

REVISED

**Minimum spectral contrast needed for vowel identification by
normal hearing and cochlear implant listeners**

Philipos C. Loizou and Oguz Poroy
Department of Electrical Engineering
University of Texas at Dallas
Richardson, TX 75083-0688

Running header: Spectral contrast for vowel identification

Address correspondence to:

Philipos C. Loizou, Ph.D.
Department of Electrical Engineering
University of Texas at Dallas
P.O. Box 830688, EC 33
Richardson, TX 75083-0688
E-mail : loizou@utdallas.edu
Phone : (972) 883-4617
Fax : (972) 883-2710

Abstract

The minimum spectral contrast needed for vowel identification by normal-hearing and cochlear implant listeners was determined in this study. In Experiment 1, a spectral modification algorithm was used that manipulated the channel amplitudes extracted from a 6-channel Continuous Interleaved Sampling (CIS) processor to have a 1-10 dB spectral contrast. The spectrally modified amplitudes of eight natural vowels were presented to six Med-El/CIS-link users for identification. Results showed that subjects required a 4-6 dB contrast to identify vowels with relatively high accuracy. A 4-6 dB contrast was needed independent of the individual subject's dynamic range (range 9 to 28 dB). Some cochlear implant (CI) users obtained significantly higher scores with vowels enhanced to 6-dB contrast compared to the original, un-enhanced vowels, suggesting that spectral contrast enhancement can improve the vowel identification scores for some CI users. To determine whether the minimum spectral contrast needed for vowel identification was dependent on spectral resolution (number of channels available), vowels were processed in Experiment 2 through n ($n=4, 6, 8, 12$) channels, and synthesized as a linear combination of n sinewaves with amplitudes manipulated to have a 1-20 dB spectral contrast. For vowels processed through 4 channels, normal-hearing listeners needed a 6-dB contrast, for 6 and 8 channels a 4-dB contrast was needed, consistent with our findings with CI listeners, and for 12 channels a 1-dB contrast was sufficient to achieve high accuracy (>80 %). The above findings with normal-hearing listeners suggest that when the spectral resolution is poor, a larger spectral contrast is needed for vowel identification. Conversely, when the spectral resolution is fine, a small spectral contrast (1 dB) is sufficient. The high identification score (82%) achieved with 1-dB contrast was significantly higher than any of the scores reported in the literature using synthetic vowels, and this can be attributed to the fact that we used natural vowels which contained duration and spectral cues (e.g., formant movements) present in fluent speech. The outcomes of Experiments 1 and 2, taken together suggest that CI listeners need a larger spectral contrast (4-6 dB) than normal-hearing listeners to achieve high recognition accuracy, not because of the limited dynamic range, but because of the limited spectral resolution.

PACs numbers: 43.71.Es, 43.71.Ky

INTRODUCTION

The vowel spectra are typically characterized by high-amplitude peaks and relatively low-amplitude valleys. Although the frequencies of the spectral peaks are considered to be the primary cues to vowel identity, the spectral contrast, i.e., the difference between the spectral peak and the spectral valley, needs to be maintained to some extent for accurate vowel identification. The importance of spectral contrast in vowel identification was investigated by Leek *et al.* (1987) using four vowel-like complexes constructed as a sum of 30 100-Hz harmonics. The amplitudes of two consecutive harmonics that defined the (formant) peaks appropriate for the vowels /i ae a U/ varied over a range of 1-8 dB above background harmonics. Results showed that normal-hearing listeners required a 1-2 dB peak-to-valley difference to identify four vowel-like harmonic complexes with relatively high (75% correct) accuracy. Alcantara and Moore (1995) showed that the minimum spectral contrast needed for vowel identification depended on, among other factors, the fundamental frequency, presentation level and the component phase (cosine vs. random) used for the synthesis of the harmonic complexes. Vowel identification was higher with cosine phase, and improved with higher presentation levels and lower fundamental frequency (50 Hz). Subjects needed a 3-dB contrast to identify six vowel-like harmonic complexes with 75% accuracy. In another study, Turner and Van Tasell (1984) showed that normal-hearing listeners could detect a 2-dB notch in a vowel-like spectrum. In summary, the above studies indicate that only a small spectral contrast is needed by normal-hearing listeners for vowel identification.

This remarkable ability of the normal auditory system to detect small amplitude changes in the spectrum was not observed in hearing-impaired listeners (Turner and Holte, 1987; Leek *et al.*, 1987). Leek *et al.* (1987) have shown that listeners with a flat, moderate hearing loss required a 6- to 7-dB peak-to-valley difference for vowel identification. This was attributed to the lack of suppression and the abnormally broad auditory filters associated with hearing loss (e.g., Pick *et al.*, 1977; Wightman *et al.*, 1977). Spectral contrast is reduced when vowels are processed through broad filters due to the shallow filter roll-off. As a result, the internal vowel representation is "blurred" leading to poorer vowel identification.

Unlike normal-hearing and hearing-impaired listeners, CI listeners have a limited spectral resolution and a limited dynamic range. Spectral contrast is reduced in cochlear implant listeners, not because of the abnormally broad auditory filters - which are bypassed with electrical

stimulation - but primarily because of the reduced dynamic range and amplitude compression. The large acoustic dynamic range is typically compressed in implant speech processors using a logarithmic function to a small electrical dynamic range, 5-15 dB. This compression results in a reduction of spectral contrast. We believe that the reduction in spectra contrast was one of the reasons that vowel recognition performance decreased in studies looking for the effect of reduced dynamic range on speech recognition (Zeng and Galvin, 1999; Loizou *et al.*, 2000). Another factor that could potentially reduce spectral contrast is the steepness of the compression function used for mapping acoustic amplitudes to electric amplitudes. A highly compressive mapping function, for instance, would yield a small spectral contrast, even if the dynamic range were large. It is therefore conceivable that a patient may have a large dynamic range, but a small effective spectral contrast because of a steep mapping function. The sensitivity setting, which affects the input gain, can also affect the spectral contrast. If a patient sets the sensitivity too high, then the acoustic amplitudes would be mapped to the high end of the compression function (above the knee point) producing a relatively flat (i.e., small spectral contrast) electrical channel amplitude pattern. Lastly, additive background noise could also reduce spectral contrast (e.g., Leek and Summers, 1996), probably to a larger degree in cochlear implant listeners compared to normal-hearing listeners due to the limited electrical dynamic range.

Given the above factors that could reduce spectral contrast in CI users and consequently affect vowel identification in quiet or in noise, then what is the minimum spectral contrast needed for vowel identification by cochlear implant listeners? The answer to this question is important for the design of CIs for two main reasons. First, it will tell us whether current speech processing strategies preserve enough spectral contrast information needed for vowel identification. Second, it will help us devise new speech processing strategies that will enhance the incoming signal to have a certain spectral contrast. Such strategies could potentially be used to enhance vowel recognition in quiet or in noise. Hawks *et al.* (1997), for instance, increased the vowel spectral contrast by narrowing the formant bandwidths, and noted improvement in vowel identification. Given that the number of channels currently supported by commercial implant processors (Loizou, 1998) varies from a low of 6 channels to a high of 22 channels, it is very important to also ask the question whether the minimum spectral contrast needed for vowel identification is dependent on spectral resolution, i.e., the number of channels available. The above questions are addressed in Experiments 1 and 2 using cochlear implant listeners and normal-hearing listeners, respectively.

In Experiment 1, six CI users fitted with a CIS processor are used to determine the minimum spectral contrast needed for vowel identification. Vowel stimuli are processed off-line and the channel amplitudes are manipulated to have a peak-to-trough ratio ranging from 1 to 10 dB. In Experiment 2, normal-hearing listeners are used to investigate possible interaction between spectral resolution (number of channels) and spectral contrast, based on the hypothesis that there might be a trade-off relationship between spectral resolution (number of channels) and spectral contrast. This hypothesis is based on the view that when speech is processed through a small number of channels the relative differences in across-channel amplitudes must be used to code frequency information. In this view, if spectral contrast was reduced, then vowel recognition ought to decline. On the other hand, when speech is processed through a large number of channels, a large spectral contrast might not be needed, since the frequency information can be coded by the channels that have energy. These questions are investigated in Experiment 2 with normal-hearing listeners, where we assess speech intelligibility as a function of number of channels and as a function of spectral contrast. Normal-hearing listeners are used because the channels and contrast manipulations can not be independently controlled with implant listeners due to the many confounding factors associated with electrical stimulation. To produce speech with varying degrees of spectral resolution and varying degrees of spectral contrast, we synthesized speech as a linear combination of sine waves and manipulated the amplitudes of the sinewaves to have a 1-20 dB peak-to-trough ratio.

