Toward the Automatic Generation of Cued Speech

Maroula S. Bratakos, Paul Duchnowski and Louis D. Braida

Abstract

Although Manual Cued Speech (MCS) can greatly facilitate both education and communication for the deaf, its use is limited to situations in which the talker, or a transliterator, is able to produce cues for the cue receiver. The availability of automatically produced cues would substantially relax this restriction. However, it is unclear whether current automatic speech recognition (ASR) technology would be adequate for producing cues automatically. To evaluate this adequacy, we measured the speech reception scores achieved by highly experienced receivers of MCS when cues were produced by one actual and several hypothesized ASR systems, as well with speechreading alone and with MCS. The systems studied modelled the effects of various types of errors and delays associated with the recognition process, and included a representative state-of-the-art, speaker-dependent phonetic recognizer. Results indicate that while current speaker-independent recognition technology probably would not provide useful cues, the cues provided by a speaker-dependent ASR would aid speech reception substantially. The benefit provided by automatically generated cues is heavily dependent on the use of an effective visual display that minimizes the effects of recognition delays so that cues are perceived in synchrony with facial actions.

Introduction

Manual Cued Speech (MCS; Cornett, 1967) is a visual display of a discrete symbolic code used in conjunction with speechreading in everyday environments. The system can be taught to very young deaf children and greatly facilitates both language learning and general education. Evaluations of deaf children who use MCS by Kaplan (1974), Ling and Clarke (1975), Clarke and Ling (1976), and Nicholls and Ling 1982) have documented the benefits provided by the system. The results reported in the latter study were particularly dramatic. Eighteen senior-class children who had used MCS for an average of 7.3 years were tested for reception of words in both Low Predictability (LP) and High-Predictability (HP) sentences similar to those used in the SPIN test (Kalikow, Stevens, & Elliot, 1977). Keyword scores improved from 25.5% (LP) and 32.0% (HP) correct for speechreading alone to 96.6% (LP) and 96.2% (HP) correct when speechreading was supplemented by cues.

In a more recent study, Uchanski et al. (1994) tested four highly experienced young-adult users of MCS. The subjects were 18-27 years of age and had acquired hearing losses before the age of two years. Test materials were videotaped Harvard sentences (IEEE, 1969), which provide few contextual cues and consequently pose considerably greater difficulty than everyday conversational speech. Using speechreading alone, the subjects were able to recognize only 25% of the keywords in the sentences. With MCS, however, keyword scores increased to 84%.

Although MCS provides enormous benefit in the education of the deaf, the assistance it provides in day-to-day communication is limited to situations in which the talker, or a transliterator, produces the cues. To overcome this limitation, Cornett and his colleagues attempted to develop an "Autocuer," a system that would derive cues similar to those of MCS automatically, by electronic analysis of the acoustic speech signal, and display them visually to the cue receiver. The following sections provide a review of the systems that resulted from this development effort, as an introduction to the research reported in this paper.

Automatically Generated Cues

In the studies to be considered, cues are derived by automatic speech recognition (ASR) technology. Therefore, the usefulness of the cues, evaluated in terms of the improvement in speechreading scores, is strongly dependent on the performance of the ASR system. In some cases this performance is reported in terms of the percentage of phonetic units (phones) in the acoustic waveform that are correctly identified by the recognizer. In computing this measure, phones identified as other phones (substitutions) and phones whose occurrence is not detected (deletions) are counted as errors. However, the phonetic transcription produced by a recognizer may also include phones not present in the speech waveform (insertions). Unlike the "correct identification" score, which ignores insertions, the number of insertions is subtracted from the number of correct identifications when computing an "accuracy" score for the recognizer. Since this terminology has become standard only relatively recently, the interpretation of some older reports of recognizer performance is problematic.
Of course, strict comparisons of performance between different systems are also very difficult to make because of differences in the databases, training regimens, and speech materials employed.

**The Gallaudet-R.T.I. Autocuer**

Under Cornett's leadership, Gallaudet University and the Research Triangle Institute undertook the development of automatic cueing systems. The best described implementation of the Gallaudet-R.T.I. Autocuer (Cornett, Beadles, & Wilson, 1977) displayed the cues as virtual images of a pair of seven-segment LED elements projected in the viewing field of an eyeglass lens. By activating the segments selectively, nine distinct symbol shapes, cueing consonant distinctions, were created at each of four positions, cueing vowel distinctions. The cue groups selected differed from those used in MCS, a concession to the difficulties of speech recognition, but, as in MCS, sounds that are difficult to distinguish through speechreading occurred in different cue groups. A block diagram of the speech analysis subsystem proposed for the wearable version of the Autocuer suggests that speech sounds would be assigned to cue groups on the basis of estimates of the voice pitch as well as measurements of zero crossing rates and peak-to-peak amplitudes in low- and high-frequency bands of speech.

To distinguish cue display limitations from speech analyzer limitations, Cornett et al. (1977) conducted speech reception tests both with "ideal" cues derived manually from spectrograms, as well as with cues derived automatically by the analyzer. Subjects included normal hearing listeners and listeners with severe to profound hearing losses. Since none of these subjects had experience with Manual Cued Speech, they received roughly 40 hours of training with the system before participating in speech reception tests. On tests of the reception of common words spoken in isolation, scores improved from 63% for speechreading alone to 84% when spectrogram-derived cues were presented. By comparison, when analyzer-derived cues were presented, scores were essentially the same as in the unaided condition. Tests of the reception of common words in short phrases, however, indicated that neither spectrogram- nor analyzer-derived cues improved scores significantly. The authors interpreted the results with isolated words as highly encouraging. They attributed the poor performance seen for words in phrases to an inability of the subjects to integrate the cues with speechreading. The authors concluded that the subjects might require as much as eight months of exposure to the cues to decode cued sentences spoken at normal speaking rates, as had been suggested by the findings of Clarke and Ling (1976).

Cornett et al. (1977) regarded the results of their study as supporting the further development of a wearable prototype of the Autocuer. Assuming that the error rate of the speech analyzer component could be improved sufficiently, the availability of such a system would permit subjects to train on automatically produced cues. They speculated that until sufficient consistent exposure to such cues could be provided to cue receivers it would be impossible to evaluate the performance of the system realistically.

Although a wearable prototype of the Autocuer has been produced, it is not in use at the present time, and development of this system does not appear to have progressed as planned. Evaluations of the most recent implementation of this Autocuer (Beadles, 1989) indicate that the phoneme identification score is roughly 54%, with a 33% deletion rate and a 13% substitution rate. Simulation studies of the reception of common words spoken in isolation (Beadles, 1989) indicate that this level of recognition performance would lead to scores of roughly 67% correct, only eight percentage points higher than for speechreading alone on the task. Thus the Autocuer's low phoneme identification score, particularly its high deletion rate, probably would not improve communication greatly relative to speechreading alone. The Cornett et al. (1977) study is discussed further on page 26.

**Potential Recognition Systems**

Although there have been many recent attempts to develop auditory (e.g., Grant, Braida, & Renn, 1991), tactile (Reed & Durlach, 1995; Boothroyd, Kishon-Rabin, & Waldstein, 1995; Reed & Delhorne, 1995; Plant, Franklin, Franklin, & Steele, 1995; Weisenberger, 1995; Weisenberger, Broadstone, & Saunders, 1989), and visual (Ebrahimi & Kunov, 1991) speechreading supplements, active research on automatic cueing systems has largely been in abeyance since the Cornett et al. (1977) study.

