

# The effects of frequency-place shift on consonant confusion in cochlear implant simulations

Ning Zhou, Li Xu,<sup>a)</sup> and Chao-Yang Lee

*School of Hearing, Speech and Language Sciences, Ohio University, Athens, Ohio 45701*

(Received 1 April 2009; revised 26 April 2010; accepted 4 May 2010)

The effects of frequency-place shift on consonant recognition and confusion matrices were examined. Frequency-place shift was manipulated using a noise-excited vocoder with 4 to 16 channels. In the vocoder processing, the location of the most apical carrier band varied from the matched condition (i.e., 28 mm from the base of the cochlear) to a basal shift (i.e., 22 mm from the base) in a step size of 1 mm. Ten normal-hearing subjects participated in the 20-alternative forced-choice test, where the consonants were presented in a /Ca/ context. Shift of 3 mm or more caused the consonant recognition scores to decrease significantly. The effects of spectral resolution disappeared when the amount of shift reached  $\geq 3$  mm. Information transmitted for voicing and place of articulation varied with spectral shift and spectral resolution, while information transmitted for manner was affected only by spectral shift but not spectral resolution. Spectral shift has shown specific effects on the confusion patterns of the consonants. The direction of errors reversed as spectral shift increased and the patterns of reversal were consistent across channel conditions. Overall, transmission of the consonant features can be accounted for by the acoustic features of the speech signal. © 2010 Acoustical Society of America. [DOI: 10.1121/1.3436558]

PACS number(s): 43.71.Es [KWG]

Pages: 401–409

## I. INTRODUCTION

Speech perception has proven to be fairly robust when the speech signal is distorted or reduced in information (e.g., Van Tasell *et al.*, 1987; ter Keurs *et al.*, 1992, 1993; Baer and Moore, 1993). It has been shown that good speech understanding can be achieved with greatly reduced spectral resolution (Shannon *et al.*, 1995). The limited spectral information can be compensated with increased temporal information in the perception of degraded speech signals (Xu *et al.*, 2005; Xu and Zheng, 2007; Xu and Pfungst, 2008). In addition, speech recognition from spectrally distorted signals that involve a number of forms of frequency-place mismatch has been studied (e.g., Dorman *et al.*, 1997; Shannon *et al.*, 1998; Fu and Shannon, 1999). In normal hearing, frequency components of an acoustic signal excite particular places in the cochlea in a tonotopic fashion. In electric hearing with cochlear implants, ideally, speech signals should be delivered to excite the appropriate places in the cochlea that match the frequency content of the acoustic signal. However, as a result of shallow insertion of the electrode array, unequal electrode-to-neuron distances, or compression of frequency maps, a number of frequency-place mismatch situations may presumably occur in cochlear implant stimulations (e.g., Dorman *et al.*, 1997; Shannon *et al.*, 1998, 2002; Fu and Shannon, 1999; Huss and Moore, 2005; Başkent and Shannon, 2003, 2004, 2005, 2006). Shallow insertion of the electrode array may result in spectral shift. In addition, in clinical mapping, the speech spectrum is typically compressively assigned to the electrode array, since the electrode array is too short to

cover the entire speech spectrum. Localized fiber loss and current spread, commonly found in implanted ears, may cause frequency warping.

In acoustic simulations of cochlear implants, the speech signal is analyzed in a number of spectral channels that are referred to as the analysis bands. The output from each band is rectified and lowpass filtered to extract the temporal envelope. The temporal envelope is used to amplitude modulate a white noise in a noise-excited vocoder or a pure tone in a tone-excited vocoder. The modulated signal is then bandpass filtered through the carrier bands. A full insertion of cochlear implant is simulated by matching the analysis bands and carrier bands in frequency. Simulation of shallow insertion is realized by shifting the carrier bands to higher frequencies relative to the analysis bands.

Research has shown that the basal spectral shift presumably resulting from shallow insertion of implants has an immediate detrimental effect on English speech recognition. Dorman *et al.* (1997) simulated four shallow insertion depths of a cochlear implant using a five-channel tone vocoder. They have shown that performance of sentence and vowel recognition progressively worsened as the simulated insertion depth became shallower. Performance of insertion depths of 22 and 23 mm (i.e., 3 to 4 mm basal shift) significantly differed from that of full insertion. In one of the experiments by Shannon *et al.* (1998), an 8-mm basal shift was simulated alone to be compared with a full insertion condition in a 4-channel noise-excited vocoder. Vowel recognition accuracy was significantly reduced and sentence recognition accuracy almost became zero. Similar results of vowel recognition were reported by Fu and Shannon (1999) that performance dropped from 80% correct in full insertion to 20% correct after the vowel spectrum was basally shifted by ap-

<sup>a)</sup>Author to whom correspondence should be addressed. Electronic mail: xul@ohio.edu

proximately 7 mm along the cochlea. A significant decrease in vowel recognition was found when the spectrum was basally shifted by 3 mm or more.

There seems to be a consensus in the literature indicating that vowel and sentence recognition are greatly susceptible to frequency-place shift as a result of shallow insertion (Dorman *et al.*, 1997; Shannon *et al.*, 1998; Fu and Shannon, 1999). As carriers of the temporal envelopes shift, so does the location of the formant frequencies that serves as a critical acoustic correlate for vowel recognition. The reduced vowel recognition in turn partially accounts for the deteriorated sentence recognition. To avoid frequency-place shift, the shallow inserted electrodes have to be mapped with tonotopically matched speech content. Faulkner *et al.* (2003) has shown that in cases of shallow insertion, tonotopic mapping does not necessarily recover speech recognition, since a large part of low frequency content in the speech signal is lost. However, the acute effects of spectral shift can be compensated to a certain extent by training (Rosen *et al.*, 1999; Fu *et al.*, 2002; Fu and Galvin, 2003; Faulkner, 2006). That is, human brains have shown to be able to learn and adapt to the frequency-shifted speech signals.

In contrast, studies that examined the effects of spectral shift on consonant recognition have yielded somewhat mixed findings. Dorman *et al.* (1997) reported that consonant recognition scores underwent smaller decreases compared to vowel and sentence recognition, but the decrease was still found to be significant with a shift of 2 mm or more. Consistent with the findings by Dorman *et al.* (1997), Rosen *et al.* (1999) reported significantly reduced consonant recognition performance as a result of basal shift, although the decrease was smaller than that for vowels and sentences. Rosen *et al.* (1999) reasoned that consonants can be recognized by the use of temporal information and gross spectral contrast, which renders consonants more immune to frequency-place shifts. Shannon *et al.* (1998), however, found in their study that while an 8-mm basal shift caused vowel recognition to greatly reduce and caused sentence recognition accuracy to essentially drop to zero, the performance of consonant recognition was nearly intact. It is possible that the discrepancies between these studies are due to the different frequency allocations or the use of different speech materials.

