Spectral and temporal cues for phoneme recognition in noise

Li Xu and Yunfang Zheng
School of Hearing, Speech and Language Sciences, Ohio University, Athens, Ohio 45701

(Received 18 June 2006; revised 28 June 2007; accepted 6 July 2007)

Cochlear implant users receive limited spectral and temporal information. Their speech recognition deteriorates dramatically in noise. The aim of the present study was to determine the relative contributions of spectral and temporal cues to speech recognition in noise. Spectral information was manipulated by varying the number of channels from 2 to 32 in a noise-excited vocoder. Temporal information was manipulated by varying the low-pass cutoff frequency of the envelope extractor from 1 to 512 Hz. Ten normal-hearing, native speakers of English participated in tests of phoneme recognition using vocoder processed consonants and vowels under three conditions (quiet, and +6 and 0 dB signal-to-noise ratios). The number of channels required for vowel-recognition performance to plateau increased from 12 in quiet to 16–24 in the two noise conditions. However, for consonant recognition, no further improvement in performance was evident when the number of channels was ≥ 12 in any of the three conditions. The contribution of temporal cues for phoneme recognition showed a similar pattern in both quiet and noise conditions. Similar to the quiet conditions, there was a trade-off between temporal and spectral cues for phoneme recognition in noise. © 2007 Acoustical Society of America. [DOI: 10.1121/1.2767000]

PACS number(s): 43.71.ES, 43.66.Ts [KWG]

Pages: 1758–1764

I. INTRODUCTION

Cochlear implants have provided hearing for thousands of patients with severe to profound sensorineural hearing loss. Speech recognition is remarkably good in most patients, with average sentence recognition better than 80% correct in quiet (see Zeng, 2004 for a review). However, reduced speech recognition in noise is one of the most common problems faced by every cochlear implant user. Skinner et al. (2002) reported that about one third of their 62 postlingually deafened cochlear implant users scored ≤ 75% correct on CUNY sentences presented in multiple-talker babble noise at a fairly moderate +10 dB signal-to-noise ratio (SNR).

Spectral cues and temporal cues in cochlear implant speech processors are represented by the number of channels and the low-pass cutoff frequency of the envelope extractor, respectively. Several previous studies have examined the effects of spectral cues or temporal cues or both on speech recognition in quiet. Fu and Shannon (2000) examined the contribution of temporal cues for phoneme recognition in normal-hearing subjects. The low-pass cutoff frequency of the noise-excited vocoder (Shannon et al., 1995) was varied, and the results demonstrated that with a low-pass cutoff frequency of 20 Hz, the subjects reached a performance plateau for both consonant and vowel recognition. These results were consistent with the data from earlier work (Van Tassel et al., 1987; Drullman et al., 1994a, b; Shannon et al., 1995).

In terms of spectral cues for speech recognition, there is converging evidence that 4 to 8 channels are sufficient to achieve good (i.e., > 85% correct) speech recognition depending on the speech materials used (e.g., vowel, consonant, or sentence) and listening conditions (Shannon et al., 1995, 2004; Dorman et al., 1997; Loizou et al., 1999; Xu et al., 2002; Zeng et al., 2005). In a recent study of English phoneme recognition in quiet, Xu et al. (2005) systematically varied both the number of channels and the low-pass cutoff frequency of the envelope extractor. Results showed that the temporal envelope information required for consonant and vowel recognition to reach a performance plateau was 16 and 4 Hz, respectively. The spectral information required for the performance plateau in consonant and vowel recognition was 6–8 channels and 12 channels, respectively. Finally, there was a trade-off between the two speech cues in quiet. In other words, in order to achieve a certain level of phoneme recognition performance, the number of channels increased as the required low-pass cutoff frequency decreased. The range over which the trade-off for consonant recognition occurred was ≤ 12 channels in the spectral domain and ≤ 32 Hz of low-pass cutoff frequencies in the temporal domain. For vowel recognition, the trade-off occurred when the number of channels was ≥ 4 and the low-pass cutoff frequency was ≤ 4 Hz.

