A SPEECHREADING AID BASED ON PHONETIC ASR

Paul Duchnowski *Louis Braida* *Maroula Bratakos* *David Lum* *Matthew Sexton* *Jean Krause*

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

ABSTRACT

Manual Cued Speech (MCS) is an effective method of communication by the deaf and hearing-impaired. We first describe our work on assessing the feasibility of automatic determination and presentation of cues without intervention by the speaker. The conclusions of this study are then applied to the design and implementation of a prototype automatic cueing system using HMM-based automatic speech recognition software to identify the cues in real time. We also describe the features of our cue display that enhance its effectiveness such as style of cue images and the timing of their transitions. Our experiments show keyword reception by experienced MCS users to improve significantly with the use of our system (66%) relative to speechreading alone (35%) on low-context sentences.

1. INTRODUCTION

Speechreading is used by virtually all listeners able to observe the speaker's face to improve their comprehension when the acoustic signal is difficult to interpret. Deaf and hearing-impaired individuals are often particularly dependent on speechreading (also sometimes referred to as lipreading) for communication. It is well known, however, that the visible articulators do not allow the observer unambiguous access to all the speech elements in most languages [8]. Manual Cued Speech (MCS) was invented in 1968 to aid in the process of visual speech reception [3].

An MCS-using speaker gestures with his/her hand to convey additional information about the identity of the phonemes that he/she is articulating. Phonemes are grouped into classes containing between two and four phonemes. The shape of the hand indicates the class of the currently spoken consonant and the position of the hand relative to the speaker's face indicates the vowel class. Each cue, i.e., a hand shape at a particular position, is thus generally associated with a consonant-vowel (CV) syllable. Special provisions are made for consonant clusters and unpaired vowels. For English MCS prescribes 8 hand shapes and 4 hand positions (Fig. 1). Phonemes that are difficult to distinguish visually[1] are assigned to different cue classes. Conversely, the phonemes assigned to the same class are easily distinguished on the lips. In this way a cue seen in conjunction with lip motion allows the receiver to identify unambiguously the spoken syllable. [VIDEO 0589_01.MPG] shows the manually cued sentence "The old castle passed from the duke to the king."

MCS is very effective in improving speech reception of its users. Studies [10, 11] show scores for keywords in low context sentences rising from roughly 30% with speechreading alone to roughly 90% with MCS, a level of reception compatible with a normal conversation. There is evidence [12] that young deaf children exposed regularly to MCS develop reading skills comparable to normal-hearing ones. To the extent that languages make use of similar phoneme sets, MCS is also relatively language-independent.

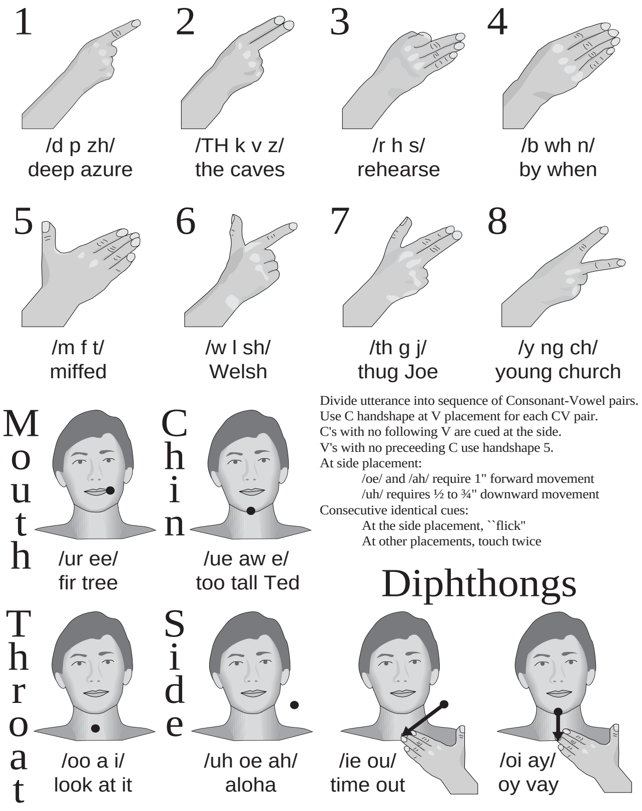
1
/d p zh/
deep azure
2
/TH k v z/
the caves
3
/r h s/
rehearse
4
/b wh n/
by when
5
/m f t/
miffed
6
/w l sh/
Welsh
7
/th g j/
thug Joe
8
/y ng ch/
young church
Mouth
/ur ee/
fir tree
Chin
/ue aw e/
too tall Ted
Divide utterance into sequence of Consonant-Vowel pairs.
Use C handshape at V placement for each CV pair.
C's with no following V are cued at the side.
V's with no preceeding C use handshape 5.
At side placement:
/oe/ and /ah/ require 1" forward movement
/uh/ requires ½ to ¾" downward movement
Consecutive identical cues:
At the side placement, ``flick"
At other placements, touch twice
Diphthongs
Throat
/oo a i/
look at it
Side
/uh oe ah/
aloha
/ie ou/
time out
/oi ay/
oy vay