Given the discrepancies found in the previous studies, the effects of spectral resolution (i.e., number of channels) and spectral shift (i.e., frequency-place shift) were further examined on consonant recognition and on consonant features. We assumed that distortion of the frequency spectrum and varying spectral resolution could have a differential effect on information transmission of different articulatory features of consonants. Data were analyzed in terms of detailed articulatory features within the broad categories of manner, place, and voicing, which have not been reported in the previous studies. It was hypothesized that features transmitted using predominant temporal information would be less affected by spectral resolution or spectral shift. On the other hand, features that rely more on spectral information would be increasingly affected as the amount of spectral shift increases or the spectral resolution becomes poorer. Further,

none of the previous studies have reported consonant error patterns in conditions that varied the degree of spectral resolution and the amount of spectral shift combined. The primary question addressed was whether consonant confusions would vary in systematic patterns with spectral resolution and spectral shift.

## II. METHODS

### A. Speech materials and signal processing

From the database recorded by Shannon *et al.* (1999), a set of digitized naturally produced consonants from one female (#3) and one male (#3) speaker was drawn. The set contained 20 consonants ( $\text{tʃ}$ ,  $\text{dʒ}$ ,  $\text{t}$ ,  $\text{d}$ ,  $\text{k}$ ,  $\text{g}$ ,  $\text{p}$ ,  $\text{b}$ ,  $\text{n}$ ,  $\text{m}$ ,  $\text{s}$ ,  $\text{z}$ ,  $\text{f}$ ,  $\text{v}$ ,  $\text{ð}$ ,  $\text{l}$ ,  $\text{r}$ ,  $\text{j}$ ,  $\text{w}$ ) produced in a /Ca/ context, resulting in 40 speech tokens in total (20 consonants  $\times$  2 speakers). The speech tokens were subjected to a noise-excited vocoder processing. Signal processing was performed in MATLAB (MathWorks, Natick, MA). The speech signals were pre-emphasized by highpass filtering at 1200 Hz (1st-order Butterworth filter, 6 dB/octave) and divided into 4, 8, 12, or 16 frequency bands. The frequency range of the analysis bands was 269–2113 Hz, covering a tonotopic location between 28 mm and 16 mm from the basal end in a 35 mm long cochlea (Greenwood, 1990). The Greenwood (1990) formula,  $F = 165.4(10^{0.06x} - 1)$  where  $x$  is the distance in mm from the apex, was used to determine the bandwidths and corner frequencies of the analysis bands. The output of each analysis band was half-wave rectified and then low-pass filtered at 160 Hz (2nd-order Butterworth, 12 dB/octave) to extract the temporal envelope. Each temporal envelope was used to amplitude modulate a white noise. The modulated signal was then bandpass filtered into either the same band where the envelope was extracted to simulate a tonotopically matched condition, or bandpass filtered into higher frequency bands to simulate frequency-place shifted conditions. The cut-off frequencies of carrier bands were also estimated by the Greenwood (1990) formula. The frequency allocation of the carrier bands was systematically manipulated to simulate a basal shift from the analysis bands over a tonotopic distance of 6 mm with a step size of 1 mm. Thus, seven frequency-place matching conditions (i.e., unshifted and 6 shifted) were created. The manipulation was repeated for all four channel conditions (i.e., 4, 8, 12, and 16). The frequency allocation for the carrier bands of 16 channels is provided in Table I. Finally, the outputs of all bands were summed up and stored on computer for acoustic presentations.

### B. Subjects and test procedure

Ten native English-speaking subjects recruited from the Ohio University student population participated in the study. The subjects were screened for normal-hearing ( $\leq 20$  dB HL) at octave frequencies between 250 and 8000 Hz. The use of human subjects was reviewed and approved by the Ohio University Institutional Review Board.

The consonant recognition test was conducted in an IAC sound booth. A graphic user interface was developed in MATLAB to present stimuli and collect responses from the subjects. The consonant stimuli were presented to the left ear of

TABLE I. Corner frequencies of the carrier bands for seven frequency-place shift conditions in a 16-band processor. The frequency allocations of the eight- and four-band processors can be derived from this table by combining adjacent two and four bands, respectively.

		Carrier bands																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
Shift (mm)	0	269	316	368	427	492	563	643	731	829	938	1058	1192	1340	1504	1686	1884	2113
	1	333	388	448	515	589	671	763	864	976	1101	1239	1393	1563	1751	1961	2193	2450
	2	407	470	539	616	701	800	900	1017	1146	1289	1447	1624	1819	2035	2276	2542	2838
	3	492	564	643	731	829	938	1058	1192	1340	1504	1686	1888	2113	2361	2637	2843	3282
	4	589	672	763	864	977	1101	1239	1393	1563	1751	1961	2173	2450	2736	3052	3404	3793
	5	701	800	900	1017	1146	1289	1447	1624	1819	2035	2276	2542	2838	3165	3529	3932	4380
	6	829	938	1058	1198	1340	1504	1686	1888	2113	2361	2537	2943	3282	3659	4076	4392	5053

the listeners via a circumaural headphone (Sennheiser, HD 265) at a most comfortable level. The subjects adjusted the soundcard output levels to their respective most comfortable levels before each session of training or test. The presentation level was approximately  $65 \pm 5$  dB SPL. The task of the subjects was to identify the consonant they had heard by clicking one of the 20 buttons labeled with the CV strings (e.g., “Ba,” “Da,” “Ga,” etc.). The subjects were trained with the tonotopically matched stimuli processed with 4, 8, 12, and 16 channels. Each training session lasted about 30 min which always started with presumably the easiest stimuli (i.e., the 16-channel condition) followed by progressively more difficult stimuli. During training, within each channel condition, the presentation of stimuli was randomized. After the subject gave a response, the button of the correct consonant flashed to provide feedback. Given the reduced bandwidth of the stimuli, and based on experiences from our previous studies and literature, training was considered adequate when recognition of the tonotopically matched stimuli reached 60% correct. On average, each subject took approximately 10 h in training before continuing with the test. In the test, the presentation of the 5600 stimuli (i.e., 20 consonants  $\times$  2 speakers  $\times$  4 channel conditions  $\times$  7 frequency-place shift conditions  $\times$  5 repetitions) was completely randomized. The test was divided into a number of sessions and took each subject approximately 5–6 h to complete.