A few studies have investigated spectral and/or temporal cues for speech recognition in cochlear implant users. In a recent study, Nie et al. (2006) covaried the number of channels and the stimulation pulse rate to evaluate the spectral and temporal cues in cochlear implant speech recognition. They confirmed the trade-off effect between spectral and temporal cues for speech recognition in cochlear implant users. However, little is known about the contributions of spectral and temporal cues to speech recognition in noise. Friesen et al. (2001) found that a larger number of channels was required in noise conditions for cochlear implant users to achieve performance equivalent to that in quiet conditions. Similar results were obtained in Fu et al. (1998) and Dorman et al. (1998). Therefore, higher spectral resolution appears to be necessary for speech recognition in noise. On the other

a) Author to whom correspondence should be addressed. Electronic mail: xul@ohio.edu
hand, the temporal envelope cue might be particularly susceptible to the interference of background noise. It is thus important to know whether the use of temporal cues differs substantially under noise conditions.

The present study was designed to examine spectral and temporal cues processed through a cochlear implant speech-processing simulation and to determine their contributions to speech recognition in noise. Spectral cues were manipulated by varying the number of channels in a noise-excited vocoder, whereas temporal cues were manipulated by varying the low-pass cutoff frequency of the envelope extractor. We attempted to test whether the relative contributions of spectral and temporal cues are similar in noise and quiet conditions and whether there is a trade-off between spectral and temporal cues for speech recognition in noise as demonstrated in quiet.

II. METHOD

A. Subjects

Ten subjects (5 males and 5 females, age 24.3±2.8, mean±s.d.) were recruited from the Ohio University student population. They were normal-hearing, native speakers of English with pure-tone air-conduction thresholds ≤20 dB HL at octave frequencies from 250 to 8000 Hz. All subjects scored ≥95% correct on English consonant and vowel recognition tests using the unprocessed original speech materials. None of subjects had a history of ear disease. The use of human subjects was reviewed and approved by the Ohio University Institutional Review Board.

B. Stimuli

Two tests (i.e., consonant test and vowel test) were conducted to assess phoneme recognition. The consonant test had 20 initial consonants that included “ba,” “cha,” “da,” “fa,” “ga,” “ja,” “ka,” “la,” “ma,” “na,” “pa,” “ra,” “sa,” “sha,” “ta,” “tha,” “va,” “wa,” “ya,” and “za” (Shannon et al., 1999). One male talker (No. 3) and one female talker (No. 3) of the multitalker cohort were used, resulting in 40 tokens (20×2=40) for each consonant test. The vowel test used 12 hVd vowels that included “had,” “haid,” “hawed,” “head,” “heed,” “hid,” “heard,” “hoed,” “hod,” “hood,” “hud,” and “who’d” (Hillenbrand et al., 1995). Two male talkers (Nos. 48 and 49) and two female talkers (Nos. 39 and 44) of the multitalker cohort were used, resulting in 48 tokens (12×4=48) for each vowel test.

C. Signal processing

A noise-excited vocoder technique (Shannon et al., 1995) was applied in stimulus signal processing, using MATLAB (MathWorks, Natick MA) software. Before signal processing, the speech signal was mixed digitally with the speech-shaped noise (Nilsson et al., 1994) with S/N of +6 and 0 dB. The rms amplitude of each speech token was calculated and the speech-shaped noise with a rms amplitude equal to 50% (+6 dB S/N) or 100% (0 dB S/N) of that of the speech token was added to the original speech. In the noise-excited vocoder (see Xu et al., 2002, 2005 for details), the original speech signals or the speech signals mixed with the noise were passed through a bank of analysis filters that varied between 2 and 32 (2, 4, 6, 8, 12, 16, 24, and 32) in number of channels. The overall bandwidth was 5350 Hz (from 150 to 5500 Hz). Half-wave rectification and low-pass filtering were used to extract the temporal envelope of each analysis band. The low-pass filter was a second-order Butterworth filter with a roll-off of 12 dB/octave. The low-pass cutoff frequency was varied between 1 and 512 Hz in 1 octave steps. A white noise bandpassed through the same analysis filters was modulated by the extracted temporal envelope of each band. Whereas the process of modulation might introduce sidebands that might overlap with adjacent bands, the effects of such on speech recognition were found negligible in our pilot experiments. Finally, the modulated noise bands were summed and then stored in the computer hard disk for later presentation.

D. Procedure

The signals were presented monaurally through a circumaural headphone (Sennheiser, HD 265) at the ear of preference in an IAC sound booth. The intensity was adjusted to the most comfortable level for each subject. For our subjects, this level was around 65 dB (A). Custom programs written in MATLAB were used to present the stimuli. In the programs, two graphical user interfaces (GUIs) were created to present the consonants and the vowels, respectively. In the GUIs, the phonemic symbols (20 initial consonants or 12 hVd vowels) were represented in alphabetical order in a grid on a computer screen. The consonant or vowel stimuli were presented acoustically in random order. The subjects were instructed to use a computer mouse to select the button that represented the consonant or vowel that they had heard.

