RESEARCH ON SPOKEN LANGUAGE PROCESSING
Indiana University

Effects of Talker Variability and Lexical Competition on Audiovisual Word Recognition by Adult Users of Cochlear Implants

Adam R. Kaiser, Karen Iler Kirk, Lorin Lachs and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

1This work was supported by NIH-NIDCD Grants K23 DC00126, R01 DC00111, and T32 DC 00012. Support also was provided by Psi Iota Xi National Sorority. We thank Marcia Hay-McCutcheon and Stacey Yount for their assistance in data collection and management. We also are grateful to Luis Hernandez and Marcelo Areal for their development of the software used for stimulus presentation and data collection.

2 DeVault Otolologic Research Laboratory, Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN

3 Also DeVault Otolologic Research Laboratory, Department of Otolaryngology-Head & Neck Surgery, Indiana University School of Medicine, Indianapolis, IN

219
Effects of Talker Variability and Lexical Competition on Audiovisual Word Recognition by Adult Users of Cochlear Implants

Abstract. The present study examined how postlingually deafened adults with cochlear implants (CIs) combine visual information from lip-reading with auditory cues in an open set word recognition task. Adults with normal hearing served as a comparison group. Word recognition performance was assessed using lexically controlled word lists presented under audio-alone (A), visual-alone (V), and combined audiovisual (AV) presentation formats. Effects of talker variability were studied by manipulating the number of talkers producing the stimulus tokens. Lexical competition was investigated using sets of lexically easy and lexically hard test words. To assess the degree of audiovisual integration above and beyond simple additive cues, an index, I, was developed. A measure of visual enhancement, R, was also used to assess the gain in performance provided in the AV condition relative to the maximum possible performance obtainable in the audio-alone format. Results showed that word recognition performance was highest for AV presentation followed by A and then V. Performance was better for single-talker lists than for multiple-talker lists, particularly under the AV presentation format. Word recognition performance was better for the lexically easy than for the lexically hard words regardless of presentation format. Visual enhancement scores were higher for single talker conditions compared to multiple talker conditions and tended to be somewhat better for lexically easy words than for lexically hard words. The pattern of results suggests that information from the auditory and visual modalities is used to access common, multimodal lexical representations in memory. The findings are discussed in terms of the complementary nature of auditory and visual sources of information that specify the same underlying gestures and articulatory events in speech.

Cochlear implants (CIs) are electronic auditory prostheses for individuals with severe to profound hearing impairment that enable many of implant users to perceive and understand spoken language. However, the benefit to an individual user varies greatly. Audio-alone performance measures have demonstrated that some users of cochlear implants are able to communicate successfully over a telephone even when lip-reading cues are unavailable (Dorman, Dankowski, McCandless, Parkin, & Smith, 1991). Other users display little benefit in open-set speech perception tests under audio-alone listening conditions, but find that the cochlear implant helps them understand speech when visual information also is available. One source of these individual differences is undoubtedly the way in which the surviving neural elements in the cochlea are stimulated with electrical currents provided by the speech processor (Fryauf-Bertschy, Tyler, Kelsay, Gantz, & Woodworth, 1997). Other sources of variability, however, may result from the way in which these initial sensory inputs are coded and processed by higher centers in the auditory system. For example, listeners with detailed knowledge of the underlying phonotactic rules of English may be able to use limited or degraded sources of sensory information in conjunction with linguistic knowledge to achieve better overall performance.

Fortunately, in everyday experience, speech communication is not limited to input from only one sensory modality. Optical information about speech obtained from lip reading improves speech understanding in listeners with normal hearing (Sumby & Pollack, 1954) as well as persons with CIs (Tyler, Parkinson, Woodworth, Lowder, & Gantz, 1997b). While lip reading cues enhance speech perception, the sensory, perceptual, and cognitive processes underlying this gain in performance are not well understood at this time. In one of the first studies to investigate audiovisual integration, Sumby and Pollack (1954) demonstrated that lip-reading cues greatly enhance the speech perception performance of
normal hearing listeners, especially when the acoustic signal is masked by noise. They found that performance on closed-set word recognition tasks increased substantially under audiovisual presentation compared to audio-alone presentation. This increase in performance was comparable to the gain observed when the auditory signal was increased by 15 dB SPL under audio-alone perception conditions (Summerfield, 1987).

Numerous studies have demonstrated that visual information from lip reading improves speech perception performance over audio-alone conditions for adults with normal-hearing (Massaro & Cohen, 1995) and for adults with mild to moderate hearing impairment (Grant, Walden, & Seitz, 1998; Massaro & Cohen, 1999). The cognitive processes by which individuals combine and integrate auditory and visual speech information with lexical and syntactic knowledge has become an important area of research in the field of speech perception. Audiovisual speech perception appears to be more than the simple addition of auditory and visual information (Bernstein, Demorest, & Tucker, 2000; Massaro & Cohen, 1999). A well-known example of the robustness of audiovisual speech perception is the “McGurk effect” (McGurk & MacDonald, 1976). When presented with an auditory /ba/ stimulus and a visual /ga/ stimulus many listeners report hearing an entirely new stimulus: a perceptual /da/. The McGurk and MacDonald study demonstrates that information from separate sensory modalities can be combined to produce percepts that differ predictably from either the auditory or the visual percept alone. However, these findings are not universal across all individuals (see Massaro & Cohen, 2000). Grant and Seitz (1998) suggested that listeners who are more susceptible to the McGurk effect also are better at integrating auditory and visual speech cues. Grant and his colleagues proposed that some listeners could improve consonant perception skills by as much as 26% by sharpening their integration abilities (Grant et al., 1998). Their findings on audiovisual speech perception may have important clinical implications for deaf and hearing-impaired listeners because consonant perception accounted for approximately half of the variance of word and sentence recognition.

Stimulus Variability and Spoken Word Recognition

In addition to differences in audiovisual integration, listeners with cochlear implants may also differ in their ability to perceive speech from a variety of different talkers and to deal with the resulting variability in the acoustic-phonetic properties of speech. Listeners with normal hearing reliably extract invariant phonological and semantic information from speech, even when the utterances are produced by different talkers using different speaking rates or dialects, different styles, or under adverse listening environments (Pisoni, 1993; Pisoni, 1996). The processes by which listeners recognize words and extract meaning from widely divergent acoustic signals is often referred to as perceptual constancy or perceptual normalization. The ability of listeners with normal hearing to normalize speech from different talkers under audio-alone presentation has been demonstrated (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Bradlow & Pisoni, 1999; Nygaard & Pisoni, 1995; Nygaard & Pisoni, 1998). However, little is known about the effects of talker variability on speech perception under audiovisual presentation conditions by normal-hearing listeners (Demorest & Bernstein, 1992; Lachs, 1996; Lachs, 1999), or by hearing-impaired listeners with cochlear implants.

One of the first studies to examine the effects of talker variability on spoken word recognition was performed by Creelman (1957). He presented lists of words consisting of tokens produced by 1 to 8 talkers to normal-hearing listeners in noise. He found poorer speech intelligibility for lists containing tokens produced by two or more talkers than lists produced by only one talker. Subsequent studies have demonstrated similar findings for normal hearing listeners (Mullennix, Pisoni, & Martin, 1989; Sommers, Nygaard, & Pisoni, 1994) and for listeners with hearing loss (Kirk, Pisoni, & Miyamoto, 1997; Sommers, Kirk, & Pisoni, 1997).
It is not clear, however, if these findings can be generalized to adult users of cochlear implants. One reason is that some talker specific attributes, such as a talker's fundamental frequency, may not be well represented in the electrical stimulation pattern provided by the current generation of multichannel cochlear implants. If the percepts elicited by changes in fundamental frequency play a role in mediating talker effects, then one might expect the effects of talker variability to be different for listeners with normal hearing than for hearing-impaired listeners with cochlear implants. Cochlear implant users may be unable to discriminate the subtle differences between similar talkers.