In 1988 Vilaclara described a speech processing system designed to recognize phonemes and produce 11 cues, called "kinemes," complementary to the speechreading of French. This system assigns consonants to one of six kineme groups and vowels to one of five kineme groups appropriate for French. When evaluated on sentences consisting of 450 consonant segments and 603 vowel segments produced by a single speaker, this system achieved an identification score for phones of 74%, corresponding to an identification score for kinemes of 79%. The significance of these results is unclear because of the relatively small amount of speech processed (roughly 50 sentences), and the lack of data on the occurrence of deletions and insertions. No evaluation of this system as a speechreading supplement has been reported.
Research on speech recognizers usually does not focus on deriving cues for aiding speechreading. In order to estimate the performance that might be achieved with current and future ASR technology, Uchanski et al. (1994) measured how well cues could be derived by a contemporary phonetic speech recognition system (Zue, Glass, Phillips, & Seneff, 1989) operated as a phonetic recognizer. The system was trained using both a multi-speaker corpus derived from the TIMIT database (Zue & Seneff, 1988) and also a single-speaker corpus of Harvard sentences (IEEE, 1969). The test materials were segmented prior to recognition, making the recognizer's task solely one of classification, so that the only possible errors were substitutions of one phone for another. The results on the single speaker corpus indicated that the recognizer distinguished consonant segments from vowel segments 94% of the time. Consonants were assigned correctly to one of six cue groups 77% of the time. The eleven monophthongs were labelled correctly 65% of the time. Although no attempt was made to assign vowels to cue groups, it is clear that they would have been assigned correctly at least 65% of the time.

Although the results obtained by both Vilaclara (1988) and by Uchanski et al. (1994) seem superior to those reported for the Gallaudet-R.T.I. recognizer, it is not clear whether this difference is significant. Indeed, it remains an open question to determine how accurate an automatic speech recognition system must be to be of benefit to speechreaders. Not all the errors made by a phonetic speech recognizer would be problematic for the cue receiver; for example, misidentification of one phoneme as another phoneme in the same cue group would have no effect since it would produce the correct cue symbol. It may also be possible for cue receivers to discern systematic patterns of cue errors with practice, and learn to deal with them.

To deal with such uncertainties, Uchanski et al. (1994) used the Post-Labelling Model of audiovisual integration (Braida, 1991) to estimate the effectiveness of recognizer produced cues in speechreading. This model has been shown to predict audiovisual confusions of consonant and vowel segments from knowledge of the confusions made separately in audio-only and visual-only presentation conditions. In this application Uchanski et al. (1994) attempted to predict how well consonant and vowel segments would be identified when speechread segments were accompanied by ASR produced cues. Confusion patterns for consonants presented visually were derived from the data of Erber (1972); for vowels from the data of Hack and Erber (1982), Wozniak and Jackson (1979), and Montgomery and Jackson (1983). Confusion matrices characteristic of speech recognition performance were derived from the data of Vilaclara and from tests of the MIT recognition system (Zue et al., 1989) operating as a phonetic recognizer. Predicted consonant identification scores for speechreading supplemented by the output of the automatic recognizers generally exceeded 75% correct, roughly 25 percentage points higher than with speechreading alone. Similar trends were obtained for vowels, although overall scores were lower and the improvements associated with the use of cues were not as large.

In order to determine whether automatically generated cues can support useful levels of reception for sentences and continuous discourse, Uchanski et al. (1994) used a mathematical model developed by Boothroyd and Nittrouer (1988) to relate phoneme scores to scores for words and whole sentences. They estimated that a level of ASR accuracy in the range of 70-80% correct for phonemes would account for the performance of trained cue receivers on a variety of sentence tests. This range is close to that achieved by the better speech recognition systems currently in existence.

### Effects of Cue Imperfections

The analyses reported by Uchanski et al. (1994) are highly encouraging with respect to the potential use of current automatic speech recognition technology for producing Cued Speech. However, they are not adequate to guide the development of an automatic cueing system. For example, Uchanski et al. (1994) considered only the effects of recognition errors on performance. Moreover, they assumed that successive recognizer outputs are statistically independent events, which they need not be, particularly when recognition techniques make use of constraints imposed by phonetic context.

In MCS, cues are generally produced with extremely high accuracy. The effect of errors in cue production on the reception of continuous speech (as opposed to syllables or words) has not been studied. The importance of understanding the effect of errors on the reception of continuous speech is underscored by the findings of the Vidvox sensory aid project (Russell, 1986). In this aid continuous speech is represented by a stream of phonetic symbols without word boundary markers. Krasner et al. (1985) demonstrated that a Hidden Markov Model (HMM) recognizer considered for the Vidvox System could achieve a phonetic recognition accuracy of 65% (19% substitutions, 3% deletions, and 13% insertions) when operating in a speaker-dependent continuous-speech mode. Huggins et al. (1986) evaluated the effect of recognition errors on reception of phone representations of the Harvard (IEEE, 1969) sentences. One hearing-impaired subject was able to read perfectly transcribed speech materials at rates in excess of 75 wpm, and five normal-hearing subjects were able to read this material at 25-35 wpm. In the same study, recognition errors were simulated and presented to a normal-hearing listener at rates equal to 0, 1/3, 2/3 and 1 times a base rate (14% substitutions, 5% deletions and 6% insertions). The subject's reading rate declined from 69 wpm (0 errors), to 26 wpm (1/3 base error rate), 21 wpm (2/3) and 14 wpm (base error rate). Moreover, word reception errors, which were nearly absent in the error-free condition, increased to roughly 30% at the full error rate. These results suggest that even a relatively small number of errors can be extremely deleterious to continuous speech reception, particularly for users who are not extensively trained. For the Vidvox aid to be effective, the automatic
recognition system must produce far fewer than 5% substitutions, 2% deletions and 2% insertions. This would require major improvements in ASR performance compared with currently available technology.

Automatically produced cues differ from those of MCS in ways other than the incidence of errors. In MCS, cues are typically produced in synchrony with the facial actions that accompany speaking because the talker prepares the hand and mouth simultaneously. By contrast, the cues that would be produced by ASR systems typically lag behind facial actions because time is required to process the acoustic signal and to make recognition decisions. These times can vary from segment to segment. Moreover, ASR accuracy is expected to improve if greater time lags can be tolerated because more sophisticated recognition algorithms can be used, such as context-dependent algorithms which base the recognition of a given segment on tentative decisions about identity of surrounding segments.

Finally, whereas the MCS code is based on a phonemic representation of what the speaker intends to say, automatic recognizers typically produce a phonetic representation of what is actually said. The phones that speakers produce often correspond to the intended phonemes only imperfectly and/or inconsistently. In principle, automatic recognizers can overcome the effects of poor or incorrect articulation by making use of higher order linguistic constraints (on word formation, grammatical structure, etc.). But this strategy requires delaying recognition until subsequent words or phrases are processed, so that cues are not produced until well after the occurrence of facial actions.

Study of Simulated Cueing Systems

In addition to the limitations on the performance of the speech recognition system, the cues presented by automatic cueing systems are likely to be displayed differently from those of MCS, even when the same cue groups are used.

Methods

A set experiments was designed to obtain greater insight into the effects that limitations on ASR performance would impose on automatic cueing systems. We generated the cues of MCS by simulating several existing or hypothesized phonetic ASR systems, displayed the cues as handshapes superimposed on a video display of the talker's face, and compared word reception in speechreading alone, with MCS, and with the automatically generated cues.

To minimize the need for training, the cue display was designed to resemble the appearance of manual cueing, the cue code was the same as that used in MCS, and all subjects were highly experienced receivers of MCS.

Because the study was intended to be exploratory in nature, it was conducted in two phases. Although the phases differed with respect to experimental conditions and subjects, both phases included three baseline conditions: SA (speechreading alone), and MCS (Manual Cued Speech), the PSC (perfect synthetic cues) condition; and five (Phase I) or six (Phase II) experimental conditions that tested the effects of cue imperfections on sentence reception. The experimental conditions are described in the next section.

