The intelligibility of speech with “holes” in the spectrum

Kalyan Kasturi and Philipos C. Loizou[a)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688

Michael Dorman and Tony Spahr
Department of Speech and Hearing Sciences, Arizona State University, Tempe, Arizona 85287

(Received 9 January 2002; revised 11 June 2002; accepted 14 June 2002)

The intelligibility of speech having either a single ‘‘hole’’ in various bands or having two ‘‘holes’’ in disjoint or adjacent bands in the spectrum was assessed with normal-hearing listeners. In experiment 1, the effect of spectral ‘‘holes’’ on vowel and consonant recognition was evaluated using speech processed through six frequency bands, and synthesized as a sum of sine waves. Results showed a modest decrease in vowel and consonant recognition performance when a single hole was introduced in the low- and high-frequency regions of the spectrum, respectively. When two spectral holes were introduced, vowel recognition was sensitive to the location of the holes, while consonant recognition remained constant around 70% correct, even when the middle- and high-frequency speech information was missing. The data from experiment 1 were used in experiment 2 to derive frequency-importance functions based on a least-squares approach. The shapes of the frequency-importance functions were found to be different for consonants and vowels in agreement with the notion that different cues are used by listeners to identify consonants and vowels. For vowels, there was unequal weighting across the various channels, while for consonants the frequency-importance function was relatively flat, suggesting that all bands contributed equally to consonant identification. © *2002 Acoustical Society of America.* [DOI: 10.1121/1.1498855]

PACS numbers: 43.71.Es, 43.71.Ky [CWT]

I. INTRODUCTION

It is generally accepted that human listeners rely on cues that exist across several frequency bands to understand speech. The question of how listeners use and combine information across several frequency bands when understanding speech is one that puzzled researchers for many decades. One of the earliest attempts to answer that question was taken by French and Steinberg (1947) with the computation of the articulation index (AI). By systematically low-pass and high-pass filtering the spectrum and measuring speech recognition, French and Steinberg (1947) determined the relative importance of various frequency bands. Although the AI method was found to be very successful in predicting speech intelligibility in many listening conditions, it has one major shortcoming. The AI method does not take into account the fact that listeners may combine and utilize speech information from multiple disjoint bands (e.g., Grant and Braida, 1991).

Although many studies investigated the intelligibility of high-passed-, low-passed- (e.g., French and Steinberg, 1947; Pollack, 1948; Kryter, 1962), and bandpassed-filtered speech (Warren *et al.*, 1995; Stickney and Assmann, 2001), not many studies have investigated the perception of bandstopped-filtered speech (i.e., speech with holes in the spectrum) or speech composed of disjoint frequency bands. Lippmann (1996) investigated the intelligibility of consonants that had a single hole in the middle of the spectrum. High consonant intelligibility (~90% correct) was maintained even after removing speech energy in the middle frequencies (800 to 4 kHz). Shannon *et al.* (2001) assessed the impact of the size and location of spectral holes with cochlear-implant and normal-hearing listeners. For the normal-hearing listeners, holes were created by dropping off 2 to 8 low-, middle-, or high-frequency bands in a 20-noise-band cochlear-implant (CI) simulation. Results showed that holes in the low-frequency region were more damaging than holes in the middle- and high-frequency regions on speech recognition. In the study by Shannon *et al.* a single hole (varying in size) in the low-, middle-, or high-frequency regions of the spectrum was introduced. A few other studies (e.g., Breeuwer and Plomp, 1984; Dorman *et al.*, 1989; Riener *et al.*, 1992) investigated speech recognition of disjoint bands of low- and high-frequency information. Synergy effects were demonstrated in the study by Riener *et al.* (1992) when subjects were presented with spectral information contained in the low- and high-frequency bands. The intelligibility of sentences through single one-third-octave bands centered around either 370 Hz or 6 kHz was roughly 23% when presented alone, but increased to 77% correct when presented simultaneously. The study by Riener *et al.* (1992), as well as those of others, demonstrated that having access to low- and high-frequency information enabled listeners to identify speech with relatively high accuracy. Listeners seemed to ‘‘fill in’’ the missing speech information.

The aforementioned studies examined speech recognition either for a single hole varying in frequency location (and size) or for a single hole in the middle of the spectrum. The scope of those studies was therefore limited in the sense that it did not consider how speech is recognized when it is composed of multiple disjoint bands involving low-, middle-, and/or high-frequency information. The present study ad-

[a)]Electronic mail: loizou@utdallas.edu

TABLE I. The first two formant frequencies (in Hz) of the male and female vowels used in this study.

		had	hod	head	hayed	heard	hid	heed	hoed	hood	hud	who'd
$F1$	Male	627	786	555	438	466	384	331	500	424	629	319
	Female	666	883	693	492	518	486	428	538	494	809	435
$F2$	Male	1910	1341	1851	2196	1377	2039	2311	868	992	1146	938
	Female	2370	1682	1991	2437	1604	2332	2767	998	1102	1391	1384

dressed this question in a systematic fashion considering all possible combinations of missing disjoint bands from the spectrum.

The answer to the question of how listeners use and combine information across frequency bands, whether isolated or disparate, is not only important for understanding speech perception but it is also important for understanding speech perception by cochlear-implant listeners or hearing-impaired listeners in general. Cochlear implants are based on the idea that there are surviving neurons in the vicinity of the electrodes. The lack of hair cells and/or surviving neurons in certain areas of the cochlea essentially creates "hole(s)" in the spectrum. The extent of the effect of holes in the spectrum on speech understanding is not well understood. It is not known, for instance, whether the spectral holes can account for some of the variability in performance among CI listeners. It is therefore of interest to first find out which set of hole pattern(s) is most detrimental for speech recognition. The answer to that question would then be useful for determining ways to somehow make up for the lost information.

The aim of this study is to examine the effect of the location and size of spectral holes on vowel and consonant recognition. Understanding this effect will provide insights as to why some CI listeners do not perform well, despite the wealth of information they receive [cochlear-implant listeners receive only a small number (4–6) of channels of frequency information, despite the fact that some implant processors transmit as many as 20 channels of information (e.g., Fishman *et al.*, 1997; Dorman *et al.*, 2000)]. In addition, we could use the data of this study to develop a frequency-importance function that takes into account the fact that listeners could combine information from disparate frequency bands in the spectrum. In experiment 1, speech was processed through six frequency bands, and synthesized as a sum of sine waves with amplitudes equal to the rms energy of each frequency band, and frequencies equal to the center frequencies of the bandpass filters. [Six channels were used as we found in previous studies (e.g., Loizou *et al.*, 1999) that six channels were enough to achieve high levels of speech understanding.] To synthesize speech with a hole in a certain frequency band, we set the corresponding sine wave amplitude to zero. We systematically created holes in each of the six frequency bands (one hole at a time) and examined vowel and consonant recognition. Similarly, speech was synthesized with two holes in the spectrum, by setting the corresponding sine wave amplitudes to zero. All possible combinations were created, including the scenarios where two holes were in adjacent frequency bands (thus making a larger hole) or where the two holes were in disjoint frequency bands. The data from experiment 1 were used in experiment 2 to derive frequency importance functions for vowel and consonant recognition.

