Paper presented to the1st International Conference on Technology for Helping People with Special Needs (ICTHP-2013). Al-Imam Mohammad Ibn Saud Islamic University, Riyadh, Kingdom of Saudi Arabia.

Facilitating Speech Understanding for Hearing-Challenged Perceivers in Face-to-Face Conversation and Spoken Presentations

Dominic W. Massaro[1,2] Michael M. Cohen[1,2], Walter Schwartz[2], Sam Vanderhyden[2], Heidi Meyer[1]

[1]Perceptual Science Laboratory, University of California, Santa Cruz, CA U.S.A.
[2]Psyentific Mind Inc, Santa Cruz, CA U.S.A.

massaro@ucsc.edu, mmcohen@ucsc.edu, walter.schwartz@gmail.com, sam.vanderhyden@gmail.com, heidicmeyer@gmail.com

Abstract

The goal of this project is to enhance the ability of hearing-challenged and deaf persons to understand conversational speech in face-to-face spoken interactions. The idea was to present certain robust phonetic properties of the incoming speech visually to supplement speech reading. We developed real-time digital signal processing of the speech and designed and trained artificial neural networks (ANNs) and Hidden Markov Models (HMMs) to learn these acoustic/phonetic properties. These properties were transformed into visual cues to supplement speechreading and whatever hearing was available. The three cues were presented as illuminations on three LEDs placed in the periphery of a lens on eyeglasses. The automated speech processing was reasonably accurate and perceivers learned to decode this information and integrate with information from the face. However, the cues could only be used deliberately for single words and phrases but not for continuous speech even with extensive practice. We subsequently turned to a new strategy that used automated speech recognition (ASR) to translate the interlocutor's speech into text. The user sees the interlocutor talk and then reads the text on a screen of a portable device such as an iPhone, iPod, or iPad. This solution, called Read What I Say, is now available in the Apple app store.

Index Terms: Hearing Impairment, Automated Speech Processing, Face-to-Face Communication, Hearing Impairment

1 Hearing Impairment

Our research and application project addresses the need for language aids for the millions of individuals who are deaf, hard-of-hearing, or have other language and speech challenges. For example, 36 million people in the U.S.A. alone live with hearing deficits and confront extraordinary difficulty participating in spoken interaction [1,2,3]. In Saudi Arabia, 13% of a sample of children ages 4-15 were hearing impaired and another 8% were were at risk of hearing impairment [4]. While many individuals rely on speechreading, cued speech, cochlear implants, or hearing aids to help them perceive spoken language, seldom do these solutions restore communication completely. Our goal is to develop another technology that can be used to enhance common face-to-face conversation. Importantly, the technology can be used as a supplement rather than simply replace other hearing and speech enhancements. The goal is to enable nearly complete understanding of face-to-face spoken conversation in a portable, low cost, and widely accessible device.

2 A Potential Solution

This research involved originally developing and testing embellished eyeglasses, which perform two simultaneous functions [5,6,7]. First, real-time acoustic analysis of an interlocutor's speech tracks several speech-relevant acoustic features: voicing, frication, and nasality. Second, these acoustic features are transformed into continuous visual cues displayed on small LEDs on the eyeglasses (see Figure 1). By integrating these visual cues with lip-reading (preferably called speechreading, because it involves more than just the lips), the user should gain much more understanding of the conversation.

Figure 1. Two illustrations of the iGlasses in use. The wearer sees the talker in central vision as well as the cues shown on LEDs in the periphery.

As shown in Figures 1 and 2, the iGlasses were envisioned to be worn as a regular pair of eyeglasses, but with two small microphones and three colored LEDs. The wearer looks at the interlocutor and the microphones deliver the interlocutor's speech to a portable processing device such as an iPhone, which processes the acoustic input. The input is analyzed for low frequency voicing information, high frequency frication energy, and nasal resonance that are associated with the acoustic/phonetic properties of voicing, frication, and nasality in English. The three properties are transformed in real-time into simple visual cues displayed on the three vertically mounted LEDs visible only to the wearer. These particular phonetic properties were chosen because they are fairly easy to track in the speech signal, and importantly, because they distinguish instances within a viseme category (a subset of phonemes that are highly confusable in speechreading). These cues also require no literacy, which is a benefit in that it widens the demographic to include pre-literate children and other non-readers.