One explanation of the effects of talker variability on spoken word recognition is that perceptual normalization increases processing demands and may divert limited cognitive resources that are normally used for speech perception (Mullennix et al., 1989; Sommers et al., 1994). This hypothesis can account for the decrease in speech perception performance for word lists produced by multiple talkers. An alternative explanation of these findings is that listeners learn to focus on the acoustic cues present in a particular individual's voice and then use these "talker-specific" cues to help them in perceiving single-talker word lists (Nygaard & Pisoni, 1998). The first proposal suggests that multi-talker lists produce a decrement in speech perception performance because cognitive resources are diverted from the normal processes of speech perception to operations needed for perceptual normalization (Mullennix & Pisoni, 1990). The second account suggests that the perceptual advantage for single-talker lists over multiple-talker lists is due to processes related to perceptual learning, attunement, and talker-specific adaptation or adjustment to an individual talker's voice (Nygaard & Pisoni, 1998).

Lexical Effects on Spoken Word Recognition

Audiovisual speech perception is a complex process in which information from separate auditory and visual sensory modalities is combined with prior linguistic knowledge stored in long-term memory. Several researchers have argued that the process of speech perception is fundamentally the same regardless of the conditions under which it is performed, implying that the nervous system processes optical and auditory cues in a similar manner using the same perceptual and linguistic mechanisms (Stein & Meredith, 1993). Recently, Grant et al. (1998) have proposed a conceptual framework that can be used to assess this proposal. Their approach combines top-down cognitive processes and bottom-up sensory processes to account for performance on audiovisual speech perception tasks. Grant et al. argue that audiovisual integration takes place prior to the influence of higher-level lexical factors. While their approach acknowledges that lexical factors may influence the perception of auditory and visual speech cues, they claim that the largest increases in audiovisual speech perception occur when the information present in the auditory and visual signals are complementary and specify the same underlying phonetic events expressed in the talker's articulation.

The Neighborhood Activation Model of Spoken Word Recognition

One way to investigate the perception of audio and visual speech cues and assess the effects of the lexicon on word recognition is to measure speech perception and audiovisual integration abilities using words that have different lexical properties. The Neighborhood Activation Model (NAM) provides a theoretical framework for understanding how spoken words are recognized and identified from sensory inputs (Luce & Pisoni, 1998). More specifically, NAM provides a theoretical basis for explaining why some words are easy to identify and other words are hard to identify. The NAM assumes that a stimulus input activates a set of similar acoustic-phonetic patterns in memory, known as a lexical neighborhood. The activation level of each word pattern is proportional to the degree of similarity between the acoustic-phonetic input of the target word and the acoustic-phonetic patterns stored in memory in a multidimensional acoustic-phonetic space. Lexical properties also strengthen or attenuate these levels of activation for particular sound patterns. In the NAM, a word's level of activation is proportional to its
word frequency, the frequency with which it occurs in the language. The probability of matching a given sensory input to a particular stored lexical pattern is based on the activation level of the individual pattern and the sum of the activation levels of all of the sound patterns selected (Luce & Pisoni, 1998).

The NAM uses information about a word’s lexical neighborhood, its acoustic-phonetic similarity space, to predict whether it will be relatively easy or relatively hard to perceive. In one version of the NAM, words are considered to be lexical neighbors, (i.e., part of the same activation set), if they differ from a target word by the addition, deletion, or substitution of a single phoneme. For example, *scat*, *at*, and *cap* are neighbors of the target word *cat*. For a given target word, the number of lexical neighbors is called the neighborhood density of the word. Words from “dense” lexical neighborhoods have many similar sounding words, whereas words from “sparse” neighborhoods have fewer similar sounding words. Neighborhood frequency is the average frequency of all the words in the neighborhood of a target word. Using these lexical characteristics and word frequency, it is possible to construct two sets of words that differ in lexical discriminability. Lexically easy words are high frequency words from low-density lexical neighborhoods with low neighborhood frequency whereas lexically hard words are low frequency words from high-density lexical neighborhoods with high neighborhood frequency. Luce and Pisoni (1998) have shown that lexically easy words are identified faster and more accurately than lexically hard words under auditory only presentation.

**Lexical and Talker Effects on Audiovisual Speech Integration**

Although there has been a great deal of research on audiovisual integration and multimodal speech perception in both normal-hearing and hearing-impaired listeners in the last few years, the contribution of the lexicon and knowledge of the sound patterns of words in the language has not been studied before and may provide important new insights into the large individual differences in outcome in patients with cochlear implants.

New knowledge about the process of audiovisual integration in deaf patients with CIs can be obtained by comparing the intelligibility of lexically easy and lexically hard words under different presentation formats. If the differences in intelligibility between lexically easy and lexically hard words are similar regardless of presentation format, this would suggest that the processes of audiovisual integration take place prior to lexical selection and contact with stored knowledge about words in long-term memory. If the differences in intelligibility between lexically easy and lexically hard words are influenced by presentation format, this would suggest that extensive interactions between sensory processing, multimodal integration, and lexical selection take place at a very early stage in the word recognition process.

It is well known that audiovisual speech perception often provides large benefits to individuals with hearing impairment, including cochlear implant recipients (Erber, 1972; Erber, 1975; Tyler et al., 1997a). In every day activities, listeners with cochlear implants perceive speech in a wide variety of contexts including television, face-to-face conversation, and over the telephone. Success in recognizing words and understanding the talker’s intended message may differ quite substantially under these diverse listening conditions. The primary goal of this study was to examine the ability of cochlear implant users to integrate the limited auditory information they receive from their implant with visual speech cues during spoken word recognition. To achieve this goal, we examined the effects of lexical and talker variability on word recognition under three presentation formats: audio-alone (A), visual-alone (V) and auditory-plus-visual (AV).
Methods