### C. Consonant feature coding

The 20 consonants were coded according to their articulatory features, which included voicing (i.e., voiced and voiceless), place (i.e., labial, alveolar, palatal, and velar) and manner of articulation (i.e., stop, fricative, affricate, nasal, and glide) (Ladefoged, 1975). The consonants were coded as shown in Table II.

TABLE II. Coding of consonants. Manner of articulation is coded from 1–5 as: stop, fricative, affricate, nasal, and glide. Place of articulation is coded from 1–4 as: labial, alveolar, palatal, and velar. Voicing is coded 1 vs. 0 for voiced vs. voiceless consonants.

		Consonants																			
		ʈ	ɖ	t	d	k	g	p	b	n	m	s	z	ʃ	f	v	ð	l	r	j	w
Voicing	0	1	0	1	0	1	0	1	1	1	0	1	0	0	1	1	1	1	1	1	1
Manner	3	3	1	1	1	1	1	1	1	4	4	2	2	2	2	2	2	5	5	5	5
Place	3	3	2	2	4	4	1	1	2	1	2	2	3	1	1	2	2	3	3	3	4

## III. RESULTS

### A. Effects of spectral shift and spectral resolution

Figure 1(a) summarizes the percent correct scores of various channel and spectral shift conditions. A two-way repeated-measure ANOVA indicated significant effects of both number of channels [ $F(3, 27)=22.84, p<0.00001$ ] and spectral shift [ $F(6, 54)=144.06, p<0.00001$ ]. Post-hoc analysis of the main factors revealed that spectral shift of  $\geq 3$  mm caused the performance to significantly decrease from the tonotopically matched condition (i.e., 28 mm) ( $p<0.05$ ). Performance with 4 channels was significantly lower than all other channel conditions ( $p<0.05$ ), while the performance of 8, 12, and 16 channels did not show significant differences from each other ( $p>0.05$ ). ANOVA showed that the interaction between the two factors was also statistically significant [ $F(18, 162)=13.39, p<0.00001$ ]. Post-hoc analysis of the interaction revealed that when the spectral shift was  $>3$  mm, the spectral resolution of the signals no longer seemed to play a role and therefore the performance of all channel conditions did not differ ( $p>0.05$ ).

### B. Information transmission analysis

Information transmission scores (%) are shown in Fig. 1(b)–1(d) for the features of voicing, manner, and place of articulation in all experimental conditions (Miller and Nicely, 1955). Three sets of two-way repeated-measure ANOVA were conducted to examine the effects of number of channels as well as the effects of place-frequency shift for all three features. For the feature of voicing, the effects of number of channels [ $F(3, 27)=5.38, p=0.005$ ] and spectral shift [ $F(6, 54)=4.64, p=0.0007$ ] were both found to be statistically significant. For the feature of manner of articulation, only the effects of spectral shift [ $F(6, 54)=4.25, p=0.001$ ], but not number of channels [ $F(3, 27)=1.31, p=0.3$ ] were

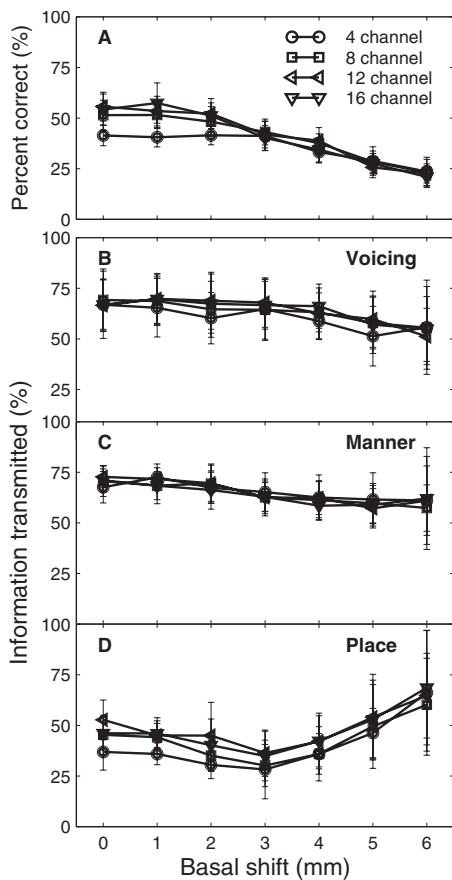


FIG. 1. Percent correct scores and information transmission scores for articulatory features ( $N=10$ ): (a) averaged percent correct scores are plotted as a function of frequency-place shift. Scores of four different channel conditions are plotted in solid lines with different symbols; ((b)–(d)) averaged information transmission scores are plotted as a function of frequency-place shift for the features of voicing, manner of articulation, and place of articulation. Scores of four different channel conditions are plotted in solid lines with different symbols. Error bars represent SDs.

found to be significant. For the feature of place of articulation, again, both main factors were found to be significant [ $F(3, 27)=17.83$ ,  $p < 0.00001$ ;  $F(6, 54)=9.08$ ,  $p < 0.00001$ ]. For all three features, the interactions between the two main effects were not statistically significant [voicing:  $F(18, 162)=1.03$ ,  $p=0.43$ ; manner:  $F(18, 162)=1.14$ ,  $p=0.32$ ; place:  $F(18, 162)=1.22$ ,  $p=0.25$ ].

Figure 2 shows the information transmitted as a function of frequency-place shift for each specific manner of articulation (i.e., stop, fricative, affricate, nasal, and glide) and each specific place of articulation (i.e., labial, alveolar, palatal, and velar). Each specific manner and place of articulation category (e.g., stop) was treated as a binary feature (e.g., stops vs. non-stops) in the derivation of their information transmission scores. Voicing is already a binary feature that has been described in Fig. 1. The effect of spectral resolution is not shown in Fig. 2, since it was not a significant factor for transmitting the feature of manner, and there was no interaction between channels and spectral shift for the feature of place of articulation. Figure 2 indicates that the non-monotonic function of place of articulation overall (Fig. 1(d)) is accounted for by the labial feature. The reason that the labial feature improved after 3 mm shift is elaborated in Discussion Section C.