The two branches of the research involved 1) the signal processing of the speech and the training of artificial neural networks (ANNs) and Hidden Markov Models (HMMs) to

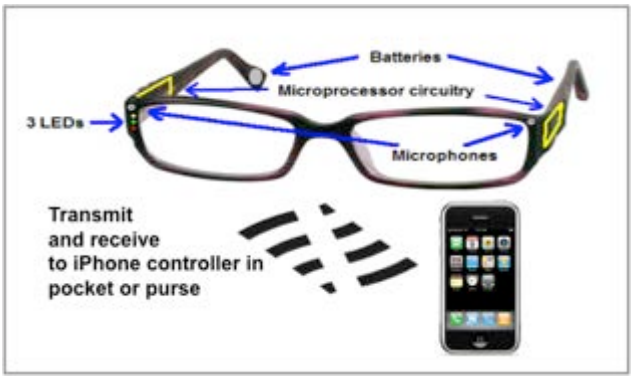

Figure 2. Schematic illustration of the iGlasses. Microphones record the talker's speech which receives real-time digital signal processing on the iPhone. Visual cues are presented on LEDs on the iGlasses.

accurately learn and track the acoustic/phonetic properties of the incoming speech and 2) how easily people can be taught to learn and use the cues in typical face-to-face settings.

2.1 Automated Speech Processing

A major milestone was the implementation and formative testing of the digital signal processing and Artificial Neural Networks (ANNs) on the iPhone. We learned that all of the iPhone and iPad models and some iPods have the computational resources to carry out the analyses in real time.

As in almost all automated speech processing systems, training on speech databases is necessary. Since there were no existing speech databases with labeled acoustic features, our team labeled two relevant databases, the Buckeye corpus and TIMIT [8,9], which were then used in the ANN training. Next, using the labeled corpora, hundreds of ANNs were trained to arrive at a configuration that met the constraints required by the real-time requirements of the intended application. The best ANNs could operate with less than a 40 ms delay, which is ideal for the real-time requirements of the iGlasses.

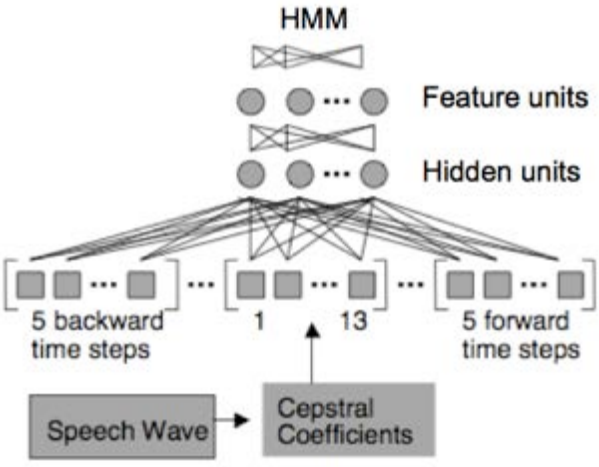

Figure 3. Illustration of the ANN/HMM used for digital processing of the incoming speech.

Figure 3 illustrates the ANN/HMM used for digital processing of the incoming speech. The ANN has a target window of 16 ms and five overlapping neighboring windows with a 8 ms overlap on each side. The Mel-Frequency Cepstral Coefficients (MFCCs) were computed on each of the five windows. Preliminary training and testing indicated that a network using just 12 MFCCs + log energy value, with 100 hidden units, gave roughly equivalent network performance relative to more complex ANNs. An independent ANN was used for each of the three features voicing, frication, and nasality, respectively. If all three of the output feature nodes are below a set threshold, the segment would be labeled as silence.