Table 1 Characteristics of the subjects tested.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Age (years)</th>
<th>Deafness</th>
<th>Etiology</th>
<th>Onset (months)</th>
<th>Past Use (years)</th>
<th>Current Use (hours/day)</th>
<th>Participation in Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>22</td>
<td>Unknown</td>
<td>3</td>
<td>12</td>
<td>1-2</td>
<td></td>
<td>I &amp; II</td>
</tr>
<tr>
<td>S2</td>
<td>21</td>
<td>Unknown</td>
<td>Birth</td>
<td>19</td>
<td>5-6</td>
<td></td>
<td>I &amp; II</td>
</tr>
<tr>
<td>S3</td>
<td>27</td>
<td>Rubella</td>
<td>Birth</td>
<td>23</td>
<td>1-2</td>
<td></td>
<td>I &amp; II</td>
</tr>
<tr>
<td>S4</td>
<td>24</td>
<td>Unknown</td>
<td>Birth</td>
<td>19</td>
<td>2-10</td>
<td></td>
<td>I</td>
</tr>
<tr>
<td>S5</td>
<td>23</td>
<td>Unknown</td>
<td>18</td>
<td>17</td>
<td>&lt;1</td>
<td></td>
<td>II</td>
</tr>
<tr>
<td>S6</td>
<td>19</td>
<td>Unknown</td>
<td>Birth</td>
<td>16</td>
<td>5-6</td>
<td></td>
<td>I</td>
</tr>
</tbody>
</table>

Subjects

Four subjects between the ages of 21-27 were employed in Phase I and five subjects between the ages of 19-27 in Phase II (see Table 1). All subjects were native English speakers who were profoundly, prelingually deafened. All were experienced cue receivers, having used MCS for 12-23 years. Their use of MCS at the time of the study ranged from 1-10
hours per day, usually with a parent or transliterator. Although some subjects used hearing aids, the testing used the video display only.

**Speech Materials**

Three sets of sentence materials were used in this research: CID (Davis & Silverman, 1970); Clarke (Magner, 1972); and Harvard (IEEE, 1969). The CID sentences, which are representative of "everyday" speech, and Clarke sentences, which are highly contextual, were used to evaluate speech reception.

These sentences contain five key words, four monosyllables and one disyllable. The Harvard sentences provide few contextual cues to word identity, e.g. "Glue the sheet to the dark blue background." so that it is difficult to identify a word correctly simply from knowing the other words in the sentence. Videotaped recordings of these sentences were made using a female speaker (teacher of the deaf) who was proficient in producing MCS. The sentences were recorded both uncued and with the production of Manual Cued Speech. Speaking rates were roughly 100 WPM for the cued sentences and 140 WPM for the uncued sentences. Four of the subjects tested (S1, S2, S5 and S6) had previous exposure to this speaker, either at school or in other tests of speech reception.

High quality audio and video recordings of the sentence materials had been prepared for earlier studies of speechreading and MCS. The sentences used for speechreading studies were recorded using a narrow angle camera lens so that the face filled roughly half the available viewing screen. The sentences used for MCS studies were recorded using a wide angle lens so that the face filled roughly one quarter of the video screen to accommodate the hand movements. Except for tests of the reception of Manual Cued Speech, which used the recordings of the cued sentences, the experiments used the recordings of the uncued sentences. High quality audio recordings of the acoustic waveform were made simultaneously with the video recordings. SMPTE (1995) time code was recorded on all materials for synchronization.

The audio recordings were segmented and labelled by phonetically trained listeners who had access to spectrographic and orthographic representations of the sentences. These listeners marked the beginnings and ends of phones in the acoustic waveform and assigned labels to the phones based on the audio signal rather than the orthographic representations. The Sensimetrics Speechstation (1994) used for segmentation typically provided 5 msec resolution. Labelling employed the code used to designate phones in the TIMIT database (Zue & Seneff, 1988). These sequences of phonetic labels were used, after possible modification to simulate phonetic speech recognizers, to determine the cues that would be produced by various automatic cueing systems.

**Synthetic Cues**

The structure of MCS assumes that consonants are typically followed by vowels. It represents each CV combination as a handshape at a specified position. Special rules are used for atypical combinations such as VC, C, CC and CVC. We developed a regular expression grammar corresponding to these rules and implemented this grammar in software as a finite state machine. Given a sequence of phonetic labels, this machine specified a sequence of handshape-position pairs according to the rules of MCS.

Images of the eight handshapes used in MCS were captured and edited from recordings of the speaker while producing MCS. These handshapes were then dubbed on a frame-by-frame basis onto the uncued narrow angle sentence recordings at appropriate positions around the face of the talker. Given the relatively large size of the face in these recordings, it was necessary to reduce the size of the synthetic handshapes so that all positions would fit on the screen. There was no fluid articulation or movement of the hand as it changed shape or position; only eight discrete shapes and five discrete positions were used.

Appearance of the cue typically began at the frame corresponding to the start of the consonant segment in the acoustic waveform and was maintained until the end of the following vowel segment (for CV combinations) or of the consonant segment (for consonants occurring alone). CV combinations that included a diphthong were displayed as a pair of cues, with the duration divided equally between the two vowels of the diphthong.

**Recognizers**

Two phonetic automatic speech recognizers were used to derive the cues used in the speech reception tests. Both recognizers were variants of Hidden Markov Model (HMM) recognizers widely under study today. The recognizers differed with respect to the types of models used to represent the phonetic units being recognized. The Context Independent Recognizer used one model for each phone, so that, for example, the same model would have been used for the vowel sounds in "bib" and "kick". The Context Dependent Recognizer used different models for the same phone, depending on the identity of the subsequent phone. In general, use of the Context Dependent Recognizer produces higher recognition
accuracy (59% correct identification of phones) than the Context Independent Recognizer (79% correct), at the cost of greater computation load. However, the lower accuracy of the Context Independent Recognizer was not important for this study, since only the pattern of recognizer errors, rather than the frequency of errors, was used in simulations of potential recognizers. By contrast, the Context Dependent Recognizer was used as an example of the performance of current state of the art recognizers. Detailed descriptions of the recognizers, together with a description of how the recognizer output was used to produce cues, is provided in the Appendix.

**Training and Testing**

During both training and testing, sentences were presented on a video monitor at a distance of roughly one meter from the subject. No audio signal was presented. Responses were recorded in writing on prepared forms.

The CID and Clarke sentences were processed and used to familiarize the subjects with each presentation condition. During training, subjects were shown each sentence orthographically after their response was recorded. The same training sentences were presented to each subject.

For testing, the Harvard sentences were used exclusively and no feedback was provided. These sentences are scored by keywords: five per sentence. Although the Harvard sentence lists have been equated for difficulty under acoustic presentation conditions, they are not necessarily of equal difficulty when presented with MCS. To equalize the average difficulty of the sentences used in the different test conditions, we used different sentence lists to represent a given test condition for different subjects. The large size of the Harvard corpus (720 sentences) generally allowed us to do this. However, subjects S2 and S6 were tested on the same lists in Phase II. Also, all subjects were tested on the same sentences in both the Speechreading Alone and Manual Cued Speech conditions.

A minimum of 40 Harvard sentences (containing a total of 200 keywords) were presented to each subject in all conditions. Some baseline conditions in Phase I utilized 60 sentences (300 key words). We estimated that scores computed from 40 sentences would reduce the standard deviation of the test score to 3.5 percentage points or less, so that differences between scores for a given subject of ten percentage points or more would generally be statistically significant.

**Conditions and Results**

The SA (speechreading alone) and MCS (Manual Cued Speech) conditions were intended to provide reference scores for all other conditions tested. Comparisons with the SA condition estimate the benefit provided by each type of cue presentation. Comparisons with the MCS condition estimate the effects of the artificial display and the degradations introduced in the cues. Comparisons with the PSC (perfect synthetic cues) condition estimate the effects of the degradations introduced in the experimental conditions.