Figure 1. Assignment of consonant sounds to hand shapes and vowel sounds to hand positions and the basic rules of Manual Cued Speech.

The assistance of Cued Speech is limited to situations in which the talker or a transliterator produces the cues. A computerized system that automatically deduced the appropriate cues from the acoustic speech and presented them to the receiver would be of potentially great benefit to Cued Speech users. It would be an attractive application of phonetic automatic speech recognition (ASR) since cues are based on the phonemes of an utterance rather than its spelling. Unlike word-based transcription of speech (automatic or not) this device would be useful to deaf individuals with low reading skills such as young children. Automatic cueing would happen in conjunction with speechreading and users might compensate to some degree for cue errors with more attentive observation of the speaker's lips. On the other hand, the failure of a text transcription would be more serious since human integration of lipreading and written words is very limited.

An attempt at an automatic cueing device, the "Autocuer", was made in the late 1970s [4]. Due in part to the limitations of available technology it was not possible to develop an effective system

[1] Such phonemes are said to form a *viseme*. Examples are /p,b,m/, /f,v/, and /iy,ih/.

that worked in real time. With the advent of improved ASR techniques and better display options we believe an automatic cueing system providing tangible benefit to the receiver is now feasible. In this paper we describe the development and initial results obtained with such a system.

2. SIMULATION STUDIES

Any automatic cueing system is likely to depart from Manual Cued Speech in three principal ways:

- The cue images will differ from naturally articulated human hands.
- The ASR system will misidentify some cues, miss some, and insert spurious ones.
- A cue cannot be recognized until after it is spoken. Uncorrected synthetic cues would always lag behind the corresponding lip motion.

We conducted simulation studies to assess the importance of each of these factors and to guide the development of the real-time system. A more detailed description may be found in [2].

2.1. Cue Display

It is evident that making the synthetic display resemble the manual system increases its chances of success with skilled MCS receivers. This strategy also minimizes training requirements - an important issue since gaining proficiency in MCS reception of conversational speech can require many months [10].

In our synthetic display images of the talker are shown on a standard television monitor. Pre-recorded still images of a human hand in one of the eight prescribed handshapes are digitally superimposed in an appropriate position near the talker’s face. The talker is generally only shown from the shoulders up, i.e. with enough space to display a cue in the “throat” position (Fig. 3).

For our simulation study the cues were *discrete* in both shape and position. The hand image is fixed in both shape and position for the duration of a cue. Shape and/or position may change instantaneously at the beginning of the next cue. Times of cue occurence are defined relative to the acoustic waveform. [VIDEO 0589_02.MPG] shows the artificially cued sentence “The loss and two wins were fair games.”

2.2. Experiments

The principal test materials consisted of the 720 low-context IEEE sentences[2] each containing 5 keywords [7], spoken by a teacher of the deaf, highly experienced in producing MCS. They were recorded in a professional studio at 30 frames/second and transferred to video disks. At least 40 sentences were used for each condition tested and no subject saw the same sentence twice.

We tested keyword reception with speechreading alone (SA), MCS, and synthetic cues. Cues were superimposed on video frames off-line by computer. The cues’ identities and time boundaries were determined from manual phonetic transcriptions of the sentences based on their acoustic waveforms. Sentences where the cues corresponded to the transcriptions exactly were considered “perfectly synthetically cued” (PSC).

To test the effect of cue imperfections we artificially introduced errors and delays. Cue errors were inserted by randomly changing phonemes in the transcription to achieve a target average phonetic accuracy (90 and 80 percent). Delays were effected by displaying the cues 1, 3, or 5 video frame (33, 99, or 165 ms) after the start time determined by the manual transcription. We also transcribed the sentences off-line with an HMM-based automatic speech recognizer using right-context-dependent phone models (see Section 3.1 for more detail on the recognition software) with a speaker-dependent phonetic accuracy of roughly 80%. Cues based on the automatic transcription (referred to as AC) also had timing errors but no systematic delay was added.