Given promising but not sufficiently accurate speech processing, we extended the signal processing in two ways. First, we programmed a Hidden Markov Model (HMM) to receive the outputs of the three ANNs (as is often the case in ASR systems). The input to the HMM was the activation of the output nodes of the ANNs corresponding to voicing, frication, and nasality, respectively. This gives 8 possible inputs to the HMM. The HMM also incorporated the transition probabilities between all pairs of the three features and silence as measured in our Buckeye and TIMIT data sets. Table 1 gives the transition probabilities among the 8 possible ANN output states.

Table 1. The transition probabilities among the 8 possible ANN output states. The rows give State n and the columns give State n +1. s = silence; v = voicing; f = frication; n = nasal

State n +1

State n	s	v	f	vf	n	vn	fn	vfn
s	0.91	0.03	0.05	0.01	0.00	0.00	0.00	0.00
v	0.04	0.90	0.02	0.02	0.00	0.03	0.00	0.00
f	0.05	0.12	0.80	0.03	0.00	0.00	0.00	0.00
vf	0.02	0.32	0.13	0.52	0.00	0.01	0.00	0.00
n	0.23	0.03	0.05	0.01	0.58	0.11	0.00	0.00
vn	0.03	0.18	0.01	0.01	0.00	0.77	0.00	0.00
nf	0.00	0.00	0.57	0.00	0.00	0.36	0.07	0.00
vfn	0.00	0.06	0.34	0.28	0.00	0.22	0.00	0.09

Second, in addition to having 3 separate networks for the three features, we programmed a three-out ANN with 8 possible output states, and used these possible states as input to the HMM. This ANN had 300 hidden units as opposed to just the 100 hidden units when a separate ANN was used for each of the three features. The best performance was given by the three-out ANN with 8 possible output states.

Table 2. The confusion matrix for the three-out ANN with 8 possible output states and with these possible states as input to the HMM. The numbers in the first column and first row correspond to the frequency of each state. s = silence; v = voicing; f = frication; n = nasal

Predicted State

State(#)	s(26823)	v(45591)	f(15794)	vf(49)	n(0)	vn(7504)	fn(0)	vfn(0)
s(28217)	0.86	0.04	0.09	0.00	0.00	0.01	0.00	0.00
v(45333)	0.03	0.89	0.02	0.00	0.00	0.06	0.00	0.00
f(12375)	0.08	0.04	0.88	0.00	0.00	0.00	0.00	0.00
vf(3833)	0.07	0.45	0.41	0.01	0.00	0.06	0.00	0.00
n(78)	0.12	0.22	0.01	0.00	0.00	0.65	0.00	0.00
vn(5895)	0.02	0.29	0.01	0.00	0.00	0.68	0.00	0.00
nf(8)	0.00	0.13	0.75	0.00	0.00	0.13	0.00	0.00
vfn(22)	0.14	0.32	0.14	0.00	0.00	0.41	0.00	0.00

Table 2 presents the confusion matrix for the three-out ANN with 8 possible output states and with these possible states as input to the HMM. As can be seen in the table, voicing plus nasality was accurately detected only about 68% of the time and misclassified as voicing about 29% of the time. Voicing was misclassified as voicing plus nasal about 6% of the time. Silence was mistakenly classified as frication about 9% or the time.

Voicing plus frication was accurately detected only about 1% of the time and usually misclassified as either just voicing or just frication. The entries for nasal alone, nasal plus frication, and voicing plus frication plus nasal are not an issue because in total they occurred only 0.1% of the time. Overall, reasonable performance was obtained with an average accuracy of .85, .89, .88, and .65 for silence, and for voicing, frication, and nasality, respectively, when the segment was labeled with just this feature.