The SA, MCS, and PSC conditions were tested at the beginning and end of each two-day session. The experimental conditions were tested in two blocks, each of which contained all of the experimental conditions. To control for the effects of learning, the order of testing the experimental conditions was randomized, with different orders used in each block. The order of testing conditions was the same for all subjects.

**Test Procedures**

In each phase, training was provided prior to testing on all conditions except for the second set of tests of the SA, MCS, and PSC conditions. In the first set of baseline tests, 10 training sentences preceded 20 testing sentences for the SA and MCS conditions. For the less familiar PSC condition, 30 training sentences were presented prior to the 20 test sentences. For the experimental conditions 10 training and 20 testing sentences were used in each block. No sentence was presented to any subject more than once. Except for S4, who was tested at home, subjects were tested individually in a sound-treated booth at M.I.T.

**Scoring and Analysis of Results**

Reception accuracy for the five keywords in each Harvard sentence was evaluated using strict scoring rules (plural nouns were not accepted for singular, past tense was not accepted for present, etc.) However, homophone substitutions were not counted as errors.

An analysis of variance (ANOVA, e.g., Winer, 1971) was applied to the arc-sine transformed scores to identify the factors that played a statistically significant role in the observed pattern of scores. Results from each phase were analyzed separately. The two factors analyzed, Conditions and Subjects, were treated as fixed effects, although conclusions
would be essentially unchanged if the Subjects factor were treated as a random effect. It was assumed that there were no differences among sentence lists.

An F test with a significance level of 0.01 indicated that there were significant differences in scores between Conditions and between Subjects, but there was no significant interaction of Conditions X Subjects. Thus the observed differences between scores in the different conditions were, on average, independent of the subject tested. These results held true in both phases.

For both phases, paired comparisons using t-tests at the two-tail 0.01 significance level were performed between scores in the PSC condition and in the experimental conditions, and between the MCS and PSC conditions, using Fisher's method. In Phase II, additional t-tests were used to compare scores in the REAL condition with those in other conditions. The results of these comparisons are discussed below.

**Phase I**

The experimental conditions in Phase I were concerned with the effects of recognizer errors and delays in the presentation of the cues. The conditions tested and average scores obtained are summarized in Table 2.

### Table 2 Conditions and keyword scores for Phase I

<table>
<thead>
<tr>
<th>Condition</th>
<th>Error Rate(%)</th>
<th>Delay ms</th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Subject 3</th>
<th>Subject 4</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>31</td>
<td>33</td>
<td>20</td>
<td>15</td>
<td>24.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCS</td>
<td>87</td>
<td>91</td>
<td>82</td>
<td>54</td>
<td>78.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSC</td>
<td>79</td>
<td>75</td>
<td>74</td>
<td>63</td>
<td>72.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E10</td>
<td>10</td>
<td>0</td>
<td>70</td>
<td>72</td>
<td>55</td>
<td>38</td>
<td>58.8</td>
</tr>
<tr>
<td>E20</td>
<td>20</td>
<td>0</td>
<td>63</td>
<td>57</td>
<td>43</td>
<td>32</td>
<td>48.8</td>
</tr>
<tr>
<td>E20D1</td>
<td>20</td>
<td>33</td>
<td>54</td>
<td>55</td>
<td>47</td>
<td>40</td>
<td>49.0</td>
</tr>
<tr>
<td>E20D3</td>
<td>20</td>
<td>100</td>
<td>46</td>
<td>39</td>
<td>36</td>
<td>28</td>
<td>37.2</td>
</tr>
<tr>
<td>E10D5</td>
<td>10</td>
<td>165</td>
<td>40</td>
<td>44</td>
<td>31</td>
<td>24</td>
<td>34.8</td>
</tr>
</tbody>
</table>

**Baseline Conditions.** The baseline conditions used unaltered recordings of the speaker. In Phase I, six lists, containing a total of 300 keywords, were used in the SA and MCS conditions, with the same lists presented to all subjects. Scores for SA (speechreading alone) fell well below 40% for all subjects and were lower than the scores obtained by each subject for all other test conditions. The range of scores is consistent with the range (16-34%) reported by Uchanski et al. (1994) for similar materials. Scores for MCS (Manual Cued Speech) were, except for S4, higher than the scores obtained in all other conditions, and were similar to those reported by Uchanski et al. for similar materials (68-93%).

The PSC (Perfect Synthetic Cues) condition was intended to evaluate a cueing system that employs an ideal speech recognizer operating on the acoustic speech waveform. For each sentence, the cues were derived from the unaltered phone sequence marked by the trained human labelers, as described on page 10. Although one might expect this sequence to be phonetically perfect, the speaker did not produce all phones exactly as specified by the sentence orthography. Some speech sounds were omitted and others were slurred. Since the transcriptions were derived from acoustic waveforms, rather than the sentence orthography, the cue sequence contained a small number of omissions and some errors unavoidably introduced by the human labelers.

Each subject was tested on four lists of sentences in the PSC condition, with different lists presented to each subject. Scores in the PSC condition were fairly high, typically only slightly lower those for MCS, except for S4, who performed better in the PSC than in the MCS condition. The difference between MCS and PSC scores was not statistically significant. If the unusual results obtained by S4 are omitted, scores in the PSC condition were slightly inferior to those in the MCS condition. In the remainder of this report, scores in the PSC condition are used as a reference to evaluate the effects of cue imperfections in other conditions.

**Errors.** The E20 and E10 conditions tested reception when random errors were introduced into 20% and 10% of the phones, as described on page 31. Scores generally decreased as the proportion of errors increased, but the difference between scores in the PSC and E10 conditions was not statistically significant. This result is encouraging because it suggests that an automatic cueing system with 10% errors could provide benefits to cue receivers almost as well as a perfect
system. However, increasing the errors by 10 percentage points (SDI, 20% errors) caused a significant decrease in keyword reception compared with PSC and E10.

Errors and Fixed Delays. Three conditions examined the effect of combining recognizer errors and fixed presentation delays. In conditions E20D1, E20D3, and E10D5, the phone sequence contained 20% errors and a 33 ms delay, 20% errors and 100 ms delay, and 10% errors and a 165 ms delay, respectively. These delays range up to roughly half the average cue duration, since there were roughly 3.5 cues per second in these materials. Errors were introduced, and delays are specified relative to the beginning of the marked acoustic segment.

Generally, as presentation delay increased, test scores decreased. However, the difference in scores between E20D1 and E20, both containing 20% errors, is small and not statistically significant, suggesting that small (33 msec) fixed delays have only minor effects. Some subjects found the 100 msec delay noticeable and bothersome and all found the 165 msec delay to be so. The decrease in scores between conditions E20D1 and E20D3 was statistically significant, but the difference in scores between E20D3 and E10D5 was not.

Phase II

Phase II testing was conducted roughly two months after Phase I and included six new experimental conditions, as well as the three baseline conditions (see Table 3).

Table 3 Conditions and scores Phase II

<table>
<thead>
<tr>
<th>Condition</th>
<th>Error</th>
<th>Delay</th>
<th>Subject</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
</tr>
<tr>
<td>SA</td>
<td>43</td>
<td>40</td>
<td>21</td>
<td>37</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>MCS</td>
<td>92</td>
<td>93</td>
<td>86</td>
<td>81</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>PSC</td>
<td>85</td>
<td>85</td>
<td>83</td>
<td>76</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>EODR</td>
<td>74</td>
<td>87</td>
<td>65</td>
<td>67</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>E10DR</td>
<td>71</td>
<td>81</td>
<td>67</td>
<td>60</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>E20DR</td>
<td>65</td>
<td>71</td>
<td>53</td>
<td>44</td>
<td>69</td>
<td></td>
</tr>
<tr>
<td>E20MK</td>
<td>61</td>
<td>61</td>
<td>48</td>
<td>46</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>E20IN</td>
<td>60</td>
<td>61</td>
<td>60</td>
<td>49</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>REAL</td>
<td>79</td>
<td>73</td>
<td>70</td>
<td>65</td>
<td>64</td>
<td></td>
</tr>
</tbody>
</table>

Baseline Conditions. In Phase II, four lists, containing 400 keywords, were used to test the SA, MCS and PSC conditions. A total of 13 lists were used to test the SA and MCS conditions, with each list presented to roughly half of the subjects in each condition.