I. EXPERIMENT 1: MINIMUM VOWEL SPECTRAL CONTRAST NEEDED BY COCHLEAR-IMPLANT LISTENERS

A. METHOD

1. Subjects

The subjects were six postlingually deafened adults who had used a six-channel CIS processor for periods ranging from three to four years. All the patients had used a four-channel, compressed-analog signal processor (Ineraid) for at least four years before being switched to a CIS processor. The patients ranged in age from 40 to 68 years and they were all native speakers of American English. Biographical data for each patient are presented in Table 1. All subjects were fitted with a 6-channel CIS processor, except for subject S1 who was fitted with a 5-channel processor.

2. Vowel stimuli

Eight monophthong vowels produced by a male speaker were used for testing. The vowels were contained in the words “heed, hid, head, had, hod, hud, hood, who’d”, and were produced by a male speaker (F0=115 Hz) randomly selected from the vowel database used by Hillenbrand et al. (1995). The vowel formant frequencies, estimated at the steady-state portion of the vowel, are shown in Table 2.

3. CIS implementation and experimental setup

The vowel stimuli were first processed off-line using the CIS strategy, saved in a file, and then presented to the CI listeners. Off-line processing was used to ensure that the channel amplitudes had the desired peak-to-trough ratio.

The CIS strategy, which involves bandpass filtering, amplitude envelope estimation and compression, was implemented in MATLAB. Signals were first processed through a pre-emphasis filter (2000 Hz cutoff), with a 3-dB/octave roll-off, and then bandpassed into 6 frequency bands using sixth-order Butterworth filters. The center frequencies of the six bandpass filters were 461, 756, 1237, 2025, 3316, and 5428 Hz. The envelopes of the filtered signals were extracted by full-wave rectification and low-pass filtering (second-order Butterworth) with a 400 Hz cutoff frequency. The six envelope amplitudes A_i ($i=1,2,\dots,6$) were mapped to electrical amplitudes E_i using a logarithmic transformation:

$$E_i = c \log(A_i) + d \quad (1)$$

where c and d are constants chosen so that the electrical amplitudes fall within the range of threshold and most-comfortable levels. The electrical amplitudes E_i were processed through a spectral contrast algorithm (see following section) which manipulated the six channel amplitudes, estimated in each cycle, to have a prescribed peak-to-trough ratio. The spectrally enhanced channel amplitudes were saved in a file, and the experimental setup shown in Figure 1 was used to load the saved channel amplitudes. The envelope amplitudes were finally used to modulate biphasic pulses of duration 40 μ sec/phase at a stimulation rate of 2100 pulses/sec. The electrodes were stimulated in the same order as in the subjects’ daily processors. For most subjects, the electrodes were stimulated in “staggered” order. The sensitivity setting on our laboratory speech processor was fixed and was identical for all subjects.

The experiments were performed on our laboratory cochlear implant processor (Poroy and Loizou, 2000) using the experimental setup shown in Figure 1. To accommodate for off-line data file processing, an I/O card (installed in the PC) was used. The six output lines of the I/O card in the PC were connected to six general-purpose I/O pins of the DSP in the laboratory speech processor, forming a 6-bit, parallel, unidirectional data bus. Since the cochlear implant was connected to the DSP during the experiments, it was necessary to isolate the PC from the rest of the circuitry, which was battery powered. This was achieved using three Burr-Brown ISO150 dual, isolated, digital coupling chips. The speech materials were pre-processed as described below and the amplitudes of the current pulses to be presented to the electrodes were stored in binary data files in the hard-drive of the PC. During the experiments, these files were downloaded to the DSP over the isolated data bus, and were read in and stored in RAM by an assembly program running on the DSP. Finally, the amplitude data was retrieved word-by-word from RAM and sent to the current sources using a serial port in the DSP. A MATLAB interface program was used for loading and “playing back” the binary data files.

4. Spectral contrast enhancement algorithm

Unlike previous studies on spectral contrast (e.g., Leek et al., 1987; Turner and Van Tassel, 1984; Alcantara and Moore, 1995) which manipulated synthetic vowels, this study manipulated naturally produced vowels. The main advantage in using natural vowels over synthetic vowels is that the natural stimuli contain both dynamic and static spectral cues commonly present in fluent speech. Manipulating the spectrum of natural vowels to have a certain peak-to-trough ratio, however, is not as simple as manipulating synthetic vowels. Simply identifying the valley, and modifying the amplitude of the valley (while fixing the peak amplitude) to have a certain peak-to-valley ratio, is not sufficient, because such a change could distort the spectrum. Likewise, identifying the peak, and modifying the amplitude of the peak (while fixing the valley amplitude) to have a certain peak-to-valley ratio, will most likely alter the shape of the spectrum as well. In addition, the latter method may introduce peak clipping, i.e., the modified spectral peak amplitude may be larger than the Most Comfortable Level (MCL), and therefore will need to be clipped to the MCL level.

A spectral contrast enhancement algorithm, which addresses the above issues (peak clipping, spectral distortion, etc.), is proposed in this study. The algorithm is implemented in the

logarithmic domain and therefore assumes that the channel amplitudes are expressed in dB units. Let E_p and E_v represent the amplitudes (in dB) of the peak and valley, respectively, of the electrical amplitudes E_i . The amplitudes E_p and E_v are estimated by finding the maximum and minimum amplitudes respectively of the first four channel amplitudes $20\log(E_i)$, $i=1,2, 3, 4$ [The first four channels cover the F1-F2 frequency region]. Then, the spectrally enhanced channel amplitudes C_i (in dB) can be obtained as:

$$C_i = \frac{E_i^* - E_v}{E_p - E_v} SR + E_p \quad i=1,2, \dots,6 \quad (2)$$

where $E_i^* = 20 \log(E_i)$, and SR is the desired spectral contrast in dB. Finally, the spectrally enhanced amplitudes C_i are converted back to the linear domain using the equation: $10^{C_i/20}$. The above equation preserves the peak amplitude and modifies not only the valley amplitude but also the other amplitudes in order to preserve the shape of the original spectrum. Figure 2 shows examples of the spectral contrast algorithm applied to the vowel /ə/. Note that the spectrally modified amplitudes, C_i , never exceed the MCL level, since the original peak amplitude is preserved (this can be verified by setting $E_i^* = E_p$ in Equation 2). By preserving the peak amplitude we avoid peak-clipping problems. There is a possibility, however, that the spectrally modified amplitudes may fall below the threshold level, and in those cases, we set the corresponding channel amplitudes to threshold. This step was necessary to ensure that the modified channel amplitudes were within the subject's dynamic range.

The above spectral contrast algorithm was applied only to the vocalic segment of the /hVd/ words. The vocalic segment was extracted from the /hVd/ words by manually removing the first and last pitch periods of the onset and offset of the vowel. Equation 2 was applied to all sets of 6-channel amplitudes computed using the CIS strategy within the vocalic segment of the word. The channel amplitudes estimated for the remaining portion (i.e., the silence and the [h],[d] segments) of the words were set to the threshold values. To avoid possible click sensations, the new onsets and offsets of the vowels were tapered off with a half Hamming window, 20-ms in duration.

5. Procedure

A total of 6 different sets of vowels was created with different spectral contrasts (1, 2, 4, 6,

8 and 10 dB) and presented to CI listeners for identification. For comparative purposes, we also presented the vowels processed through the CIS strategy, but were not modified. There were 9 repetitions of each vowel, presented in blocks of 3 repetitions each. The 7 sets of vowels were completely randomized within each block. The test session was preceded by one practice session in which the identity of the vowel was indicated to the listeners.

