iGlasses: An Automatic Wearable Speech Supplement in Face-to-Face Communication and Classroom Situations

Dominic W. Massaro, Miguel Á. Carreira-Perpiñán, and David J. Merrill

Psychology, University of California, Santa Cruz

Electrical Engineering & Computer Science, University of California, Merced

Media Lab, Massachusetts Institute of Technology

Cite as: “Automatic Wearable Speech Supplement in Face-to-Face and Classroom Situations". Chapter 20 in Carol LaSasso, Jacqueline Leybaert, and Kelly Lamar Crain (eds.): "Cued Speech and Cued Language Development of Deaf and Hard of Hearing Children", Plural Publishing Inc., 2009.Send Correspondence to

Dr. Dominic W. Massaro

Department of Psychology

University of California

Santa Cruz, CA 95064 USA

work: 831-459-2330

FAX: 831-459-3519

email: massaro@ucsc.edu

URL: http://mambo.ucsc.edu/psl/dwm

Running Head: Wearable Speech Supplement

This book on Cued Speech is representative of a fairly recent paradigm shift in spoken language processing. Traditionally, speech was viewed as solely an auditory phenomenon. Research manipulating multiple sources of potential information, however, indicates that speech perception is most productively viewed as multimodal and sensitive to a variety of inputs in addition to the auditory speech input. This ability to exploit multiple modalities and multiple sources of information is a godsend to almost all individuals at some time in their lives. Although the auditory input alone is insufficient for adequate communication for many individuals and/or many situations, lipreading (also known as speechreading because it involves more than just the lips) allows individuals to perceive and understand oral language and even speak. Speechreading seldom disambiguates all of the spoken input, however, and other techniques have been used to create a richer input. Cued Speech, a solution to having limited auditory input that consists of hand gestures while speaking, provides the perceiver with information that disambiguates linguistic cues seen on the face. We have built on the conceptual framework underlying Cued Speech to design *iGlasses*, an automated wearable computer, to supplement face-to-face speech with additional information. Analogous to Cued Speech, this language aid can facilitate speech perception and language understanding for persons who have limited auditory input. Before addressing the needs for language aids and the challenges they provide, we summarize evidence for viewing speech perception as a pattern recognition problem involving multiple sources of information from multiple modalities.

/H1/ Multiple Sources of Information in Speech Perception.

One of the fundamental principles underlying the effectiveness of Cued Speech is that language perceivers easily learn to naturally integrate visual gesture cues with auditory and visible speech input. In contrast to this principle, speech science evolved as the study of a unimodal auditory channel of communication because speech was traditionally viewed as primarily auditory (e.g., Denes & Pinson, 1963). There is no doubt that the voice alone is usually adequate for understanding and, given the popularity of mobile phones, might be the most frequent medium of communication today. However, there are many deaf and hard-of-hearing individuals who require other sources of language input. The face is valuable as a source of language input even for hearing individuals because many environments in which communication occurs involve a noisy auditory channel, which degrades speech perception and recognition. Speech should be viewed as a multimodal phenomenon because the human face presents visual information during speaking that is critically important for effective communication. Experiments indicate that our perception and understanding of language are influenced by a speaker's face as well as the actual sound of speech (Bernstein, 2005; Massaro, 1987, 1998; Summerfield, 1987).

There are several reasons why the use of auditory and visual information in face-to-face interactions is so successful and why it holds so much promise for language communication (Massaro, 1998). These include a) the information value of visible speech, b) the robustness of visual speech, c) the complementarity of auditory and visual speech, and d) the optimal integration of these two sources of information. We will review evidence for each of these properties and begin by describing an experiment illustrating how facial information improves recognition and memory for linguistic input.

/H2/ Information Value of Visible Speech.

The value of visible speech is demonstrated by the results of a series of experiments in which 71 college students reported the words of sentences presented against a backdrop of noise (Jesse et al., 2000/2001). On some trials, only the acoustic sentence was presented (unimodal condition). On some other trials, the acoustic sentence was appropriately aligned with a highly realistic computer-animated face known as Baldi (bimodal condition). Baldi's presence facilitated performance for all participants. Performance accuracy more than doubled for participants

performing particularly poorly when given acoustic speech alone. Although a unimodal visual condition was not included in the experiment, based on previous research, we believe that participants would have performed much more poorly under such a condition than under the unimodal acoustic condition (Massaro, 2004). Thus, the combination of acoustic and visual speech is often described as synergistic because their combination can lead to a level of performance significantly higher than performance using either modality alone.