Participants

Forty-one adults served as listeners in this study and were paid for their participation. Twenty were postlingually deafened adult users of cochlear implants who were recruited from the clinical population at Indiana University (Table 1). All of these listeners had profound bilateral sensorineural hearing losses and had used their cochlear implant for at least six months. Their mean age at time of testing was 50 years. The comparison group consisted of 21 listeners with self-reported normal hearing. They were recruited from within Indiana University and the associated campuses through newspaper and e-mail advertisements and announcements. These participants averaged 42 years of age. All of the listeners in the comparison group had pure tone thresholds below 25 dB HL at 250, 500, 1000, 2000, 3000, and 4000 Hz and below 30 dB HL at 6000 Hz. Each participant was reimbursed for travel to and from testing sessions and was paid $10.00 per hour of testing.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Age at Test (years)</th>
<th>Onset of Deafness (years)</th>
<th>CI Use (Months)</th>
<th>Implant Type</th>
<th>Processor</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI1</td>
<td>69</td>
<td>65</td>
<td>30</td>
<td>N22</td>
<td>Spectra</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI2</td>
<td>43</td>
<td>27</td>
<td>24</td>
<td>MedEl</td>
<td>Combi40</td>
<td>n-of-m</td>
</tr>
<tr>
<td>CI3</td>
<td>51</td>
<td>50</td>
<td>8</td>
<td>N24</td>
<td>Sprint</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI4</td>
<td>71</td>
<td>36</td>
<td>6</td>
<td>MedEl</td>
<td>Combi40+</td>
<td>CIS</td>
</tr>
<tr>
<td>CI5</td>
<td>35</td>
<td>30</td>
<td>48</td>
<td>N22</td>
<td>Spectra</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI6</td>
<td>42</td>
<td>39</td>
<td>6</td>
<td>N24</td>
<td>Sprint</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI7</td>
<td>59</td>
<td>38</td>
<td>108</td>
<td>N22</td>
<td>Spectra</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI8</td>
<td>62</td>
<td>56</td>
<td>60</td>
<td>N22</td>
<td>MSP</td>
<td>MPEAK</td>
</tr>
<tr>
<td>CI9</td>
<td>36</td>
<td>34</td>
<td>6</td>
<td>N24</td>
<td>Sprint</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI10</td>
<td>59</td>
<td>50</td>
<td>107</td>
<td>N22</td>
<td>Spectra</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI11</td>
<td>45</td>
<td>40</td>
<td>60</td>
<td>N22</td>
<td>Spectra</td>
<td>MPEAK</td>
</tr>
<tr>
<td>CI12</td>
<td>19</td>
<td>5</td>
<td>96</td>
<td>N22</td>
<td>Spectra</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI13</td>
<td>49</td>
<td>41</td>
<td>6</td>
<td>N24</td>
<td>Sprint</td>
<td>ACE</td>
</tr>
<tr>
<td>CI14</td>
<td>34</td>
<td>30</td>
<td>36</td>
<td>Clarion</td>
<td>Clarion</td>
<td>CIS</td>
</tr>
<tr>
<td>CI15</td>
<td>66</td>
<td>57</td>
<td>63</td>
<td>Clarion</td>
<td>Clarion</td>
<td>CIS</td>
</tr>
<tr>
<td>CI16</td>
<td>74</td>
<td>65</td>
<td>34</td>
<td>N22</td>
<td>Spectra</td>
<td>SPEAK</td>
</tr>
<tr>
<td>CI17</td>
<td>68</td>
<td>58</td>
<td>12</td>
<td>Clarion</td>
<td>Clarion</td>
<td>CIS</td>
</tr>
<tr>
<td>CI18</td>
<td>40</td>
<td>9</td>
<td>18</td>
<td>Clarion</td>
<td>Clarion</td>
<td>CIS</td>
</tr>
<tr>
<td>CI19</td>
<td>37</td>
<td>3</td>
<td>57</td>
<td>Clarion</td>
<td>Clarion</td>
<td>CIS</td>
</tr>
<tr>
<td>CI20</td>
<td>44</td>
<td>43</td>
<td>12</td>
<td>Clarion</td>
<td>Clarion</td>
<td>CIS</td>
</tr>
</tbody>
</table>

Table 1. Demographics of patients with cochlear implants.

Stimulus Materials

The stimulus materials used in the present investigation were drawn from a large database of digitally recorded audiovisual speech tokens (Sheffert, Lachs, & Hernández, 1996). This database contains 300 monosyllabic English words produced by five male and five female talkers. For the present study, we created six equivalent word lists that would allow us to examine the effect of presentation format, talker variability, and lexical competition on spoken word recognition. Each test list contained 36
words. On each list, half of the words were lexically easy, and half were lexically hard. Lexical density was calculated for each word by counting the number of lexical neighbors using the Hoosier Mental Lexicon database (Nusbaum, Pisoni, & Davis, 1984). The word frequency values represented the number of times each target word occurred per one million words of text (Kucera & Francis, 1967). Two versions of each of the six original word lists were produced: one version contained tokens produced by a single talker. The second version contained tokens produced by six different talkers. This arrangement enabled us to administer a single-talker or multiple-talker version of each test list.

**Balanced Word List Generation.** The specific audiovisual stimulus tokens used in the 12 word lists were selected from the digital database using intelligibility data obtained from undergraduate psychology students at Indiana University in two earlier investigations (Lachs & Hernández, 1998; Sheffert et al., 1996). In these intelligibility studies, different groups of students listened to the words produced by each of the talkers in the database and typed the word they perceived into a computer. Separate groups of listeners were used for each talker under each presentation mode (A, V, and AV). The average intelligibility of each word produced by each talker was computed separately under each of the three presentation formats. In creating the final word lists, 216 words were selected from seven of the talkers using a customized computer program. This program generated equivalent word lists within a given presentation format regardless of lexical discriminability. Thus, the average intelligibility of the lexically easy words and the lexically hard words was equivalent across the six lists used under the three presentation formats. Paired t-tests revealed no significant differences in the speech intelligibility scores between any of the lists under a given presentation format.

Because one goal of this study was to investigate the effects of talker variability on word recognition, the lists were also balanced for talker effects. To accomplish this, the talker with the average visual-alone speech intelligibility score was chosen as the talker for the single-talker lists. Visual-alone intelligibility scores were used to select the single talker because the audio-alone and audiovisual intelligibility scores, which were obtained from normal-hearing listeners, were near ceiling. Once the speaker for the single-talker lists was chosen, the intelligibility scores for the tokens produced by the single talker and the remaining six talkers were used to evaluate intelligibility of the single- and multiple-talker lists respectively. This selection process was based on the audio-alone, visual-alone, and audiovisual intelligibility data for each token. Following this procedure, all six word lists were equally intelligible under a given presentation format regardless of talker condition.

**Procedure**

Testing was conducted in a single-walled sound treated IAC booth (Model #102249). The digitized audiovisual stimuli were presented to participants using a PowerWave 604 (Macintosh compatible) computer equipped with a Targa 2000 video board. All listeners were tested individually. The experimental procedures were self-paced. Video signals were presented with a JVC 13U color monitor. Speech tokens were presented via a loudspeaker at 70 dB SPL (C weighted) for participants using CIs. Each participant was administered three single talker and three multiple talker lists. Within each talker condition, one list was presented using an audio-alone format, one using a visual-alone format, and one using an auditory plus visual format. Visual-alone conditions were achieved by attenuating the loudspeaker and audio-alone conditions were achieved by turning off the video display monitor.

For the conditions where auditory stimulation was present in the stimulus, normal hearing participants were tested using a -5 dB signal to noise ratio (SNR) in speech spectrum noise at 70dB SPL relative to the 65 dB SPL speech tokens. This SNR was chosen during preliminary testing to prevent most of the participants with normal hearing from attaining ceiling performance on the task. All of the
participants were asked to verbally repeat the word that was presented aloud. The experimenter subsequently recorded their responses into computer files on-line. No feedback was provided.

Results and Discussion

Analysis of Raw Scores

The data from all 41 subjects were submitted to a 4-way repeated-measures ANOVA, with the factors of Presentation Mode (Visual-alone, Audio-alone, vs. Audiovisual), Talker Variability (Single vs. Multiple), and Lexical Competition (Easy vs. Hard) treated as within-subjects variables, and Group (normal hearing vs. cochlear implant) as a between-subjects variable.

Table 2 presents a summary of the raw scores obtained by the two participant groups as a function of presentation format, lexical competition and talker variability. It should be noted that, with the exception of the visual only presentation format, direct comparisons of the raw scores between the cochlear implant and normal hearing control groups are not valid. Recall that in formats where auditory speech information was presented, the cochlear implant group was tested in the quiet while the normal-hearing comparison group was tested in white noise to reduce performance below ceiling levels. It is the pattern of performance within each listener group that can be compared, not absolute scores between groups. It is interesting to note, therefore, that several commonalities between the two comparison groups emerged with respect to the manipulated factors.