As in Phase I, different lists were presented to each subject in the PSC condition. Scores for these three conditions averaged roughly 10 points higher than in Phase I, and were somewhat higher than those reported by Uchanski et al. (1994). The three subjects who participated in both phases all obtained higher scores on these conditions than when tested on comparable conditions in Phase I.

Averaged across subjects, scores in the PSC condition were the second highest of all conditions, and these scores were not statistically different from those of the MCS condition. The three subjects who participated in both Phases I and II improved more on the PSC condition than on the MCS or SA conditions, suggesting that they were not fully practiced with this novel display in the Phase I tests.

Errors and Random Delays. These conditions evaluated the effects of combining errors (introduced as described on page 31) with the random presentation delays that are likely to occur in a real-time recognition system. Conditions EODR, E10DR, and E20DR, contained 0%, 10% and 20% errors respectively, and cues were presented with delays of 0, +1 and -1 frame (with equal probability) relative to the frame corresponding to the start of the acoustically labelled phone.

The difference between scores obtained in the MCS and EODR conditions was not significant, nor was the difference between PSC and EODR. This result is consistent with that for the E20D1 condition in Phase I, and suggests that one frame (33 ms) of fixed or random delay has little effect. The differences between PSC and the other random delay conditions (E20DR and E10DR) were significant, as was that between EODR and E20DR, but the difference between scores in the E20DR and E10DR conditions was not. For the most part, as with the fixed delay conditions, speech reception scores decreased as the error rate increased.
Marked Errors. Some speech recognizers are capable of both identifying the spoken phone and estimating whether the recognition decision is likely to be correct. The E20MK condition explored whether such indications could assist the cue receiver to deal with recognition errors. Phonetic transcripts were prepared in which 20% of the phones were substituted or deleted. On half of the substitution errors, the associated cue handshape was outlined with a red square to indicate to the subject that it might be wrong. Since the recognizer would be expected to report erroneously that some correctly identified phones were incorrect, the handshapes corresponding to an equal number of correctly recognized phones were also outlined.

Most of the subjects obtained scores in the E20MK conditions that were somewhat lower than those in the E20DR condition, which had the same error rate, but also included jitter in time of presentation. This suggests that the subjects may have treated all marked cues as erroneous. It seems likely that subjects did not receive adequate training to develop successful strategies for making optimum use of the error indications.

Insertions. The error conditions described thus far simulated the tendency of automatic recognizers to fail to detect the presence of speech elements (deletion errors) and to confuse speech elements for one another (substitution errors), but not the tendency to insert elements where they do not occur in the phonetic sequence (insertion errors). The sum of deletion and insertion rates is often roughly constant for a particular type of recognizer (Schwartz, Chow, Kimball, Roucos, Krasner, & Makhoul, 1985). To examine the effect of such erroneous insertions, in condition E20IN the deletion rate for each input phone was halved and insertions were introduced at the same rate as deletions. The identity of the inserted phone was selected among the remaining phones with equal probability of occurrence. The duration of the inserted phone was arbitrarily set equal to 70% of the average duration of that phone in the corpus. To accommodate the insertion, the preceding and following phones were shortened. Additional details are provided in Bratakos (1995).

Scores in the E20IN conditions were roughly equal to those in the E20MK condition, i.e. somewhat lower than those in the E20DR condition. Two factors seem likely to account for the potency of insertion errors. First, the erroneous insertion of a phone altered the duration and time of occurrence of adjacent phones, and delayed the presentation of the following phone. Second, whereas substitutions and deletions are likely occur naturally in the presentation and reception of MCS, insertions represent a new phenomena for cue receivers. Also, as with the E20MK condition, subjects may not have had adequate training to develop strategies to deal with the insertion of erroneous cues.

Context-Dependent Recognizer. In this condition (REAL) cues were produced by a state-of-the-art phonetic recognizer trained on the talker who produced the materials used in the speech reception tests. This recognizer (described in the Appendix) operated off-line rather than in real-time, and produced a phone sequence (including time markings) that included substitution, deletion, and insertion errors. This phone sequence was used to produce a cue sequence according to the rules described on page 11, using the times of occurrence of phones estimated by the recognizer rather than the times included in the phonetic transcriptions of the acoustic waveforms. No attempt was made to delay the presentation of the cues in accordance with the time required by the recognition process. Scores for the REAL condition were fairly similar to those observed in the E10DR condition which simulated a 10% error rate and included 33 ms of cue presentation jitter.

Discussion

Comparison of Phases

Three subjects (S1, S2, and S3) participated in both Phases I and II, and three conditions (SA, MCS and PSC) were tested in both of these phases. Scores in all three conditions were higher in Phase II than in Phase I for all three subjects. Since each subject received the same sentences in the SA and MCS conditions as the other subjects, it is not possible to separate out the effects of training from variations in sentence difficulty. In the PSC condition, on the other hand, subjects received different sentences from one another, so the improvements in the scores in the PSC condition can be attributed to the effects of training, specifically to increased familiarity with the cue display. Indeed, these improvements were substantially larger than those observed in the MCS condition.

Effectiveness of the Synthetic Cues

The synthetic cues in this study were intended to resemble the cues of Manual Cued Speech so that deaf subjects experienced in MCS could perform the task without extensive training. Although this display was a simplified version of the MCS display, subjects appeared to have little difficulty adapting to the synthetic cues.

It should be noted that the synthetic cues were different from the Manual Cues in four ways. First, the synthetic cues were presented at 140 wpm, a rate similar to that a real-time system will encounter, rather than the 100 wpm rate for MCS. Subjects commented on the increased speed when seeing the synthetic cues for the first time. Most human cue
producers reduce their speaking rates when cueing, probably because of the mental and physical demands of cueing. Although the effect of presentation rate on the reception of the Harvard sentences via Cued Speech has not been studied, it has been studied for the reception of the CID sentences via American Sign Language (Fischer, Delhorne, & C.M. Reed, 1997). At the normal presentation rate of 2.2 signs per second, scores averaged 88% correct. Although large reductions in scores were seen for presentation rates in excess of 4.4 signs per second, there were only small reductions (to 86%) in scores when the rate was 1.4 times normal. This suggests that the difference in speaking rates may not be a major contributor to the small difference between the scores in MCS and PSC conditions.

The second difference was that the synthetic cues were discrete in both shape and position and lacked fluid transitions between cues. Additional research is needed to determine whether this difference is an important one. At the cost of a more complex display, synthetic cues could be made to appear more fluid by interpolating handshapes at intermediate positions. On the other hand, the discreetness of cue position and shape may improve cue reception by compensating for the increased blurring expected at normal speaking rates. With discrete cues, the handshapes remain at the vowel positions longer, so that the cue receiver has steady views of the handshape and vowel position. As a result, it is perhaps not surprising that only one subject (S6) complained about the lack of transitions. Further study is required to determine whether the use of more fluid handshapes would improve reception sufficiently to justify the increased cost of a more complex display system.

Third, as discussed on page 10, the synthetic handshapes were somewhat smaller, in relation to the talkers face, than the natural handshapes. Many subjects found this bothersome, and expressed difficulty distinguishing between cues at the throat position, as well as distinguishing between the throat and chin positions. This difficulty may account for some of the difference in scores between PSC and MCS.