II. EXPERIMENT 1: HOLES IN THE SPECTRUM

The intelligibility of speech having either a single hole in various bands or having two holes in disjoint or adjacent bands in the spectrum was assessed with normal-hearing listeners. The extent of the effect of the location, size, and pattern of spectral holes on vowel and consonant recognition was evaluated.

A. Method

1. Subjects

Twenty normal-hearing listeners (20 to 25 years of age) participated in this experiment. All subjects were native speakers of American English. The subjects were paid for their participation. Eleven of the subjects were tested at the University of Texas at Dallas and the remaining nine subjects were tested at Arizona State University.

2. Speech material

Subjects were tested on consonant and vowel recognition. The consonant test used 16 consonants in /aCa/ context taken from the Iowa consonant test (Tyler *et al.*, 1987). All the syllables were produced by a male speaker.

The vowel test included the words: "heed, hid, hayed, head, had, hod, hud, hood, hoed, who'd, heard" produced by male and female talkers. A total of 22 vowel tokens was used for testing, 11 produced by 7 male speakers and 11 produced by 6 female speakers [not all speakers produced all 11 vowels]. The stimuli were drawn from a set used by Hillenbrand *et al.* (1995). The first two formant frequencies (as estimated by Hillenbrand *et al.*) of the vowels used for testing are given in Table I.

3. Signal processing

Speech material was first low-pass filtered using a sixth-order elliptical filter with a cutoff frequency of 6000 Hz. Filtered speech was passed through a pre-emphasis filter with a cutoff frequency of 2000 Hz. This was followed by bandpass filtering into six different frequency bands using sixth-order Butterworth filters with center frequencies of 393, 639, 1037, 1685, 2736, and 4444 Hz, respectively. The frequency boundaries of the six bands are given in Table II. The filters were designed to span the frequency range from 300 to 5500 Hz in a logarithmic fashion. The output of each channel was passed through a rectifier followed by a second-order Butterworth low-pass filter with a center frequency of 400 Hz to obtain the envelope of each channel output. Cor-

TABLE II. The 3-dB frequency boundaries of the six bands with the corresponding center frequencies (Hz) of each band.

Band	Lower frequency (Hz)	Upper frequency (Hz)	Center frequency (Hz)
1	300	487	393
2	487	791	639
3	791	1284	1037
4	1284	2085	1685
5	2085	3388	2736
6	3388	5500	4444

responding to each channel a sinusoid was generated with frequency set to the center frequency of the channel and with amplitude set to the root-mean-squared (rms) energy of the channel envelope estimated every 4 ms. The phases of the sinusoids were estimated from the fast Fourier transform (FFT) of the speech segment. The sinusoids of each band were finally summed and the level of the synthesized speech segment was adjusted to have the same rms value as the original speech segment.

To create a hole in frequency band N ($1 \leq N \leq 6$), we set the amplitude of the sinusoid corresponding to frequency band N to zero. Speech was synthesized using the remaining five channel amplitudes. Similarly, two holes were created by setting the amplitudes of the sinusoids in frequency bands M and N to zero. Speech was synthesized using the remaining four channel amplitudes.

Vowel and consonant stimuli were created for six single-hole conditions and 15 two-hole conditions as shown in Table III. All possible combinations of removing two out of the six frequency bands were considered. For comparative purposes, we also created a baseline condition in which we did not remove any frequency bands. Overall, subjects were tested with a total of 22 conditions.

TABLE III. The 22 test conditions considered in this study. The zeroth condition corresponds to the baseline condition. The channel(s) removed in each condition are indicated with a zero.

Condition	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5	Channel 6
0	1	1	1	1	1	1
1	0	1	1	1	1	1
2	1	0	1	1	1	1
3	1	1	0	1	1	1
4	1	1	1	0	1	1
5	1	1	1	1	0	1
6	1	1	1	1	1	0
7	0	0	1	1	1	1
8	0	1	0	1	1	1
9	0	1	1	0	1	1
10	0	1	1	1	0	1
11	0	1	1	1	1	0
12	1	0	0	1	1	1
13	1	0	1	0	1	1
14	1	0	1	1	0	1
15	1	0	1	1	1	0
16	1	1	0	0	1	1
17	1	1	0	1	0	1
18	1	1	0	1	1	0
19	1	1	1	0	0	1
20	1	1	1	0	1	0
21	1	1	1	1	0	0

4. Procedure

The experiments were performed on a PC equipped with a Creative Labs SoundBlaster 16 soundcard. Stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones. The words were displayed on a computer monitor, and a graphical user interface was used that enabled the subjects to indicate their response by clicking a button corresponding to the word played. No feedback was given during the test.

At the beginning of each test the subject was presented with a practice session in which the vowels or consonants were processed through six channels—no holes were introduced (baseline condition). After the practice session, the subjects were tested with the various spectral hole conditions. Two groups of subjects were used, 11 from University of Texas—Dallas and 9 from Arizona State University. The 11 subjects at The University of Texas at Dallas were tested with the 14 test conditions labeled as, 0, 1, 2, 3, 4, 5, 6, 7, 9, 15, 16, 18, 20, and 21 in Table III. The zeroth condition corresponded to the baseline condition in which all six channels were present. The nine subjects at Arizona State University were tested with the fifteen conditions labeled as 0, 1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 14, 17, and 19 in Table III. Note that both groups of subjects were tested with the baseline condition. The order in which the conditions were presented was partially counterbalanced between subjects to avoid order effects. In the vowel and consonant tests, there were nine repetitions of each vowel and each consonant. The vowels and the consonants were completely randomized.