<table>
<thead>
<tr>
<th>Talker Condition</th>
<th>Lexical Discrim.</th>
<th>CI (N=20)</th>
<th>NH (N=21)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Presentation Format</td>
<td>Presentation Format</td>
</tr>
<tr>
<td></td>
<td></td>
<td>V</td>
<td>A</td>
</tr>
<tr>
<td>Single-Talker</td>
<td>Easy</td>
<td>23.9</td>
<td>34.4</td>
</tr>
<tr>
<td></td>
<td>Hard</td>
<td>8.9</td>
<td>29.4</td>
</tr>
<tr>
<td>Multiple-Talker</td>
<td>Easy</td>
<td>21.7</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td>Hard</td>
<td>9.4</td>
<td>23.9</td>
</tr>
</tbody>
</table>

Table 2. Mean percent correct word recognition performance by condition.

Effects of Presentation Format

Figure 1 displays the overall performance of the two participant groups under the three presentation formats, averaged across talker condition and lexical competition. A significant main effect of Presentation Mode was observed, $F(1,39) = 352.6, p < 0.001$. Regardless of group membership, performance in the visual-alone condition ($M = 13.81$) was worse than in the audio-alone condition ($M = 39.18$), which was even worse than in the audiovisual condition ($M = 68.20$). Presentation format also interacted with the Group variable. Simple effects analyses revealed that this interaction was supported by differences in performance between the normal-hearing and cochlear implant groups in the visual-alone, $F(1,39) = 5.40, p = 0.03$, and the audio-alone, $F(1,39) = 9.73, p = 0.003$, presentation conditions. As shown in Figure 1, cochlear implant users obtained higher scores in the visual-alone condition than their normal-hearing counterparts. In contrast, normal hearing subjects tested in noise obtained higher scores in the audio-alone condition than cochlear implant users tested in the quiet.
Figure 1. Percent correct word recognition performance of the CI users and the normal hearing participants under the three presentation formats averaged over talker and lexical variables. Listeners with CIs were tested in quiet and normal-hearing comparison participants were tested in noise at −5 dB SPL.

Correlations. One way to evaluate the relationship and to assess the underlying similarities among visual-alone, audio-alone, and audiovisual speech perception skills is to correlate performance for each presentation format. Previous studies have demonstrated significant correlations among visual-alone, audio-alone, and audiovisual performance on consonant perception measures (Grant & Seitz, 1998) and between visual-alone and audio-alone performance on word perception measures (Watson, Qiu, Chamberlain, & Li, 1996). Grant et al. found that for sentence materials only audio-alone and audiovisual performance were correlated. However, correlations with visual-alone scores failed to reach significance.

To examine the relations among the presentation formats, speech intelligibility scores obtained under each presentation format were correlated separately for each group of listeners. Figure 2 shows each subject’s performance in the audio-alone (filled circles) and visual-alone (open circles) presentation conditions plotted as a function of audiovisual speech perception performance. The top panel shows this data for the CI group while the bottom panel shows the data for the NH group. The data are collapsed over talker and lexical variables. Significant correlations were observed between audio-alone performance and audiovisual performance for both groups of listeners, $r(20) = +0.81, p<.001$, for CI listeners and $r(21) = +0.67, p<.001$, for NH listeners. However, the correlations between visual-alone performance and audiovisual performance were not significant for either group. Additional correlations were computed between the audio-alone and visual-alone performance for each group of listeners. None of these correlations was significant either. Inspection of Figure 2 shows that the individual scores for listeners using cochlear implants varied over a somewhat greater range than the individual performance for listeners with normal hearing tested in noise. Audiovisual speech perception scores for listeners with
normal hearing varied from 60% to 92% and performance for listeners using cochlear implants ranged from 53% to 88%, when one poorly performing listener was excluded from each group. Audio-alone performance varied between 22% and 74% for the normal hearing group, and between 10% and 61% for the cochlear implant group. Ranges in the visual-alone condition were more restricted, with performance between 3% and 22% for the normal hearing group and between 6% and 31% for the cochlear implant group. It is very likely that the lack of correlations involving visual-alone performance is due to floor effects and the absence of variability for performance in the visual-alone presentation condition. Performance in this condition ranged from 0% to 44.4%.

Figure 2. Scatterplot of individual participants' performance in a unimodal presentation format (audio-alone or visual-alone) and their performance in the audio-visual presentation format. Performance is shown separately for CI listeners or listeners with normal hearing. Data were averaged across talker and lexical variables. Regression lines are shown separately for correlations between AV vs. A as well as for AV vs. V performance.
Effects of Lexical Competition

The omnibus 4-way ANOVA also revealed a significant main effect of Lexical Competition, $F(1,39) = 158.89, p < 0.001$. Easy words ($M = 45.83$) were recognized better than Hard words ($M = 34.96$). Interestingly, this factor also interacted with the group variable, as shown in Figure 3. A simple effects analysis of this interaction showed that it was due to a difference between groups in accuracy of the Hard words $F(1,39) = 5.5, p = 0.024$, but not the Easy words $F(1,39) = 1.38$, n.s. As shown in Figure 3, normal-hearing subjects identified Hard words better than cochlear implant users.

![Graph showing percent correct word recognition by Group (CI vs. NH) for Easy and Hard words.]

Figure 3. The percentage of words correctly identified by listeners with CIs and the control participants with normal hearing as a function of lexical competition of the stimulus words

Effects of Talker Variability

The main effect of Talker Variability was also significant, $F(1,39) = 13.69, p = 0.001$. Overall, single talker lists ($M = 42.01$) were identified better than multiple talker lists ($M = 38.78$). Talker Variability also interacted with Presentation Mode, $F(1,39) = 4.73, p = 0.01$. Figure 4 illustrates this interaction. Simple effects analyses revealed that the source of this two-way interaction was the difference between single and multiple talker lists in the audiovisual presentation condition only, $F(1,39) = 16.76, p < 0.001$; single talker lists were identified better than multiple talker lists.

Finally, Figure 4 shows the significant three-way interaction between Presentation Format, Talker, and Lexical Competition, $F(1,39) = 4.33, p = 0.011$. For ease of comparison, this figure has been split so that each hearing group is represented separately. The top two panels show the interaction for the
cochlear implant group, and the bottom two panels show the interaction for the normal-hearing group. Tests of simple effects showed a complex pattern of results. For easy words (the left panels for both groups), the difference in performance between single and multiple-talker lists did not differ in any presentation condition, although there was a marginal effect of talker in the audiovisual presentation condition, $F(1,39) = 3.4, p = 0.073$. For hard words (the right panels for both groups), however, the simple effects of talker were significant in both the audiovisual presentation condition, $F(1,39) = 18.06, p< 0.001$, and the audio-alone condition, $F(1,39) = 5.96, p = 0.019$. In addition, the analysis revealed a marginal effect of talker in the visual-alone presentation condition, $F(1, 39) = 3.4, p = 0.073$. As shown in the figure, whenever there was a significant effect of talker, performance on single talker lists was better than performance on multiple talker lists, except in the marginally significant case of visual-alone performance for lexically hard words.

![Graphs showing percent correct word recognition for easy and hard words, with data for cochlear implant users and normal hearing groups, showing interaction effects for talker and presentation format.]

