Auditory-Visual Speech Processing
(AVSP'98)
Terrigal – Sydney, Australia
December 4-6, 1998

AUTOMATIC GENERATION OF CUED SPEECH FOR THE DEAF: STATUS AND OUTLOOK

Paul Duchnowski *Louis Braida* *David Lum* *Matthew Sexton* *Jean Krause* *Smriti Banthia*

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

ABSTRACT

Manual Cued Speech is a system of hand gestures designed to help deaf speechreaders distinguish among ambiguous speech elements. We have developed a computerized cueing system that uses automatic speech recognition to determine and display cues to the cue receiver. Keyword scores of 66% in low-context sentences have been obtained with this system, almost double the speechreading-alone scores. We describe the design issues with the largest impact on the cuer's performance, concentrating on the characteristics of cue display and cue timing. Enhancing cue images with color is found to improve their discriminability and may lead to improved speech reception by cue receivers.

1. INTRODUCTION

Listeners with hearing loss typically rely heavily on speechreading. However, even the most skilled speechreaders typically miss more than one-third of the words spoken due to the similarity in appearance of certain speech elements (e.g., /p/, /b/, and /m/) which leads to the inherent ambiguity of speechreading. Manual Cued Speech (MCS) was invented in 1968 to aid in the task of resolving this ambiguity thus improving communication [3].

In MCS, the speaker gestures with his/her hand to resolve ambiguities among speech elements that are often confused by speechreaders. MCS prescribes 8 hand shapes to distinguish among English consonants and 4 hand positions relative to the face to distinguish among English vowels, with small variations for other languages. Thus a hand shape placed in one of the MCS positions corresponds to consonant-vowel (CV) syllable. Special rules cover cueing of consonant clusters and vowels without a consonant. Figure 1 shows the assignment of specific phonemes to cue classes and the basic rules of manual cueing. The members of a particular cue class are generally easily distinguished from each other on the lips. On the other hand, phonemes that are ambiguous visually are placed in different classes. Simultaneously seeing the motion of the speaker's lips and the corresponding cue allows the cue receiver to resolve most ambiguities in speechreading. [VIDEO AVSP17_1.MPG] shows the manually cued sentence, "The lawn grew around the flower bed."

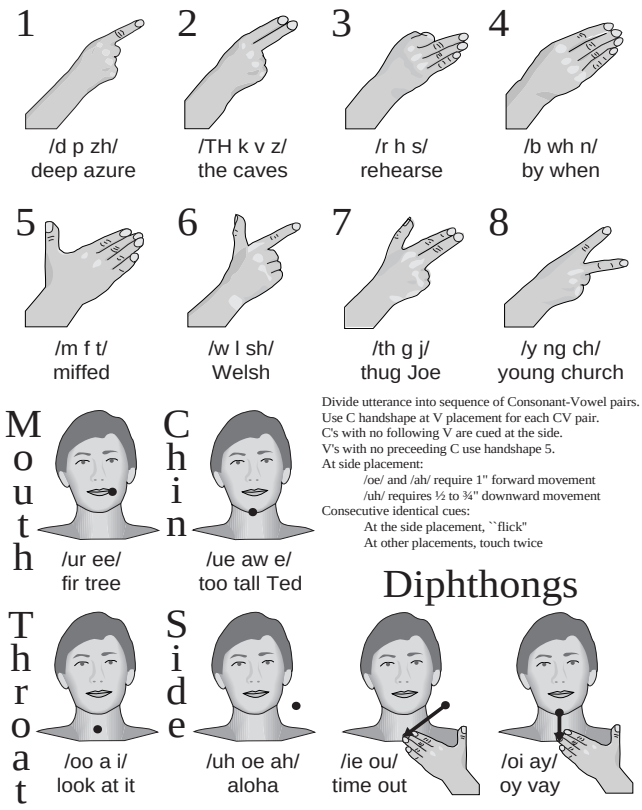

Figure 1. Assignment of consonant sounds to hand shapes and vowel sounds to hand positions and the basic rules of Manual Cued Speech.