B. Results

The mean percent-correct scores for the single-hole conditions are shown in Fig. 1. A one-way ANOVA with repeated measures showed a significant main effect of the location of the spectral hole [$F(6,60)=9.7, p<0.0005$] on consonant recognition. *Post hoc* tests according to Tukey (at alpha=0.05) showed that the scores obtained with channels 4, 5, or 6 off were significantly lower than the baseline score. The average scores of the baseline condition were not significantly different ($p=0.313$) between the two groups of subjects. The scores obtained with channels 1, 2, or 3 off were not significantly different from the baseline score.

The consonant confusion matrices were analyzed in terms of percent information transmitted as per Miller and Nicely (1955). The feature analysis is shown in Fig. 2. A one-way ANOVA showed a nonsignificant effect [$F(6,60)=14.6$, $p=0.484$] for the feature ‘‘manner’’ and a nonsignificant effect for the feature ‘‘voicing’’ [$F(6,60)=2.7, p=0.061$]. The feature ‘‘place’’ was significantly [$F(6,60)=15.6, p<0.0005$] affected. *Post hoc* Tukey tests showed that conditions in which channel 4, 5, or 6 were removed were significantly different from the baseline condition.

For the vowel data, a one-way ANOVA showed a significant main effect [$F(6,54)=14.5, p<0.0005$] of the location of the spectral hole on vowel recognition. A *post hoc*

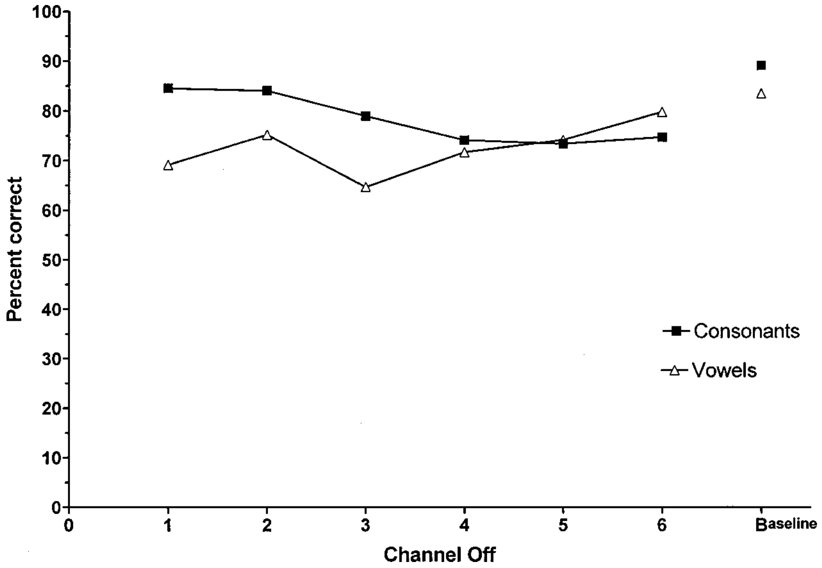

FIG. 1. Mean percent scores for vowel and consonant recognition as a function of the location of the spectral hole. The holes were centered around the channel center frequencies. In the baseline condition, all channels were present.

Tukey test showed that the scores obtained with channels 1, 3, or 4 off were significantly different from the baseline score $(p<0.05)$. The score obtained when channel 2 was off was not significantly different from the baseline score. The fact that channels 1, 3, and 4 were found to have a significant effect on vowel recognition was not surprising since those channels cover the $F1-F2$ frequency range.

The mean percent-correct scores for the two-hole conditions are shown in Fig. 3. The mean scores dropped significantly when a second hole was introduced in the spectrum. The baseline score for consonant recognition dropped from 89.06% to an average (across all conditions) of 69.6%. A one-way ANOVA showed a significant main effect on consonant recognition when two holes were introduced in the spectrum $[F(15{,}120)=6.4, p<0.0005]$. *Post hoc* Tukey tests showed that several channel pair combinations significantly affected consonant recognition: (1,2), (1,4), (1,6) (2,3), (2,6), (3,4), (3,6), (4,5), (4,6), and (5,6). The drop in performance when both channels 1 and 2 were removed was due to the low scores obtained for nasal (/m/,/n/) and labial-stop consonant (/b/,/p/) recognition. Overall, we found that the scores obtained with channel pairs that included channels 4, 5, or 6 were significantly lower than the baseline score $(p<0.05)$. This seems to be consistent with the single-hole conditions, and reinforces the message that channels 4, 5, and 6 are very important for consonant recognition.

The consonant confusion matrices were analyzed in terms of percent information transmitted. The feature analysis is shown in Fig. 4. A one-way ANOVA with repeated measures showed a significant effect $[F(15{,}120)=5.5$, $p<0.0005]$ for the feature "manner," a significant effect for the feature "voicing" $[F(15{,}120)=3.5, p<0.0005]$, and a significant effect $[F(15{,}120)=6.7, p<0.0005]$ for the feature "place." *Post hoc* Tukey tests showed that the manner score obtained with channel pair (1,2) removed was significantly lower $(p<0.0005)$ than the baseline score. The voicing scores obtained with channel pairs (1,2) and (1,4) removed were significantly lower $(p<0.005)$ than the baseline score. All place scores were significantly $(p<0.005)$ lower than the baseline score.

For vowel recognition, a one-way ANOVA showed a significant main effect $[F(15{,}75)=6.9, p<0.0005]$ when

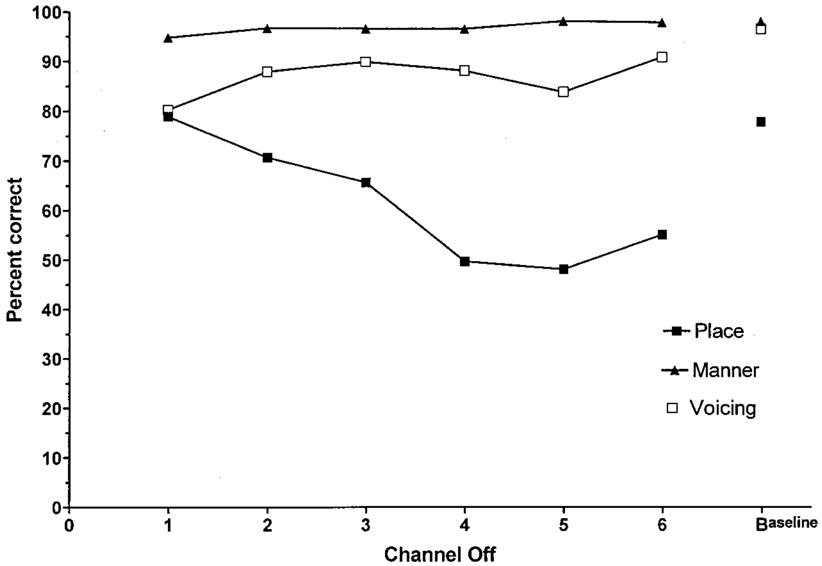

FIG. 2. Percent information transmitted for the features place, manner, and voicing as a function of the location of the spectral hole.