**Figure 4.** The percent of words correctly identified in each presentation format as a function of talker variability for lexically easy words and lexically hard words. The top two panels show data for cochlear implant users and the bottom two panels show data for the normal-hearing group.
Audiovisual Integration (I)

Conceptually, audiovisual integration means that listeners can do more than simply add the perceptual cues obtained from the auditory and visual modalities to increase their speech perception scores. One way to quantify the extent of this integration process is to compare the observed audiovisual performance to an estimate of the performance that a listener could achieve assuming no integration at all. As an example, consider an individual who receives a short open-set word recognition task consisting of the four words: lace, wife, short, and long. Under audio-alone presentation, this hypothetical listener correctly perceives lace and short (50% correct). Under visual-alone presentation, this individual correctly perceives wife and short (50% correct). The best the listener could do then without any integration would be to correctly identify lace, wife, and short in the audiovisual modality (75% correct). Any additional gain above this level of performance would provide evidence for some form of integration.

Unlike the hypothetical case, however, the listeners in the present experiment did not receive the same tokens under each of the three presentation formats, so the observed responses could not be compared in this way. Only the raw percentage scores could be used for this purpose. We assumed that the upper bound estimate used in this experiment was simply the sum of visual-alone and audio-alone performance. The example above can be used to demonstrate that this is a more conservative estimate. Comparing the auditory and visual responses above, one would expect audiovisual performance of 75% correct (lace, wife, short). On the other hand, if only percentage scores were used, one would expect audiovisual performance to be 100% correct (50%+50%). Limiting the process of AV integration to simple addition would limit perception to this upper bound score.

Individual Data

Figure 5 shows individual subjects’ performance as a function of the sum of audio-alone and visual-alone performance. The top panel shows the data for listeners in the cochlear implant group while the panel shows the data for listeners in the normal-hearing group. In both figures, the diagonal represents the audiovisual score that would be obtained by each listener if audiovisual performance were simply equal to the sum of the audio-alone and video-alone scores. The observed scores in Figure 5 demonstrate that almost all of the listeners were able to perform above the prediction expected from simple additive integration. Audiovisual performance in excess of this (i.e., performance above the diagonal) provides support for non-additivity of audiovisual integration.
Figure 5. Scatterplot of individual participants' performance in the AV presentation mode plotted against the sum their auditory-only and visual-only scores. Scores are averaged over lexical competition and talker variability. Scores above the diagonal represent AV scores greater than the simple sum of performance in the two unimodal conditions.
Group Analysis

An index of audiovisual integration (I), can be calculated as the difference between observed audiovisual performance and the predicted performance estimated by the simple addition of visual-alone and audio-alone scores (equations 1 and 2). Again, this is simply the amount by which the observed performance exceeded that predicted by an additive model of audiovisual speech perception; it is also the distance from the diagonal to the observed score in Figure 5. These values can be used to estimate the portion of audiovisual speech intelligibility that is attributable to the processes of integration. Due to ceiling effects, this analysis can only be used when the sum of audio-alone and visual-alone performance is less than 100%. As an upper bound performance estimate, the index “I” was calculated for each participant using equation 1. Only participants whose audio-alone + visual-alone performance was less than 90% in every condition were used in this analysis. This criterion was chosen to minimize ceiling effects in the calculation of “I”. The final analysis included 17 participants with cochlear implants and 17 participants with normal hearing.

\[ I = AV_{measured} - AV_{estimate} \]  

where

\[ AV_{estimate} = A_{measured} + V_{measured} \]

The “I” scores from these 34 subjects were used as the dependent variable and submitted to a three-way repeated-measures ANOVA using Talker, Lexical Competition and Group as independent variables. The analysis revealed that the index of integration, “I”, differed between the two groups, \( F(1,34) = 6.49, p = 0.016 \). Overall, the integration score “I” was higher for cochlear implant users (M = 20.43) than for normal hearing listeners (M = 11.84). In addition, the integration score “I” was larger: for lists of lexically hard words (M = 21.03) than for lists of lexically easy words (M = 11.24), \( F(1,34) = 8.68, p = 0.006 \). A marginally significant difference in the integration score was also found for the Talker factor. Performance on multiple talker lists (M = 18.57) was significantly higher than on single talker lists (M = 13.70), \( F(1,34) = 3.65, p = 0.07 \). None of the other interactions reached significance.

The overall pattern of results suggests that audiovisual integration benefit (I) is greatest in the conditions where the auditory or visual portion of the stimulus alone is ambiguous or underspecified in some way, causing performance in the single modality presentation conditions to suffer. As a result, the combined information increases performance above and beyond the level that would be expected from simple addition of information from the two separate input modalities. It is interesting to note, however, that this principle applies whether unimodal stimulus ambiguity arises due to properties of the perceiver (i.e., the effect of group), properties of the stimulus item itself (i.e., the effect of lexical competition), or properties of the environment in which those stimuli are presented (i.e., the marginal effect of talker).

Visual Enhancement (Rv)

In their pioneering study of audiovisual speech perception, Sumby and Pollack (1954) developed a quantitative metric to evaluate the gains in speech intelligibility performance due to the addition of visual information from seeing a talker’s face. Because speech perception scores have a theoretical maximum (i.e., perfect performance), the measure was developed to show the extent to which additional visual information about speech improved performance relative to the amount by which audio-alone performance could possibly improve. Their metric, \( R_v \), can be used to assess the extent of visual enhancement for an individual perceiver in our study. To assess visual enhancement, \( R_v \) was calculated
for all 41 participants based on the recognition scores obtained in the audiovisual and audio-alone conditions using Equation 3, from Sumby and Pollack (1954). In the equation, "AV" is performance in the audiovisual presentation condition, and A is performance in the audio-alone condition. $R_a$ was calculated separately for lexically easy and lexically hard words in each of the two talker conditions. The $R_a$'s resulting from this analysis are reported in Table 3 and displayed graphically in Figure 6.

$$R_a = \frac{AV - A}{1.0 - A}$$

<table>
<thead>
<tr>
<th>Talker Condition</th>
<th>CI</th>
<th>NH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total R</td>
<td>Easy R</td>
</tr>
<tr>
<td>Single-Talker</td>
<td>.56</td>
<td>.64</td>
</tr>
<tr>
<td>Multiple-Talker</td>
<td>.43</td>
<td>.50</td>
</tr>
</tbody>
</table>

Table 3. Mean visual enhancement ($R_a$) by listener group (CI or NH), list type and presentation format.

![Figure 6](image.png)

Figure 6. Visual enhancement ($R_a$) for the CI users and the comparison normal-hearing participants calculated separately in each of the two talker conditions. $R_a$ represents the gain in word recognition performance due to the addition of visual speech information.
An ANOVA was used to analyze the visual enhancement scores (R_s) in each condition. Listener group was a between subject factor while talker variability and lexical competition were within subject factors. Overall, R_s was larger for single talker (M = 0.50) conditions than for multiple talker conditions (M = 0.43), F(1,39) = 4.78, p = 0.04. This result suggests that listeners realized more of the potential benefit from the addition of visual information to auditory information under single talker presentation conditions than under multiple talker conditions.

The interaction between Talker and Group was also significant, F(1,39) = 4.05, p = 0.05. Figure 6 shows R_s scores for the relevant cells of this interaction. Simple effects analysis revealed that the interaction was due to a difference in visual enhancement for single vs. multiple talker lists for cochlear implant users, F(1,39) = 8.58, p = 0.006, but not for normal-hearing participants, F(1,39) < 1, n.s.

There was also a marginal main effect of Lexical Status, F(1,39) = 3.82, p = 0.06. R_s scores for lexically easy words (M = 0.51) were higher than the scores for lexically hard words (M = 0.42). This result indicates that listeners obtained somewhat greater visual benefit from words that have less competition than from words that have more competition. All other main effects and interactions were not significant.