The effectiveness of MCS in improving speech reception of its users is well documented (e.g. [8, 9]). Experienced receivers of MCS achieve nearly perfect reception of everyday connected speech and children who grow up communicating by MCS develop reading skills comparable to their normal-hearing peers [10]. Since the system is phoneme-based it is also adaptable to many languages.

In spite of these advantages, the widespread use of MCS is restricted by the small number of individuals who are trained to produce the cues. Given the recent advances in automatic speech recognition (ASR) and display technologies, we believe a computerized sys-

tem that identifies the cues and presents them to the receiver is now feasible.

The ideal system would require little or no intervention by the speaker and would be compatible with the manual system. These characteristics will increase acceptance of the automatic cuer, re-inforce the learning of cue reception, and simplify the development of the device. Proficiency in MCS reception can require well over a year of practice. Training time of such length would make research on an automatic cuer all but impractical. In this paper we describe the development of an automatic cueing system and the design factors that affect its success, especially the appearance and timing of the synthetic cues.

2. AUTOMATIC CUEING SYSTEM

Figure 2 shows the current configuration of our prototype automatic cueing system. The talker sits facing a video camera and wearing a microphone. PC1 preprocesses the acoustic waveform and handles capture of images of the talker. PC2 (an AlphaStation 500) performs phonetic recognition and produces the best-match cue sequence. The digital images are stored in PC1 memory for two seconds prior to superposition of a hand image corresponding to the cue indicated by PC2 and playback on a monitor for the cue receiver. This delay period allows more than enough time for the cue to be identified by the recognizer (we estimate that one second or less should be adequate.) The artificially cued talker, as seen by the cue receiver, is thus delayed by two seconds relative to the talker's actions but is displayed as smooth, full-motion video.

Our phonetic recognition software is based on the HTK suite of programs [6] which implement a Hidden Markov Model (HMM) ASR system and is currently operated in speaker-dependent mode. Each 20 ms-long frame of the input speech is represented by a vector of 25 cepstral coefficients, differences, and frame delta energy. A phone is modeled with a three-state HMM with parameters described by mixtures of six Gaussian densities. We are currently using generalized context-dependent models and a modified Viterbi beam search for recognition. Phonetic accuracy is roughly 74% for live speech. More detailed description of the automatic cuer's implementation and of the recognizer may be found in [5].

2.1. Effectiveness

We have tested the current automatic cueing system, as well as versions developed earlier, using primarily

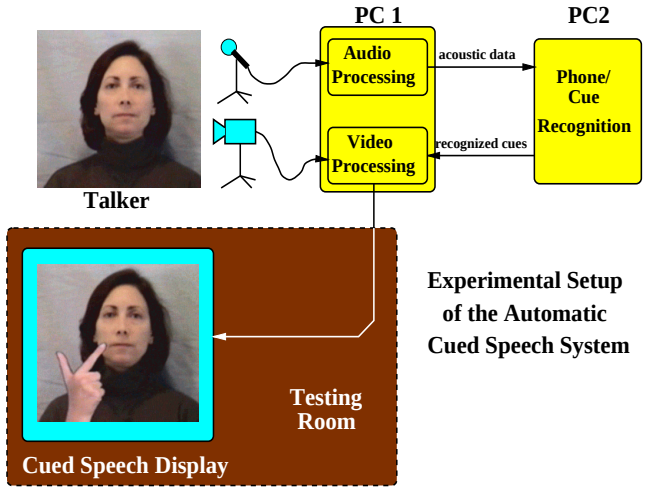

Figure 2. Current ACS system with one computer devoted to cue recognition and the other primarily handling image buffering, cue superposition, and display. The talker and cue receiver are placed in separate rooms.

sentences as experimental materials. In particular, we have used the low-context IEEE sentences [7] as well as IEEE-like sentences that we constructed ourselves using the same keywords and grammatical structure.[1] Each sentence was presented only once to a given subject. Subjects have been young adults with at least ten years of MCS experience.