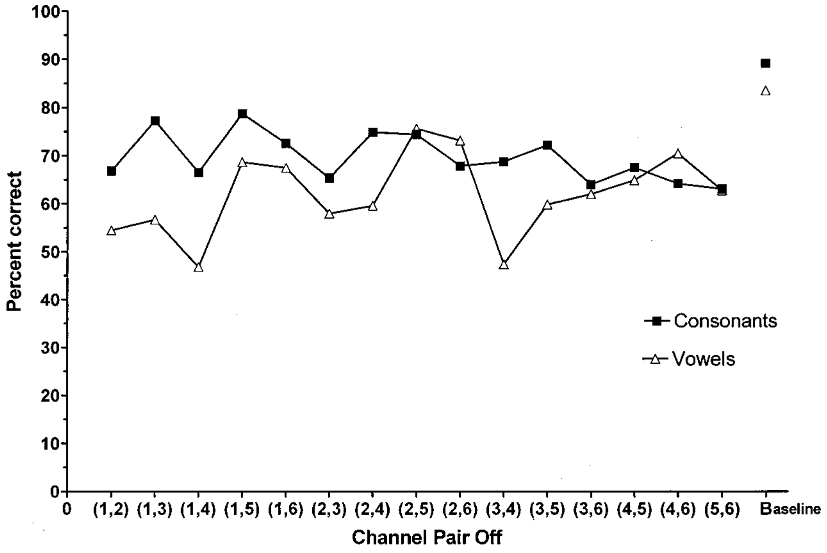

FIG. 3. Mean percent scores for vowel and consonant recognition as a function of the location of the pair of spectral holes. The holes were introduced at frequencies centered at the channel pairs indicated. In condition (1,4), for instance, channels 1 and 4 were removed from the spectrum. In the baseline condition, all channels were present.

two holes were introduced in the spectrum. *Post hoc* Tukey tests showed that several channel pair combinations were significantly affected on vowel recognition: (1,2), (1,3), (1,4),(2,3), (2,4), (3,4), (3,5), and (5,6). The drop in vowel performance when both channels 5 and 6 were removed was due to the low scores obtained for the vowels in "heed," "hid," and "hayed." *Post hoc* Tukey tests showed that the scores obtained with channel pairs that included channel 1, 3, or 4 were significantly lower from the baseline scores ($p < 0.05$), consistent with the outcome in the single-hole conditions. More specifically, the lowest scores on vowel recognition were obtained with channel pairs (1,2), (1,3), (1,4), and (3,4).

C. Discussion

The above results suggest that vowel and consonant recognition suffer when holes are introduced in the spectrum. The degree of degradation in recognition performance as well as effect of the location of the spectral holes was different for vowels and consonants.

1. Effect of location of spectral holes

For vowels, statistical analysis showed a significant drop in performance when either channels 1, 3, or 4, centered at 393, 1037, and 1685 Hz, respectively, were removed. It is safe to assume that channel 1 codes $F1$ information, and channels 3 and 4 code $F2$ information for most vowels (Table I). Channel 3 may also code $F1$ information for some female vowels (i.e., vowels in "hod" and "hud"). Depending on how high the $F2$ frequency is for some speakers, channel 5 (and, indirectly, channel 6) may also be important for the recognition of some vowels. Channel 5 may, for instance, code $F2$ information for some vowels (i.e., heed, hid, hayed) produced by female speakers or children who generally have a high $F2$ frequency. Indeed, close examination of the individual vowel's scores indicated that the identification of the female vowels in "heed," "hid," and "hayed" dropped significantly when both channels 5 and 6 were removed.

It is interesting to note that vowel recognition performance was not significantly affected when channel 2 (centered at 639 Hz) was removed. Channel 2 most likely codes

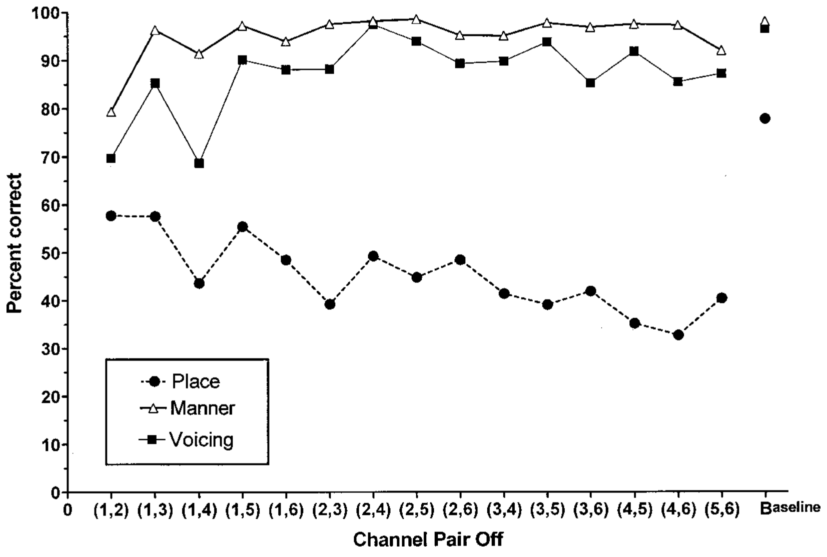

FIG. 4. Percent information transmitted for the features place, manner, and voicing as a function of the location of the pair of frequency bands removed.

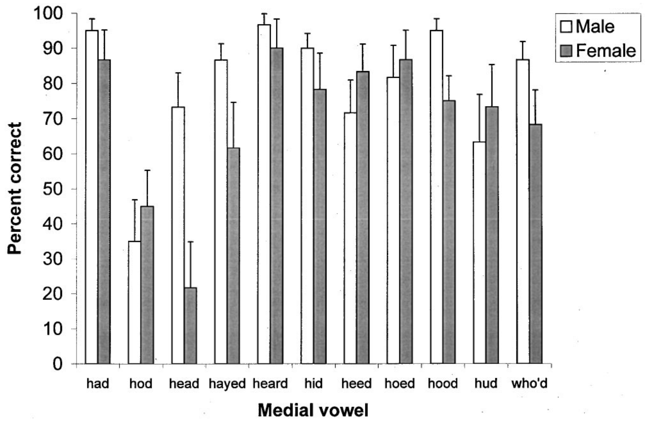

FIG. 5. Mean percent scores on individual vowel recognition for the condition in which channel 2 was removed from the spectrum ($n=20$). The dark and white bars give the scores obtained with vowels produced by female and male speakers, respectively. Error bars indicate standard errors of the mean.