For both consonant and vowel tests, there were 80 combinations of number of channels and low-pass cutoff frequency (8 numbers of channels×10 low-pass cutoff frequencies) and each combination was tested in three conditions (quiet, and S/N of +6 and 0 dB). Thus, in the consonant test there were 9600 responses collected from each subject (20 tokens×2 talkers×80 combinations×3 conditions), and in the vowel test there were 11520 responses (12 tokens×4 talkers×80 combinations×3 conditions). All subjects participated in a training session of about 3 h by listening to the processed speech materials in order to familiarize themselves with the experiment. Four combinations [i.e., (1) 24 channels, 128 Hz low-pass cutoff; (2) 12 channels, 32 Hz low-pass cutoff; (3) 6 channels, 8 Hz low-pass cutoff; and (4) 2 channels, 2 Hz low-pass cutoff] were used during the training in such a sequence. Feedback was provided during training. During experimental testing, no feedback was provided. The order of the 80 combinations of number of channels and low-pass cutoff frequency and the 3 noise conditions were randomized. The experiment was scheduled for blocks of 1, 2, or 3 h, and each subject took approximately 25 h to complete the experiment.


L. Xu and Y. Zheng: Spectral and temporal cues in noise

1759
III. RESULTS

In the present study, there were three independent variables (i.e., number of channels, low-pass cutoff frequency, and noise level) and two dependent variables (consonant and vowel recognition scores). For the results of both consonant and vowel tests, a three-way ANOVA with repeated measures revealed that all three main factors (number of channels, low-pass cutoff, and stimulus type) were significant (all p=0.000). The grand mean scores for both consonant and vowel recognition averaged across all 80 combinations of number of channels and low-pass cutoff frequency showed a decrease of 14.7 percentage points from quiet to +6 dB S/N, and a decrease of 9.7 percentage points from +6 to 0 dB S/N.

Figure 1 shows the group-mean phoneme recognition score as a function of number of channels for the consonant (upper panels) and vowel (lower panels) tests under the three conditions (quiet, S/N of +6 and 0 dB). The data were the mean scores across all ten subjects. In each panel, each line represents data using one of the ten low-pass cutoff frequencies as indicated by the legend in the upper-left panel.

![Figure 1](image)

Figure 1. Consonant and vowel recognition as a function of number of channels under three conditions (quiet, S/N of +6 and 0 dB). The data were the mean scores across all ten subjects. In each panel, each line represents data using one of the ten low-pass cutoff frequencies as indicated by the legend in the upper-left panel.

number of channels beyond 12 did not consistently improve speech recognition scores in noise conditions.

Figure 2 shows the group-mean phoneme recognition score as a function of low-pass cutoff frequency for the consonant (upper panels) and vowel (lower panels) tests under the three conditions (quiet, S/N of +6 and 0 dB). Each line represents one of the eight numbers of channels. In general, phoneme recognition improved as a function of low-pass cutoff frequency. The low-pass cutoff frequencies required to reach performance plateau, in both quiet and noise conditions, were about 16 and 4 Hz for consonants and vowels, respectively. Thus, the contributions of temporal cues to phoneme recognition in quiet and in the two noise conditions did not appear to be significantly different from each other.

Figure 3 shows the phoneme recognition score as a function of both number of channels and low-pass cutoff frequency for the consonant (left column) and vowel (right column) tests under the three conditions (quiet, S/N of +6 and 0 dB). Each gray-scale shaded area represents a phoneme recognition score for a particular range of numbers of channels and low-pass cutoff frequencies.

An exponential function, \( y=ae^{b(x+c)}+d \), where \( y \) is the percent correct scores and \( x \) is the number of channels (or low-pass cutoff frequency) was used to fit each of the group-mean score curves shown in Figs. 1 and 2. The values of the parameters of the exponential function were derived based on a method of ordinary least squares and were found to be in the following ranges: \( a (−1.5 \text{ to } −0.9) \), \( b (−0.6 \text{ to } −0.1) \), \( c \)
From each of the fitting curves, the knee point, defined as the number of channels or low-pass cutoff frequency at which 90% of the performance plateau is reached, was then obtained. Those data are also presented in Fig. 3 by the two lines that overlie the contour plots. The vertical line represents the number of channels at which the recognition performance became asymptotic. The number of channels at which the consonant recognition reached plateau was around 12 for both quiet and noise conditions, indicating that an increase of spectral information beyond 12 channels did not improve consonant recognition in noise. On the other hand, the number of channels required to reach vowel recognition plateau was around 12 in quiet and between 16 and 24 in the noise conditions, indicating that an increase of spectral information beyond 12 channels did not improve vowel recognition in noise. The horizontal line represents the low-pass cutoff frequency at which the recognition performance became asymptotic. Note that for both quiet and noise conditions, the low-pass cutoff frequencies for the performance plateau were between 8 and 16 Hz for consonant recognition and around 4 Hz for vowel recognition (Fig. 3).