Our subjects were generally young adults with at least ten years of experience with MCS. They were tested in two phases with at least four subjects in each phase. They viewed the stimuli on monitors in sound-treated rooms. No audio signal was presented.

2.3. Results

Figure 2 shows keyword scores for selected test conditions from the two experimental phases. Each symbol represents the correct recognition percentage of a subject under a cued condition as a function of that subject’s unaided recognition percentage. Symbols falling above the solid curve indicate an improvement in performance sufficient to allow reasonable conversation.

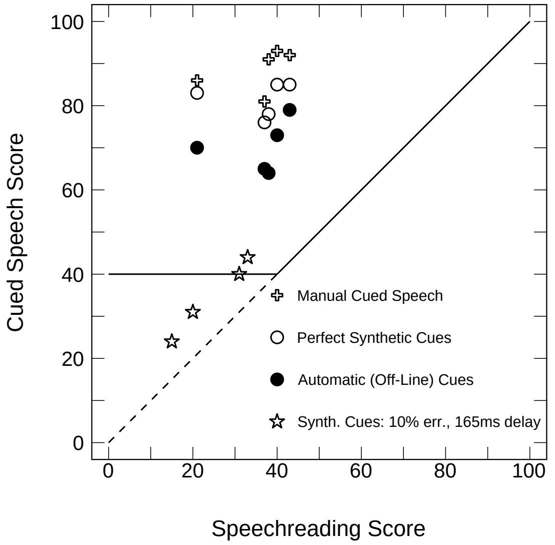

Figure 2. Selected keyword reception scores (aided vs. speechreading alone) for subjects in the simulation study.

As expected, the lowest scores were obtained in the SA condition (roughly 30%) ; the highest with MCS (89%). Scores in the PSC condition (81%) were slightly lower than in MCS reflecting the effects of differences in speaking rates (100 wpm for MCS, 140 for PSC), cue display (articulated vs. discrete), and cue timing (in the PSC condition, the time reference for the display of cues was derived from the acoustic waveform; in MCS the shapes and positions of the cuer’s hands often change before there is detectable sound).

Both simulated recognizer errors and delays in the display of cues reduced scores. When 10% of the phones were in error, scores decreased by 14 percentage points; a 20% rate of errors reduced scores by 24 points (not shown in Figure 2). In combination with a delay of 165 ms, the effect of a 10% error rate was

[2] A representative sentence might be: “Glue the sheet to the dark blue background.”

even larger, reducing scores by 38 points. The effect of a 20% error rate was also increased by a delay of 99 ms but delays of 33 ms, whether fixed or random, did not affect scores significantly. When the cues were derived from the phone sequence produced by the HMM recognizer scores were roughly the same as as for the errors simulated at 10%. It appears that the recognizer's tendency to cluster errors rather then distribute them uniformly results in more words free of cue errors for the same phone error rate. The subjects' scores in the AC condition averaged over 70%.

The results of the simulation study are encouraging, suggesting that an ASR system can produce cues that aid speechreading. On the other hand, they also revealed the deleterious influence of recognizer delay on the effectiveness of these cues.

3. REAL-TIME AUTOMATIC CUEING SYSTEM

The results of the simulation study guided the design of our real-time cueing system. Figure 3 shows a block diagram of the resulting prototype.

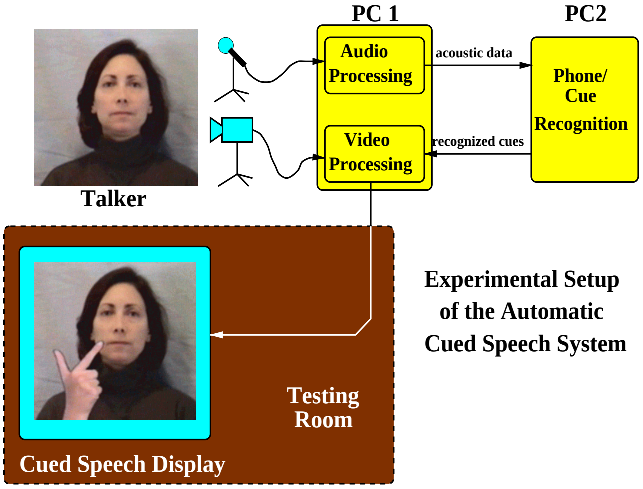

Figure 3. Current ACS system with one computer devoted to cue recognition and the other primarily handling image buffering, cue superposition, and display. The talker and cue receiver are placed in separate rooms.