2.2 Perceptual Learning of the Cues

A series of learning and test exercises was developed and given to subjects to learn to use the visual cues in speech processing. These exercises were implemented on the iPhone, iPod, and iPad. These devices are inexpensive and convenient and several of our participants had their own devices and could practice and test at their own discretion. We developed an application, BaldiExp, that allows different learn, test, and evaluation exercises and can be used and modified by our experimenters and participants without any programming experience. Initially, a set of learn and test exercises was aimed at whether reasonably trained participants could use the simlulated LED cues in combination with the face to adequately perceive the face-to-face speech with no sound. These perceptual learning sessions simulated a perfectly accurate speech processing so that the accuracy of performance could be directly evaluated. Figure 4 illustrates three successive views from the utterance of the word *fan* to show the occurrence of the appropriate cues during the three phonetic segments.

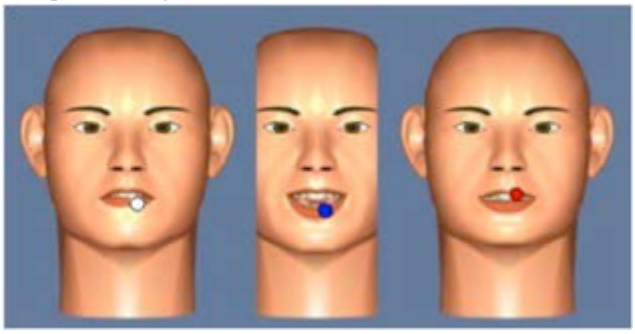

*Figure 4. Three successive views from the utterance of the word **fan** to show the occurrence of the appropriate cues during the three phonetic segments.*

Subjects: There were six participants. All subjects were young adults, approximately in their early twenties. All subjects were familiarized with the project and received sufficient background training on the cues and speechreading. All subjects received monetary compensation for their participation. (In retrospect, it might have been somewhat more successful if the subjects were given bonuses for improvement and very good performance.)

Materials: In all cases, Baldi, our computer-animated talking face developed to produce realistic speech, was used to simulate the face in the face-to-face scenario. Baldi and text to speech synthesis are used to eliminate the tedious content preparation for training and testing that would have been required if real faces were used. The exercises were performed on the iPhone, iPod Touch or iPad, using the BaldiExp application. We developed our exercise protocol on these devices because of the ease of programming exercises and their availability, low cost, responsiveness, and user-friendliness. A user manual is available and is constantly updated as enhancements are made to the program.

Procedure: Subjects were familiarized with the cues by performing a number of exercises. For several months beforehand, subjects progressed through a changing protocol using formative evaluation to optimize learning. The participants were exposed to a variety of different learning and test presentations in the form of syllables, words, and short phrases and sentences. The protocol consisted of each set of test items being made into a Learn, a Test, and an Evaluation exercise.

A learn exercise with test words consists of Baldi presenting each word in a test set by speaking aloud simultaneously with the cues, followed by Baldi mouthing the word with no sound simultaneously with cues. The subject pronounces the word in synchrony with Baldi's mouthing of the word. The idea is that this will connect the cues with the mouth movements in a more integrative way, which will increase proficiency in using cues.

A test exercise presents each test word with Baldi mouthing the test word simultaneously with the cues, followed by a response request. There are two types of response options: the first is a two alternative force choice or 2AFC, the second has a list of all possible alternatives in the test set. For 2AFC there will be two options to choose from: the correct answer, and an incorrect one. In the second task, subjects are given a list of the possible test items. Word lists are usually relatively short, between 10 and 30 words. In both types of tests, after the answer has been selected Baldi will say the correct word aloud simultaneously with cues. There is also the option to have the test word repeated once more with Baldi mouthing the word along with the cues as an additional reinforcement. When the test is completed there is a results screen that gives the score and the option of emailing the results for archival purposes.

