

AUTOMATIC PRODUCTION OF CUED SPEECH

Louis Braida and Paul Duchnowski

Research Laboratory of Electronics
 Massachusetts Institute of Technology
 Cambridge, MA 02139, USA

ABSTRACT

In Manual Cued Speech, the talker produces a sequence of hand gestures that resolve many of the ambiguities faced by a speechreader who is unable to hear the talker's voice. Since only a small number of talkers are able to produce cues fluently, attempts have been made to develop automatic systems that derive the cue sequence by analyzing the acoustic speech waveform. A system that produces cues in real-time is described, and its performance as an aid to speechreading is analyzed. Possible methods for increasing the effectiveness of automatic cueing systems are considered.

1. INTRODUCTION

Although most individuals who are deaf make use of speechreading, the ability to communicate by speechreading alone is severely limited because many acoustic distinctions between speech sounds are not manifest visually. In Manual Cued Speech [2] (MCS, Fig. 1), a visual supplement to speechreading, the talker produces hand signals while speaking to resolve these ambiguities. Eight hand shapes are used to distinguish between groups of consonants and four hand positions are used to distinguish between groups of vowels. The set of sounds in each group are readily distinguished through speechreading alone. Sounds that are difficult to distinguish through speechreading alone are placed in different groups.

The MCS system has been successfully taught to very young deaf children, who have subsequently used the system to facilitate communication, language learning, and general education. The benefits MCS provides to speech reception have been carefully documented: experienced receivers of MCS tested at our laboratory typically obtain keyword scores in low-context sentences of 84% for MCS compared to 30% for speechreading alone.

Research supported in part by the National Institute of Deafness and Other Communicative Disorders.

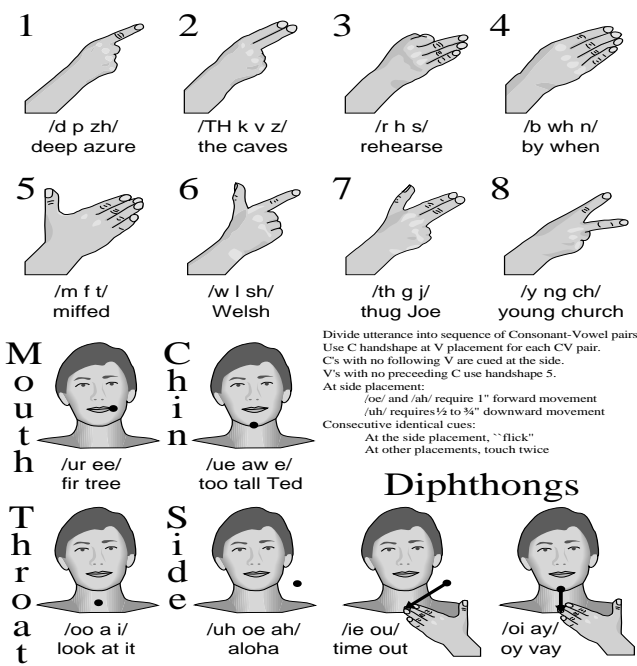


Figure 1: Composition of cue groups for consonant sounds, associated with hand shapes, and vowel sounds, associated with hand positions, in Manual Cued Speech.

2. PREVIOUS RESEARCH

The use of MCS in everyday communication is largely limited by the relatively small number of individuals who are proficient at producing MCS.

To overcome this limitation, Cornett and his colleagues at Gallaudet University and the Research Triangle Institute attempted to develop an "Autocuer" [3] that would derive cues similar to those of MCS automatically. A microprocessor was used to analyze the acoustic speech signal, with results displayed as virtual images projected near the face of the talker via modified eyeglasses.