Similar results are found when noise-free speech is presented to persons with limited hearing (Erber, 1972). Adolescents and young adults who were either profoundly deaf or had severely-impaired hearing benefited more from face-to-face speech than they benefited from acoustic speech alone. Perceivers with severely impaired hearing (having a hearing loss between 75 and 90 dB) experienced the largest performance gain, exhibiting nearly perfect performance in the bimodal condition relative to either of the unimodal conditions (Massaro, 1998, p. 159; Massaro & Cohen, 1999).

/H2/ Robustness of Visual Speech.

Empirical findings indicate that the ability to obtain speech information from the face is robust; that is, perceivers are fairly good at speechreading in a broad range of viewing conditions. To obtain information from the face, the perceiver does not have to fixate directly on the talker's lips, but can be looking at other parts of the face or even somewhat away from the face (Smeele et al., 1998). Furthermore, accuracy is not dramatically reduced when the facial image is blurred (for example, because of poor vision); when the face is viewed from above, below, or in profile; or when there is a large distance between the talker and the viewer (Massaro, 1998; Munhall & Vatikiotis-Bateson, 2004; Munhall et al., 2004). These findings indicate that speechreading is highly functional in a variety of suboptimal situations. The robustness of visible speech is particularly important in the context of our research and development because perceivers combine speechread information with additional visual cues.

/H2/ Complementary Auditory and Visual Speech.

Complementary sources of information occur in circumstances where one source of information is most informative when the other source is weakest. In auditory/visual speech, two segments that are easily distinguished in one modality are relatively ambiguous in the other modality (Massaro & Cohen, 1999). For example, the difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. The complementary nature of two sources of information makes their combined use much more informative than if the two sources were redundant (Massaro, 1998, Chapter 14, pp. 424-427). In our application of these principles, our goal is to make linguistic information that is particularly difficult to see on the face visible.

/H2/ Optimal Integration of Sources of Information.

The final advantage afforded by having both auditory and visual sources of information is that perceivers tend to combine or integrate them in an optimally efficient manner (Massaro, 1987; Massaro & Cohen, 1999; Massaro & Stork, 1998). There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them such that both sources are used but the least ambiguous source has a greater influence on interpretation. Perceivers integrate information available from each modality efficiently according to the tenets described by the Fuzzy Logical Model of Perception (FLMP). This model of perception is mathematically equivalent to Bayes' theorem, a statistical method that is optimal for combining two sources of evidence to test among hypotheses (Massaro, 1998, Chapter 4). The FLMP assumes that visible and audible speech signals are each evaluated (independently of the other source) to determine how much they support each alternative. The integration process optimally combines these support values to determine how much their combination supports various alternatives. The perceptual outcome for the perceiver is a function of the relative degree of support among the competing alternatives. The most compelling evidence for the FLMP comes

from an important experimental manipulation that systematically varies the ambiguity of each source of information (Massaro, 1998).

We have found that, like adults, typically-developing children (Massaro, 1984; 1987; 1998), deaf and hard-of-hearing children (Massaro, 1999, 2004, 2006; Massaro & Cohen, 1999), and autistic children (Massaro & Bosseler, 2003; Williams et al., 2004) integrate information from both the face and the voice. This optimal integration occurs even if the auditory and visual speech are not perfectly synchronous (up to at least 100 ms), a finding that is critical to the requirements of our research application. Finally, analogous to the benefits found with Cued Speech, the results described below indicate that individuals can easily learn to integrate facial information with supplementary visual features of speech.

We now discuss the need for automatically supplementing spoken language with visual features of speech and how our approach to speech perception can motivate the development of technology to provide additional sources of information in language processing.

/H1/ Need for Language Supplements

There are millions of individuals with language and speech disabilities who require additional support for language comprehension and acquisition. In California alone, there are almost 200,000 deaf, hard-of-hearing, and speech/language impaired children who receive special education services (http://www.cde.ca.gov/re/pn/sm/index.asp). As an example of a specific need addressed by these services, most deaf and hard-of-hearing children have significant deficits in both spoken and written vocabulary knowledge (Breslaw et al., 1981; Holt, Traxler, & Allen, 1997). A similar situation exists for autistic children, who lag behind the expected trajectory in language acquisition (Tager-Flusberg, 2000). Currently, however, these needs are frequently underserved. One problem that the people with these disabilities face is that there are not enough qualified teachers, interpreters, and other professionals to give them the one-on-one attention that many of them need to successfully learn how to communicate using language.