An evaluation exercise randomly presents all words in a test set twice – once with the visual cues and once without the cues. This allows a direct comparison of performance when Baldi mouths the word with the presence of the cues to the condition when Baldi mouths the word with no cues. Response options are the same as for a test. We also evaluated the effect of the voicing cue being on during vowels compared to the voicing cue being off during vowels, and found no significant difference. Figure 5 shows Baldi on the iPad with a display of the voicing cue.

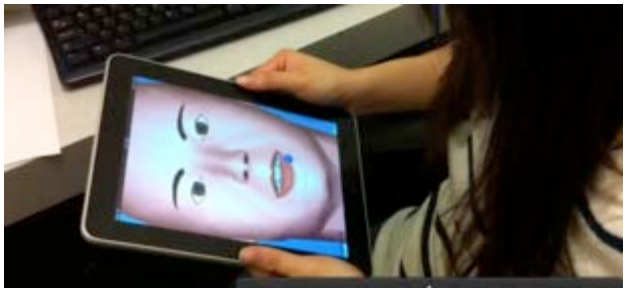

Figure 5. Baldi on the iPad with a display of the voicing cue.

Results: After 3 or 4 months of training and testing about 3-7 hours a week, subjects consistently performed better with than without the visual cues. This was true for 5 of the 6 subjects. The sixth subject spoke English as a second language and had great difficulty in the task overall. Table 3 gives the

average proportion correct as a function of blocks of training and testing for the two conditions.

Table 3. Average proportion correct for all six subjects in the initial learning and test exercises, with no cues or with cues across 8 blocks of trials.

Block:	1	2	3	4	5	6	7	8
Condition								
no cues	0.403	0.439	0.475	0.496	0.518	0.546	0.583	0.591
with cues	0.531	0.610	0.662	0.696	0.733	0.772	0.794	0.831

At the end of this training, subjects were tested in an expanded factorial design with three types of trials: face alone, cues alone, and both face and cues. As described next, quantitative model tests indicated that the subjects were able to integrate the face and visual cues, as described by the Fuzzy Logical Model of Perception [6].

2.3 Ability to Integrate the Face and Cues

Although there are several paradigms to study auditory-visual integration in speech perception, we employ a simple task in which test alternatives are presented with just the cues, with just the face, and both together (called a both condition).

Subjects: The same subjects just described in the initial learning test study participated.

Procedure: Subjects were tested on two lists of 42 one-syllable words each [6]. Test A included the words Path, Pang, Pad, Bath, Bang, Bad, San, Sad, Sat, Zan, Zad, Zat, Fan, Fad, Fat, Van, Vad, and Vat. The words in Test B were Pan, Pat, Pad, Ban, Bat, Bad, Man, Mat, Mad, Tap, Tab, Tam, Dap, Dab, Dam, Nap, Nab, and Nam. Each subject repeated the test ten times over the course of five weeks, for an hour a session. Most tests were done in the lab in a sectioned off room with few distractions.

Results. As can be seen in Table 4, the participants performed much more accurately given both the face and the cues than with either one alone. This synergy has been observed in previous studies of integration of the face and the voice, and the results indicate a true integration of the face and the cues. This outcome is a critical stage in proving the feasibility of the iGlasses innovation.

Table 4. Number of Sessions (#) and proportion of correct for the cue, the face, and for both the cue and the face for each of the 5 subjects for Texts A and B.

Subject	# Test A	Cue	Face	Both
E	13	0.33	0.23	0.91
H	11	0.39	0.23	0.97
K	14	0.52	0.43	0.77
S	4	0.29	0.25	0.96
T	10	0.39	0.19	0.97
Subject	# Test B	Cue	Face	Both
E	12	0.48	0.09	0.86
H	11	0.52	0.14	0.89
K	11	0.59	0.31	0.80
S	5	0.47	0.10	1.00
T	10	0.56	0.14	0.96

The question motivating this study was whether persons could integrate newly-learned cues about frication, voicing, and nasality with visible speech in speechreading. These results were well-described by the fuzzy logical model of perception [6], which predicts optimal or maximally efficient integration. This result reinforces our goal of developing technology to translate acoustic characteristics of speech into visual cues that can be used to supplement speechreading when hearing is limited.