General Discussion

The results from this study of audiovisual word recognition suggest that similar factors affect the process of spoken word recognition in normal hearing listeners and postlingually deafened adults with cochlear implants. Overall, we observed only a marginal difference in the mean performance levels of both groups (p = 0.07), indicating that our goal of equating performance across both listener groups was met with the signal-to-noise ratio we picked for use with the normal-hearing participants. It is important to note here that using broad-band noise to degrade speech for the normal-hearing listeners is not the same type of auditory degradation experienced by cochlear implant users. However, overall similarities between the groups in terms of the effects of our manipulated factors provide some new insights into the perceptual and linguistic processes at work in both groups of listeners during spoken word recognition.

Analysis of Raw Scores

Presentation Format. Presentation Format affected both groups of listeners in similar ways: visual-alone performance was consistently below audio-alone performance. In addition, performance was always best when both auditory and visual sources of information were available for speech perception. The significant cross-over interaction between Group and Presentation Format revealed that normal-hearing listeners performed better than cochlear implant users in the audio-alone condition but that CI users performed better than normal hearing listeners in the visual-alone condition. This finding is consistent with a recent report by Bernstein, Auer, and Tucker (2001) who found reliable differences in the performance of normal-hearing and hearing-impaired speechreaders on a visual-alone speech perception task. The pattern of results observed in the present study may be due to the way lip-reading skills were acquired in these patients. The CI users in our sample were all progressively deafened post-lingually. It is possible that over long periods of time, a gradual reliance on lip-reading eventually leads to greater use of the visual correlates of speech when the auditory information in the speech signal is no longer sufficient to support word recognition. Further work on the time-course of learning speechreading skills in post-lingually deafened adults prior to implantation is needed before any definitive conclusions can be drawn.
We also found that the two groups of listeners achieved roughly the same level of performance in the audiovisual condition, even though they differed in the extent to which they were able to perceive speech from either sensory modality alone. This result illustrates the complementary nature of auditory and visual information about speech (Summerfield, 1987): when the information available in one sensory modality (e.g., audition) is noisy, degraded, or impoverished, information available in the other modality (e.g., vision) can "make up the difference" by providing complementary cues that combine to enhance overall word recognition performance in a particular task.

According to the event-based theory of speech perception (Fowler, 1986), the complementarity of the two sources of sensory stimulation about speech arises because auditory and visual sources of information in speech are structured by a unitary, underlying articulatory event. That is, when a person speaks, their articulatory patterns and gestures simultaneously shape both auditory and optic patterns of energy in very specific, lawful ways. The relations between the two modalities, then, is specified by the information relating each pattern to the common, underlying, dynamic vocal tract gestures of the talker that produced them. It is precisely this time-varying articulatory behavior of the vocal tract that has been shown to be of primary importance in the perception of speech (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Remez, Fellowes, Pisoni, Goh, & Rubin, 1998; Remez, Rubin, Pisoni, & Carrell, 1981; Remez, Rubin, Berns, Pardo, & Lang, 1994).

With this conceptualization in mind, the sensory and perceptual information relevant for speech perception is modality-neutral or amodal, since it can be carried by more than one sensory modality (Fowler, 1986; Gaver, 1993; Remez et al., 1998; Rosenblum & Saldaña, 1996). The amodal nature of phonetic information is demonstrated convincingly in studies showing that perceptual information obtained via the tactile modality, in the form of Tadoma, can be used and integrated across sensory modalities in speech perception (Fowler & Dekle, 1991), albeit with limited utility.

Although the information necessary for speech perception may be modality-neutral, the internal representation of speech appears to be based on an individual's experience with perceptual events and actions in the physical world (Lachs, Pisoni, & Kirk, 2001). Thus, awareness of the intermodal relations between auditory and visual information is contingent upon experience with more than one sensory modality. Because our CI participants were all post-lingually deafened adults, they had all had prior experience with the auditory properties of speech and had acquired knowledge of the lawful correspondences between auditory and visual correlates of speech. As the present findings demonstrate, the CI participants were able to make use of this experience under audiovisual presentation and were able to recognize isolated words at levels comparable to our normal hearing participants.

**Lexical Competition Effects.** We also found robust effects of Lexical Competition in this study. For both groups of listeners, lexically easy words were recognized better than lexically hard words, indicating that normal-hearing and cochlear implant listeners organize and access words from memory in fundamentally similar ways. Thus, phonetically similar words in the mental lexicons of CI users compete for selection during word recognition. This process is also affected by word frequency such that higher frequency words are more apt to win out among phonetically similar competitors. The finding that lexical competition affected our CI group is not surprising because the participants in this group were all post-lingually deafened and had no evidence of any central nervous system involvement prior to or after the onset of deafness. Presumably, they developed extensive lexical representations when they had normal hearing and retained some form of this information over time after their hearing loss.

Despite these overall similarities, several differences in the effects of lexical competition were found between the groups. CI users performed more poorly on lexically hard words than did normal hearing listeners. However, performance for both groups on lexically easy words was statistically
equivalent. This interaction suggests that the CI users were less able to make the fine acoustic/phonetic distinctions among words that are needed to distinguish lexically hard words from their phonetically similar neighbors. Although the cochlear implant appears to provide enough auditory information to recognize words when only gross acoustic cues are sufficient, in many patients the implant may not be sufficient to provide the more fine-grained phonetic information necessary to discriminate between very similar lexical candidates.

**Talker Variability Effects.** Across both groups of listeners, we observed an interesting three-way interaction between talker variability, lexical competition and presentation format. For lexically easy word lists, talker variability did not play a role in word recognition performance, although a marginal difference in performance was observed between single-talker and multiple-talker lists under audiovisual presentation. However, the effects of talker variability were robust for lists of lexically hard words. Performance on single talker lists was better than performance on multiple talker lists in both the audiovisual and audio-alone presentation conditions. The difference was only marginally significant in the visual-alone condition. The results on the effects of talker variability are consistent with the proposal that repeated exposure to a single talker allows the listener to encode voice-specific attributes of the speech signal. Once internalized, voice-specific information can improve word recognition performance (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1995).

The “single talker advantage” appears to be most helpful when there is a great deal of lexical competition among words and fine phonetic discrimination is required, as there is with lexically hard words. Talker-specific information appears to be used in conditions where a detailed perceptual representation of the acoustic/phonetic input can serve to more clearly disambiguate multiple word candidates from within the lexicon. The reduced magnitude of the talker effect in visual-alone conditions may be because the perceptual input provided by the optical display of speech itself is insufficient to specify a set of lexical candidates small enough that fine-grained phonetic information can improve word recognition. Also, it is very likely that detailed talker-specific information would be difficult to obtain for visual-alone presentation of short words presented in isolation (see Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). Nygaard et al found that it was much more difficult for participants to learn novel voices from isolated words than for short sentences. Apparently, participants are able to make use of several different acoustic correlates of talker-specific information from sentences that are not present in isolated words.

**Analysis of Perceptual Errors**

Up to this point, only the accuracy scores for correct word recognition have been presented. However, it is also possible to examine the errors made by participants in order to determine if their responses were randomly distributed, or if they followed a pattern that made use of partial stimulus information and some form of sophisticated guessing strategies. By determining whether confusions and error responses were lexical neighbors of the intended target words, we can assess the extent to which Presentation Format and Talker Variability affected the information available to the perceiver under the various conditions. Our earlier discussion suggested that audiovisual presentation influences lexical competition by providing more robust perceptual information as input to the word recognition process. If this prediction is correct then we would expect that even in cases when the ultimate output of the word recognition system was incorrect, the incorrect response would be perceptually more similar to the target word in those cases where the input to the system was more clearly specified.