The stimuli were presented directly to the subjects through our laboratory processor at a comfortable listening level. To collect responses, a graphical interface was used that allowed the subjects to identify the vowels they heard by clicking on the corresponding button on the graphical interface.

B. RESULTS AND DISCUSSION

The results, scored in percent correct, for the different spectral contrasts are shown in Figure 3. Repeated measures analysis of variance indicated a significant main effect of peak-to-trough ratio [$F(6,30)=10.49$, $p<0.005$] on vowel recognition. Performance increased monotonically as the peak-to-trough ratio increased from 1 to 4 dB, and leveled off thereafter. Post-hoc analysis (according to Fisher's LSD) showed that the scores obtained at 4 and 6 dB were not significantly different ($p=0.784$). Neither were the scores obtained at 4 dB and the unenhanced condition significantly different ($p=0.593$). The scores obtained at 2 and 4 dB were not significantly different ($p=0.173$), but the scores obtained at 2 and 1 dB were significantly different ($p < 0.05$).

The individual subjects' performance on vowel recognition is shown in Figure 4. The subjects' performance varied considerably as a function of peak-to-trough ratio. Most subjects (S2, S4, S5, S6) achieved maximum performance at 6 dB peak-to-trough ratio, one subject (S3) achieved maximum performance at 4 dB, while another subject (S1) achieved maximum performance at 8 dB peak-to-trough ratio. Vowel recognition performance declined for subjects S3 and S4 when the peak-to-trough ratio became larger than 4 dB. We suspect that this was due to the fact that the dynamic range of some electrodes was smaller than 10 dB for some subjects. For instance, the average dynamic range of electrodes 5 and 6 for subject S4 was 6 dB, i.e., it was smaller than the tested peak-to-trough ratio. In this case, over enhancing the channel amplitudes might have the same effect as turning off individual electrodes, since enhanced amplitudes smaller than the threshold levels were set to the threshold levels. Subject S1 needed an 8 dB peak-to-trough ratio to reach asymptotic performance. We suspect that this is may be due to the fact that she was

fitted with a 5-channel processor, compared to the other subjects who were fitted with 6-channel processors. This outcome suggests the possibility that a larger spectral contrast is needed for subjects receiving a small number of independent channels of stimulation. This hypothesis is investigated further in Experiment 2.

The outcome that subjects achieved maximum vowel recognition performance at different levels of spectral contrast led us to wonder whether that was related to the subject's dynamic range, which ranged from a low of 9 dB for some subjects to a maximum of 28 dB for others. That is, were the subjects with the larger dynamic range the ones requiring larger spectral contrast to achieve maximum levels of performance? This was based on the assumption that subjects with a wide dynamic range should have a slow growth of loudness; hence they should require a larger spectral contrast for the same loudness difference. Similarly, were the subjects with the smaller dynamic range the ones requiring smaller spectral contrast? To answer these questions, we performed correlation analysis (Figure 5) between the average (across all electrodes) dynamic range and the amount of spectral contrast needed to achieve maximum performance. The resulting correlation (Pearson's) coefficient between dynamic range and spectral contrast was very weak ($r=0.334$) and non-significant ($p=0.517$). As shown in Figure 5, subject S6 who had a large dynamic range (26 dB) required the same amount of spectral contrast to achieve maximum performance as subject S5 who only had a 10 dB dynamic range. This outcome suggests that the amount of spectral contrast needed for vowel identification is independent of the dynamic range, and therefore may be dependent on other factors. Experiment 2 investigates the possibility that spectral resolution might be one of the factors affecting the amount of spectral contrast needed to reach asymptotic performance.

As shown in Figure 4, not all subjects reached an asymptote in performance as the peak-to-trough ratio increased. Performance for some subjects reached a peak at 6 dB and then declined slightly thereafter. We expected that the subjects' performance would asymptote at the same level as that obtained using the original (un-modified) vowels. That was not the case, however. In fact, some of the spectrally modified vowels were more easily identified than the original vowels. Figure 6 shows the average scores for each vowel for the original, the 4-dB and the 6-dB contrast conditions. The majority of the vowels benefited from spectral contrast modification with the largest benefit obtained for the vowels /a i u ?/. The fact that the spectrally modified vowels (to have a 4 and 6 dB peak-to-trough ratio) were more easily identified than the original vowels

suggests that some vowels had originally smaller spectral contrast. Indeed, we found out that the spectral contrast of some vowels was smaller than 6 dB before enhancement. Figure 7 shows, as an example, the histogram of peak-to-trough ratios of the channel amplitudes of the vowel /a/ processed through subject S2's processor, i.e., computed after bandpass filtering, envelope detection and logarithmic compression. The peak-to-trough ratio of the original (un-modified) vowel /a/ varied from a low of 0.3 dB to a high of 4 dB, with an average of 1.9 dB. It was therefore not surprising that subject S2's performance on identification of the vowel /a/ jumped from 11% correct for the original vowels to 78% correct for the vowels enhanced to 6 dB spectral contrast.

Vowel /a/ (un-enhanced) was the most difficult vowel to identify (Figure 6), consistent with previous findings by Loizou *et al.* (1998) on vowel identification by CI users. Close analysis of the well identified and the poorly identified tokens of "hod" in the Loizou *et al.* (1998) study showed that the poorly identified tokens lacked the distinct peak in the channel amplitude spectrum characteristic of the well identified tokens. The poorly identified tokens of "hod" were characterized by a more diffuse distribution of energy across channels 4-6, and had therefore smaller spectral contrast. Increasing the spectral contrast of the vowel /a/ made the peak in the channel amplitude spectrum more distinct and perceptually more salient, leading to a significant improvement in identification. As shown in Figure 6, not all vowels benefited from spectral contrast enhancement. This is because some vowels have inherently larger spectral contrast than others, with the front vowels having the largest spectral contrast (Fant, 1973). So, no improvements were obtained when the original (un-enhanced) vowels had a spectral contrast larger than 4-6 dB.

Subject S2 was not the only subject that benefited in vowel recognition from spectral contrast enhancement. As shown in Figure 4, subjects S1, S3, S4 also benefited. Subject S3's scores improved from 76% correct using the original vowels to 94% using vowels modified to have a 4-dB spectral contrast. Subject S4's scores improved from 47% correct using unenhanced vowels to 64% using vowels enhanced to 6-dB contrast. These results are encouraging as they suggest that post-processing the channel amplitudes (estimated using the CIS strategy) through a spectral-contrast enhancement algorithm can improve the vowel recognition performance of some CI listeners.

In addition to the improvement in vowel identification, enhancing the spectral contrast may also potentially improve consonant identification. Dorman and Loizou (1996) showed that the identification of the consonants /p t k/, which were responsible for the majority of the consonant

confusion errors, can be improved by enhancing the peak of the consonant spectra at the onset. To improve the identification of /ka/ for example, Dorman and Loizou (1996) low-pass filtered the consonant using a cutoff frequency just below the frequency of channel 5. The low-pass filtering reduced the energy in channels 5 and 6, thereby emphasizing the “mid-frequency” peak characteristic of velars. Low-pass filtering improved the spectral contrast of /k/ and consequently improved recognition, much like the spectral contrast algorithm in this study improved the contrast of the vowel /a/ and consequently improved recognition.

The results of this experiment not only tell us about the minimum spectral contrast needed for vowel identification by CI listeners, but they also tell us about the absolute minimum dynamic range needed for vowel identification. For subjects fitted with 6-channel cochlear implant processors, a minimum of 6-dB dynamic range is needed for vowel identification. And this is a very conservative estimate, because it does not account for the compression of the acoustic amplitudes to electric amplitudes. The (logarithmic) compression maps the input signal to a small portion of the output dynamic range, and it rarely, if ever, covers the whole dynamic range. It is possible, as shown in Figure 4 in Loizou *et al.* (2000), for a signal to be mapped to a 24-dB dynamic range, and have less than 10 dB of spectral contrast. Having therefore a dynamic range larger than 6-dB increases the probability that the resulting spectral contrast will be at least 6 dB.