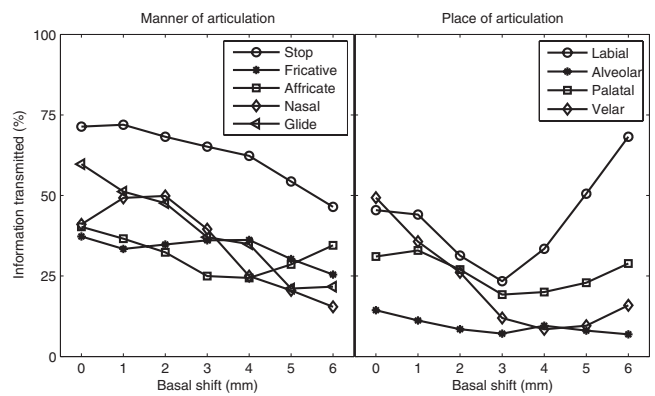


FIG. 2. Information transmission for particular manner and place of articulation ( $N=10$ ). Left panel: Information transmitted for sub-categories of manner of articulation is plotted as a function of frequency-place shift in different symbols. Right panel: Information transmitted for sub-categories of place of articulation is plotted as a function of frequency-place shift in different symbols.

### C. Confusion analysis

Confusion matrices of data pooled across all channel conditions are shown in Fig. 3 with different panels for different spectral shift conditions. Note that the quantitative analysis of the error patterns described below was conducted for each spectral shift and number of channel condition in order to elucidate the effects of both factors on confusions. In fact, such analysis revealed that the confusion error patterns at different spectral resolutions were similar (see below).

In Fig. 3, the confusion matrices were organized based on features of manner and place of articulation. First, the consonants that are of the same manner of articulations were grouped together. Boundaries between manners of articulation are indicated by the white squares. Within each group of manner of articulation, the consonants were then sorted following the order of alveolar, palatal, velar, and labial, if applicable. This place order reflects the frequency region of the acoustic correlates that are the most relevant to the perception of place of articulation (Stevens, 1998; Pickett, 1999; Johnson, 2003).

Specifically, one of the primary acoustic correlates of place distinction of stops is the spectral dominance of the short-term spectrum at consonantal release (Fant, 1973; Stevens and Blumstein, 1978). Such acoustic cues are usually associated with specific articulatory actions. For example, alveolar stops are produced with a relatively short front cavity. Therefore, they have the spectral peak located in a relatively high frequency range. In contrast, velar stops are produced with a longer front cavity with spectral prominence in a lower frequency range. The production of labial stops involves virtually no front vocal cavity, since the constriction is made at the lips. Therefore, their spectra show a diffused pattern, with acoustic energy distributed over the low frequency range. For nasals, the frequencies of anti-formants are associated with place distinctions. Anti-formants are local spectral energy minimum that arises as a result of the oral cavity absorbing acoustic energy at specific frequency ranges from the nasal resonance system. The frequencies of the anti-formants are associated with the length of the oral cavity. For

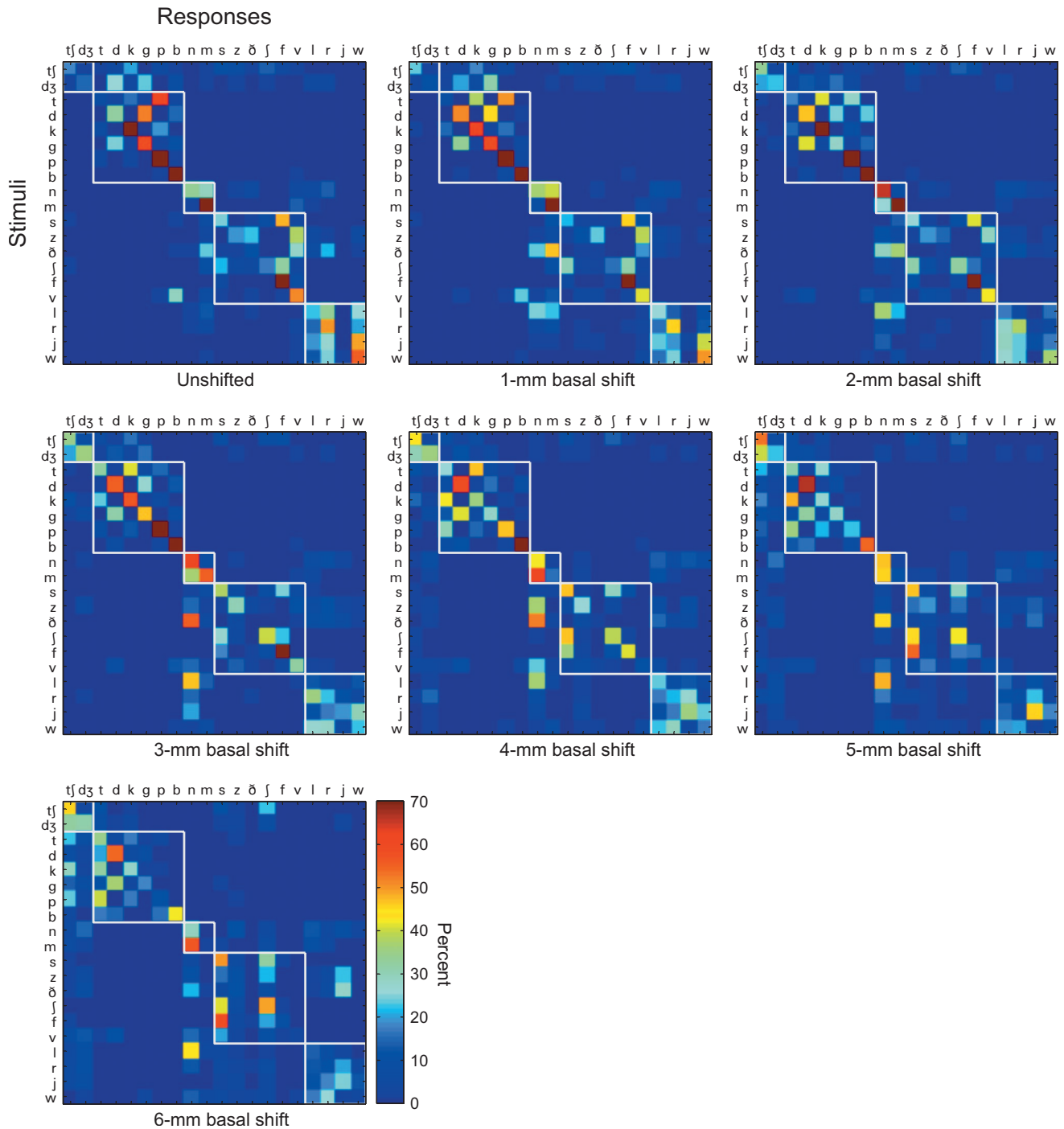


FIG. 3. Confusion matrices using data pooled across channel conditions ( $N=10$ ). The white squares indicate boundaries of manner of articulation. Within each manner of articulation, the consonants were sorted, whenever appropriate, based on place of articulation following the order of alveolar, palatal, velar, and bilabial. The color scale in each cell represents the value in it with reference to the color bar on the right of the bottom panel. The value in the cell of row  $j$  and column  $k$  is the percent of times stimulus  $j$  was recognized as  $k$  ( $j=1:20$ ;  $k=1:20$ ).