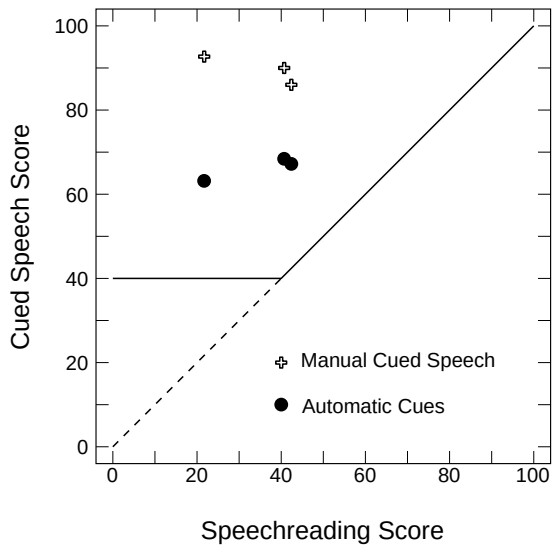

Figure 3. Keyword reception scores (aided vs. speechreading alone) for subjects using MCS and our most advanced automatic cueing system.

Figure 3 shows keyword reception scores achieved by three subjects using MCS and the current version of the automatic cuer. Each symbol plots the aided

[1] A representative IEEE-like sentence may be "The sheep wandered into the barn of the new farm."

#	Recognizer Type	Recognizer Phone Acc. (%)	Cue Display	Benefit to Cue Receiver
1	ctxt-ind.	65	smooth	18
2	ctxt-dep.	72	discrete	28
3	ctxt-dep.	72	dynamic	32
4	ctxt-dep.	74	dynamic+	57
5	ctxt-dep†	80	discrete	65
6	perfect	100	discrete	86

Table 1. Improvement in cue receivers' keyword scores for successive versions of the automatic cueing system and two simulations (systems 5 and 6). †This recognizer operated off-line.

score of a particular subject vs. his or her score with speechreading alone (SA). Scores above the main diagonal thus indicate an improvement in speech reception. The solid curve is the lower bound of the region where a given speechreading aid raises the scores to a level useful for communication.

Word scores for MCS averaged 90%, for automatic cueing (AC) 66%. The latter were almost double the average SA score of 35% and show a clear benefit to the cue receivers. These results were supported by subjective comments of the users who felt that the automatic system clearly improved their comprehension.

2.2. Effect of Cue Display

While improving cue accuracy is expected to improve the effectiveness of the automatic cuer, we found the characteristics of cue display also play a significant role. Table 1 shows the improvement in the cuer's benefit[2] afforded by successive versions of the system as related to the cue recognition accuracy and display style.

Systems 5 and 6 refer to simulated, off-line cueing systems that we tested to establish feasibility and benchmarks for the real time cuer. They used a *discrete* display style - the hand image is fixed in both shape and position for the duration of a cue and changes instantaneously at the beginning of the next cue. [VIDEO AVSP17_2.MPG] shows the synthetically cued sentence, "The low power made every light go dark," using discrete cues. Our simulation experiments are described in detail in [2].

[2]We define benefit as the percentage of the difference between SA and MCS keyword reception captured with the automatic cuer. E.g., for SA score of 30% and MCS score of 90% an AC score of 60% gives a benefit of 50.

Systems 1 through 4 are characterized by increasing accuracy of the phonetic recognizer and improvements to the cue display. The *smooth* display in system 1 showed the hand image always transitioning between target positions. The accuracy of the associated ASR system was likely too low to give much benefit regardless of display style. Improving cue accuracy in system 2 enhanced the automatic cuer's effectiveness. But it was the changes to cue display in systems 3 and 4 that clearly increased the benefit to the cue receiver.

The *dynamic* display uses heuristic rules to apportion cue display time between time spent at target positions and time spent in transition, i.e., at intermediate locations between these targets. The handshape images themselves are not articulated - they change shape instantaneously between frames. For the *dynamic+* display we studied tapes of human cuers to better understand their cue timing strategies. We found that cues are often formed before the corresponding sound is produced. To approximate this effect we adjusted the start times of cues to begin 100 ms before the boundary determined from acoustic data by the cue recognizer. We also experimented with the timing of the conversion from one hand shape to the next and attained good results when cues change halfway through the transition. [VIDEO AVSP17_3.MPG] shows the synthetically cued sentence, "Just use it up and order more later," using dynamic+ cues.