The final difference, discussed in Section 2.4, was that the phonetic transcriptions of the PSC condition contained a small number of missing and incorrect phonetic events, while the MCS condition had been carefully recorded in order to be free of such imperfections. It seems unlikely that much of the difference between MCS and PSC scores can be attributed to these errors.

Scores in the PSC condition were, on average, roughly 8 points lower than in the MCS condition. This difference in scores is likely due to the combined effects of the differences in cues described above and the unfamiliarity with the cue display used in PSC. The scores of the three subjects who were tested in both Phases of the study improved by 8.3 points in the PSC condition in Phase II, but only 3.7 points in the MCS condition, suggesting that additional practice with the synthetic cues might further reduce this difference.

**Fine-Grained Analysis**

Ten conditions explored the effect of errors on the reception of Cued Speech. Three conditions had nominal phone error rates of 10%, and six conditions had nominal phone error rates of 20%. Measured phone error rates in keywords differed slightly from nominal values because the errors occurred randomly in the phone stream. For most conditions, phone error rates were close to, but slightly higher than, nominal rates (see Table 4).

<table>
<thead>
<tr>
<th>Condition</th>
<th>E_p</th>
<th>E_w</th>
<th>Condition</th>
<th>E_p</th>
<th>E_w</th>
</tr>
</thead>
<tbody>
<tr>
<td>E10</td>
<td>10</td>
<td>11.5</td>
<td>E10DR</td>
<td>10</td>
<td>12.0</td>
</tr>
<tr>
<td>E20</td>
<td>20</td>
<td>22.8</td>
<td>E20DR</td>
<td>20</td>
<td>23.2</td>
</tr>
<tr>
<td>E20D1</td>
<td>20</td>
<td>23.5</td>
<td>E20IN</td>
<td>20</td>
<td>27.0</td>
</tr>
<tr>
<td>E20D3</td>
<td>20</td>
<td>23.5</td>
<td>E20IN</td>
<td>20</td>
<td>30.0</td>
</tr>
<tr>
<td>E10D5</td>
<td>10</td>
<td>11.5</td>
<td>REAL</td>
<td>20.8</td>
<td>36.8</td>
</tr>
</tbody>
</table>

**Table 4 Measured Error Rates.** The percentage of all phones with errors $E_p$. The percentage of keywords with at least one phone in error is $E_w$.
However, for the E20MK and E20IN conditions, the measured error rates for phones in keywords were 27% and 30%, nearly half again as large as the nominal rates. The cause of these unexpectedly high rates is unknown. The measured error rate for the REAL condition was 21%.

In conditions that introduced phone errors, each phone was treated as an isolated entity. Errors occurred equally often in the phones in all words in the sentence. Furthermore, the occurrence of multiple phone errors in words was governed by the principle of statistical independence. For example, when the nominal error rate is 10%, a word such as "canoe", which has four phones, will have no phone errors with probability 0.65, one error with probability 0.29, and two or more errors with probability 0.06. When the nominal error rate is 20%, these probabilities are changed to 0.40, 0.40, and 0.20, respectively, and when the nominal rate is 30%, they are 0.24, 0.41, and 0.35. Generally, the observed distribution of word errors is consistent with this pattern, although there appear to be fewer than expected word errors in the E20MK condition.

It is important to note that the occurrence of a phone error within a keyword does not imply that the word will be incorrectly identified. The error may be a substitution within the same cue group, the associated cue may be misperceived by the cue receiver, or word formation constraints may allow the cue receiver to correct the error. Except for conditions in which cues were substantially delayed relative to the phone boundaries marked in the acoustic waveform, more than 40% of words that contained phone errors were identified correctly (Table 5).

Table 5 Measured Word Identification Rates. C1 is the percentage of keywords with phone errors that were identified correctly; Cc is the percentage of keywords without phone errors that were identified correctly.

<table>
<thead>
<tr>
<th>Condition</th>
<th>C1</th>
<th>Cc</th>
<th>Condition</th>
<th>C1</th>
<th>Cc</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSC</td>
<td>73.2</td>
<td>58.8</td>
<td>PSC</td>
<td>81.8</td>
<td></td>
</tr>
<tr>
<td>E20</td>
<td>43.8</td>
<td>64.2</td>
<td>EODR</td>
<td>53.8</td>
<td>72.6</td>
</tr>
<tr>
<td>E10</td>
<td>51.2</td>
<td>55.8</td>
<td>E10DR</td>
<td>62.8</td>
<td>73.4</td>
</tr>
<tr>
<td>E20131</td>
<td>45.2</td>
<td>49.8</td>
<td>E20MK</td>
<td>48.2</td>
<td>68.4</td>
</tr>
<tr>
<td>E20D3</td>
<td>29.5</td>
<td>36.0</td>
<td>E20IN</td>
<td>53.2</td>
<td>69.6</td>
</tr>
<tr>
<td>E10D5</td>
<td>32.5</td>
<td>36.0</td>
<td>REAL</td>
<td>58.2</td>
<td>77.2</td>
</tr>
</tbody>
</table>

On the other hand, keywords that were free of phone errors were sometimes identified incorrectly, as in the MCS and PSC conditions. Generally, the misidentification rate for such words increased as the overall phone error rate increased, as in the PSC, E10, and E20 conditions (Table 5). This implies that, even for materials such as the Harvard sentences that provide few contextual cues, reception of a given word can be affected by the reception of other words in the sentence.

Two conditions had phone error rates higher than the nominal 20%. The E20MK condition had an average phone error rate of 27%. Although this rate was slightly higher than those in the E20, E20DR, E20D1, and E20D3 conditions, the 64% keyword error rate for the E20MK condition was consistent with keyword error rates on the other conditions with the same nominal phone error rate. Since the increase in phone error rate did not change the keyword error rate, subjects' scores were probably not affected by this increase. The E20IN condition, on the other hand, had a phone error rate of 30% and a keyword error rate of 74%, roughly 10 percentage points above that of the other conditions with a nominal phone error rate of 20%. Surprisingly, the higher than expected phone error rate did not reduce scores in the E20IN condition a great deal.

**Effect of Delays**

Delays between the occurrence of a speech sound and the display of the associated cue were introduced in six conditions. Table 5 shows that when words were free of phone errors receivers are not affected by random delays that are short (33 msec) compared to the average duration of a cue (285 ms). Phase I tests revealed that a systematic delay of 33 msec had little effect on the ability of cue receivers to identify words when the nominal phone error rate was 20%. Larger systematic delays (165 msec) had dramatic effects, even at modest (10%) phone error rates. Although we did not include conditions that tested reception with error-free cues and long delays, it should be noted that such delays had effects on the
identification of words in which no phonetic errors had been introduced as well as words that contained such errors. In the 10% phone-error condition such delays reduced the reception of error-free words by 28 percentage points.

**Performance of the Real Recognizer**

The E20 and E20DR conditions, which simulated a 20% phone error rate without and with a 33 ms delay were nominally similar to the REAL condition. Although the average phone error rate in the simulated conditions (23%) was close to that in the REAL condition (21 %), identification scores in the E20 (49%) and E20DR (60%) conditions were substantially inferior to those in the REAL condition (70%). This discrepancy was unexpected and appears to reflect differences in the patterns of errors between the simulated error conditions and REAL. Several differences are noteworthy.

Although the phone error rate in the REAL condition is roughly 20% (Table 4), the rate of occurrence of errors in keywords is more similar to that seen in conditions with a nominal phone error rate of 10% (e.g., E10 and E10DR) than to conditions with a phone error rate of 20% (e.g., E20 and E20DR). Phone errors were introduced in the simulated conditions according to the principle of statistical independence, and the observed relationship between phone and word error rates is consistent with this assumption. The relation between phone and word errors in the REAL condition is that imposed by the HMM-based phonetic recognizer, rather than by the principle of statistical independence. Apparently this recognizer was less likely to introduce errors in keywords (generally content words) than in the other words in the sentence (19.7% vs 23.7%/a). Also, phone error rates were found to be higher at the beginnings of sentences, but keywords rarely occur at the beginnings of sentences.