The talker sits facing a video camera and wears a lapel microphone. PC1 digitizes and pre-processes the microphone signal. PC2 (an AlphaStation 500) performs phonetic recognition and derives a sequence of identified cues. Simultaneously, PC1 framegrabs video images of the talker and stores them in a memory buffer for two seconds. This gives PC2 time to identify the cue corresponding to each video frame. At the end of the storage period the identified cue (a handshape at a specified position) is superimposed on the talker's image and the composite image is displayed on a television monitor. The storage interval was chosen conservatively and is probably at least twice as long as needed. Note that, since all frames are stored for the same amount of time, the cued output appears as normal, full-motion video, albeit delayed relative to the live speech by two seconds.

3.1. Automatic Cue Recognition

The acoustic speech waveform was sampled at 10 kHz and divided into 20 ms-long frames with 10 ms overlap. Each frame was parameterized by a 25-element vector with 12 mel-frequency cepstral coefficients, 12 difference cepstral coefficients, and the difference between frame energies. To improve robustness of this representation we applied RASTA processing [6] to the parameter vectors.

The phonetic recognition programs were based on the HTK software from Entropic [5] and operated in speaker-dependent mode. We used three-state Hidden Markov phone models. The output probability densities were estimated with mixtures of six Gaussian densities, found to be optimal in our pilot studies for our available training data of about 1100 sentences per speaker. We initially trained 46 context-independent phone models similar to those used in [9]. Recognition was performed with the HTK Viterbi beam search decoder, modified to produce a continuous phone sequence.

The context-independent recognizer achieved phonetic accuracy[3] of 71% off-line and about 65% in live experiments. To improve recognition accuracy we explored several sets of context-dependent models. Best results were obtained with triphone models with generalized contexts and 13 context classes, achieving off-line accuracy of over 80%. We further modified the HTK decoding routines by implementing a faster search similar to the Forward-Backward algorithm described in [1] and adaptive beam computation to assure real-time performance. The average live-speech accuracy of the resulting recognizer was roughly 74%.

The phone sequence produced by the recognizer is converted to cue codes according to the rules of MCS and transmitted, including the cue start times, to the cue displaying computer PC1.

3.2. Cue Display

Some of the participants in the simulation study commented unfavorably on the discrete nature of the handshapes in our cue display. We experimented with several variations of the display, all designed to suggest a smooth motion between positions. We also studied tapes of human cuers to gain insight into their cue-timing strategies.

The currently most successful display uses heuristic rules to allocate cue display time between time spent at target positions and time spent in transition, i.e., intermediate positions, between these targets. We observed that human cuers often begin to form a cue before producing the corresponding audible sound. To approximate this effect we adjusted the start times of the cues to begin 100 ms before the boundary determined by the cue recognizer. We also found the timing of the conversion from one handshape to the next to be broadly optimal when cues change halfway through the transition. The handshape images themselves remain static - the change is instantaneous, with no intermediate shapes. [VIDEO 0589_03.MPG] shows the artificially cued sentence "The kite may fly on this windy day" using the improved display style.

3.3. Experiments

We tested the real-time cue generating system using a new speaker and experienced cue receivers. Some of the subjects had participated in our simulation study, but care was taken not to repeat any of the stimuli seen in that study. Speech materials consisted of IEEE sentences (Sec. 2.2) and a new set of IEEE-like

[3] Accuracy scores account for substitutions, deletions, and insertions.

sentences constructed by us, using the same keywords and grammatical structure.

The experiments were conducted live, with the speaker seated in a sound-proof booth and the cue receivers watching the automatically cued speech on a TV screen in another. A testing session generally lasted 2 to 3 hours with sentences presented in groups of 40 to 60 with breaks in between. Each condition was tested with at least 50 sentences (250 keywords). Each session began with 40 to 60 practice sentences but beyond that no training of the subjects was done.

3.4. Results

All versions of the automatic cueing system that we tested resulted in at least a small benefit to the cue receiver relative to speechreading alone. As expected, improving the accuracy of the cue recognizer improved the keyword scores of the subjects. The switch to dynamic cue display had a surprisingly large positive effect on the keyword scores: an average increase of 14 percentage points.