These results were initially very promising, but even with additional training the subjects did not progress to a level of expertise that would be necessary for success in face-to-face communication. The subjects learned the cues explicitly and could deliberately integrate them with the visible speech from the face. However, they did not seem to be able to utilize them automatically as in everyday conversations. It is generally accepted that 5000-10000 hours are required to achieve expertise, and we soberly accepted that this time commitment was not feasible for our projected application.

3 An Alternative Solution

Our newest research results and relevant results from our continuing research showed:

1) Learning to use the three-LED system's cues proved difficult, even with training, especially for the older users who comprise the bulk of the hard of hearing population. Although most of our testers were able to learn the cues after many hours of practice and were able to use the cues in combination with the face, none of them developed an "automatic" response to interpreting the cues.

2) The LED cues plus speechreading would provide only about 50% accuracy in interpreting typical speech, according to our tests and research: 25% by speech reading and 25% by the LED cues. Our neural network output is only 85% accurate, increasing the difficulty for the listener to learn and interpret the LED signals.

3) We considered and implemented a new system to increase the potential accuracy to 90% by adding two LEDs to report on the "manner" of speech by indicating vowels, stop consonants, fricatives (as opposed to frication) and liquids, but 5 LEDs would have been more complex for users to interpret and to learn to use than just 3 LEDs. In addition, our speech processing algorithms proved to be much less accurate for these cues and difficult to deliver in real-time, as required by application. These three results indicated that the original proposal is not feasible.

One of the primary motivations for our initial approach was that we believed that full-blown ASR was not accurate enough to translate speech into text for the hearing-impaired listener [5]. Since that time, however, ASR has become much more accurate and we believe that appropriate employment of an ASR system can be accurate enough to translate spoken communication into written text that can be read by the hearing-impaired listener. The quality of the SIRI recognizer on the iPhone and the quality of the ASR on the Jibbigo speech-to-speech translation app on the iPhone [10] demonstrate that ASR should be accurate enough to aid hearing-impaired listeners.

Because of the dramatic improvement with ASR and the difficulties our perceivers had with learning the LED cues, we concluded that a different technology pairing, ASR with easily-readable text, would be a more effective solution to the problem. Since text, rather than simple LEDs, would be required for the output, simple eyeglasses with LEDs are no longer feasible for the output at this stage. The technology for presenting text on eyeglasses and the cost are prohibitive at the current time. Thus, our research direction changed to implement a different solution.

3.1 Read What I Say

Our initial product, called Read What I Say, is now for sale in the Apple App store (see Figure 6). The app is installed on an iPad, iPhone, or iPod which would be held or worn on a vest by a person talking to a hearing-impaired listener. This solution would also allow two hearing-impaired listeners to communicate. This solution requires no learning and simply requires the hearing-impaired or deaf listeners to be able to read. The app uses the ASR system of Mobile Technologies LLC [11] to carry out ASR. Their ASR system incorporates thousands of person hours and millions of dollars of research and application funding, which cannot be approximated by a new effort [12]. Most importantly, the ASR in our app [13] runs locally on the portable devices without access to a large server. This feature is essential for situations when an Internet connection is not available.

The goal of the Read What I Say [14] is to make it possible for all of us to engage in face-to-face conversations, even though we have limited hearing and/or the conversation is occurring in very noisy conditions. After the talker records a sentence or two, the words will occur on the iPad screen. For best accuracy, the talker should read clearly and with emphasis. She or he makes the recording by touching the screen just before and just after talking. The words will then appear on the iPad screen. The user should make sure that he or she has a good view of the screen for reading. Figure 6 illustrates the use of the application with a person with hearing aids.