$F1$ information either together with channel 1 or alone. Information about $F1$ is captured by channel 2 alone when the first formant frequency of the vowel falls near the center frequency of channel 2. In that case, a peak in the channel spectrum is observed at channel 2, and consequently removing channel 2 will significantly reduced performance. This is demonstrated in Fig. 5, which shows the listeners' individual vowel performance when channel 2 was removed. Vowels /ɛ/ and /a/ were the only vowels that were significantly affected because the $F1$ frequency of those vowels happened to be near the center frequency of channel 2. For the remaining vowels, however, as evident from Fig. 5, listeners seemed to infer $F1$ information from channel 1 when they did not have access to channel 2 information. This suggests that having a rough estimate of $F1$ is sufficient for the recognition of most vowels. That was not the case with $F2$, since removing either channels 3 or 4 affected vowel recognition.

For consonants, statistical analysis showed a significant drop in performance when either channels 4, 5, or 6 were removed. This outcome is consistent with the conventional view that high-frequency cues are important for recognition of place. So, the drop in consonant recognition performance was primarily due to a reduction in information transmitted for place (Fig. 2).

Removing any of the low-frequency channels (1–4) affected vowel recognition, and removing any of the high-frequency channels (4–6) affected consonant recognition. Interestingly enough, channel 4, which had a center frequency of 1685 Hz, was found to be important for both vowel and consonant recognition. The frequency (1685 Hz) corresponding to channel 4 is close to the well-known crossover frequency[1] estimated in articulation index studies. Depending on the speech material used, the crossover frequency was found in articulation index studies to be in the range of 1550 to 1900 Hz (Studebaker *et al.*, 1987; Hirsh *et al.*, 1954; French and Steinberg, 1947).

Overall, with the exception of channel 2 which did not significantly affect either vowel or consonant recognition, removing single channels caused a modest, but significant reduction in performance in vowel and consonant recognition. Consonant recognition was less affected than vowel recognition. It should be noted that the drop in performance, although statistically significant, was not dramatic for either consonant or vowel recognition. Even in the worst-case conditions, vowel and consonant recognition remained about 70% correct. So, relatively high vowel and consonant recognition performance can be maintained even with a single hole in the spectrum. This outcome is consistent with the data reported by Shannon *et al.* (2001) with cochlear-implant listeners. Shannon *et al.* artificially created single holes by turning off a number of (apical, middle, or basal) electrodes in CI listeners who were fitted with the 22-electrode Nucleus device. Holes were created that were 2–8 electrodes wide corresponding to 1.5–6.0-mm width. High vowel and consonant recognition was maintained even when as many as four electrodes were turned off either in the low-, middle-, or high-frequency regions.

2. Effect of size and pattern of spectral holes

In 5 out of the 15 conditions tested, the size of the hole or equivalently, the width of the notch in the spectrum, doubled, since in these conditions [i.e., channel pairs (1,2), (2,3), (3,4), (4,5), and (5,6)] the channels that were removed were adjacent to each other. This caused a large drop in vowel recognition performance, and only a moderate drop in performance for consonant recognition. Vowel recognition dropped in some cases to as low as 47% correct. The lowest performance occurred when $F1$ information was missing [e.g., pair (1,2)], when $F2$ information was missing [pairs (2,3), (3,4)], or when both $F1$ and $F2$ information was missing [pair (1,4)].

Consonant recognition was only mildly affected by the location of the pairs of frequency bands removed. The decrease in consonant identification was due primarily to the loss of place information (Fig. 4). The manner and voicing features were significantly affected only when information about $F1$ was missing. Overall, consonant identification remained robust and hovered around 70% for most conditions. Even when the middle- and high-frequency bands were absent, consonant recognition remained around 70% correct. This outcome is consistent with the data reported by Lippmann (1996), who evaluated consonant recognition by presenting a low-pass band below 800 Hz and a high-pass frequency band with cutoff frequency varying from 3.15 to 8 kHz. He observed a high score of 91% correct when the high-pass cutoff frequency was 3.15 kHz. This corresponded to the case where channels 4 and 5 were removed in our study. The score for that condition was 67.5% correct. The difference in scores between our study and Lippmann's can be attributed to the fact that our listeners only had access to four channels (two channels were removed) of frequency information. Similar findings were reported by Dorman *et al.* (1989) with CI listeners fitted with a four-channel processor. No significant difference was found between the consonant identification score obtained with only channels 1 (low frequency) and 4 (high frequency) activated and the score obtained with all four channels activated. Our study extended Dorman's and Lippmann's findings to show that high consonant recognition can be maintained even in the absence of not only middle frequencies but also low-high, low-middle, low-high, and middle-high frequency information.

Overall, we can say that vowel recognition seems to be sensitive to the size and pattern of holes in the spectrum. This was not surprising, since it is known that listeners rely primarily on spectral cues to identify vowels. In contrast, listeners make use of both temporal-envelope cues and spectral cues to identify consonants. In the absence of sufficient spectral cues, listeners probably rely more on temporal cues to identify consonants. As shown in this experiment (Fig. 4), these temporal cues did not seem to be affected by the frequency location of the pair of bands removed [except when channels (1,2) were removed]. We believe that is the reason that consonant recognition remained relatively high (~70% correct) even when two holes were introduced in the spectrum. The above results have certain implications for cochlear implants. The finding that the location and pattern of holes affects mostly vowel recognition suggests that in cochlear implants, neuron survival (responsible for the holes in the spectrum) ought to account for some of the variability in vowel recognition performance among CI listeners.