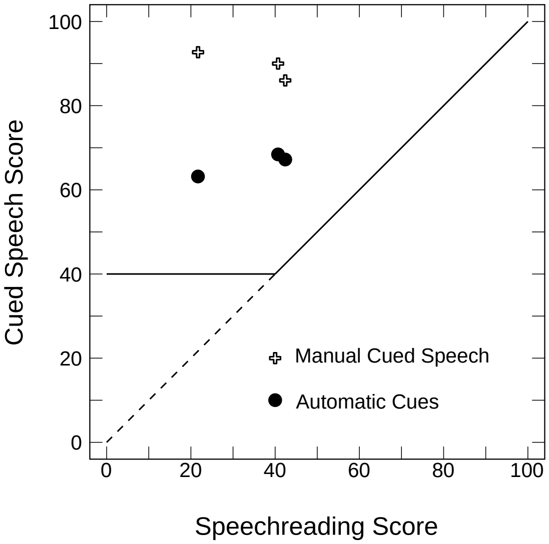

Figure 4. Keyword reception scores (aided vs. speechreading alone) for subjects using MCS and our most advanced automatic cueing system.

Figure 4 shows the keyword reception scores of the three subjects tested with our best-performing system, which used generalized-context triphone models for recognition and dynamic, timing-adjusted cue images for display. The subjects' performance under MCS remains high at roughly 90%. The automatic cues result in average keyword reception of 66%. This represents almost a doubling of the speechreading-alone scores.

These results were supported by subjective comments of the cue receivers who reported a clear benefit from the automatic system and are consistent with the results of subsequent, informal tests.

4. CONCLUSION

We have demonstrated, through simulation studies and live, real-time experiments that automatic generation of Cued Speech is feasible with current ASR and display technologies. We have designed and implemented a prototype cueing system that can automatically determine and display cues for speakers unfamiliar with MCS. Skilled receivers of MCS correctly receive almost twice as many keywords in low-context sentences using our system as they do with speechreading alone. This benefit is achieved with virtually no training on the part of the receivers.

Efforts are underway to improve the accuracy of our cue recognition system which would clearly contribute to the effectiveness of the cues. We are also working on improving its robustness, in anticipation of field trials in a classroom or lecture setting. The unexpectedly significant influence of the display style also suggests that further gains may be achieved with appropriate modifications. We are studying the cue timing issue to determine a better strategy for synchronizing the cues to the visible facial actions of the speaker. We are also exploring techniques for improving the discriminability of the hand images such as the use of color. This could prove beneficial for automatically cued speech where cues are about 30% shorter than in MCS.

5. ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health. The authors thank Joe Frisbie for generously providing the cue chart of Figure 1.

6. REFERENCES

[1] S. Austin *et al.* "The Forward-Backward Search Algorithm," *Proc. 1991 Int. Conf. Speech, Acoust., Sig. Proc.*, 1:697–700.

[2] M.S. Bratakos *et al.* "Towards the Automatic Generation of Cued Speech," To appear in *The Cued Speech Journal.*

[3] R.O. Cornett. "Cued Speech." *Am. Annals Deaf*, 112:3–13, 1976.

[4] R.O. Cornett *et al.* "Automatic Cued Speech." *Proc. Res. Conf. on Speech-Proc. Aids for the Deaf*, Gallaudet College, 224–239, May 1977.

[5] Entropic Research Laboratories, Inc. HTK: Hidden Markov Model Toolkit V1.5. December 1993.

[6] H. Hermansky and N. Morgan. "RASTA Processing of Speech," *IEEE Trans. Speech Audio Proc.*, 2(4):578–589, Oct. 1994.

[7] IEEE. "IEEE Recommended Practice for Speech Quality Measurements," *Technical Report No. 297*, June 1969.

[8] P.L. Jackson. "The Theoretical Minimal Unit for Visual Speech Perception: Visemes and Coarticulation," *The Volta Review*, 90(5):99–115, Sept. 1988.

[9] K.-F. Lee and H.-W. Hon. "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. Acoust., Speech, and Sig. Proc.*, 37(11):1641–1648, Nov. 1989.

[10] G. Nicholls and D. Ling. "Cued Speech & the Reception of Spoken Language," *J. Speech Hearing Res.*, 25:262–269, 1982.

[11] R.M. Uchanski *et al.* "Automatic Speech Recognition to Aid the Hearing Impaired. Prospects for the Automatic Generation of Cued Speech," *J. Rehab. Res. & Dev.*, 31:20–41, 1994.

[12] J.E. Wandel. "Use of internal speech by hearing and hearing-impaired students in oral, total communication, and Cued Speech programs," PhD Dissertation, Columbia University, New York, 1989.