Read What I Say [13], an iPhone app using ASR technology, is a proof-of-concept for facilitating face-to-face communication. The hard of hearing individual has his/her interlocutors talk into an iPhone or iPad, which will present the text of what they said. The user first sees the person talking and then sees the corresponding text so that both of these can be used together to help understand the message.

Figure 6. The use of the application Read What I Say with a person with hearing aids.

3.2 Alternatives to Read What I Say

As mentioned earlier, ASR has improved significantly in the last few years and there are surprisingly good systems available. SIRI on the iPhone, for example, gives very good automated speech recognition performance (even if the question-answering component is sometimes disappointing). Apple does not permit other applications to access SIRI, however, and therefore its ASR is not available. Our ASR system could be improved with access to the internet and we will soon implement this embellishment [11].

3.3 Alternatives to Read What I Say

Our current iPad application Read What I Say bypasses the problem of a poor speech signal from the talker because the talker has the iPad or iPhone device and can talk directly into it with a near microphone, or simply into the device's microphone. The written text from the ASR is shown on the same device and viewed by the hard-of-hearing person.

4 Current Goals

The iPad app serves as an intermediate goal of facilitating face-to-face communication in a seamless manner. There are three aspects of our research currently underway to improve the quality of this type of assistive technology. The first is to enhance the signal processing from remote microphones and the second is to present the recognized text on eyeglasses in the form of a head up display (HUD).

4.1 Enhanced Signal Processing on Remote Mics

The longer range goal is to implement the original idea of microphones mounted on eyeglasses and the text presented on eyeglasses. The hard-of-hearing person using the iGlasses will capture the speech from a remote distance, such as a few feet or yards, in face-to-face interactions. Thus, it is critical to get a good quality speech signal from the talker. The user can look directly at the talker and the signal processing can take advantage of having one microphone on each of the two temples of the glasses to better isolate the speech from the talker. This design also benefits for the use of the head as a shadow when irrelevant speech or noise comes from the periphery. Currently, we are researching and implementing techniques to do noise cancelation and beam forming and signal enhancement to provide a clean signal to the ASR system.

We are researching the implementation of Voice Tracker II™ Array Microphone, from Acoustic Magic [14], which has both the original implementation of their patented, automatic and electronic steering, "listening beam" technology plus a high-quality acoustic echo cancellation (AEC) algorithm. The second product is the Andrea Electronics' [15] system P-C1-1021450-100 Pure Audio® USB-SA with Free Array-2SA, which has an external digital sound card with patented noise reduction technology and SuperBeam® array microphone bundle. The third possibility, from Li Creative Technologies, Inc. [16], has a highly directional, USB plug and play array microphone that provides crisp, clear, noise-reduced speech. We plan to modify the best system so that it can be adapted to microphones on eyeglasses, as we originally planned.

We have just recently evaluated several microphone arrays with noise cancellation and beam forming software, and compared their performance with the built-in microphones on the

iPhone and iPad. Surprisingly, we could find no significant improvement in ASR performance with these enhanced systems relative to the default microphone(s). Apple appears to have created a very good input system that doesn't require embellishment for most applications. If our users are very close to the talkers they want to understand, it might be sufficient to use the iPhone's input rather than add on additional microphones.

4.2 Head Up Display (HUD)

The technology of the head up display (HUD) has also advanced considerably since we began our project. Lumus [17] has developed transparent eyeglasses that allow text to be superimposed on the wearer's normal view. They can also be easily placed over regular eyeglasses. Our microphones could therefore be easily placed on these eyeglasses and connected to

Figure 7. Schematic illustration of how a child with a HUD can read a transcription of what the mother is saying.

the iPhone. Given our ASR implemented on the iPhone, we can perform the automatic speech recognition and transmit the resulting text to these eyeglasses.

4.3 Digital Signage and Robots as Companions

We are also exploring the idea of presenting the text on a display in the user's environment. Digital signage is becoming more pervasive and the written linguistic information could be presented on various screens in the room or even outdoors. There is also a concerted effort to develop robots as companions [18]. In this case, the hearing challenged person would have by their side a robot with microphones, an ASR system, and a visual display. The user could then look to the robot for the written transcription of what is being spoken.