The two lines also divide each of the contour plots in Fig. 3 into four quadrants. The upper-right quadrant represents the number of channels and the low-pass cutoff frequency at which the performance plateau was reached. The performance was essentially the same at any point in that spectral and temporal space. The upper-left quadrant represents the number of channels before the performance reached plateau but the low-pass cutoff frequency after the performance reached plateau. Therefore, the performance in that space depended only on the number of channels. On the contrary, the lower-right quadrant represents the low-pass cutoff frequency before the performance reached plateau but the number of channels after the performance reached plateau. Therefore, the performance in that space depended only on the low-pass cutoff frequency. Of particular interest is the lower-left quadrant. It represents the number of channels and low-pass cutoff frequency before the performance plateau. Thus, the performance depended on both the number of channels and the low-pass cutoff frequency in that space. It was also the space where there was a trade-off between the spectral and the temporal cues for phoneme recognition. It was evident that the trade-off existed for consonant and vowel recognition in both quiet and noise conditions. Note that the range over which the trade-off occurred was different for consonant and vowel recognition. The trade-off for consonant recognition occurred with the number of channels ≤12 and the low-pass cutoff frequency ≤16 Hz. Correspondingly, the trade-off for vowel recognition occurred with the number of channels ≥4 and the low-pass cutoff frequency ≥4 Hz (Fig. 3).

**IV. DISCUSSION**

Noise has a detrimental effect on speech recognition. In the present study, the mean scores for both consonant and
vowel recognition using the vocoder processed speech tokens showed a decrease of about 15 percentage points from quiet to +6 dB S/N, and a decrease of 10 percentage points from +6 to 0 dB S/N. These results are in close agreement with those of Fu et al. (1998), who tested four normal-hearing subjects listening to unprocessed and vocoder processed speech sound under ten different noise conditions from +24 to −15 dB S/N. With the unprocessed stimuli, the subjects' performance started to decrease when the S/N decreased to 0 dB. When the subjects listened to the vocoder processed stimuli, their performance started to decrease at about +12 dB S/N. The performance score for consonant and vowel recognition decreased about 15 percentage points from S/N of +24 to +6 dB, and about 10 percentage points from +6 to 0 dB S/N. A study of sentence recognition by Stickney et al. (2004) showed similar results. In addition, they also found that a single-talker masker exerted a stronger masking effect than did the steady-state noise with eight
channels. This might be because the single-talker masker introduced more informational masking (central competition between the speech signal and the masker).

In the present study, we found that an increase in the number of channels in noise conditions beyond the point of plateau performance in quiet improved vowel recognition but not consonant recognition. As shown in Figs. 1 and 3, the knee points (the points where performance began to plateau) for consonant recognition were almost the same (i.e., 12 channels) in both quiet and noise conditions. For vowel recognition, however, the number of channels required to reach the performance plateau increased in the noise conditions (16–24 channels) compared to that in quiet (12 channels). This might be due to the fact that vowel recognition relies more on spectral cues than consonant recognition does (Peterson and Barney, 1952). Parikh and Loizou (2005) performed acoustic analysis of the vowels in noise and showed that F1 could be detected more reliably than F2, suggesting that listeners had access to correct F1 information but vague F2 information. Therefore, a higher spectral resolution as a result of a larger number of channels might help resolve the formants better in noise conditions. It is worth noting that the observed benefit of an increased number of channels for vowel recognition in noise might be due to the ceiling effects on vowel recognition in quiet. Although the highest mean scores of all subjects were 88.5% correct with the vocoder processed vowels (lower than the 95–100% correct scores with unprocessed vowels), we could not exclude the possibility that ceiling effects had caused the vowel recognition to reach performance plateau at a lower number of channels. On the other hand, consonants inhabit an auditory/acoustical space that is more multidimensional than that inhabited by vowels. Some features of consonants, such as the place of articulation, rely more on spectral cues, as shown by Xu et al. (2005). Due to a limited number of subjects and experiment repetitions, the analysis of information transmission was not performed in the present study. In future studies, it would be interesting to evaluate the effects of number of channels and low-pass cutoff frequency on transmission of various phonemic features in noise.