example, the first anti-formant for the labial nasal /m/ is lower than the first anti-formant for the alveolar /n/ because the former involves a longer oral cavity (Fujimura, 1962). For fricatives, the association between spectral prominence and place of constriction has also been identified (Heinz and Stevens, 1961). The spectral dominance is located at the highest frequencies for alveolars followed by palatals and bilabials. In sum, the acoustic dominance or attenuation associated with place of articulation is generally located at the highest frequency for alveolar, followed by palatal, velar,

and labial consonants. A spectral analysis of the original stimuli used in this study confirmed the differential spectral characteristics associated with place of articulation as suggested by the acoustical theory of speech production.

The consonant recognition data showed that error patterns changed systematically as the spectral shift increased (Fig. 3). In the tonotopically matched condition, the confusions were to a large extent confined within each manner of articulation boundary. As the spectral shift increased, confusions started to cross the boundaries of manner of articula-

tion. More interestingly, a reversal of error patterns was observed as a result of spectral shift. For example, in the matched condition, stimulus /t/ frequently produced error responses of /k/. As the shift increased, there were still /t-/k/ errors, but the direction of the confusion was reversed. That is, /k/ started to induce error response of /t/. For almost all confusions that occurred in the tonotopically matched condition, the direction of the confusion was gradually reversed as the shift increased. If each confusion matrix was separated into two triangles along the main diagonal, the errors seemed to be largely confined within the upper triangle in the matched condition and mostly to the lower confusion triangle as the spectral shift increased.

The reversal of the errors could best be described as the transposition of the confusion matrix and could be quantified by correlation analyses. A correlation analysis was performed for each of the four channel conditions (Fig. 4, top four panels) and repeated for data pooled across channels (Fig. 4, bottom panel). For a given channel condition, the errors in the tonotopically matched condition were correlated with the errors in each of the shifted conditions. Therefore, six correlation coefficients were derived (Fig. 4, untransposed all). The descending function described that the error pattern existing in the matched condition gradually disappeared in the shifted conditions. Further, the confusion matrices of the six shifted conditions were transposed. The transposition exchanged the abscissa and ordinate of the elements in a matrix so that visually, the matrix flipped over. Again, the errors of the matched condition were correlated with the ones of the shifted conditions post transposition, yielding six correlation coefficients (Fig. 4, transposed all). The increase of the correlation reflected that the error pattern in the matched condition was gradually replaced with its reversal.

Since each consonant pair involved in the error reversal was the two consonants primarily of the same manner of articulation, the same correlation analyses described above were conducted again only for the errors that occurred within the categories of manner of articulation (i.e., errors within the squared boundaries). The resulting functions (Fig. 4, untransposed within manner and transposed within manner) in most cases showed steeper slopes than those that used all error elements in the matrices.

Note that the four functions were overall consistent across channel conditions, which indicated that the systematic error pattern was not resulted from spectral resolution, but primarily from spectral shift. The panel that plotted the correlation coefficients for the pooled data reflected the effects of spectral shift on consonant confusion alone.

## IV. DISCUSSION

### A. Effects of spectral shift and spectral resolution

The results of overall percent correct scores were in agreement with previous studies reporting significant effects of frequency-place shift (e.g., Dorman *et al.*, 1997; Fu and Shannon, 1999) and significant effects of spectral resolution (e.g., Xu *et al.*, 2005; Xu and Zheng, 2007; Fu and Shannon, 1999) on consonant recognition. Recognition accuracy was

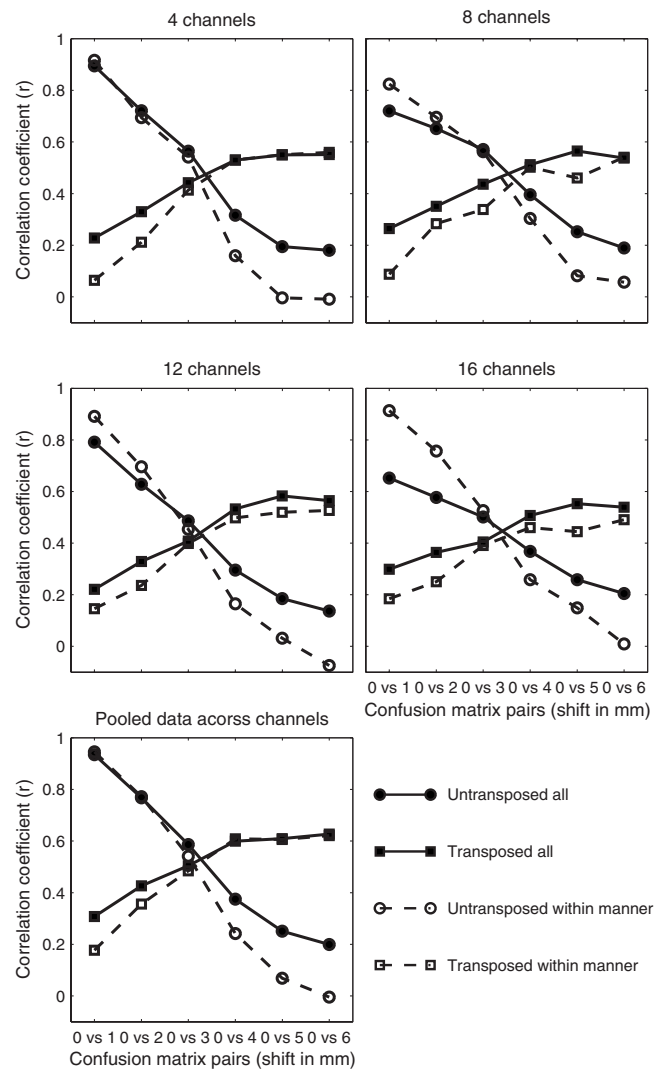


FIG. 4. Correlation between the confusion matrices ( $N=10$ ). The squares represent the correlation coefficient values between the confusion matrix in the tonotopically matched condition with the transposed confusion matrices of the six shifted conditions for all errors (filled symbols in a solid line) and for errors within manner of articulation (open symbols in a dashed line). The circles represent the correlation coefficient values between the confusion matrix in the tonotopically matched condition with the original confusion matrices of the six shifted conditions for all errors (filled symbols in a solid line) and for errors within manner of articulation (open symbols in a dashed line). The upper four panels plot correlations for each of the four channel conditions, and the bottom panel plots correlations for the pooled data across the channel conditions.