These display and timing changes, coupled with a small improvement in recognizer accuracy led to a significant improvement in the cueing system's effectiveness (system 5). These results suggest that benefit obtained with the simulated system 6 may be achievable with ASR accuracy of well under 100% if the cue display style and synchronization are refined further.

3. ENHANCING CUE RECEPTION

In theory, the combination of information conveyed by the speaker's visible facial actions and the information conveyed by cues should result in near perfect speech reception. This is, in fact, the case with syllables [9]. Nevertheless, highly experienced MCS receivers make errors on perfectly-cued low-context sentences spoken at 100 wpm, suggesting that 10–20% of the segments are perceived incorrectly [9]. These errors may be due to: 1) difficulty in distinguishing between similar hand shapes or positions, 2) difficulty in accurately perceiving facial actions of the speaker, 3) inability to integrate the cues with speechreading.

Preliminary analysis of the responses of cue receivers suggests that more than one-quarter of the word errors can be attributed to incorrect reception of cues for segments in words [9]. One way to reduce the number of such errors may be to alter the appearance of the surface or the outline of the hand. Two easily confused shapes could be displayed differing in brightness, size, or color. In recent experiments we focused on cue coloration to improve discrimination.

3.1. Experimental Design

Our experiments tested the discriminability of hand shapes before and after some of the hand images were digitally colored. We used pre-recorded consonant-vowel-consonant (CVC) syllables spoken by one female talker. There were 24 initial consonants with eight tokens of each for a total of 192 distinct syllable recordings. The vowel was always /ah/ while the final consonant was drawn at random from a pool of eight and did not play a role in these experiments.

Our subjects were normal-hearing adults, none of whom had any familiarity with MCS. Prior to the experiments they were given about ten hours of training in identifying the eight hand shapes and speechreading CVCs (shown separately). They learned to identify hand shapes shown for as little as 66 msec almost perfectly and scored between 50 and 60% on initial consonant identification. Two types of experiments were then conducted.

Experiment 1 A single hand shape was superimposed on a CVC recording for the duration of one video frame (33 msec). This frame was temporally located about 100 msec after the start of the initial consonant. Four distinct hand shapes were overlaid on each token, each in a different but random MCS position. A different set of hand shapes was associated with each CVC token but in such a way that every initial consonant was paired equal number of times with each hand shape, resulting in 768 distinct stimuli.

During an experimental session 192 randomized tokens were presented to the subjects who were asked to identify the hand shape and the initial consonant. No audio was presented. Subjects were not required to identify the position of the cue.

After several sessions we analyzed the subjects' responses and identified the three hand shapes with the lowest average identification scores. These three hand images were digitally colorized blue, red, and green while texture and contrast of the image were preserved to the extent possible. New stimuli were made using these and the five uncolored hand images using the procedure above. The subject tests were then repeated.

Stimulus Identified	Condition/Session Block (number of sessions in block)			
	U (4)	C (2)	U (2)	C (2)
Consonants	43.2	45.0	43.4	44.5
Handshapes	83.6	91.6	88.3	94.1

Table 2. Average percent correct scores for five subjects in Experiment 1.

Stimulus Identified	Condition/Session Block (number of sessions in block)			
	U (2)	C (2)	U (1)	C (1)
Consonants	43.0	43.7	43.3	42.7
Handshapes	65.2	75.5	73.3	79.3

Table 3. Average percent correct scores for three subjects in Experiment 2.

Experiment 2 A sequence of three hand shapes was superimposed on the CVC images. The middle hand shape, the *target*, lasted 6 video frames (198 msec) beginning at the same frame as the single hand shape in Experiment 1. The same strategy of assigning four target hand shapes to each CVC token was followed although different combinations were used. The surrounding cues were chosen randomly with the restriction that they had to be of different shape and appear in positions distinct from the target. Each surrounding cue was shown for 2 video frames (66 msec). The subjects only had to identify the target hand shape.