III. EXPERIMENT 2: FREQUENCY-IMPORTANCE FUNCTIONS

Several investigators have used the AI method to determine frequency-importance functions. The AI method uses a quantity between 0 and 1 to represent the proportion of speech information available in a specific frequency band to the listener. This information is then multiplied by a frequency-importance or ‘‘weighting’’ function, which is obtained using a rather time-consuming process of low-pass and high-pass filtering speech. The AI method assumes that the information contained in each band is independent of the information contained in other bands and does not take into account the fact that listeners may combine speech information from multiple disjoint bands. This was first demonstrated by Kryter (1962), who evaluated recognition of passband speech, and showed that the AI could not adequately predict intelligibility of passband speech. Similar findings were also reported by Grant and Braida (1991). Several methods were proposed in the literature to circumvent this shortcoming, including the correlation method by Doherty and Turner (1996) and a recent method based on statistical decision theory by Musch and Buus (2001). In this experiment, we use the data from experiment 1 to derive a frequency-importance function based on a least-squares approach. Unlike the AI method, the proposed least-squares approach makes use of the listener’s scores on perception of vowels and consonants composed of disjoint frequency bands.

A. Least-squares approach

Our approach to obtain the importance of each frequency band follows the method proposed by Ahumada and Lovell (1970). We used the results from experiment 1 to predict the importance or perceptual ‘‘weight’’ of each channel.

We calculated the weight w_i of each channel by predicting the responses of the subject as a linear combination of the strength of each channel, i.e.,

$$R_k = \sum_{i=1}^{6} w_i E_{ik}, \tag{1}$$

where R_k is the mean percent-correct score for condition k and E_{ik} is the strength of the ith channel corresponding to condition k. The strength of each channel is a binary value that can be either 0 or 1 depending on whether the channel is off or on, respectively. The value of k ranges from 1 to 22 spanning all channel combinations (Table III). Forming the prediction error e_k as

$$e_k = R_k - \sum_{i=1}^{6} w_i E_{ik}, \tag{2}$$

we can estimate the channel weights by minimizing the sum of all the squared errors with respect to w_i. Alternatively, Eq. (1) can be written in matrix form as

$$R = EW, \tag{3}$$

where R is a 22-dimensional vector containing the mean percent-correct scores for conditions 1 to 22, E is the data matrix (22×6) consisting of the strengths of each channel (Table III), and $W=[w_1,w_2,\ldots,w_6]$ is a 6-dimensional vector consisting of the desired channel weights.

The above set of equations represents an overdeterministic system of equations since we have 6 unknowns (the channel weights) and 22 equations (one for each condition). We calculated the weights W by solving the matrix equation given by (3) using a least-squares approach

$$W = (E^T E)^{-1} E^T R. \tag{4}$$

After obtaining the solution from Eq. (4), we normalized the weights so that the sum of all the weights was equal to 1.

B. Results and discussion

The relative weights of the various channels are shown in Fig. 6 for the vowel and consonant stimuli. As can be seen, the shape of the weighting function was different for vowels and consonants. For vowels, there was unequal weighting across the various channels, suggesting that each channel contributed differently in understanding these vowel tokens. Channels, 1, 3, and 4, centered at 393, 1037, and 1685 Hz, respectively, received the largest weight. This outcome was consistent with the listener’s reduction in performance in experiment 1 when those channels were removed. Also, consistent with our data from experiment 1, channels 2, 5, and 6 received the lowest weight.

The weighting function for the consonants was relatively flat. This suggests that for consonant recognition all channels are equally important. This outcome is consistent with the data reported recently by Mehr *et al.* (2000). Mehr *et al.* estimated the frequency-importance function of nonsense syllables using the correlational method. Speech was divided into six frequency bands, and a randomly chosen level of filtered noise was added to each channel on each trial. Channels in which the signal-to-noise ratio was more highly cor-

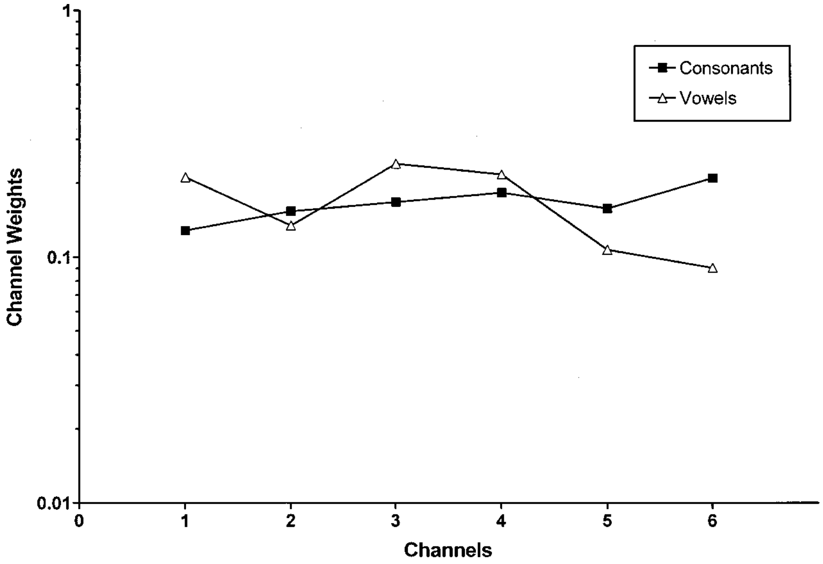

FIG. 6. Frequency-importance functions for vowels and consonants.

related with performance had a larger weight, and channels with smaller correlations had lower weights. Their results showed a flat weighting function for normal-hearing listeners. Unequal weighting functions, accompanied with a large variability among subjects, was noted for the CI listeners in their study.

The individual listener's weighting functions are given in Fig. 7 for vowel and consonant recognition. Weighting functions are given for 6 of the 20 subjects, 2 subjects with the highest vowel scores (Fig. 7, panels a and b), 2 with the middle vowel scores (panels c and d), and 2 with the lowest vowel scores (panels e and f). Most listeners had a relatively flat weighting function for consonants with a small variability. There was a larger variability among subjects in the

