297–1308.
Parikh, G., and Loizou, P. C. (2005). "The influence of noise on vowel and consonant cues," *J. Acoust. Soc. Am.* **118**, 3874–3888.
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., 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.

- 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., Fitzgerald, M. B., Sagi, E., and Glassman, E. K. (2015). "Bilateral cochlear implants with large asymmetries in electrode insertion depth: Implications for the study of auditory plasticity," *Acta-Oto-Laryngol.* **135**, 354–363.
- Svirsky, M. A., Sagi, E., Meyer, T. A., Kaiser, A. R., and Teoh, S. W. (2011). "A mathematical model of medial consonant identification by cochlear implant users," *J. Acoust. Soc. Am.* **129**, 2191–2200.
- Swanepoel, R., Oosthuizen, J. J., and Hanekom, J. J. (2012). "The relative importance of spectral cues for vowel recognition in severe noise," *J. Acoust. Soc. Am.* **132**, 2652–2662.
- Syrdal, A. K., and Gopal, H. S. (1986). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. Acoust. Soc. Am.* **79**, 1086–1100.
- Winn, M. B., Chatterjee, M., and Idsardi, W. J. (2012). "The use of acoustic cues for phonetic identification: Effects of spectral degradation and electric hearing," *J. Acoust. Soc. Am.* **131**, 1465–1479.
- Winslow, R. L., and Sachs, M. B. (1987). "Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise," *J. Neurophysiol.* **57**(4), 1002–1021.
- Won, J. H., Tremblay, K., Clinard, C., Wright, R., Sagi, E., and Svirsky, M. (2016). "The neural encoding of formant frequencies contributing to vowel identification in normal-hearing listeners," *J. Acoust. Soc. Am.* **139**, 1–11.
- Young, E. D., and Sachs, M. B. (1979). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Am.* **66**, 1381–1403.
- Zahorian, S. A., and Jagharghi, A. J. (1993). "Spectral-shape features versus formants as acoustic correlates for vowels," *J. Acoust. Soc. Am.* **94**, 1966–1982.