Another important finding from the present study is that temporal cues contributed to phoneme recognition similarly in both quiet and noise conditions. As illustrated in Figs. 2 and 3, the knee points of the low-pass cutoff frequencies were almost constant in the three conditions for consonant (16 Hz) and vowel (4 Hz) recognition. Fu and Shannon (2000) examined phoneme recognition of both cochlear implant listeners and normal-hearing listeners in quiet, and found that when the low-pass cutoff frequency was at 20 Hz, subjects reached a performance plateau. Xu et al. (2005) demonstrated similar results in quiet: The low-pass cutoff frequencies required for performance plateau were 16 and 4 Hz for consonant and vowel recognition, respectively. Therefore, when the low-pass cutoff frequency is higher than 16 or 20 Hz, the auditory system might not be sensitive to the differences with added temporal cues for speech recognition. There are several possible reasons for this. First, noise may obscure temporal cues, which makes the cues less effective. It should be pointed out that we used a steady-state speech-shaped noise. Conceivably, if a temporal-modulated noise or a single-talker noise (Qin and Oxenham, 2003; Stickney et al., 2004) were used, the interference of the noise on the temporal envelopes would be greater, rendering the temporal cues even less effective. Second, there is a trade-off between temporal and spectral cues in both quiet and noise conditions as shown in Fig. 3. Thus, when the spectral cues are sufficient, only a limited low-pass cutoff frequency is required for recognition task. Third, the auditory system might not be able to utilize additional temporal cues even if they are available. Psychoacoustical experiments by Bacon and Viemeister (1985) showed a low-pass characteristic in the temporal modulation transfer function (TMTF). Although the cutoff of the TMTF was found to be around 50–60 Hz, human listeners are most sensitive to the temporal modulation at an even lower frequency.

There are a number of implications for auditory prostheses based on the results of the present study. We showed that increased spectral information can improve vowel recognition in noise whereas increased temporal information does not improve phoneme recognition in noise. Current cochlear implant systems provide limited frequency resolution. A typical cochlear implant user may have only 6–10 functional channels (Fishman et al., 1997; Garnham et al., 2002; Kong et al., 2004). Therefore, increasing the effective number of channels might be a key to improving speech recognition in noise for cochlear implant users. Current-steering (or virtual-channel) technique, coupled with reduced channel interaction, might provide more effective channels for cochlear implant users. Another implication is related to presentation of the temporal information in the cochlear implant system. We showed that a low-pass cutoff frequency of 16 Hz was sufficient to achieve a speech recognition performance plateau both in quiet and in noise. Current cochlear implant devices typically use a low-pass cutoff frequency of around 400–500 Hz. Such a relative high low-pass cutoff might not be optimal. If the low-pass cutoff frequency is set at a lower value, a lower stimulation pulse rate will be sufficient to represent the temporal envelope information. Middlebrooks (2005) studied temporal information transmission in the guinea pig cortex and found that a low pulse rate instead of a high pulse rate resulted in significantly better temporal envelope detection. Perceptual studies in cochlear implant users also showed that a low rate stimulation yielded a better modulation detection threshold than a higher rate stimulation (Galvin and Fu, 2005; Pfingst et al., 2007). Therefore, we speculate that a low low-pass cutoff frequency and a low rate stimulation might achieve better speech recognition in quiet and noise for cochlear implant users. However, this speculation requires testing in cochlear implant users.

V. CONCLUSIONS

In this study, we systematically varied the number of channels and the low-pass cutoff frequency of the envelope extractor to examine the relative contributions of spectral and temporal cues to phoneme recognition in noise. Results demonstrated that noise reduced speech recognition performance, and an increase in the number of channels beyond...
that required for plateau performance in quiet could only improve vowel recognition in noise. Temporal cues contributed to phoneme recognition similarly in both quiet and noise conditions. There was a trade-off between temporal and spectral cues for speech recognition in noise as well as in quiet. It is suggested that future cochlear implant design focuses more on increasing the spectral resolution so as to provide better speech recognition in noise for implant users.

ACKNOWLEDGMENTS

We are grateful to Bryan Pfingst, Fan-Gang Zeng, Associate Editor Ken Grant, and an anonymous reviewer for their constructive comments. This research was funded by a Scholarly Activity Award from the College of Health and Human Services at Ohio University.