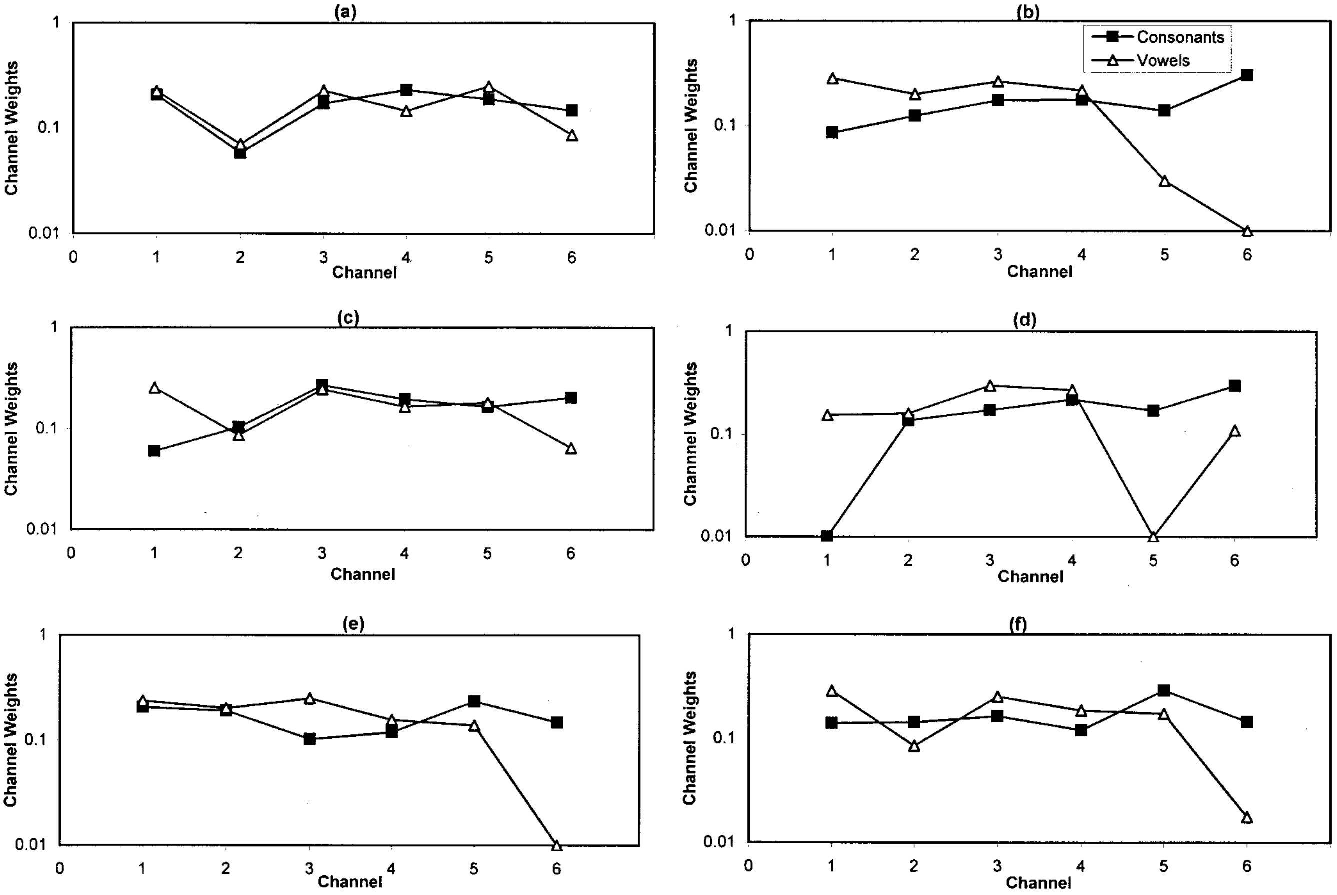

FIG. 7. Individual listener's frequency-importance functions for vowel and consonant recognition. Panels (a) and (b) show the frequency-importance functions for two subjects with the highest vowel scores, panels (c) and (d) show the functions for two subjects with middle scores, and panels (e) and (f) show the functions for two subjects with the lowest vowel scores.

shape of the weighting functions for vowels, suggesting that subjects used different listening strategies for vowel recognition.

The fact that the weighting functions for vowels and consonants were different suggests that subjects were using different listening strategies to identify vowels and consonants. For vowel identification, listeners rely primarily on spectral cues and therefore place more emphasis or more ''weight'' on the channels that code $F1$ and $F2$ information. For consonant identification, listeners rely on both temporal-envelope and spectral cues, which are distributed across all channels. Hence, all frequency bands contributed equally to consonant identification, at least for the filter spacing used in this study. The data from experiment 1 (Fig. 3) are consistent with this conclusion. The fact that consonant recognition remained relatively constant, around 70% correct, regardless of which pairs of channels were removed, clearly demonstrated that all channels contributed equally to consonant recognition. Had the listeners placed more emphasis on certain channels or pairs of channels, we would have seen a dramatic decrease in performance at those channel(s), as we did with the vowels. We suspect that, in general, the frequency-importance function must be dependent, among other factors, on the speech material and the frequency spacing used. Studebaker *et al.* (1987), for instance, showed that the shape of articulation index function and the crossover frequency depended on the speech material.

We did not vary the frequency spacing in this study, but rather used the logarithmic spacing typically implemented in current cochlear-implant processors (Loizou, 1998). According to the data obtained in this experiment, logarithmic spacing provided an equal amount of speech information in each frequency band for consonant identification. This outcome has important implications for cochlear implants. Logarithmic spacing would be desirable assuming that CI listeners are able to extract information from *all* their electrodes. As shown by many investigators (e.g., Fishman *et al.*, 1997; Dorman *et al.*, 2000; Zwolan *et al.*, 1997), that is not the case. This suggests that the frequency spacing should be customized for each CI subject in such a way that their resulting frequency-importance function has larger weights on the functional electrodes and smaller weights on the not-so-functional electrodes.

Despite the differences between the least-squares approach used in this study and the correlational method used by Mehr *et al.* (2000), we obtained a similar (almost identical) weighting function for nonsense syllables. The testing process involved in deriving the weighting functions is time consuming, and therefore both methods are impractical for clinical applications. Another drawback of the correlational method is that it is dependent on the number of trials used for testing. As many as 1200 trials were required in some cases to get significant raw correlations (Mehr *et al.*, 2000; Turner *et al.*, 1998). Our method is not largely dependent on the number of trials, but requires an adequate number of conditions. In our study, we needed to run a total of 22 conditions, which is considerably less than the 135 conditions needed for articulation index studies (e.g., Studebaker *et al.*, 1987) to estimate the frequency-importance function. In brief, the least-squares approach proposed in this study is another viable approach for obtaining frequency-importance functions.

IV. SUMMARY AND CONCLUSIONS

(i) When a single hole was introduced in the spectrum, vowel and consonant recognition decreased. The degree of degradation in performance depended on the location of the hole or, equivalently, the frequency band removed. For vowels, there was a significant drop in performance when either of the frequency bands, 1, 3, and 4 centered around 393, 1037, and 1685 Hz, respectively, were removed. For consonants, there was a modest, yet significant, drop in performance when either of the frequency bands 4, 5, and 6, centered around 1685, 2736, and 4444 Hz, respectively, were removed.