More importantly, there was a tendency in the REAL condition for multiple errors to cluster within keywords at a rate higher than expected on the basis of the principle of statistical independence. For example, when the recognizer processed keywords consisting of four phones, the phone error rate was found to be 16.8%. If the principle of statistical independence had applied, 47% of these words would have been free of errors, whereas 57% were found to be error-free. Similarly, while 14% of four-phone keywords were predicted to contain two or more phone errors, 19% were found to have pairs of errors. As a result of this tendency, a larger proportion of error-free keywords contained no errors in the REAL condition than in simulated conditions, such as E10 and E10DR with the same overall phone errors rate. Since words that were free of phone errors were identified more accurately than words that contained phone errors, better performance in the REAL condition is to be expected.

It should also be noted that conditions with the same nominal phone error rate as the REAL conditions had different distributions of error types. In particular, the phone errors in the E20, E10, E20DR, and E10DR conditions contained roughly equal numbers of substitutions and deletions. The REAL condition, by contrast, contained 11 % substitutions, 4% deletions, and 6% insertions.

**Other Conditions**

It is difficult to make definitive statements about the effect of marking potential errors (E20MK) or of introducing insertion errors (E20IN) because the measured error rates for these conditions were unexpectedly high. If one extrapolates from the E20DR condition, which has nearly the same word error rate as the E20MK condition, no net improvement in scores was achieved by marking potential substitution errors. Rather there were small reductions both for words without errors (some of which would be marked as erroneous) and for words with errors. The failure for the marks to improve performance may also be due to the distracting nature of the marks themselves (red boxes surrounding handshapes). Rather than integrating the marks into their decisions, some receivers reported ignoring marked cues and focusing on the lips of the speaker. The benefits of providing feedback to cue receivers on potential errors should be reexamined in future research. Less distracting marks, e.g. a change in screen background, could be considered. Moreover the potential benefit provided by marks when recognition errors cluster within words deserves attention.

The E20IN condition had the highest measured phone and keyword error rates of any condition tested, in particular higher than the E20DR condition to which it was similar. However, it is noteworthy that the word identification rates for the E20IN and 20DR conditions were roughly equal. Evidently some of the insertion errors had little effect on word identification. One would expect this to be true for closures, for example, since not all consonants can follow a closure. However, all consonant phones were treated as equally likely insertions. Since most closures are easily detected visually, many closure related insertions may have been largely ignored by the receivers.
Choice of Cue Groups

Each Manual Cued Speech cue group contains several phonemes, with groups chosen to minimize speechreading confusions. The cue groups of MCS were chosen arbitrarily; other equally effective cue groups may be better suited to the performance of real ASR systems (Cornett et al., 1977; Uchanski et al., 1994).

In this study, the occurrence of a phone error resulted in the production of an incorrect cue 95% of the time. The only phone substitutions that did not cause cue changes in the 160 test sentences of the REAL condition were the phone pairs /I/-/w/, which are both in the I-handshape group, and the pairs /N/-/a/, which are both in the side-front vowel position group. Based on these results, it would seem that different cue groups might be better matched to the confusions made by the recognizer used, so that frequently confused phones would fall into the same cue group. Of course, cue receivers would have to be trained on the new groups.

Comparison to Other Studies

Vidvox

Certain aspects of the Vidvox Study (Russell, 1986) are similar to those of the current study. Both tested the reception of sentences based on the presentation of discrete symbols produced by a simulated speaker-dependent HMM speech recognizer. Fairly similar techniques were used to introduce simulated errors into the symbol stream. However, whereas in the Vidvox Study subjects read these symbols as text, in the current study the symbols supplemented speechreading. Both studies tested subjects who were highly competent in English and neither provided much training on degraded reception conditions. In contrast to the current study, however, subjects in the Vidvox Study were provided 4-9 hours of training on the task under error-free conditions, incomparable to the years of training of the receivers of MCS.

Despite the similarities, the results of the two studies are dramatically different. In terms of reading rates for error-free material, most of the Vidvox subjects achieved rates of 30 wpm (one exceeded 70 wpm) with negligible reception errors, while the MCS receivers tested in the current study received speech at 100 wpm (83% accuracy) and 140 wpm (77% accuracy). When tested on a simulated recognizes with 75% phone accuracy, reception of the Vidvox code proved nearly impossible for most subjects. Similar tests in the current study reduced word reception, but only to roughly 60% correct.

The Vidvox study evaluated the effect of marking insertion and deletion errors by presenting a non-phonetic symbol instead of an erroneous phonetic symbol. When the overall error rate was roughly 8% and errors were identified perfectly, only two of seven subjects were able to read at rates in excess of 30 wpm, the estimated lower limit of useful rates. Simulations of a 75% success rate in identifying phone errors, a rate believed to be "rather ambitious," resulted in dramatic loss in readability. In the current study, marking errors that were identified with a 50% success rate resulted in a modest (5 percentage point) reduction in reception when the phone error rate was 20%.

Although the ability of subjects to read the Vidvox code with relatively little training were regarded as encouraging by the authors of the Vidvox study, the effects of even a small number of recognizes errors on this ability were causes for great pessimism. Not only did the presence of a small number of errors appear to change a reading task to one of puzzle solving, but also the subjects found the task aversive and were unwilling to attempt to train on the reception of material containing errors. By contrast, highly trained receivers of MCS were only slightly affected by even greater phone error rates. This probably reflects the fact that the cues derived by the recognizes constitute only one component of the speech display: The speechread component is unaffected by these errors. Moreover, the vast amounts of experience available to subjects tested in the present experiment may have developed a resistance to the effects of imperfect cues. It is possible that the conclusions drawn by the Vidvox study would have been markedly different if their subjects had been provided with comparable amounts of experience.

Gallaudet-R.T.I. Autocuer

It is difficult to compare the results of the current study directly to studies conducted in conjunction with the development of the Autocuer (Cornett et al., 1977; Beadles, 1989) because of differences in subjects and test materials.

The cue receivers tested in the current study typically had many years experience using English for communication via Manual Cued Speech. Their high scores in the MCS and PSC conditions demonstrate their great competence in using this speechreading supplement system. By contrast the subjects tested in the Autocuer studies were undergraduate students at Gallaudet University who had little experience either with spoken English or with Manual Cued Speech. It was necessary to provide these subjects with extensive training before they could make use of error-free cues to identify consonants (20 hours) or isolated words (40 hours). This training was inadequate for them to obtain significant benefit from the cues when the same words were embedded in short phrases or sentences.
The test materials used in the present study were monoand disyllabic words in low-context sentences, with an essentially open vocabulary, spoken by a single speaker. Although some of the subjects were familiar with this speaker, they were unfamiliar with the test materials: The limited number of items used for training were never used for testing. By contrast, the 1977 Autocuer studies used isolated common words drawn from a small (500 word) fixed vocabulary. In that study, subjects were presented each possible recorded stimulus item three times during training before testing. Although these words were produced by four different speakers, the resulting tests were far less challenging than those used in the current study; in the speechreading alone condition scores were roughly 60% in the Autocuer studies, but only 25-36% in the current study.

In the Autocuer studies, cues were presented with a fixed delay of 150 ms relative to the syllable onset to simulate the processing time required for recognition. As a result, one condition in the Autocuer study is fairly similar to the E20DR condition in the study. Both had roughly equal substitution and deletion error rates, that combined to roughly 20%, and both presented cues with comparable delays (150 and 100 ms). However, word scores in the Autocuer study (77%) were substantially higher than in the E20D3 condition of the current study (37%). Since scores for error-free cues were fairly similar (86% for the simulated Autocuer and 78-87% in the PSC condition of the current study), a 20% error rate produced only a modest decline in scores in the Autocuer study, but produced dramatically lower scores in the current study.