For the purposes of this error analysis, we considered a response to be a lexical neighbor of the original target word if it differed from the target word by the insertion, deletion or substitution of a single phoneme (Luce & Pisoni, 1998). For example, if the target word was “cat”, the response “sat” would be
considered a lexical neighbor by substitution, "at" would be considered a neighbor by deletion, and "scat" would be considered a lexical neighbor by insertion.

To carry out the error analysis, we converted the target words and the incorrect responses into phonetic representations. The target words and responses were transformed into a computer-readable phonetic notation using DECtalk (DCT03) Text-to-Speech System. Each target word-response pair that was incorrect was then analyzed in greater detail to determine the nature of the phonetic and lexical confusions. The proportion of errors that were in the lexical neighborhood of the target word ("neighborhood errors") was then computed. This measure may be thought of as an index of how similar an incorrect response was to the original target word. Because of the significant effects of talker variability found earlier, the data were analyzed separately for single talker and multiple talker lists. Results from both sets of analyses are reported separately below.

Single-Talker Confusions. We first present the results of the error analysis using data from the single-talker conditions. The proportion of incorrect responses that were neighbors of the target word was computed and then submitted to a 2 (Presentation: audiovisual or audio-alone) x 2 (Lexical Competition: Easy or Hard) x 2 (Group: normal hearing or cochlear implant) ANOVA. Responses in the visual-alone condition were not included in the error analysis because the criterion for neighborhood membership in visual-alone conditions is still an unresolved issue awaiting further investigation (however, see Auer & Bernstein, 1997 for preliminary work on this topic).

The ANOVA revealed a significant main effect of Presentation Format, $F(1,39) = 4.713$, $p = 0.036$. The top panel of Figure 7 shows the proportion of incorrect responses that were lexical neighbors of the target word for both groups of participants under the audiovisual and audio-alone conditions. For both groups of listeners, it can be seen that a higher proportion of errors were lexical neighbors of the target word under audiovisual presentation than under audio-alone presentation. However, this factor did not interact with listener group, indicating that a participant's hearing status did not affect this pattern. In general, the addition of visual information to auditory information helps to narrow down or constrain the set of possible lexical candidates competing for selection during word recognition.

The bottom panel of Figure 7 shows the proportion of errors that were neighbors of the target word for both groups of participants separately by lexical competition. The figure shows that incorrect responses were more often selected from the neighborhoods of hard words than easy words. This pattern was confirmed by the presence of a significant main effect of lexical competition, $F(1, 39) = 30.0$, $p < 0.001$. The interaction between lexical competition and listener group was not significant, indicating again that the pattern was similar across both groups of listeners. None of the remaining interactions was significant. Once again, the results indicate that the combination of auditory and visual information about speech produces a more detailed perceptual representation that places additional constraints on the set of possible response alternatives in the mental lexicon.

For the single talker conditions, it is apparent that both normal-hearing listeners and listeners with cochlear implants display very similar error patterns and lexical confusions. Relative to audio-alone presentation, audiovisual presentation increased the likelihood that an incorrect response would be selected from within the lexical neighborhood of the target word. This pattern of errors is consistent with the proposal that the additional visual information contained in an audiovisual signal serves to refine and constrain the sensory and perceptual information needed for lexical discrimination and selection. The additional visual information under audiovisual presentation influenced and controlled the participants' responses, even when they failed to identify the target word correctly.
The lexical properties of the target word also had an effect on the nature and number of errors that came from the same similarity neighborhood. Specifically, more errors were observed in the neighborhoods of lexically hard target words than lexically easy target words. This is not a surprising result, because lexically hard words, by definition, generate more competition during recognition and have more neighbors with which they might be confused than do lexically easy words. Because participants in this study were required to respond with only English words in an open-set response format, the response confusions that they generated were simply more likely to produce an error in the neighborhood for hard words than they were for easy words.
Multiple-Talker Confusions. The pattern of incorrect responses for the multiple-talker lists was similar to the results observed for the single talker lists. As in the previous analysis, the proportion of incorrect responses that were in the neighborhood of the target word was used as the dependent variable in a 2 (Presentation Format) x 2 (Lexical Competition) x 2 (Group) ANOVA. The results showed a main effect of Presentation Format, $F(1, 38) = 49.50, p < 0.001$. Audiovisual presentation was better than audio-alone presentation. Figure 8a shows this main effect, with the two listener groups displayed separately for ease of comparison. The interaction between Presentation Format and listener Group was only marginally significant, $F(1, 38) = 3.093, p = 0.087$. Inspection of the top panel of Figure 8 shows that participants with cochlear implants received greater benefit from audiovisual presentation than the normal hearing comparison group; once again audiovisual presentation increased the number of response confusions within a lexical neighborhood.

![Bar chart showing percentage of neighbor responses for CI and NH groups with auditory-alone and audiovisual conditions.](chart1.png)

**Figure 8.** Percentage of error responses that were within the lexical neighborhood of the target word for the multiple-talker conditions. Error responses are shown separately for listeners with CIs and the normal hearing participants as a function of presentation format and lexical competition of the target word.
As in the single talker analysis, we also found a main effect of Lexical Competition, $F(1, 38) = 28.446, p < 0.001$. The proportion of confusions within a lexical neighborhood was much higher for lexically hard words than for lexically easy words. The bottom panel of Figure 8 shows the proportion of errors that were neighbors of the Easy and Hard words, separately for each listener group. Again, the interaction between Lexical Competition and listener Group was marginally significant in the multiple talker analysis, $F(1, 38) = 3.054, p = 0.089$. Examination of the figure shows that this result was due to an increase in the number of incorrect responses from the neighborhoods of hard words relative to easy words in the normal-hearing group. The increase in neighborhood responses for hard words in the cochlear implant group was smaller.

The analysis of the incorrect responses from the multiple-talker lists replicates the patterns observed in the single talker lists, although the differences were somewhat smaller and only marginally significant. Audiovisual presentation increased the number of lexical confusions that came from within the same lexical neighborhood as the target word. In addition, the added task load of recognizing words from multiple talkers seems to have produced smaller differences in the extent to which these factors affect participants with either normal hearing or cochlear implants. However, the overall pattern of the responses is similar across talker conditions and listener groups.

**Audiovisual Integration ("I")**

We also assessed the extent to which our participants combined auditory and visual information. Our measure of integration, "I", reflected the amount by which the observed audiovisual performance differed from the scores predicted by a simple, additive model of audiovisual integration. Overall, the two hearing groups differed on this measure; the cochlear implant users benefited more from combined audiovisual inputs than the normal hearing group. Although the CI group were better speechreaders than the NH group, the magnitude of the difference in performance in the visual-alone condition was not as great as the advantage the NH group had in the audio-alone condition. Thus, on the whole, the CI group performed worse in the unimodal conditions than the NH group. However, no difference in performance was observed in the audiovisual condition between the two groups. The results show that CI users were better able to combine the redundant, multimodal information than the NH group.

We also observed effects of lexical competition on the integration scores. Specifically, the measure of audiovisual integration was better for lexically hard words than for lexically easy words. Hard words require the perceiver to make fine phonetic distinctions for accurate identification. If these distinctions cannot be made based on auditory information in the acoustic signal, then the addition of visual speech information can serve to disambiguate the competing lexical entries. In contrast, easy words are much more distinct and discriminable and thus acoustic information alone may be adequate to identify these words. The addition of visual cues to highly discriminable audible patterns contributes little to the recognition of lexically easy words.

**Visual Enhancement (Rv)**

The second measure of audiovisual integration that we examined was Rv. This is the gain in speech intelligibility due to the presentation of combined audiovisual information relative to audio-alone presentation (Sumby & Pollack, 1954). This measure of integration was used because gains above audio-alone performance due to audiovisual presentation are necessarily limited by the theoretical maximum: 100% performance. We found that talker variability only affected visual enhancement scores for the CI users in the single talker condition. There was no effect of talker variability on visual enhancement for normal-hearing listeners. This finding does not mean that NH listeners were unaffected by talker
variability. However, talker variability did not affect the degree to which normal-hearing listeners could combine audiovisual information. The present findings suggest that CI users are better able to extract idiosyncratic talker information from audiovisual displays than NH listeners are, because they rely more on visual speech information to perceive speech in every day situations. With repeated exposure to audiovisual stimuli spoken by the same talker, the CI users exhibited a gain above and beyond that observed in normal hearing listeners. Cochlear implant users may be able to acquire more detailed knowledge of the cross-modal relations between audition and vision for a particular talker. Because NH listeners can successfully process spoken language by relying entirely on auditory cues, they may not have learned to utilize visual cues as successfully (Bernstein et al., 2001). For NH listeners, combined audiovisual information from a single talker may not provide any additional information about that talker than the cues provided by audio-alone presentation. This is especially true in a short-term laboratory experiment like the present one. The normal-hearing listeners have little prior exposure to visual-alone stimuli. If normal-hearing listeners came back repeatedly to the lab and were forced to listen to degraded speech in noise, they might develop a greater awareness of the inherent relationships between auditory and visual speech and show these effects after a short period of time.

Limitations of Present Findings

Although this is the first detailed study of audiovisual word recognition in postlingually deafened adult patients following cochlear implantation, there are a number of limitations in the experimental design that are worth mentioning here. First, we want to emphasize that direct comparisons of the raw scores between the two groups of listeners used in this study need to be made with some degree of caution. The data obtained from the normal-hearing listeners were collected under masking conditions using white noise while the data obtained from the CI patients were collected in the quiet. Masking noise was used with the normal-hearing listeners simply to reduce scores from ceiling levels of performance. The nature of the degradation resulting from the presentation of speech in noise to normal-hearing listeners is not equivalent to the transformation of speech that is processed by a cochlear implant and then presented as an electrical signal to a hearing-impaired listeners' auditory system. The two forms of signal degradation are not commensurate despite the fact that there was no overall statistical difference in the audiovisual condition between the two groups of listeners in the global analysis of variance of the main effects. Equivalent levels of performance on the word recognition task do not imply that the signals were encoded and processed in the same manner by both groups of listeners.

While there were similarities in performance across the two groups as a function of the variables under study, a number of the comparisons revealed small and consistent differences within groups. It is these differences that are informative and provide some new insights into how patients with CIs recognize spoken words and make optimal use of the available auditory information provided by their implant under different listening conditions. Both groups of listeners combine and integrate auditory and visual information about spoken words, but the CI patient appears to make somewhat better use of the visual information in more difficult listening conditions when there is ambiguity about the talker, or when they are forced to make fine phonetic discrimination among acoustically confusable words. The deaf listeners combine visual information with the available auditory information conveyed by their cochlear implant to support open-set word recognition, but they do so in somewhat different ways than the normal-hearing listeners.

It is not surprising that the same basic underlying lexical selection processes are used by both groups of listeners in this study. After all, the patients with cochlear implants were all postlingually deafened and had acquired language and lexical knowledge normally prior to the onset of their hearing loss. There is little reason to expect any large differences in the effects of lexical competition in these patients (Luce & Pisoni, 1998). Both groups showed sensitivity to the lexical manipulations used here.
The pattern of their word recognition scores demonstrates that they recognize spoken words “relationally,” in the context of other words they know and have in their mental lexicons. Both groups are sensitive to frequency and phonetic similarity. The differences between the two groups of listeners occur earlier during perceptual analysis, when the initial sensory information in encoded prior to lexical selection.

In future studies using normal-hearing listeners, it may be better to use other methods of signal degradation, such as noise-band speech, to simulate the nature of the hearing loss and signal transformation produced by a cochlear implant (Dorman, Loizou, Fitzke, & Tu, 1998; Dorman, Loizou, Kemp, & Kirk, 2000; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). The use of isolated words as stimulus materials in this study also presents several limitations in terms of providing reliable and useful information about variability and the dynamics of a talker’s vocal tract. Earlier studies by Nygaard et al. (1995) and Nygaard and Pisoni (1998) have shown that it is much more difficult to learn to identify novel voices from isolated words than longer sentences. Although greater in length, sentences provide additional information about prosody and timing. This may help to increase familiarity with a novel voice in a shorter period of time with less exposure.

Although we did not observe any consistent individual differences in performance across any of the experimental conditions, it is well known that listeners with cochlear implants display a wide range of individual differences in performance on various outcome measures. The patients selected for this study were all good users who were able to derive large benefits from their cochlear implants. All of them were able to respond appropriately in an open set word recognition task given the limited auditory input provided by their implant. Examination of the other implant patients who do more poorly under these conditions may reveal a wider range of scores, a different pattern of audiovisual integration skills and different levels of reliance on visual information about speech.

Clinical Implications

With recent advances in cochlear implant technology, many postlingually deafened adults are now able to achieve very high levels of spoken word recognition through listening alone (Kirk, 2000). Other patients may derive substantial benefit from a cochlear implant only when the auditory cues they receive are combined with visual information from a talker’s face. Like listeners with normal hearing, most cochlear implant recipients benefit from the presence of visual speech cues under difficult listening situations, such as when the talker is speaking rapidly or has an unfamiliar dialect, or when listening in the presence of background noise.

Structured aural rehabilitation activities with a sensory aid (either a cochlear implant or a hearing aid) often rely on highly constrained and organized listening activities intended to enhance the ability to discriminate or recognize various acoustic cues in speech. Words or sentences are usually presented in the audio-alone format by a single clinician. There has been little systematic application of the findings from recent studies on variation and variability in speech perception and multi-modal perception to therapy and rehabilitation with clinical populations. The findings from the present study suggest that it may be fruitful to apply some of the knowledge gained recently about audiovisual speech perception to clinical problems associated with intervention and aural rehabilitation after a patient receives a sensory aid. Exposure to multiple talkers and a wide range of speaking styles in both audio-alone and auditory-visual modalities may provide patients with a greater range of stimulus variability during the first few months of use after receiving an implant; this in turn may help patients develop more robust perceptual strategies for dealing with speech in real world listening conditions that exist outside the clinic and research laboratory. Specially-designed word lists can be developed and used for training materials under different presentation formats to emphasize difficult phonetic contrasts that may be hard to recognize under audio-
alone conditions but easy to identify under audiovisual conditions. Similarly, activities such as connected discourse tracking using audiovisual stimuli may promote the development of robust multimodal speech representations and enhance spoken language processing. Auditory training activities using multimodal stimuli may enhance both the perception of audio-alone and visual-alone speech cues. As noted above, not all cochlear implant recipients can recognize speech through listening alone. For many of these patients, the cochlear implant serves as a sensory aid to improve lip reading skills they already have and use routinely in processing spoken language.

References


Lachs, L. (1999). A voice is a face is a voice, *Research on Spoken Language Processing No. 23*. Bloomington, IN: Speech Research Laboratory, Indiana University Bloomington.