Although a wearable prototype of the Autocuer was produced, the system is not in use at the present time, with no further development planned. This failure appears to reflect three limitations faced by this early development effort: poorly understood technical requirements, primitive speech recognition technology, and inadequate computational resources. Design goals were established in simulation studies, but no experienced users of cued speech were available at that time who could serve as subjects. Because only relatively inexperienced subjects were used, test materials were restricted to isolated words and short phrases. As a result the design goals derived for the Autocuer may not have been realistic. Unlike the HMM recognition techniques in present use, which are based on statistical models of sequences of speech characteristics, the ASR component of the Autocuer used heuristic decision rules based on estimates of easily extracted speech parameters, such as fundamental frequency, zero crossing rates, and peak-to-peak amplitudes in low- and high-frequency bands of speech. Speech processing was performed by a wearable/portable system based on an 8-bit microprocessor that provided less than 1% of current computation rates.

Although Cornett’s attempt did not prove successful, the three factors that limited his efforts are much less restrictive today. Theoretical analyses based on the measured performance of state of the art phonetic recognizers [4] have suggested that cues derived via ASR systems might improve speechreading substantially. These predictions have been confirmed by a simulation study [1] that tested highly experienced receivers of MCS. These subjects achieved keyword scores averaging 70% using cues produced by an HMM recognizer operating offline on recorded speech, compared to 36% for speechreading alone.

The simulation study also explored the effects of varying the time at which cues were displayed relative to the speech waveform. Whereas error-free cues produced keyword scores of 81% when aligned to the speech waveform, jitter of only 1 display frame (33 ms) reduced scores by five percentage points. Delaying the cue display by 166 ms, less than the duration of many CV syllables, reduced scores by 23 percentage points for cues produced with phone error rates comparable to those of the HMM recognizer tested.

3. REAL-TIME CUEING SYSTEM

The results of the simulation study were sufficiently encouraging to warrant a new attempt to develop the real-time automatic cueing system shown in Fig. 2. For convenience, the three major functions: parameterizing

the acoustic speech signal, speech recognition to identify cues, and display of cues to the cue receiver were distributed between two computers and a floating-point DSP board. Developments in computer hardware should soon make it possible to integrate these functions in a single laptop PC.

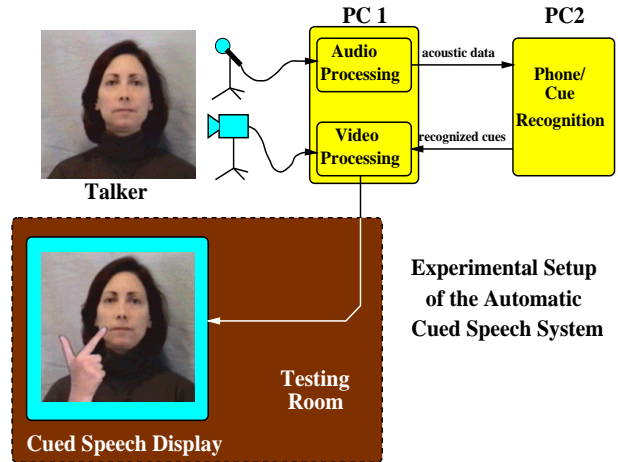


Figure 2: Automatic cueing system with one computer (PC1) assigned to speech acquisition, parameterization, image buffering, cue superposition, and display while the other (PC2) is assigned to cue recognition. The talker and cue receiver are situated in separate rooms, as in the evaluation tests.

3.1. Parameterization

The speech waveform was sampled at 10 kHz, high frequency pre-emphasized by a first-order filter with a cutoff frequency of 150 Hz and divided into 20 ms-long frames with 10 ms overlap. For each frame a vector of 25 parameters was derived including 12 mel frequency cepstral coefficients, 12 differences of cepstral coefficients across frames, and the difference between frame energies. RASTA processing was applied to the parameter vectors to improve robustness. Parameter vectors were time-stamped to allow for subsequent synchronization with video images of the talker.

3.2. Recognition

The recognition subsystem was a speaker-dependent HMM-based phonetic recognizer that used three-state left-right models based on mixtures of six diagonal-covariance Gaussian densities. Static and difference parameters formed distinct “streams” with different probability densities. Models were trained on roughly 60 minutes (1000 sentences) of speech data. Recognition

was performed using a Viterbi beam search modified to accept continuously arriving data vectors and to decode the corresponding phone sequence in real time.