II. EXPERIMENT 2: MINIMUM VOWEL SPECTRAL CONTRAST NEEDED BY NORMAL-HEARING LISTENERS

In Experiment 1 we found that most cochlear implant listeners who were fitted with a 6-channel processor needed at least a 4-6 dB peak-to-trough ratio for accurate vowel recognition. In this experiment, we investigate whether this outcome holds when speech is processed through a larger (or smaller) number of channels. We hypothesize that there is a trade-off between spectral resolution (number of spectral channels) available and spectral contrast needed. This hypothesis was partially motivated by the finding that one of our CI users (S1), who was fitted with a 5-channel CIS processor, needed a larger spectral contrast for vowel identification compared to the other CI users (see Fig. 4).

To produce speech with varying degrees of spectral resolution, speech was filtered through 4-12 frequency bands, and synthesized as a linear combination of sinewaves with amplitudes extracted from the envelopes of the bandpassed waveforms, and frequencies equal to the center

frequencies of the bandpass filters. The spectral contrast algorithm presented in Experiment 1 was applied to the sinewave amplitudes to produce vowels with varying degrees of spectral contrast, ranging from 1 to 20 dB. The intelligibility of vowels was assessed as a function of spectral resolution and as a function of spectral contrast, using normal-hearing listeners as subjects.

A. METHOD

1. Subjects

Nine graduate students from the University of Arkansas at Little Rock¹ served as subjects. All of the subjects were native speakers of American English and had normal hearing. The subjects were paid for their participation.

2. Speech material

The same vowel stimuli used in Experiment 1 were used.

3. Signal Processing

Signals were first processed through a pre-emphasis filter (2000 Hz cutoff), with a 3 dB/octave roll-off, and then bandpassed into n frequency bands ($n=4, 6, 8, 12$) using sixth-order Butterworth filters. Logarithmic filter spacing was used for $n<8$ and mel spacing was used for $n \geq 8$. The center frequencies and the 3-dB bandwidths of the filters can be found in Loizou *et al.* (1999). The envelopes of the signal were extracted by full-wave rectification, and low-pass filtering (second-order Butterworth) with a 400 Hz cutoff frequency. The envelope amplitudes were estimated by computing the root mean-square (rms) energy of the envelopes every 4 msec. The spectral contrast algorithm presented in Experiment 1 was used to modify the peak-to-trough ratio of the estimated envelope amplitudes to Q dB ($Q=1, 2, 4, 6, 8, 10, 15, 20$). Sinewaves were generated with amplitudes equal to the spectrally-enhanced envelope amplitudes, and frequencies equal to the center frequencies of the bandpass filters. The phases of the sinusoids were estimated from the FFT of the speech segment (Loizou *et al.*, 1999). 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.

In addition to the spectrally enhanced vowels, we also processed vowels as described above but without enhancing the envelope amplitudes. We used this condition for comparative reasons and refer to it as the “unenanced” condition.

4. Procedure

The experiment was performed on a PC equipped with a Creative Labs SoundBlaster 16 soundcard. The subjects listened to the speech material via closed ear-cushion headphones at a comfortable level set by the subject. A graphical interface was used that allowed the subjects to select the vowel they heard using a mouse.

Before each condition, subjects were given a practice session with examples of vowels processed through the same number of channels and the same peak-to-trough ratio in that condition. A sequential test order, starting with speech material processed through a large number of channels ($n=12$) and continuing to speech material processed through a small number of channels ($n=4$), was employed. We chose this sequential test design to give the subjects time to adapt to listening to altered speech signals. The test order for the different peak-to-trough ratios in each channel condition was counterbalanced between subjects.

B. RESULTS AND DISCUSSION

The results, scored in percent correct, are shown in Figure 8. A two-factor (channels and peak-to-trough ratio) repeated measures analysis of variance (ANOVA) showed a significant main effect of number of channels [$F(3,24)=8.73$, $p<0.0005$], a significant effect of peak-to-trough ratio [$F(8,64)=67.37$, $p<0.0005$], and a significant interaction between number of channels and peak-to-trough ratio [$F(24,192)=3.73$, $p<0.0005$].

For vowels processed through 4 channels, normal-hearing listeners needed at least a 6-dB peak-to-trough ratio to identify vowels with greater than 80% accuracy. Post-hoc analysis, according to Tukey, showed that the vowel scores obtained at 6dB were not significantly different ($p=0.9$) from the scores obtained at 20 dB. The scores obtained at 10 dB were not significantly different ($p=1.0$) from the scores obtained at 20 dB. For vowels processed through 6 or 8 channels, normal-hearing listeners needed a 4 dB peak-to-trough ratio to identify vowels with the same accuracy. This is

consistent with our findings in Experiment 1 with cochlear implant users fitted with 6-channel processors. The scores obtained with 4-dB contrast using 6 or 8 channels were not significantly different ($p > 0.5$, Tukey post-hoc) from the scores obtained at 20 dB. Finally, for vowels processed through 12 channels, normal-hearing listeners needed only a 1 dB peak-to-trough ratio to identify vowels with greater than 80% accuracy. Post-hoc analysis (Tukey) showed that the score obtained at 2 dB was only marginally different ($p = 0.044$) from the score obtained at 20 dB.

The above results obtained with 4 channels confirm our original hypothesis that when the spectral resolution is poor, a comparatively larger spectral contrast is needed for vowel identification. A larger spectral contrast is needed, because we suspect that listeners must be using amplitude differences across channels to infer the frequency content (e.g., formant locations, etc.) of the signal when the spectral resolution is poor. Conversely, when the spectral resolution is fine (12 channels), a small spectral contrast (1 dB) is sufficient. The results in Experiment 1 with CI patients fitted with 6-channel processors showed that a 4-6 dB amplitude difference between the peak and the valley needs to be maintained for accurate vowel recognition. Consistent with the above hypothesis and the findings of Experiment 2, subject S1, who was fitted with a 5-channel processor, needed a larger spectral contrast (8 dB) to achieve maximum performance on vowel recognition. Judging from the subject's low scores on open set recognition (Table 1), it seems likely that subject S1 may be receiving a small number (probably less than 5) of independent channels of stimulation. The results of Experiment 2 suggest that if we could somehow provide at least 12 channels of stimulation to CI listeners, then a small spectral contrast (1-2 dB), and consequently, a small dynamic range (at least 2 dB) would be sufficient for vowel recognition.

The results obtained with 12 channels are consistent with those reported in the literature (Turner and Van Tassel, 1984; Leek *et al.*, 1987; Summerfield *et al.*, 1987; Alcantara and Moore, 1995) that only a 1-2 dB spectral contrast is needed to identify vowel-like harmonic complexes with 70-75% correct accuracy. Note that the subjects in the Alcantara and Moore (1995) study needed a 3-dB contrast to achieve 75% correct accuracy (six vowel-like harmonic complexes were used in their study, whereas Leek *et al.* used four vowel-like harmonic complexes). Our study showed that high vowel recognition performance ($> 80\%$ correct) can be achieved even with 1-dB spectral contrast. This vowel identification threshold is the same as the psychophysical threshold needed to detect a change in the amplitude spectrum of a complex signal. Green *et al.* (1983)

showed, for instance, that normal-hearing listeners can detect 1-dB increments added to one component of a complex signal.

The mean scores obtained in this study with 1-dB contrast were considerably higher than any of the scores reported in the literature on a similar experiment. For a 1-dB contrast, the subjects of Leek *et al.* (1987) achieved 55% accuracy, the subjects of Alcantara and Moore (1995) achieved 35% accuracy, while our subjects achieved 82% accuracy. Higher performance was achieved in this study even though we represented the vowel spectra with 12 frequency components as opposed to 30 harmonics in the Leek *et al.* study, and used a larger number of vowels (8 vowels in our study vs. 4 vowels in the Leek *et al.* study and 6 vowels in the Alcantara and Moore study). We believe that a higher performance in vowel recognition was obtained in our study because we used natural vowels. Our vowel stimuli contained most of the spectral cues present in naturally produced vowels, including F0 variation and formant movements. In addition, the listeners had access to duration cues. We do not believe that the high performance obtained with our stimuli was primarily because of duration cues, because a recent study by Hillenbrand *et al.* (2000) with normal-hearing listeners showed that the vowel duration had a small overall effect on vowel identification.