Such differences in results between the Autocuer and current studies might have arisen even if the subjects had similar skills in lipreading and cue receiving; we believe the differences are due to the easier task of speechreading isolated words compared to speechreading words in sentences having little semantic context. The conditions tested in the studies of the Autocuer appear to be more appropriate for evaluating systems used in vocabulary building exercises than in everyday communication. Based on the amount of training provided to the cue receivers tested in the Autocuer study, we believe their ability to understand words in low-context sentences spoken at normal rates would have been extremely limited, even with cues.

Conclusions

Based on the results of our experiments it seems clear that current ASR technology can derive cues from the acoustic signal that would greatly improve communication by speechreaders. Experienced cue receivers were able to deal with the effects of ASR errors under conditions that were extremely difficult for the subjects who attempted to use the Vidvox display and that led to no improvement over speechreading for the subjects who were tested with the simulated Autocuer. These findings support the theoretical analysis of Uchanski et al. (1994), but also reveal important factors not considered in that study.

Our optimism about the potential of ASR-derived cues to aid speechreaders is based on the performance of commercially available software for a speaker-dependent speech recognition system. Although we did not implement a realtime recognition system having a comparable level of performance for this study, it seems clear that such a system could be realized using present-day technologies. Currently, sufficient accuracy can be attained in acoustically benign environments, provided that the problem is restricted to pre-specified talkers.

On the other hand, our results also suggest that current multi-speaker recognition technology could not produce cues accurate enough to provide substantial benefit to speechreaders. The prospect for a portable, wearable, automatic cueing system that can deal with the speech of unfamiliar talkers seems as remote now as it was when the Autocuer was developed.

Since all speech recognizers are likely to make some phone errors, it is essential that errors in the reception of cues be minimized if such cues are to benefit the speechreader. The experiments we have reported indicate that achieving this goal requires solving two additional problems. First, the cues must be displayed distinctively. Static handshapes displayed at fixed positions about the face may not be the best way to convey cues. It may be possible to improve cue reception by manipulating other display variables, e.g., size, orientation, intensity, hue, and saturation. Since reception of cues in the MCS (and PSC) condition is not perfect, even though the cues are free of errors, ample opportunity exists to improve the effectiveness of an automatic cueing system beyond improving the recognizer.

Second, the cues must be presented with minimal delay relative to speech production. Unfortunately, this consideration conflicts with the need for high recognizer accuracy; at the present time ASR accuracy improves when the recognizer is allowed to delay decisions about a given phone until subsequent phones are processed. Achieving a synchronized display may require that the cues be super-imposed on delayed images of the face. For cue receivers who use hearing aids, it would likewise be important to delay the acoustic signal to maintain synchrony with the cues. Providing such synchrony is technically feasible at present, but at the cost of increased complexity. As a result, it is likely that the first automatic cueing systems would be restricted to home and school environments, rather than used in everyday activities.
Simulation studies such as those reported in this manuscript can be reliable guides to the design of real-world systems if they fairly represent the situations to be encountered and if the outcome measures are appropriate to the problem being solved. The pessimistic conclusions reached by the authors of the Vidvox study may reflect, in part, the inadequate training provided to the subjects. On the other hand, the overly optimistic conclusions of the Autocuer Project may be the result of unrealistic simulations and unrepresentative measures of effectiveness. Hopefully, the simulations used in the current study will prove more useful to the future developers of speechreading supplements.

Appendix

**Context Independent Recognizer**

The recognizer used in the majority of the tests had been developed previously in an attempt to exploit the redundancy in the speech waveform (Duchnowski, 1993). It performs phonetic recognition on each of four non-overlapping frequency bands of the input and combines the results to produce the overall phone estimate. Each of the four band-recognizers was of the HMM type. The input (in each band) was divided into 20ms frames with 10ms overlap. Each frame was parameterized with four mel-frequency cepstrum coefficients, four delta coefficients, and four energy-measuring parameters. Each of these parameter vectors was quantized from separate codebooks with 128 prototypes. Each of 48 phones, including voiced and unvoiced closures, and epinthetic and normal silences, was modelled as a three state sequence with discrete observation probabilities. These phone models were context-independent. The recognizer was trained on 3696 TIMIT sentences from 462 speakers and tested on 840 sentences from 168 different speakers. It identified 54% of the phones correctly (37% substitutions and 9% deletions). If confusions among acoustically similar phones, such as closures and silences, are ignored, the identification score for the reduced 39-phone set is increased to 59%.

**Cue Specification**

The relationship between the manually produced phone transcription and the sequence of phones, specified by the recognizer was used to construct a "confusion matrix" whose rows corresponded to the transcribed phones, and whose columns corresponded to the recognized phones. Each entry in this matrix specified the number of times a given phone in the TIMIT transcription was recognized as itself, deleted, or confused with another phone. As such, this matrix characterized the pattern of correct responses and of substitution and deletion errors for this recognizer. However, since the number of errors made by this recognizer was too large for the automatic cueing application, the confusion matrix was modified to achieve a higher identification score while retaining the error pattern. This modification involved increasing the number of correct identifications of each input phone while leaving the number and distribution of errors unchanged.

The modified confusion matrix specified the relative frequency with which each phone would be correctly recognized, deleted, or incorrectly identified as one of the other phones. Given the modified confusion matrix and the manually produced phonetic transcription of the sentence, the output of a simulated recognizer was determined according to the principle of statistical independence. For each phone in the sequence, the output phone was determined randomly in accordance with the frequencies specified by the row corresponding to the phone in the modified confusion matrix. Successive phones were treated independently, and the random selection of recognized phones occurred without replacement. As a result, the proportion and pattern of errors in the output phone sequence corresponded stochastically, rather than exactly, to that specified by the modified confusion matrix.

**Context Dependent Recognizer**

As discussed above, in the majority of the tests the ASR was simulated by modifying the phone confusion matrix produced by a real recognizer that employed context-independent phone models and discrete observation probabilities. To evaluate the effectiveness of cues that would be derived by a state-of-the-art phonetic ASR, we developed a second recognition system based on commercially available software (Woodland & Young, 1993; Young Woodland, & Byrne, 1993). This recognizer employed context-dependent phone models and continuous observation probability densities (e.g., Bahl, Jelinek, & Mercer, 1983; Rabiner, 1989). Recognition was based on 25 parameters computed every 10 msec from the acoustic waveform (12 mel-frequency cepstral coefficients, 12 delta coefficients and one measure of normalized energy, all computed over a 20ms frame). These parameters were assigned to a static vector and a dynamic vector. The distribution of each vector was modeled with a mixture of six Gaussian densities.

The recognizer used 1151 right-context dependent phone models derived from 960 training sentences (a mixture of CID, Clarke and Harvard) that had been spoken by the test speaker under non-cued conditions. The parameters of a bigram phone language model were also estimated from the training data. The Viterbi algorithm was used to recognize the
phones in 180 testsentences. Phone recognition accuracy was 79% (with 10.7% substitutions, 4.0% deletions, and 6.2% insertions).

References


**Author Note**

This study would not have been possible without the efforts of Dr. S. L. Goldman, M. Tuyo, S. Jones, and J. Wozniak concerned with the preparation of segmented and labeled speech materials, and with software for creating simulations of Cued Speech. The authors are grateful to Dr. R. C. Beadles for providing data from tests of simulations of the Autocuer performed subsequent to the study of Cornett et al. (1977). The research reported in this paper was supported by the National Institute on Deafness and Other Communications Disorders and by the H. E. Warren Professorship at M.I.T.

**Permission to copy**: The National Cued Speech Association, Inc, the copyright owner of this document allows single copies to be made for personal use, or multiple copies for classroom use. This consent does not extend to other kinds of copying including (but not limited to) copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale.