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

iGlasses: Improving Speech Understanding in Face-to-Face Communication and Classroom Situations

Dominic W. Massaro, Michael M. Cohen, Sam Vanderhyden, Heidi Meyer, Tracy Stribling, & Cass Sterling

Department of Psychology
University of California, Santa Cruz
Santa Cruz, California 95064 U.S.A.
1-831-229-1666
1-831-459-2330
FAX 1-831-459-3519
massaro@ucsc.edu

Summary

There are 36 million people nationwide who are deaf or hard of hearing, and confront extraordinary difficulty participating in spoken interaction. We are developing and testing embellished eyeglasses, which perform real-time processing of several speech-relevant acoustic features, including voicing, frication, and nasality, and transform these acoustic features into visual cues displayed on LEDs on the eyeglasses. By integrating these visual cues with lipreading, the user gains nearly full perceptual access to the conversation. A series of learning and testing exercises have been developed for the iPhone, iPod, and iPad. With just several hours of practice, participants of different genders, ages, and hearing ability are able to learn to benefit from these features. Digital signal processing and Artificial Neural Networks (ANNs) are implemented on the iPhone to process the speech and to transform the acoustic features into visual cues. The iGlasses have great promise for making spoken language accessible to all individuals.

The need for language aids is pervasive in today's world. There are for example, 36 million people nationwide living with hearing deficits that confront extraordinary difficulty participating in spoken interaction (Kochkin et al., 2007). While many individuals rely on lipreading, cued speech, cochlear implants, or hearing aids to help them perceive spoken language, none of these restore communication completely or are not beneficial in all situations. The goal of this research is to develop technology to enhance common face-to-face conversation for the millions of individuals who are deaf, hard-of-hearing, or have other language and speech challenges. Importantly, our technology improves on previous speech aids by enabling nearly complete understanding of face-to-face spoken conversation in a widely accessible device.

Our research and product development involve embellished eyeglasses, which perform two simultaneous functions. The first function is real-time acoustic analysis of an interlocutor's speech that tracks several speech-relevant acoustic features: voicing, frication, and nasality. The second is that these acoustic features are transformed into continuous visual cues displayed on LEDs on the eyeglasses. By integrating these visual cues with lipreading, the user gains nearly full perceptual access to the conversation.

The iGlasses are worn as a regular pair of eyeglasses, but contain two small microphones and three colored LEDs (see Figure 1). The wearer looks at the interlocutor and the microphones deliver the interlocutor's speech to an iPhone, which processes the acoustic input. The input is analyzed for low frequency voicing information, high frequency frication energy, and nasal resonances that are associated with the acoustic/phonetic properties of voicing, frication, and nasality. The three properties are then transformed in real-time into simple visual cues displayed on the three vertically mounted LEDs, on the iGlasses. These particular phonetic properties were chosen because they are fairly easy to track in the speech signal, and more importantly, because they distinguish instances within a viseme category (i.e., subsets 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.

Given that we do not have a complete functional prototype of the iGlasses that would be available to all learners, we simulated the application by pairing visual cues with a computer-animated face, referred to as Baldi, which produces accurate visual speech (Massaro, 1998). We have shown that when the cues are learned, this learning can be transferred to a prototype of the iGlasses. A series of exercises to train the visual

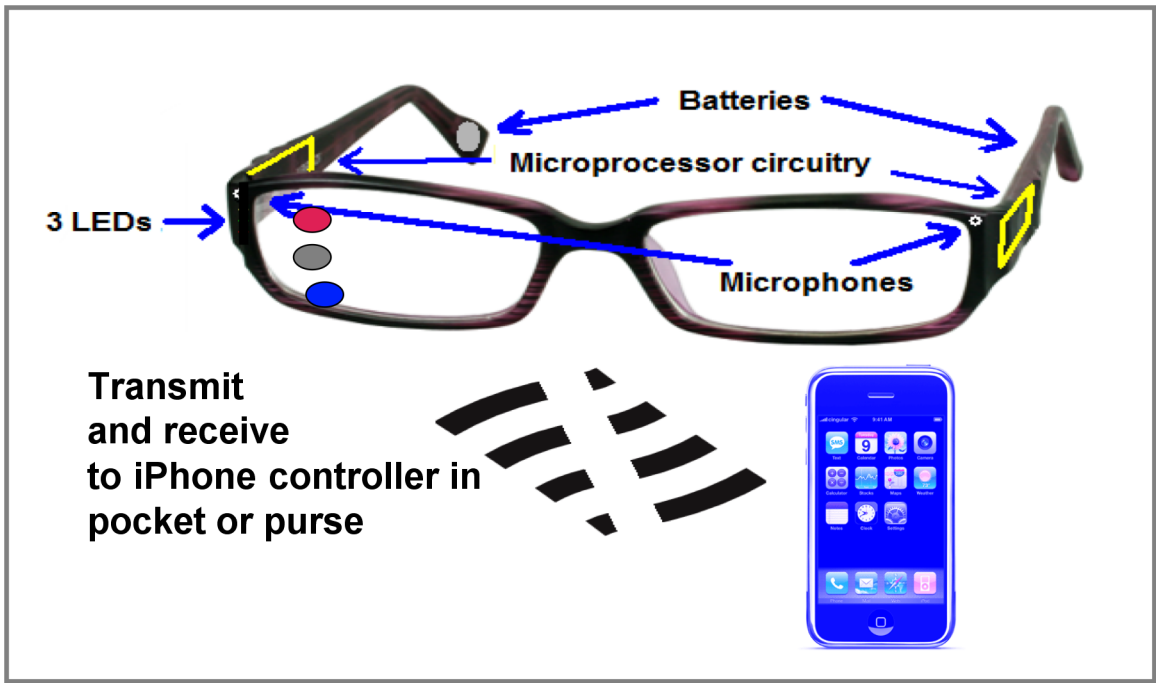

Figure 1. Prototype of the iGlasses setup. Microphones receive the speech input, which is transmitted to the iPhone, which carries out the signal analysis to determine the acoustic features that are then transformed to visual cues sent to LEDs on the iGlasses.

cues has been developed to be used on the iPhone, iPod, and iPad. These devices are inexpensive and can be always available so that participants can practice at their convenience. We developed an iPhone application, BaldiExp, that implements different learn, test, and evaluation exercises and can be used and modified by our experimenters and participants without any programming experience. Figure 2 shows an illustration of three frames of Baldi saying the word fan along with the visual cues. In the current implementation, the visual cue for nasality is signaled by a top red disk; frication is signaled by the middle white disk; and voicing is signaled by the bottom blue disk.

We created a lesson plan with detailed instructions and training trials to illustrate the value of these cues in supplementing speechreading. One engaging learning exercise begins with Baldi saying the word aloud and then mouthing it with the corresponding visual cues. The participant repeats the word aloud in synchrony with Baldi's mouthing of the words. This putative type of embodied participation should reinforce the appropriate learning of the cues. In the test exercises, Baldi mouths the word with the cues and participants have to indicate which word was said. Feedback allows the participants to track their performance and to continue learning the cues. The evaluation exercises are similar to the test exercises except that Baldi also mouths the

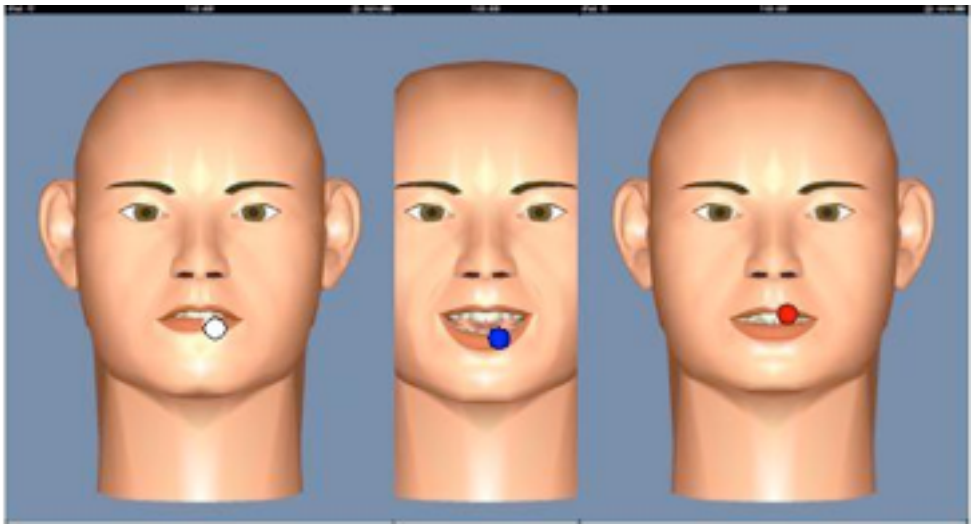

Figure 2. Illustration of three frames during the word fan. The white dot indicates frication during the /f/, the blue dot voicing during the vowel /ae/, and the red dot during the /n/.

test words without the cues. Differences in performance with and without the cues is a direct index of the participants learning and utilization of the visual cues.

The learning and test exercises begin with words but evolve through phrases and sentences as learning progresses. With just several hours of practice, participants of different genders, ages, and hearing ability were able to learn to benefit from the presence of the cues (Masaro et al., 2009). Their understanding of consonants, words, and sentences improved significantly with the iGlasses relative to the appropriate control condition. In addition, the results indicated that when the cues are learned with Baldi, this learning can be transferred to a prototype of the iGlasses.

We also implemented digital signal processing and Artificial Neural Networks (ANNs) on the iPhone. Since there were no existing speech databases with labeled acoustic features, we had to develop these databases ourselves. Our team successfully labeled two relevant databases, the Buckeye corpus and TIMIT, 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 real-time constraints so that the visual cues could be presented simultaneously with the facial information. We learned that the iPhone has the computational resources to carry out the analyses in real time. The ANNs appear to give fairly good performance when presented with novel speech on the iPhone. Our

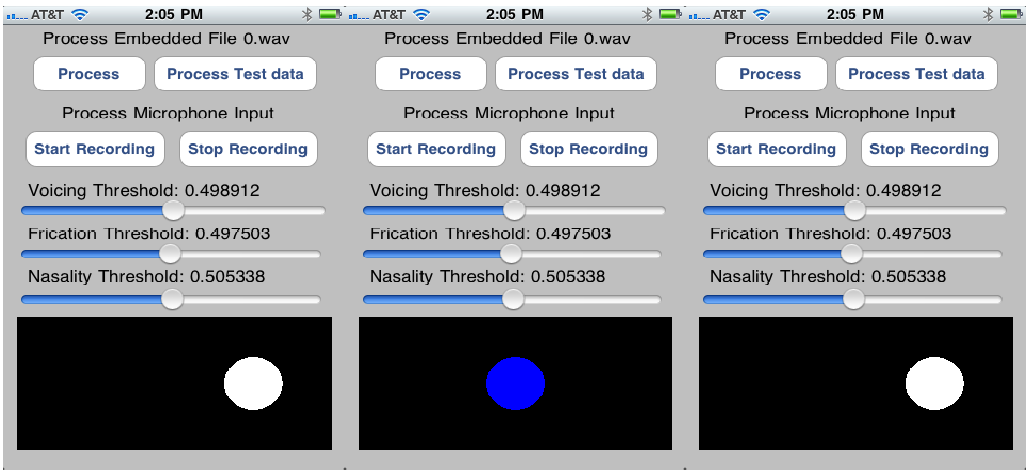

Figure 3. Three successive outputs of the trained ANNs for the word sash. The white output indicates frication and the blue output indicates voicing.

iPhone program processes the incoming speech and outputs the three feature values on the screen, as shown in Figure 3. Given this application and display, we serendipitously discovered another learning procedure. Speaking and watching the outputs of the visual cues, not only indicates whether the signal processing is accurate, it also reinforces the learning of the cues. The learner expects certain cues when she utters a word or phrase and immediately observes whether these observations are correct.

The iGlasses device holds much promise for a variety of reasons. It has been demonstrated that people naturally integrate several sources of speech information, and we have recently shown that they integrate the face and visual cues (Massaro et al., submitted). In addition, the proposed system does not replace auditory information with the supplementary cues but rather supplements the auditory speech that is normally available to the listener. The system we propose is naturally available to all individuals who can wear a pair of eyeglasses. The device does not require literate speakers because no written information is presented as would be the case in a captioning system. It is also age-independent in that it can be used by toddlers, adolescents, and throughout the life span. The phonetic basis for the speech driven cues should also reinforce an understanding of the phonology of the language. This transparency should also facilitate the learning of English by persons in the Deaf community. One of the major advantages of our proposed system is that it is language independent because all languages share the same fundamental acoustic characteristics. The iGlasses could also benefit persons with hearing aids and cochlear implants, which now provide significant help for many individuals.

References

Kochkin, S., Luxford, W., Northern, J.L., Mason, P., Tharpe, A.M., 2007. Are 1 Million Dependents with Hearing Loss in America Being Left Behind? Hearing Review, 14:10, (http://www.hearingaidtaxcredit.org/).

Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, Massachusetts: MIT Press.

Massaro, D.W., Carreira-Perpinan, M.A. & Merrill, D.J. (2009). Optimizing Visual Perception for an Automatic Wearable Speech Supplement in Face-to-Face Communication and Classroom Situations. Proceedings of the 42nd Hawaii International Conference on System Sciences-2009.