Several studies have shown that hearing-impaired listeners need a larger spectral contrast compared to normal-hearing listeners to achieve high vowel recognition performance (e.g., Leek *et al.*, 1987). This was attributed to the wider-than-normal auditory filters. The situation with CI listeners, however, is quite different, since the auditory filters are bypassed with electrical stimulation. The results from Experiment 2 suggest that cochlear implant listeners need a larger spectral contrast than normal-hearing listeners not because of the limited dynamic range, but because of the reduced spectral resolution.

CONCLUSIONS

- ?? Cochlear implant listeners fitted with 6-channel CIS processors need at least a 4-dB spectral contrast to identify natural vowels with high accuracy. Most subjects achieved the highest performance on vowel recognition with a 6-dB spectral contrast, while one subject needed 8 dB.
- ?? Increasing the vowel spectral contrast to 6 dB benefited most subjects in vowel recognition. Some subjects' vowel scores improved by about 20 percentage points when the vowels were enhanced to 6-dB. These results are encouraging as they suggest that we can improve vowel

recognition for CI users, simply by post-processing the CIS channel amplitudes through a spectral contrast enhancement algorithm. The proposed spectral contrast enhancement algorithm used in this study is relatively easy to implement and is amenable for real-time implementation.

- ?? The results of Experiment 2 with normal-hearing listeners indicated that the minimum spectral contrast needed for vowel identification was dependent on the spectral resolution, i.e., the number of channels of frequency information available. For vowels processed through 4 channels, normal-hearing listeners needed at least a 6-dB peak-to-trough ratio to identify vowels with greater than 80% accuracy, while for vowels processed through 6 or 8 channels, normal-hearing listeners needed a 4 dB peak-to-trough ratio to identify vowels with the same accuracy, consistent with our findings with CI users. For vowels processed through 12 channels, normal-hearing listeners needed only a 1 dB peak-to-trough ratio to identify vowels with greater than 80% accuracy.
- ?? The above findings with normal-hearing listeners are consistent with our hypothesis that when the spectral resolution is poor, a larger spectral contrast is needed for vowel identification. Conversely, when the spectral resolution is fine, a small spectral contrast (1 dB) is sufficient.
- ?? For vowels processed through 12 channels, a 1-dB contrast was sufficient to reach high performance (> 80% correct) on vowel recognition. The high scores achieved with 1-dB contrast were significantly higher than the scores reported in the literature (55% correct in the Leek *et al.* study and 33% correct in the Alcantura and Moore study). The high performance obtained in our study can be attributed to the fact that we used naturally produced vowels.
- ?? The outcomes of Experiments 1 and 2, taken together suggest that CI listeners need a larger spectral contrast (4-6 dB) than normal-hearing listeners to achieve high recognition accuracy, not because of the limited dynamic range, but because of the limited spectral resolution.

FOOTNOTES

¹ The authors were previously affiliated with the University of Arkansas at Little Rock before joining the University of Texas at Dallas.

ACKNOWLEDGMENTS

We would like to thank the reviewers for providing valuable suggestions to the manuscript.

This research was supported by Grant No. R01 DC03421 from the National Institute of Deafness and other Communication Disorders, NIH.

REFERENCES

Alcantara, J. and Moore, B. (1995). "The identification of vowel-like harmonic complexes: Effects of component phase, level, and fundamental frequency," *J. Acoust. Soc. Am.*, 97, 3813-3824.

Dorman, M. and Loizou, P. (1996). "Improving consonant intelligibility for Ineraid patients fit with Continuous Interleaved Sampling (CIS) processors by enhancing contrast among channel outputs," *Ear and Hearing*, 17, 308-313.

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

Hawks, J., Fourakis, M., Skinner, M., Holden, T. and Holden, L. (1997). "Effects of formant bandwidth on the identification of synthetic vowels by cochlear implant recipients," *Ear and Hearing*, 18(6), 479-487.

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

Hillenbrand, J., Clark, M. and Houde, R. (2000). "Some effects of duration on vowel recognition," *J. Acoust. Soc. Am.* **108**, 3013-3022.

Green, D., Kidd, G., and Picardi, M. (1983). "Successive versus simultaneous comparison in auditory intensity discrimination," *J. Acoust. Soc. Am.* , 73, 639-643.

Leek, M., Dorman, M. and Summerfield, Q. (1987). "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, 81(1), 148-154.

Leek, M. and Summers, V. (1996). "Reduced frequency selectivity and the preservation of spectral contrast in noise," *J. Acoust. Soc. Am.*, 100, 1796-1806.

Loizou, P. (1998). "Mimicking the human ear: An overview of signal processing techniques for

converting sound to electrical signals in cochlear implants,” *IEEE Signal Processing Magazine*, **15(5)**, 101-130.

Loizou, P., Dorman, M. and Powell, V. (1998). “The recognition of vowels produced by men, women, boys and girls by cochlear implant patients using a six-channel CIS processor,” *J. Acoust. Soc. Am.* **103(2)**, 1141-1149.

Loizou, P., Dorman, M. and Tu, Z. (1999). “On the number of channels needed to understand speech,” *J. Acoust. Soc. Am.* **106(4)**, 2097-2103.

Loizou, P., Dorman, M. and Fitzke, J. (2000). “The effect of reduced dynamic range on speech understanding: Implications for patients with cochlear implants,” *Ear Hearing*, 21(1), 25-31.

Pick, G., Evans, E. and Wilson, J. (1977). “Frequency resolution of patients with hearing loss of cochlear origin,” in *Psychophysics and Physiology of Hearing*, edited by E. Evans and J. Wilson (Academic, London).

Poroy, O. and Loizou, P. (2000). “Development of a speech processor for laboratory experiments with cochlear implant patients,” *IEEE International Conference on Acoustics Speech and Signal Processing*, 6, 3626-3629.

Summerfield, Q., Sidwell, A. and Nelson, T. (1987). “Auditory enhancement of changes in spectral amplitude,” *J. Acoust. Soc. Am.*, 81(3), 700-708.

Turner, C. and Van Tassel, D. (1984). “Sensorineural hearing loss and the discrimination of vowel-like stimuli,” *J. Acoust. Soc. Am.*, 75, 562-566.

Turner, C. and Holte, L. (1987). “Discrimination of spectral-peak amplitude by normal and hearing-impaired subjects,” *J. Acoust. Soc. Am.*, 81, 445-451.

Wightman, F., McGee, T. and Kramer, M. (1977). "Factors influencing frequency selectivity in normal and hearing-impaired listeners," in *Psychophysics and Physiology of Hearing*, edited by E. Evans and J. Wilson (Academic, London).

Zeng, F-G. and Galvin, J. (1999). "Amplitude mapping and phoneme recognition in cochlear implant listeners," *Ear Hear.* **20**, 60-74.

Table 1. Biographical data of the six cochlear-implant users who participated in this study.

Subject	Gender	Age (years) at detection of hearing loss	Age at which hearing aid gave no benefit	Age fit with Ineraid	Age at testing	Etiology of hearing loss	Score on H.I.N.T sentences in quiet	Score on NU-6 words in quiet
S1	F	10	46	47	55	unknown	44	20
S2	F	7	31	33	40	unknown/ hereditary	100	80
S3	F	23	48	51	57	unknown	100	71
S4	M	5	43	48	58	unknown	92	43
S5	M	20	46	63	68	unknown	88	46
S6	M	19	19	29	41	Cogan's syndrome	100	93

Vowel	F1 (Hz)	F2 (Hz)	F3 (Hz)
(h)a(d)	647	1864	2561
(h)o(d)	871	1204	2595
(h)ea(d)	555	1851	2624
(h)i(d)	441	2080	2721
(h)ee(d)	367	2390	2777
(h)oo(d)	468	1115	2492
(h)u(d)	629	1146	2643
(wh)o('d)	366	919	2378

Table 2. The formant frequencies of the vowels used in this study.