Humans can learn and use language successfully without adequate auditory input. Sign language, a form of language without an auditory component, parallels spoken language in acquisition, use, and communication. Moreover, even oral language can subserve communication when auditory input is degraded or even absent. Lipreading (also known as speechreading because it involves more than just the lips) allows hearing-impaired individuals to perceive and understand oral language and even, in some cases, to speak (Bernstein, Demorest, and Tucker, 2000; Kisor, 1990; Mirrelles, 1947). Speechreading seldom completely disambiguates spoken input, however, and other techniques have been used to create a richer input. For example, Cued Speech, a solution to limited auditory input, consists of hand gestures made while speaking that provide the perceiver with information that can potentially disambiguate linguistic cues seen on the face. Very few people are proficient in Cued Speech or have the motivation to learn it, however, so many individuals with limited auditory speech input are faced with insufficient input in a variety of face-to-face and classroom-like environments.

Building on the value of Cued Speech and the innovative idea of Upton (1968), a solution Michael Cohen and Dominic Massaro proposed (Massaro, 1998) is to establish the technology required to design a device that would perform acoustic analysis of speech and transform several acoustic features into visual features that the speechreader would use in conjunction with visual cues from the speaker's face. This device would transform acoustic features associated with important linguistic information that is not directly observed on the face into visual cues intended to enhance intelligibility and ease of comprehension. We will review research that shows that people can learn to integrate these linguistic features with incomplete visual information to achieve enhanced language comprehension. Furthermore, similar to Cued Speech, the users of this device would have the advantage of gaining additional phonological awareness through the use of these linguistic features. We now discuss research that illustrates the value of providing additional visual cues to supplement auditory speech input.

/H1/ Supporting Research on Supplementing Visible Speech.

As illustrated throughout this book, Cued Speech has become an accepted form of communication for deaf and hard-of-hearing individuals. Cued Speech was designed as a means for supplementing lipreading by providing manual cues to phoneme identity to replace information not perceivable from the talker's face. Properties of Cued Speech include: 1) hand gestures that can be learned, 2) structure based on the phonemes of the spoken language, and 3) functionality at the earliest stages of language acquisition. One drawback to Cued Speech, however, is that both communicating parties need to know the system of cues for it to be effective. Although being deaf or hard of hearing and their family members and friends might be motivated to learn a system of cues, we cannot expect other individuals to be similarly motivated. Thus, a solution that does not require any specialized skills to supplement communication would be ideal.

Cornett's (1967) idea was to provide supplementary visual cues for language based on the realization that speechreading does not provide sufficient detail to distinguish all of the phonemes in a language, but only subsets of phonemes, such as /b/, /p/, /m/ versus /f/, /v/. Cued Speech hand gestures were therefore designed to denote different subsets of phonemes so that both subsets together would indicate a single phoneme. For example, when the hand gesture with the index finger extended that signals the subset of phonemes /d/, /p/, /zh/ is used in conjunction with the speechread phonemes /b/, /p/, /m/, the combination of cues denotes the phoneme /p/. The linguistic and psychophysical structure of Cued Speech categories is not ideal, however, which probably makes their learning and understanding more difficult than necessary. Meaningful categories, such as birds, fish, and chairs, share prominent perceptual and conceptual properties (Rosch & Lloyd, 1978). Thus, the supplementary feature solution we propose is more noticeably perceptually-based and conceptually-based than Cued Speech and also provides continuous information indicating the degree to which a feature is present.

In Upton's (1968) seminal automatic language feature-signaling device, relatively simple circuitry extracted three features from the acoustic signal: voicing, frication, and stop. These three simple features, plus two combination features: voiced fricative and voiced stop, were conveyed to the user via five tiny lamps cemented to the lens of a pair of glasses in order to appear near the mouth of the talker being viewed. After a "considerable" training period, the viewer (Upton) was able to use the transformed acoustic cues to disambiguate speechread information. Although Upton's initial paper includes only his subjective report, later papers (Pickett, Gengel, Quinn & Upton, 1974; Gengel, 1976) documented positive empirical results (using somewhat modified versions of the original device). Supporting research focusing on laryngeal, nasal, and total intensity feature information presented in the tactile modality (Miller, Engebretson & DeFillipo, 1974) as well as voicing and stop features presented in the visual and tactile modalities (Martony, 1974) was reported at about the same time.

Two other attempts have been made to design an automatic cueing system that accomplishes the same outcome as Cued Speech. Acoustic cues are extracted from the speech input and these are transformed into visual cues. The first of these attempts, which began in 1969, was termed the Autocuer (Cornett, 1977). The Autocuer consists of a pair of eyeglasses through which a virtual image of seven LED (Light Emitting Diode) elements is projected. Linguistic cues are presented to the user in visual form via patterns in the LED array. Evaluation of the Autocuer occurred without automatic extraction of the acoustic cues from the speech but rather ideal (hand-extracted) acoustic cues were used. The task was performed with normal-hearing as well as deaf and hard-of-hearing listeners. Recognition of isolated words was tested after a considerable training period consisting of 40 hours and was shown to increase from 63% with speechreading-only to 84% with speechreading plus use of the Autocuer (Cornett, 1977). . Later evaluations of the original Autocuer reported a less substantial increase (8%) when the extraction of the acoustic cues from the speech were generated by a real-time recognizer (Duchnowski et al., 2000). This weaker result was most likely due to poor accuracy (54%) of the cue recognizer.