not affected by spectral shift until the shift was larger than 3 mm. Performance with higher spectral resolution (i.e., 8, 12, and 16 channels) was equivalent to each other across the spectral shift conditions. Interaction between spectral resolution and spectral shift was found to be significant. The interaction was also reported by Zhou and Xu (2008) that examined lexical tone recognition under similar experimental conditions. Performance with higher spectral resolution (i.e., 8, 12, and 16 channels) was better than the performance of 4 channels in the range of a small amount of shift (i.e.,  $\leq 3$  mm).

The overall accuracy of the shifted condition was lower than what was reported in previous studies. This was likely due to the overall reduced frequency range used in the cur-

rent study. Moreover, the consonants used (Shannon *et al.*, 1999) included affricates and more voiced fricatives that by nature are not only articulatory but also perceptually challenging. Another possibility was that the greater number of consonants tested here compared to the previous studies (Dorman *et al.*, 1997, Shannon *et al.*, 1998, Rosen *et al.*, 1999) increased the decision space and reduced the chance level.

## B. Information transmission for articulatory features

Overall, among the three broad articulatory features, voicing and manner were transmitted relatively well and were less affected by both spectral resolution and shift than place of articulation. This was probably because the transmission of voicing and manner is dependent more heavily on temporal information that is relatively better preserved in vocoder processing than spectral information. This result was consistent with previous studies (e.g., Xu *et al.*, 2005; Xu and Zheng, 2007).

In particular, the transmission of the voicing feature to a large extent is facilitated by temporal information of duration, which is minimally affected by the reduction of spectral resolution or distortion of place-frequency shift. The delay between stop release and the onset of voicing, known as the voice onset time, is a reliable temporal cue for voicing distinction of stop consonants (Lisker and Abramson, 1967). Furthermore, the duration of closure is longer for voiceless stops than voiced stops. Differences in duration of constriction also contribute to the voicing distinction of stops. In the present study, consonants were produced in a syllable initial position (i.e., CV). In natural speech, when the consonant is in an intervocalic position, time coordination of the vocal tract will result in a longer vowel and much shorter stop closure, if the stop is voiced rather than voiceless (Slis and Cohen, 1969). Voicing distinction of fricative consonants has similar time related acoustic correlates as stop consonants.

Although transmission of voicing primarily depends on temporal cues, there are spectral cues for voicing distinction such as low frequency energy and correlated periodicity. It is therefore reasonable to assume that better frequency resolution to a certain degree helps the transmission of voicing feature. This idea was supported by data that showed a small but significant effect of number of channels on voicing feature (Fig. 1(b)). Periodicity in the temporal envelopes, however, may not be well utilized when they are coded in shifted bands. In fact, Oxenham *et al.* (2004) suggested that the perception of temporal pitch largely depends on the tonotopic match between the temporal envelope and the places being excited. In addition, in shifted conditions, errors that typically occurred were that voiced consonants being recognized as voiceless in the manner categories of stop and fricative. This was probably because the low-frequency energy in voiced consonants disappeared as their spectra were shifted to higher frequency ranges.

As for manner, transmission was not dependent on spectral resolution at all, although spectral shift did play a role (Fig. 1(c)). It was hypothesized that information transmission would be better for specific features of manner that depend

less on spectral information. Of all, stop consonants had an overall high information transmission score (Fig. 2, left panel). The production of stops involves time events of a closure and release, which produces salient amplitude-time variation different from the other manners. In fact, based on a multidimensional scaling analysis, Van Tasell *et al.* (1987) showed that the burst amplitude of a broadband signal contributes to the categorization of stop consonants. Again, if the stop consonants are perceived in an intervocalic context, the relative silence during the closure is a cue that may contribute to some release from the spectral distortion. Van Tasell *et al.* (1992) also showed that amplitude envelope of a broadband signal distinguishes sonorant consonants from others, because the average amplitude of the sonorant consonants is higher than any other consonants including the voiced ones. Although the differences in manner of articulation have been primarily described in terms of the degree and area of constriction, it has been also suggested that the gesture differences in manner of articulation may also be accompanied by distinctions in constriction duration (Romero, 1993).

Dorman *et al.* (1990) compared the use of pure temporal cue (i.e., voicing, amplitude, and burst envelope) for consonant recognition in better versus poorer cochlear implants users since pure temporal cues are readily available in the cochlear implants stimulations. It was found that the better subjects were able to receive and make use of the temporal features more efficiently than the poorer subjects. More importantly, in cases of shallow insertion, the pure temporal cues including amplitude-time variation patterns and duration should not be affected by frequency-place shift. Therefore, they should provide some information for manner distinctions.

Our data indicated that transmission of place of articulation was weaker than that of manner and voicing. Place of articulation for some manners may be perceived based on temporal features (Ohde and Stevens, 1983). Nonetheless, acoustic cues for place of articulation are primarily spectral, including formant transition and spectral prominence in specific frequency ranges (Liberman, 1957; Stevens and Blumstein, 1978). Therefore, the number of channels influenced the perception of place distinction to a much greater degree than the perception of voicing and manner features (Fig. 1(d)). It is worth noting that transmission of place of articulation did not show a linear decrease as a function of frequency-place shift as the voicing and manner features did. As the shift increased, the transmission score reduced to the lowest level at 4 mm and returned to a higher level at 5 and 6 mm. This is accounted for primarily by the non-monotonic information transmission function of the labial consonants as shown in Fig. 2 (right panel) and the reason is elaborated in the following section. The alveolar consonants were perceived poorly in all conditions.

## C. Confusion matrices and reversed error patterns

Careful examination of the consonant confusions revealed an interesting and systematic error pattern associated with the shift of spectral energy of the consonants. Recall that in each confusion matrix, within each manner category,

the consonants were sorted again following their formant or anti-formant frequencies from high to low (Fig. 3). The re-organization of the confusion matrices allowed a clearer presentation of the error patterns. In the matched condition, confusions were primarily limited to consonants that were of the same manner and voicing. Within each manner category, it was almost always the case that the stimuli with relatively high formant (or anti-formant) frequencies were recognized as the ones with the lower formant (or anti-formant) frequencies. As a result of this re-organization of the matrix, errors in the matched condition appeared only in the upper confusion triangle of the matrix. In the tonotopically matched condition, the frequency range transmitted (i.e., 269–2113 Hz) appeared to be too narrow to include the formant frequencies of alveolar consonants, in particular. The alveolar and in a few cases, palatal consonants, reasonably induced error responses that have an overall similar acoustic features (i.e., same manner and voicing) but have formant frequencies that are located at the lower spectral regions. Specifically for example, /t/ was often confused as /k/, /d/ as /g/, and /z/ as /v/. As the frequency-place shift was introduced, the direction of the confusion occurred in the matched condition started to reverse. The reversal of the errors could be well explained by the nature of the spectral shift. The process of spectral shift virtually moved the location of the spectral peaks of the consonants to gradually higher frequency places. In these conditions, the stimuli with shifted spectra were likely to be perceived as consonants with relatively high formant (or anti-formant) frequencies. In the 6-mm shift condition, the spectral peaks of low formant consonants were shifted to that of the alveolar stop and voiceless fricative, which were in fact the most frequent error responses. Further shift of the spectrum may not necessarily induce more salient patterns of error reversal, since it may cause the spectrum to be shifted to exceed the frequency region of speech. Notice that the reversed errors were only confined to the consonant pairs that are of the same voicing and manner features. The reversal of the errors was the most obvious for stop and fricative consonants. These are the consonants that have a consistent relation between the formant frequency and the place of constriction.

The reversal of the errors was quantified by the correlation analyses as shown in Fig. 4. The reversal of the errors was quantified for each of the channel conditions to elucidate the role of spectral resolution in the patterns of error reversal. The quantification was repeated for the pooled data across channels (Fig. 4, bottom panel). The correlation coefficient using the transposed matrices demonstrated a linear increase with the spectral shift (i.e., Fig. 4, transposed all). It quantified the degree to which the transposed confusions of the shifted conditions resemble that of the matched condition. The increasing function suggested that errors present in the matched condition were gradually replaced with the reversed ones with spectral shift. The increase was even more robust when the correlation analyses were limited to the errors within the same manner of articulation (i.e., Fig. 4, transposed within manner). This means that error reversal involved primarily pairs of consonants within the same manner. The patterns of the error reversal (i.e., the increasing

functions) were consistent across the number of channel conditions. These patterns suggested that spectral resolution did not have an impact on the pattern of how the errors reversed. The correlation analyses using the untransposed matrices quantified the degree to which the confusion in the match condition persisted as a function of spectral shift. The decreasing functions of the pooled data and data of each channel conditions indicated that the error present in the matched condition gradually disappeared (i.e., untransposed all). This pattern also held true for each of the channel conditions within manner (i.e., untransposed within manner).

The information transmission curve of the labial consonants can now be explained. The labials were recognized the best in the matched condition, because the place-related spectral characteristics located at low frequency regions were covered in the limited frequency range transmitted (269–2113 Hz, see Table I). In the shifted conditions, as the spectra of the stimuli were shifted to higher frequencies, the stimuli were less likely to induce responses of low frequency labial consonants resulting in high rate of correct rejection of labials. This, in turn, explains the increasing information transmission scores for labials as a function of shift. The alveolar consonants, however, tended to be always involved in confusions across all conditions. In the matched condition, the alveolar consonants tended to be mis-identified because of frequency truncation, while in the shifted conditions they became false alarms of low frequency consonants. This accounted for the consistently low scores of the alveolar in all conditions.

#### D. Implications for cochlear implants

Current cochlear implant systems well preserve the temporal information of the signals. Nonetheless the spectral cues of speech can be distorted in many forms due to insertion depth, current spreading, and frequency compression. Spectral distortion may cause the neural firing profiles to greatly deviate from the spectrum of the acoustic signal. The results of this simulation study indicate that for consonant recognition, cochlear implant patients might have difficulties in perceiving place features of consonants, compared to manner and voicing features as a result of frequently-encountered shallow insertion of the electrode arrays. The error patterns observed in the simulation data, however, do not necessarily reflect the error patterns observed in cochlear implant users. Current interference as well as many other forms of spectral distortion besides spectral shift might cause perceptual errors. Caution must be exercised, therefore, when generalizing data from simulation studies to performance in the cochlear implant users.

#### V. CONCLUSION

Consonant recognition was affected by both frequency-place shift and spectral resolution in cochlear implant simulations. Broad features of voicing, manner, and place of articulation were all subject to the effects of frequency shift. However, the manner of articulation was not affected by spectral resolution. Pure temporal cues for articulatory features, such as duration, and amplitude-time variations, may

provide release from the effects of frequency-place shift as well as poor spectral resolution. Place of articulation was transmitted relatively poorly because of its heavily weighted spectral cues. Interesting patterns of error reversal were observed as a function of spectral shift. High frequency truncation may cause the consonants with high frequency spectral characteristics to be confused with the consonants of the same voicing and manner but with lower frequency spectral characteristics. On the other hand, basal spectral shift may cause the consonants with low frequency spectral characteristics to be confused with those with higher frequency spectral characteristics. Such systematic error reversal patterns were found to be consistent across channel conditions. Overall, the results support the notion that temporal and spectral features of the acoustic signals can reasonably account for consonant recognition performance in normal-hearing listeners.

## ACKNOWLEDGMENTS

Zinny Bond, Ken Grant (Associate Editor), and two anonymous reviewers provided constructive comments on an earlier version of the manuscript. Heather Schultz provided technique assistance in the preparation of the manuscript. The study was supported in part by NIH NIDCD Grant Nos. R15-DC009504 and F31-DC009919.

Baer, T., and Moore, B. C. J. (1993). "Effects of spectral smearing on the intelligibility of sentences in noise," *J. Acoust. Soc. Am.* **94**, 1229–1241.

Başkent, D., and Shannon, R. V. (2003). "Speech recognition under conditions of frequency-place compression and expansion," *J. Acoust. Soc. Am.* **113**, 2064–2076.

Başkent, D., and Shannon, R. V. (2004). "Frequency-place compression and expansion in cochlear implant listeners," *J. Acoust. Soc. Am.* **116**, 3130–3140.