Figure Captions

Figure 1. Block diagram of the experimental setup.

Figure 2. Example of spectral modification of the vowel /?/ to 2-10 dB contrast. The original, unenhanced, channel amplitudes are shown in the dotted line.

Figure 3. Mean performance of cochlear-implant listeners on vowel recognition as a function of spectral contrast. Error bars indicate \pm standard errors of the mean.

Figure 4. Individual cochlear implant subject's performance on vowel recognition as a function of spectral contrast.

Figure 5. Correlation between average (across all electrodes) electrical dynamic range and the amount of spectral contrast needed to achieve maximum vowel recognition performance.

Figure 6. Mean performance of CI listeners for the un-enhanced, the 4-dB and the 6-dB contrast conditions for each vowel. Error bars indicate standard errors of the mean.

Figure 7. Histogram of peak-to-trough ratios of the channel amplitudes of the vowel /a/ processed through subject S2's processor.

Figure 8. Mean performance of normal-hearing listeners on vowel recognition as a function of spectral contrast and number of channels.

FIGURE 1

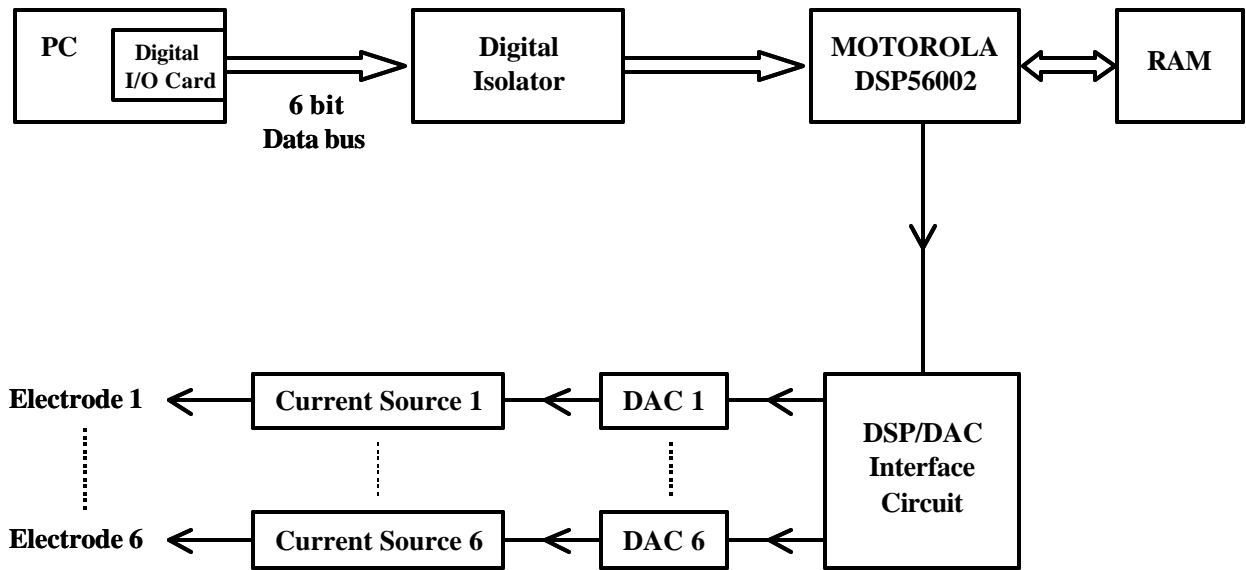


FIGURE 2

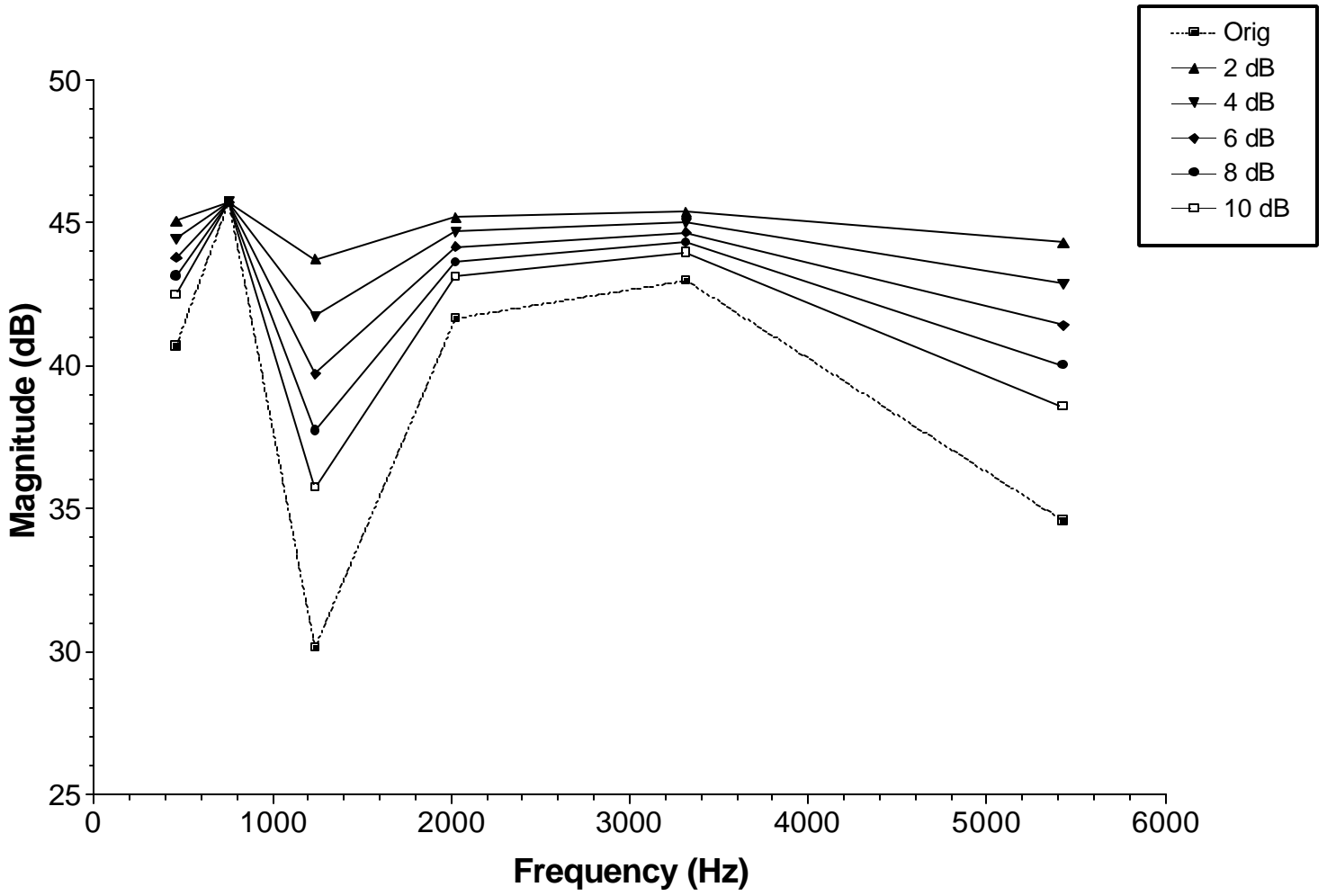


FIGURE 3

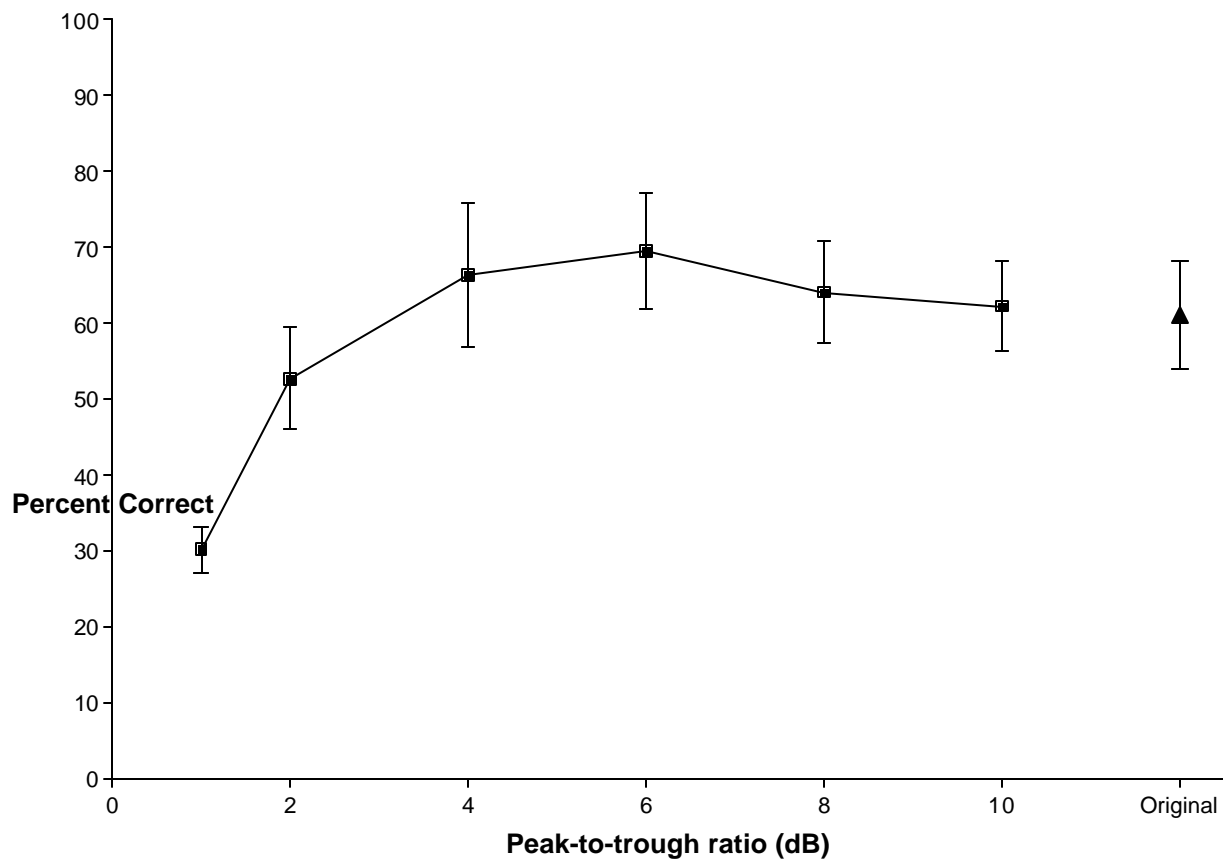


FIGURE 5

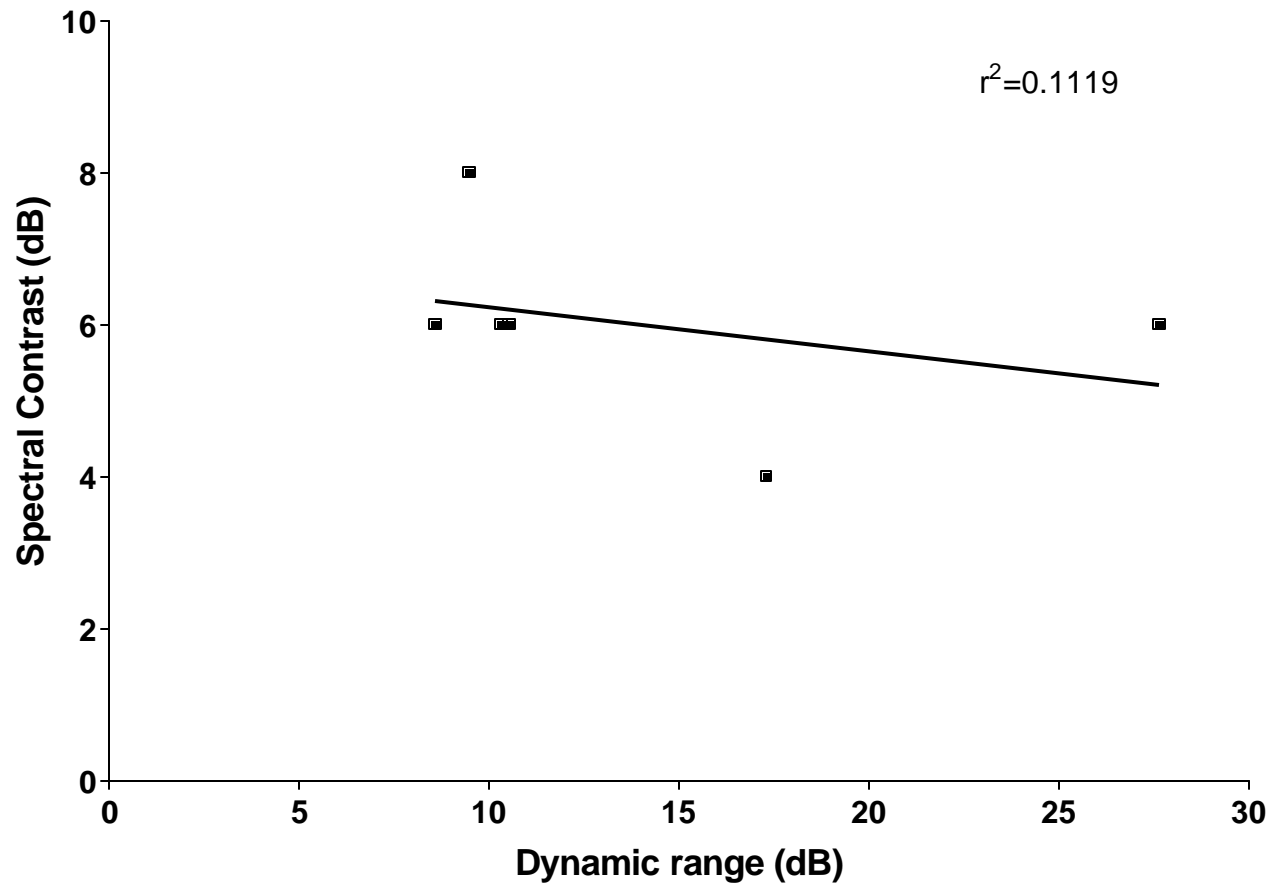


FIGURE 4

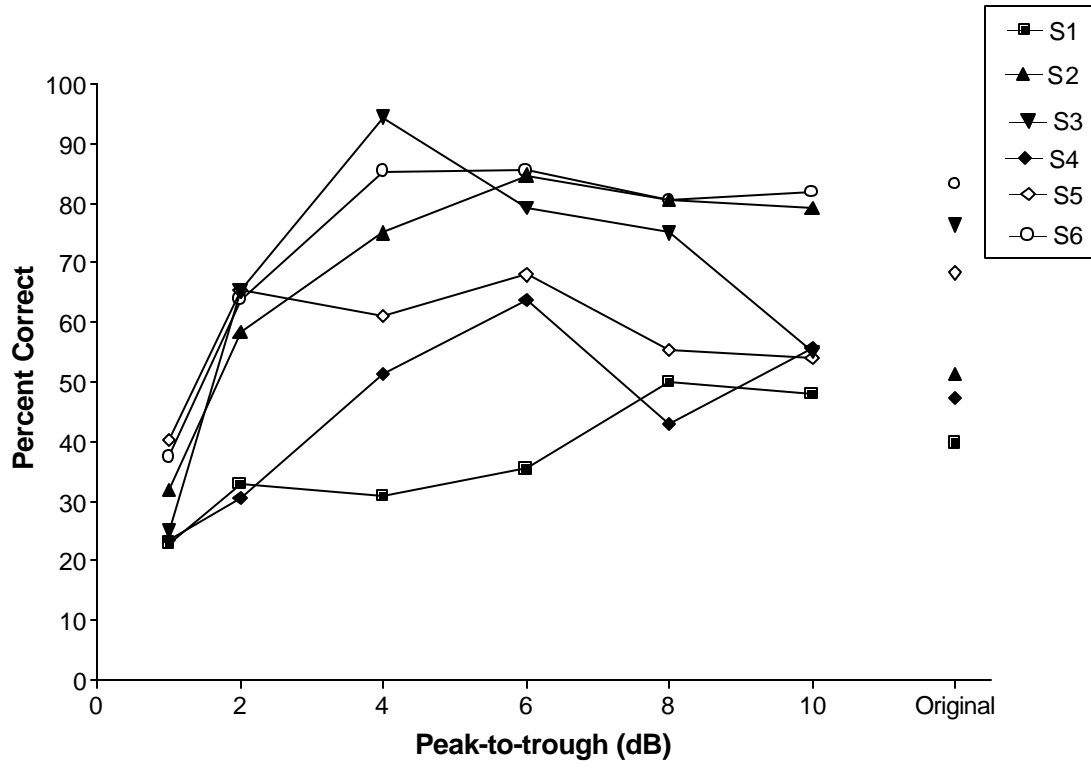


FIGURE 6

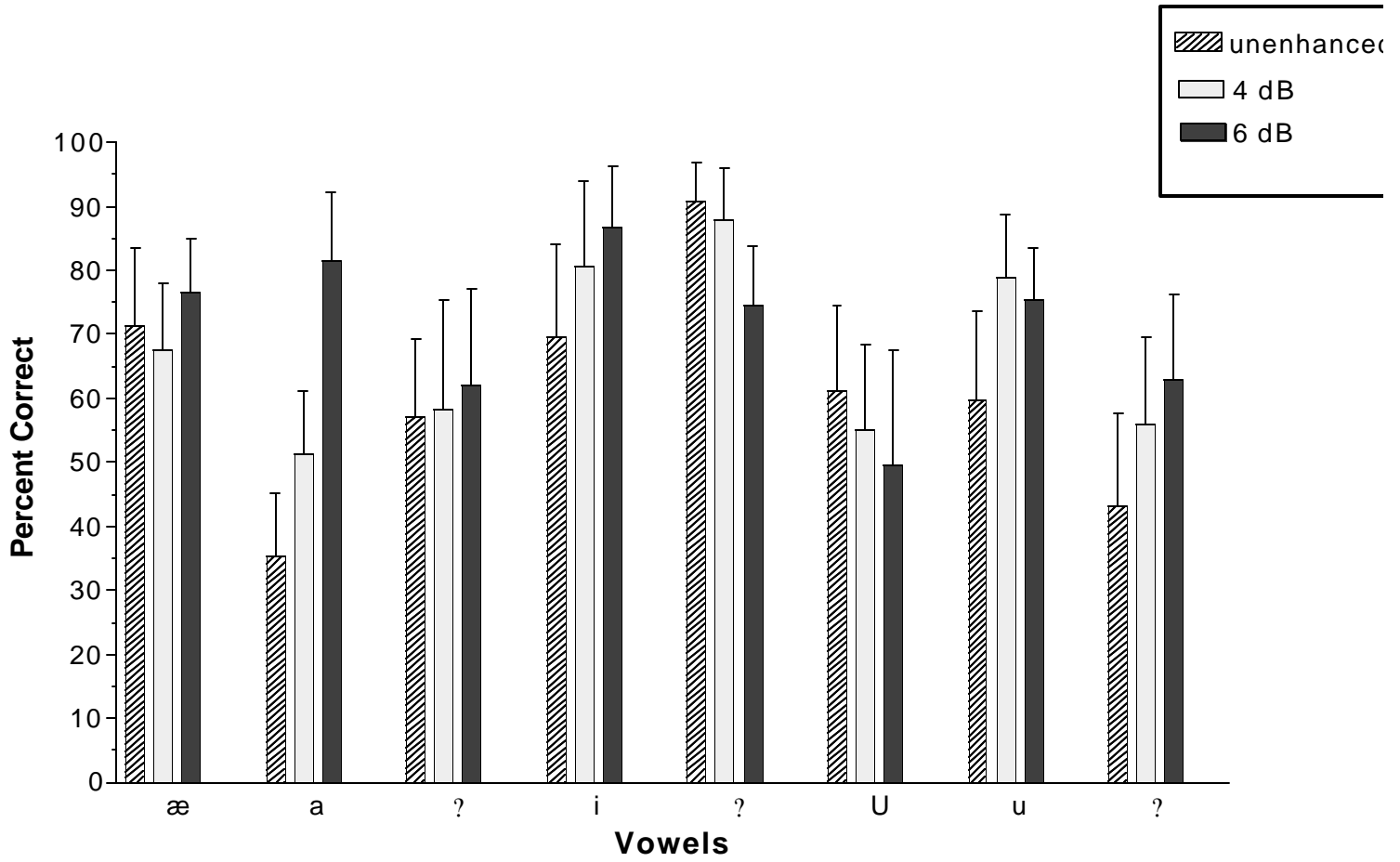


FIGURE 7

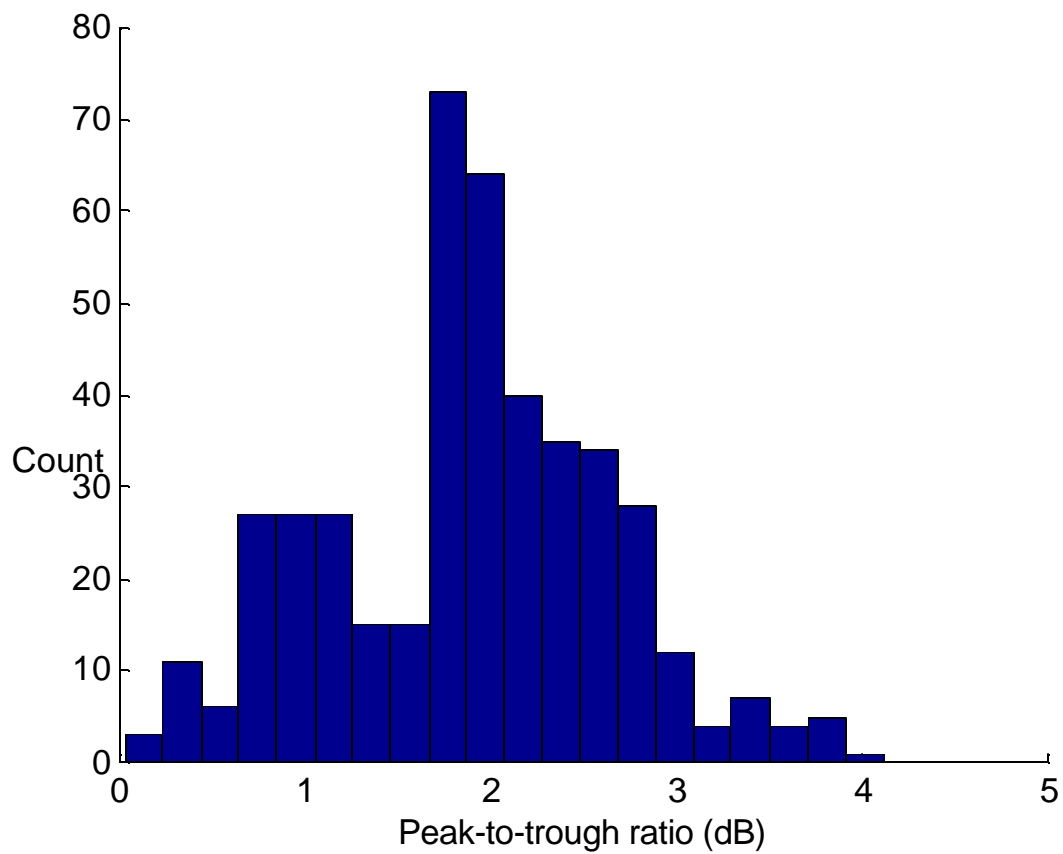


FIGURE 8