Again, after several sessions, we identified the most often confused hand shapes. In addition to the three hand images of Experiment 1 we colorized two more images using yellow and violet (see discussion below). When re-recording the stimuli we used this new set of hand images for the target as well as the surrounding images in the three cue sequence.

3.2. Results - Identification Scores

Five subjects participated in Experiment 1. Of these, three completed the entire Experiment 2. A fourth subject completed part of Experiment 2 - her scores are not reported here although they are consistent with the others.

Experiment 1 The experiment consisted of four sessions of uncolored (U) stimuli, followed by two sessions of colored (C) ones, two U, and two C. Table 2 shows the average consonant and hand shape

scores Results are given separately for each U and C session block chronologically.

We performed an ANOVA to evaluate the significance of the differences in scores. At the 0.01 level we found no significant changes in the consonant recognition between U and C stimuli or session blocks. On the other hand, the improvement in hand shape recognition when color was added was significant. There were also differences in scores across subjects but no significant differences in the pattern.

Experiment 2 In this experiment two U sessions were followed by two C sessions, one U, and one C. Table 3 shows the average scores. As in Experiment 1 ANOVA showed no significant effects of conditions, sessions, or subjects on consonant scores. On the other hand, all three variables had an effect significant at the 0.01 level on hand shape identification scores.

Taken together, experiments 1 and 2 indicate that selectively coloring the hand images may increase hand shape discrimination by the receivers. The improvement is not large but statistically significant. It is encouraging that consonant scores do not change when the hand shapes are colored.

3.3. Confusion Patterns

We analyzed confusion matrices obtained by pooling all subjects' responses over all sessions for a given condition (two matrices per experiment). The combined matrices were fit by a metric multidimensional scaling procedure described in [1]. For each confusion matrix, this procedure was used to derive the coordinates of a set of points that represent the means of two-dimensional Gaussian distributions that characterize the perceptual properties of the cues used to identify the handshapes. Distances between points are inversely related to the rate at which the handshapes were confused, with the unit of distance corresponding to the common standard deviation of the Gaussian distributions.

Experiment 1 The configurations derived from the confusions observed in the single-frame tests are shown in Fig. 4. For the uncolored handshapes confusions are generally determined by the number of extended fingers (e.g, 1, 2, 3, and 4; 5, 6, and 7) and by the extension of the thumb (e.g, 1 and 6, 2 and 7, & 4 and 5). The effect of coloration is to reduce the occurrence of confusions between handshape 3 (blue) and handshapes 2 and 4, as seen in the increased separation between the corresponding points. Similarly the increased separation between the points corresponding to handshape 7 and handshapes 5 and 6 indicates that confusions between these stimuli were reduced by coloration.