Phone models were dependent on both the preceding and subsequent phone context. To limit the number of models, generalized contexts were used. Thirteen context classes, each containing between 2 and 6 phones, were constructed. Of the 7774 possible models only those most frequently seen in training data (roughly half) were used.

To achieve real-time operation a decoding procedure similar to the Forward-Backward search was used. Context-independent models were used in the first (backward) pass to determine which phone hypotheses exceed a likelihood threshold. This “pruning” reduces the number of phone models (the “beam width”) whose match to the acoustic data was then evaluated in the second (forward) pass using context-dependent models.

The accuracy of this recognizer was 80% off-line and 74% for live speech. The recognized phones were converted into a sequence of codes for cues via a finite-state grammar. These codes and start/stop times were then sent to the cue display subsystem.

3.3. Display

The cue display subsystem captured images of the speaker’s face, superimposed the appropriate cues on these images, and displayed the resulting sequence of images to the cue receiver. As discussed above, our simulation study had underscored the importance of synchronizing the presence of cues to the talker’s speech. Since cues cannot be identified before the speaker has uttered the corresponding syllable, the video image of the talker’s face was stored for delayed playback while the cue to be displayed was recognized. The delay time was 2 seconds, a period that was more than adequate to allow the cue to be recognized. Each stored frame was then retrieved, one of the hand shapes was superimposed at the appropriate location, and the composite image was displayed on a monitor. The artificially cued talker, as seen by the cue receiver, was thus delayed by two seconds relative to the real talker, but was displayed continuously, using smooth, full-motion video.

The display used heuristic rules to apportion time spent at target positions (corresponding to vowels) and time spent in transition between these positions. Typically 150 ms was allocated to the transition provided the hand could remain at the target position for at least 100 ms. Because human cuers often start to form cues before producing audible sound, the time at which cues were displayed was advanced by 100 ms relative to the start time determined by the recognizer. The change

from one hand shape to the next occurred halfway through a transition.

4. EVALUATIONS

The automatic cueing system was evaluated by three experienced users of MCS. Test materials consisted of keywords spoken in low-context phonetically-balanced IEEE-like sentences spoken by a single female talker who was a skilled producer of MCS. Test conditions included speechreading alone (SA), Manual Cued Speech (MCS) and automatically produced cued speech (ACS). Responses were scored as the percentage of keywords recognized correctly, using strict scoring rules.

Results indicated that the automatically generated cues produced scores averaging 66.3% correct compared to scores of 35.0% for SA and 89.6% for MCS. Thus the cues generated by the real time system provided roughly half the improvement in keyword recognition score that was obtained with MCS. All three subjects reacted favorably to these artificially generated cues, indicating that they afforded appreciable aid relative to speechreading alone.

5. IMPROVED CUEING SYSTEMS

The results of our evaluation of the effectiveness of automatically generated cues are highly encouraging and confirm the outcomes of our simulation studies. Since we used low-context sentences as test materials, it is likely that trained receivers of MCS could achieve even higher levels of speech reception in realistic, full discourse situations.

Nevertheless we believe that even greater assistance can be provided to speechreaders by improved automatic cueing systems. Some improvements are likely to arise from advances in automatic speech recognition systems, from increases in computation rates, and from their interaction. For example, the accuracy of the recognition system depends on the size of the beam width, which is largely determined by computation speed in a real-time recognition system. Increased computation speed should reduce the number of correct phone alternatives that are deleted by pruning.

The benefits provided by automatically generated cues can also be increased by improvements in the way the recognized cues are presented to the receiver. These improvements in the display should reduce the difficulty cue receivers experience in distinguishing between similar hand shapes and improve their ability to integrate the cues with speechreading.

5.1. Cue Discriminability

Analysis of the performance of experienced cue receivers on speech reception tests suggests that more than one-quarter of the word errors can be attributed to incorrect reception of cues for segments in words [4]. Even in perfectly cued materials, 10–20% of the segments are perceived incorrectly by experienced cue receivers.

It may be possible to reduce the number of such errors by altering the appearance of the surface or the outline of the hand. The results of two recent experiments indicate that discrimination between highly confused shapes can be improved, by coloring the handshapes redundantly, without interfering with speechreading,

In both experiments listeners with normal hearing who were unfamiliar with cued speech identified the consonant spoken and the handshape displayed. Analysis of confusions between different handshapes indicated that they were largely determined by the number of extended fingers (e.g, in Fig. 1, shapes 1, 2, 3, and 4; 5, 6, and 7) and by the extension of the thumb (e.g, shapes 1 and 6, 2 and 7, & 4 and 5). Coloring the three shapes that were least well identified reduced the error rate from 14% to 7% when cues were displayed tachistoscopically. Coloring the five shapes that were least well identified reduced error rates from 31% to 23% when cues were displayed for 200 ms surrounded temporally by 66 ms displays of conflicting shapes.

These results indicate that selectively coloring the hand shapes used to display the cues reduces confusions among consonants. Changes in the display of cue shapes are easy to introduce, but it is unclear how much training will be required for cue receivers to benefit from them. Experiments to determine whether such coloration improves the reception of cued speech are currently under way.

5.2. Cue Integration

It may be possible to improve the ability of cue receivers to integrate displayed cues with speechreading by increasing the correspondence between the actions of the synthetic hand and those of a human hand. Although the timing of hand and facial actions is specified in MCS, the manner in which a cuer should change the shape of the hand during cue transitions is not. However, cue receivers are sensitive to both the timing of hand movements and the shape changes that occur as the hand moves from one target position to the next.

The synchronization of hand and facial actions has proven to be a key variable in determining the intelligibility of automatic cued speech. Phonetic recogniz-

ers align their transcriptions to acoustic events, which usually (but not always) lag the corresponding visible facial actions. Direct comparison of the displayed synthetic cues with the hand motions of human cuers indicated that this delay averaged roughly 100 ms. By advancing the presentation of cues by this amount (three video frames) relative to the recognized acoustic events, we were able to improve speech reception substantially.

This simple synchronization adjustment may not be optimal, however. Additional research is required to determine whether the required advance is phone-specific and/or context-dependent. It may also be possible to improve speech reception improvements by exaggerating the timing differences employed by human cuers.

6. CONCLUSION

The work described here suggests that automatic cueing systems are likely to have more limited applications than originally envisioned by those who attempted to develop an Autocuer nearly a quarter century ago. Even expected advances in technology are unlikely to make it possible to extract useful cues automatically from real speech waveforms in arbitrary acoustic environments. Moreover, the need to provide accurate temporal registration between the displayed cues and facial actions makes it unlikely that a system that displays cues on the live face of the talker, as in the Autocuer, can be successful. In more restricted environments, such as homes and classrooms, it should be possible to provide automatic cueing systems similar to the one described here that provide tangible assistance to deaf speechreaders.

REFERENCES

- [1] M.S. Bratakos, P. Duchnowski, and L. D. Braida. "Towards the Automatic Generation of Cued Speech," *The Cued Speech Journal* vol. 6, June 1998, 1–37.
- [2] R.O. Cornett. "Cued Speech." *Am. Annals Deaf*, 112:3–13, 1976.
- [3] R.O. Cornett *et al.* "Automatic Cued Speech." *Proc. Res. Conf. on Speech-Proc. Aids for the Deaf*, Gallaudet College, 224–239, May 1977.
- [4] R.M. Uchanski, L.A. Delhorne, A.K. Dix, C.M. Reed, L.D. Braida, and N.I. Durlach. "Automatic Speech Recognition to Aid the Hearing Impaired. Prospects for the Automatic Generation of Cued Speech," *J. Rehab. Res. & Dev.*, 31:20–41, 1994.