5 Summary

We have traveled a somewhat circuitous path in our quest to improve communication in face-to-face conversations. We began with an application design grounded in the belief that full blown automated speech recognition (ASR) was not fast enough or accurate enough to meet the goal of real-time communication. After just a few years into the project, ASR had improved significantly to make its employment feasible. More generally, assistive technology is a rapidly developing field and the goal of our research and applications is to extend the range of mind and behavior by using behavioral science principles to guide the use of the developing technology.

6 Acknowledgements

The research was supported by grants from the Center for Information Technology Research in the Interest of Society (CITRIS) and the National Science Foundation (SBIR awarded to Animated Speech Corporation, now TeachTown). The ASR system used in Read What I Say is from Jibbigo/Mobile Technologies, LLC (http://www.jibbigo.com/website/about-us/about-us). Thanks also to my dear friend Bill Rowe for the artistry in Figure 7.

7 References

[1] Kochkin S. (2005). MarkeTrak VII: Hearing loss population tops 31 million people. The Hearing Review. 12(7), 16-29.

[2] Kochkin, S. (2006). Better Hearing Institute MarkeTrak VII™ Semi-Annual Hearing Aid Market Survey, Better Hearing Institute, Alexandria, VA.

[3] NIDCD, National Institute on Deafness and Other Communication Disorders (2008), Heath Information: Statistics on Voice, Speech, and Language, Bethesda, MD http://www.nidcd.nih.gov/health/statistics/vsl.asp

[4] Jamal, T. S., Daghistani, K. J., Zakzouk, S. M. (2001). Speech Abnormality among Saudi Arabian Children With Hearing Impairment. Bahrain Medical Bulletin, Vol. 23, No.1, March 2001.

[5] Massaro, D.W., Carreira-Perpinan, M.A. & Merrill, D.J. (2010). iGlasses: An Automatic Wearable Speech Supplement in For Individuals' Speech Comprehension in Face-to-Face and Classroom Situations. In C. J. LaSasso, K. L. Crain, & Leybaert, J. (Eds.) Cued Speech and Cued Language for Deaf and Hard of Hearing Children (pp. 503-530). San Diego, CA: Plural Publishing Inc.

[6] Massaro, D. W., Cohen, M. M., Meyer, H., Stribling, T., & Sterling, C., & Vanderhyden, S. (2011a) Integration of Facial and Newly Learned Visual Cues in Speech Perception. American Journal of Psychology, 124, 341-354.

[7] Massaro, D. W., Cohen, M. M., Vanderhyden, S., Meyer, H., Stribling, T., & Sterling, C. (2011b). iGlasses: Improving Speech Understanding in Face-to-Face Communication and Classroom Situations. 26th Annual International Technology & Persons with Disabilities Conference. San Diego, CA, May 14-19, 2011

[8] Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd release) http://www.buckeyecorpus.osu.edu. Columbus, OH: Department of Psychology, Ohio State University (Distributor).

[9] Garofolo, J. S., et al. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia.

[10] Jibbigo (2011). http://www.jibbigo.com/website/about-us/about-us Accessed November 18, 2011.

[11] Mobile Technologies LLC (2011) http://www.jibbigo.com/website/about-us/about-us. Accessed November 18, 2011.

[12] Sphinx (2011). http://speech.cs.cmu.edu Accessed November 18, 2011.

[13] Read What I Say (2012). http://itunes.apple.com/si/app/read-what-i-say/id495703423?mt=8

[14] Acoustic Magic (2012). http://www.acousticmagic.com/

[15] Andrea Electronics (2012). http://www.andreaelectronics.com/

[16] Li Creative Technologies, Inc. (2012). http://www.licreative.com/products

[17] Lumus (2012). http://www.lumus-optical.com/index.php

[18] http://www.robotcompanions.eu/