(ii) Vowel recognition was affected the most, with the lowest performance (60% correct) obtained when channel 3, responsible for coding $F2$ information, was removed. Consonant recognition remained relatively high at around 70% correct even when high-frequency channels were removed. Feature analysis indicated that the drop in consonant performance was primarily due to loss of place information. The manner and voicing features were not affected by the location of the hole in the spectrum.

(iii) When two holes were introduced in the spectrum, vowel recognition decreased even further, and consonant recognition remained constant around 70% correct (the same as in the single-hole condition).

(iv) Vowel recognition performance was dependent on the frequency location of the pairs of bands removed. In particular, removing pairs of bands that contained $F1$ and/or $F2$ information caused a significant drop in performance.

(v) In contrast, consonant recognition was only mildly affected by the location of the pair of frequency bands removed. Consonant recognition remained robust at 70% correct, even when the middle- and high-frequency speech information was missing. This outcome is consistent with Lippmann's (1996) findings that accurate consonant recognition can be maintained even when the middle frequencies in the spectrum are absent. Our study extended Lippmann's findings to show that high consonant recognition can be maintained even in the absence of disjoint frequency bands involving low-, high-, and/or middle-frequency information.

(vi) The shapes of the frequency-importance functions, derived in experiment 2 using a least-squares approach, were different for vowels and consonants. This is in agreement with the notion that different cues are used by listeners to identify consonants and vowels.

(vii) For vowels, there was unequal weighting across the various channels. Channels 1, 3, and 4 received the largest weight. The frequency-importance function for consonants was relatively flat, suggesting that all channels contributed equally to consonant identification, at least for the logarithmic filter spacing used in this study. This has important implications for cochlear implants. For CI listeners who are not able to extract useful information from *all* their electrodes, the logarithmic filter spacing might not be the optimal filter spacing.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC03421 from the National Institute of Deafness and other Communication Disorders, NIH. This project was the basis for the Master's thesis of the first author (K.K.) in the Department of Electrical Engineering at the University of Texas—Dallas.

[1] The crossover frequency is the frequency which divides the frequency spectrum into two parts contributing equally to intelligibility. Estimated in articulation index studies, it is the frequency at which the performance with high-passed speech and the performance with low-passed speech is the same.

Ahumad, Jr., A., and Lovell, J. (**1990**). "Stimulus features in signal detection," J. Acoust. Soc. Am. **49**, 1751–1756.

Breeuwer, M., and Plomp, R. (**1984**). "Speechreading supplemented with frequency-selective sound-pressure information," J. Acoust. Soc. Am. **76**, 686–691.

Doherty, K. A., and Turner, C. W. (**1996**). "Use of a correlational method to estimate a listener's weighting function for speech," J. Acoust. Soc. Am. **100**, 3769–3773.

Dorman, M., Dankowski, K., McCandless, G., and Smith, L. (**1989**). "Consonant recognition as a function of the number of channels of stimulation by patients who use the Symbion cochlear implant," Ear Hear. **10**(5), 288–291.

Dorman, M., Loizou, P., Fitzke, J., and Tu, Z. (**2000**). "The recognition of NU-6 words by cochlear implant patients and by normal-hearing subjects listening to NU-6 words processed in the manner of CIS and SPEAK strategies," Ann. Otol. Rhinol. Laryngol. Suppl. **109**(12), Suppl. 185, 64–66.

Fishman, K. E., Shannon R. V., and Slattery, W. H. (**1997**). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor," J. Speech Hear. Res. **40**(5), 1201–1205.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **18**, 90–119.

Grant, K., and Braida, L. (**1991**). "Evaluating the articulation index for auditory-visual input," J. Acoust. Soc. Am. **89**, 2952–2960.

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

Hirsh, I., Reynolds, E., and Joseph, M. (**1954**). "Intelligibility of different speech materials," J. Acoust. Soc. Am. **26**, 530–538.

Kryter, K. (**1962**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1702.

Lippmann, R. P. (**1996**). "Accurate consonant perception without midfrequency speech energy," IEEE Trans. Speech Audio Process. **4**, No. 1, 66–69.

Loizou, P. (**1998**). "Mimicking the human ear: An overview of signal processing techniques for converting sound to electrical signals in cochlear implants," IEEE Signal Process. Mag. **15**(5), 101–130.

Loizou, P., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech," J. Acoust. Soc. Am. **106**, 2097–2103.

Mehr, M. A., Turner, C. W., and Parkinson, A. (**2000**). "Channel weights for speech recognition in cochlear implant users," J. Acoust. Soc. Am. **109**, 359–366.

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

Musch, H., and Buus, S. (**2001**). "Using statistical decision theory to predict intelligibility. I. Model structure," J. Acoust. Soc. Am. **109**, 2896–2909.

Pollack, I. (**1948**). "Effects of high-pass and low-pass filtering on the intelligibility of speech in noise," J. Acoust. Soc. Am. **20**, 259–266.

Riener, K., Warren, R., and Bashford, J. (**1992**). "Novel findings concerning intelligibility of bandpass speech," J. Acoust. Soc. Am. **91**, 2339.

Shannon, R., Galvin, J., and Baskent, D. (**2001**). "Holes in hearing," J. Assoc. Res. Otolaryngol. **3**, 185–199.

Stickney, G., and Assmann, P. (**2001**). "Acoustic and linguistic factors in the perception on bandpass-filtered speech," J. Acoust. Soc. Am. **109**, 1157–1165.

Studebaker, G., Pavlovic, C., and Sherbecoe, R. (**1987**). "A frequency importance function for continuous discourse," J. Acoust. Soc. Am. **81**, 1130–1138.

Turner, C., Kwon, B., Tanaka, C., Knapp, J., and Doherty, K. (**1998**). "Frequency importance functions for broadband speech as estimated by the correlational method," J. Acoust. Soc. Am. **104**, 1580–1585.

Tyler, R., Preece, J., and Lowder, M. (**1987**). The Iowa audiovisual speech perception laser videodisc. *Laser Videodisc and Laboratory Report*, Dept. of Otolaryngology, Head and Neck Surgery, University of Iowa Hospital and Clinics, Iowa City.

Warren, R. M., Riener, K. R., Bashford, Jr., J. A., and Brubaker, B. S. (**1995**). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," Percept. Psychophys. **57**, 175–182.

Zwolan, T., Collins, L., and Wakefield, G. (**1997**). "Electrode discrimination and speech recognition in postlingually deafened adult cochlear implant subjects," J. Acoust. Soc. Am. **102**, 3673–3685.