Başkent, D., and Shannon, R. V. (2005). "Interactions between cochlear implant electrode insertion depth and frequency-place mapping," *J. Acoust. Soc. Am.* **117**, 1405–1416.

Başkent, D., and Shannon, R. V. (2006). "Frequency transposition around dead regions simulated with a noiseband vocoder," *J. Acoust. Soc. Am.* **119**, 1156–1163.

Dorman, F. M., Soli, S., Dankowski, K., Smith, M. L., McCandless, G., and Parkin, J. (1990). "Acoustic cues for consonant identification by patients who use the Ineraid cochlear implant," *J. Acoust. Soc. Am.* **88**, 2074–2079.

Dorman, M. F., Loizou, P. C., and Rainey, D. (1997). "Simulating the effect of cochlear implant electrode insertion depth on speech understating," *J. Acoust. Soc. Am.* **102**, 2993–2996.

Fant, G. (1973). *Speech Sounds and Features* (MIT Press, Cambridge, MA).

Faulkner, A. (2006). "Adaptation to distorted frequency-to-place maps: Implications of simulations in normal listeners for cochlear implants and electroacoustic stimulation," *Audiol. Neuro-Otol.* **11**, 21–26.

Faulkner, A., Rosen, S., and Stanton, D. (2003). "Simulations of tonotopically mapped speech processors for cochlear implant electrodes varying in insertion depth," *J. Acoust. Soc. Am.* **113**, 1073–1080.

Fu, Q.-J., and Galvin, J., III (2003). "The effects of short-term training for spectrally mismatched noise-band speech," *J. Acoust. Soc. Am.* **113**, 1065–1072.

Fu, Q.-J., Shannon, R., and Galvin, J., III (2002). "Perceptual learning following changes in the frequency-to-electrode assignment with the Nucleus-22 cochlear implant," *J. Acoust. Soc. Am.* **112**, 1664–1674.

Fu, Q.-J., and Shannon, R. V. (1999). "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," *J. Acoust. Soc. Am.* **105**, 1889–1900.

Fujimura, O. (1962). "Analysis of nasal consonants," *J. Acoust. Soc. Am.* **34**, 1865–1875.

Greenwood, D. D. (1990). "A cochlear frequency-position function for several species—29 years later," *J. Acoust. Soc. Am.* **87**, 2592–2605.

Heinz, J. M., and Stevens, K. N. (1961). "On the properties of voiceless fricative consonants," *J. Acoust. Soc. Am.* **33**, 589–596.

Huss, M., and Moore, C. J. B. (2005). "Dead regions and pitch perception," *J. Acoust. Soc. Am.* **117**, 3841–3852.

Johnson, K. (2003). *Acoustic and Auditory Phonetics*, 2nd ed. (Blackwell Publishing, Oxford).

Ladefoged, P. (1975). *A Course in Phonetics*, 5th ed. (Thomson/Wadsworth, Boston, MA).

Lieberman, A. M. (1957). "Some results of research on speech perception," *J. Acoust. Soc. Am.* **29**, 117–123.

Lisker, L., and Abramson, A. (1967). "Some effects of context on voice onset time in English stops," *Lang. Speech* **10**, 1–28.

Miller, G. A., and Nicely, P. (1955). "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.* **27**, 338–352.

Ohde, N. R., and Stevens, N. K. (1983). "Effect of burst amplitude on the perception of stop consonant place of articulation," *J. Acoust. Soc. Am.* **74**, 706–714.

Oxenham, J. A., Bernstein, J. G. W., and Penagos, H. (2004). "Correct tonotopic representation is necessary for complex pitch perception," *Proc. Natl. Acad. Sci. U.S.A.* **101**, 1421–1425.

Pickett, J. M. (1999). *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology* (Allyn & Bacon, Boston, MA).

Romero, J. (1993). "Duration aspects in manner of articulation distinctions," *J. Acoust. Soc. Am.* **94**, 1764.

Rosen, S., Faulkner, A., and Wilkinson, L. (1999). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," *J. Acoust. Soc. Am.* **106**, 3629–3636.

Shannon, R. V., Galvin, J. J., III, and Baskent, D. (2002). "Holes in hearing," *J. Assoc. Res. Otolaryngol.* **3**, 185–199.

Shannon, R. V., Jansvold, A., Padilla, M., Robert, M. E., and Wang, X. (1999). "Consonant recordings for speech testing," *J. Acoust. Soc. Am.* **106**, L71–L74.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.

Shannon, R. V., Zeng, F.-G., and Wygonski, J. (1998). "Speech recognition with altered spectral distribution of envelope cues," *J. Acoust. Soc. Am.* **104**, 2467–2476.

Slis, I. H., and Cohen, A. (1969). "On the complex regulating the voiced-voiceless distinction (parts I and II)," *Lang. Speech* **12**, 80–102, 137–155.

Stevens, K. N. (1998). *Acoustic Phonetics* (MIT, Cambridge, MA).

Stevens, K. N., and Blumstein, S. E. (1978). "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Am.* **64**, 1358–1368.

ter Keurs, M., Festen, J. M., and Plomp, R. (1992). "Effect of spectral envelope smearing on speech reception. I," *J. Acoust. Soc. Am.* **91**, 2872–2880.

ter Keurs, M., Festen, J. M., and Plomp, R. (1993). "Effect of spectral envelope smearing on speech reception. II," *J. Acoust. Soc. Am.* **93**, 1547–1552.

Van Tasell, J. D., Greenfield, G. D., Logemann, J. J., and Nelson, A. D. (1992). "Temporal cues for consonant recognition: Training, talker generalization, and use in evaluation of cochlear implants," *J. Acoust. Soc. Am.* **92**, 1247–1257.

Van Tasell, J. D., Soli, S. D., Kirby, V. M., and Widin, G. P. (1987). "Speech waveform envelope cues for consonant recognition," *J. Acoust. Soc. Am.* **82**, 1152–1161.

Xu, L., and Pflingst, B. E. (2008). "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hear. Res.* **242**, 132–140.

Xu, L., Thompson, C. S., and Pflingst, B. E. (2005). "Relative contributions of spectral and temporal cues for phoneme recognition," *J. Acoust. Soc. Am.* **117**, 3255–3267.

Xu, L., and Zheng, Y. (2007). "Spectral and temporal cues for phoneme recognition in noise," *J. Acoust. Soc. Am.* **122**, 1758–1764.

Zhou, N., and Xu, L. (2008). "Lexical tone recognition with spectrally mismatched envelopes," *Hear. Res.* **246**, 36–43.