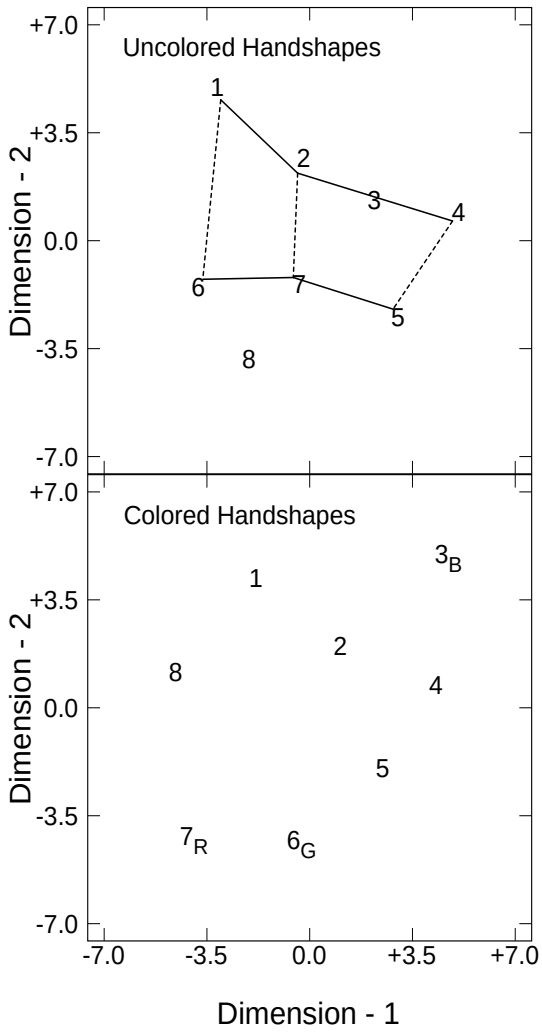

Figure 4. Means of two-dimensional distrbutions characterizing confusions among the eight hand shapes in Experiment 1. The solid curves in the upper panel connect hand shapes differing mainly in the number of outstreched fingers. The dashed lines connect hand shapes differing only by the presence of the thumb. Subscripts in the lower panel indicate the color assigned to the particular hand shape.

Experiment 2 The pattern of errors was similar to that of Experiment 1 and hand shape means showed configurations similar to those of Figure 4. The distances between points for these configurations were generally smaller than the corresponding distances in Fig. 4, consistent with the reduced identification accuracy in these tests. The similarity in the structure of the configurations for the uncolored handshapes in the single-frame and triple-cue tests suggests that subjects used roughly the same visual cues (number of fingers

extended and thumb extension) in both experiments.

The two experiments show that, with relatively little practice, color information can be integrated by human observers to improve discrimination of confusable hand shapes. The analysis of confusion matrices agrees with intuitive prediction of likely misidentifications and demonstrates how colorization can resolve many of these. We are planning to conduct experiments with artificially cued sentences to determine how well MCS users can be trained to make use of the colored cues in connected speech. [VIDEO AVSP17_4.MPG] shows the sentence, "It takes little sugar to sweeten grapes," artificially cued with hand images of Experiment 1.

4. DISCUSSION

In an early attempt at an "Autocuer" [4], a portable microprocessor-based device analyzed the acoustic input and used heuristic rules to identify speech sounds and assign them to cues. The cues were then coded as patterns of illuminated LED segments projected for the receiver onto his/her eyeglasses. No adjustments were made to correct for the time required to recognize the cue - the cues were always delayed relative to the start times of the corresponding phonemes. It did not prove possible to develop an effective system that worked in real time.

We believe that artificial cues resembling manual cues as closely as possible have a better chance of success. While improved cue recognition accuracy clearly increases the cues' benefit, results of Section 2.2 indicate a significant contribution by the characteristics of the display. The issue of cue synchronization to the talker's visible facial actions seems crucial and was not considered for the Autocuer.

The flexibility of our cue display offers potential improvements in cue reception not possible with either MCS or the Autocuer's display. Any changes in cue appearance designed to enhance their discriminability should retain the basic compatibility with MCS. Maintaining the similarity of the artificial cue display to natural MCS cues not only improves speech reception but greatly alleviates the problem of training. Users of our system trained in MCS needed minimal practice to understand the display.

5. CONCLUSIONS

We have successfully demonstrated a prototype automatic cueing system that provides a significant benefit to the cue receivers. Low-context keyword scores almost double relative to speechreading alone to 66%. The system's development suggests that appropriate cue display can be as important as the accuracy of the automatic cue recognition in determining the effectiveness of the system. In particular, we have found that the characteristics of cue transitions and synchronization to the visible facial actions of the speaker play a significant role.

Our preliminary investigation of cue discrimination suggests that cue receivers may commit hand shape identification errors that could be reduced by enhancing the contrast among synthetic cues. Selectively coloring the hand images improves their discrimination by subjects untrained in MCS. We plan to test the effect of colored cues on the reception of sentences by experienced MCS users. We also intend to refine and test the timing of the cue display.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] L. Braida. "Crossmodal integration in the identification of consonant segments," *Q. J. Expt. Psych.*, 43A(3):647–677, 1991.

[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] P. Duchnowski *et al.* "A Speechreading Aid Based on Phonetic ASR," *Proc. Int. Conf. Spoken Lang. Proc.*, Dec. 1998.

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

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

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

[9] 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.

[10] 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.