The second of the attempts at an automatic cueing system used a sophisticated, speaker-dependent, speech recognizer to derive individual phones for conversion into a time-aligned set of cues for display (Duchnowski et al., 2000). The cues were designed to look identical to Cued Speech so that those already familiar with this system of communication would have no trouble interpreting its display. Keyword scores for low-context sentences increased by 31% over speechreading alone (Duchnowski et al., 2000) although the scores fell well below Cued Speech controls. (Keyword scores using Upton's device for 6 months yielded an increase of 18% relative to sentences in the speechreading alone condition.) Despite this achievement, Duchnowski et al. (2000) expressed doubt that a portable version of this device was feasible. One reason for this pessimistic conclusion is that speech recognition was used to generate the cues. Our approach bypasses all-inclusive speech recognition because automatic speech recognition (ASR) recognizes words, not acoustic features, and requires extensive computational resources, limiting the process to less than real-time performance. (Optimal performance occurs with at least a 3 GHz processor when a complete sentence is available. Successful systems carry out an acoustic signal analysis that evaluates about 60-90 spectral features (which are unrelated to linguistically-relevant features). All three of these limitations preclude our use of ASR because the requirements for our approach are the tracking of acoustic features at close to real-time performance and a lightweight portable device with limited computing power. Our proposed alternative is to create a device that is capable of quickly detecting a few robust acoustic features of speech, mapping them into visual cues, and conveying them to the viewer to facilitate speech interpretation.

To compensate for the delay required for all-inclusive speech recognition, Duchnowski et al. (unpublished) recorded a video of a talking face and replayed it to the listener in tandem with Cued Speech by a 2-second delay. This solution would be functional in a televised broadcast or played on a monitor such as a video iPod. However, Duchnowski's system would be impractical in face-to-face encounters whereas our envisioned system would be highly functional in most foreseeable applications.

In summary, widespread use of Cued Speech and research with visual cueing systems show that automatically supplementing speech with visual features is a worthwhile goal. Our current research is pursuing the development of a successful system of augmented communication that satisfies requirements such as light footprint for a wearable device, operation in near-real time, accurate tracking of acoustic speech features, learnable visual features representing auditory speech features, and ability of the user to integrate these features with auditory and visual features of speech.

/H1/ Research from the Perceptual Science Laboratory

Our research has investigated how to automatically supplement talking faces with information that is ordinarily conveyed by auditory means. This research consists of two areas of inquiry, which will be discussed in the next two sections: 1) developing a neural network to perform real-time analysis of selected acoustic features for visual display, and 2) determining how quickly participants can learn to use these selected cues and how much they benefit from them when combined with speechreading.

/H2/ Acoustic Feature Analysis.

The goal of feature analysis is to track certain acoustic features in real time and to transform them into continuous visual displays. We developed and trained a neural network to recognize three auditory speech characteristics: nasality, voicing, and frication. The training database was a sample of 23 words, containing 2,607 analysis frames from within the Bernstein & Eberhardt (1986) corpus. Each frame was 7.8 ms (Hanning windowed), and a new frame was sampled every 1.6 ms. The phoneme segments and their durations were determined using Verterbi alignment (a speech recognition algorithm when the phonemes are known). In previous research, we developed a computer-animated talking head trained on real speech to produce accurate synthetic speech with appropriate coarticulation (Massaro et al., 2005). To improve speech perception for hard-of-

hearing individuals, Massaro et al. (2007) patented a set of supplementary visible speech features, such as vibrating the throat to signify voicing, to provide additional information not seen on the face, and these features were shown to be effective in training speech perception and production in hard-of-hearing children (Massaro & Light, 2004). Baldi could now be aligned with the natural speech in the training database to give subphonemic features describing the moment-to-moment changes in voicing, frication, and nasality.

The neural net included 22 input units, 8 hidden-layer units, and 3 output units. A fast Fourier transform (FFT) computed the amount of energy in each of 20 Bark frequency bands (the Bark scale is nonlinear to match the properties of the peripheral auditory system). These measures, together with overall amplitude and number of zero-crossings, gave a 22-valued input vector. The feedback to the three output nodes were the subphonemic values computed in the alignment process. The weights on the connections among the units in the neural net were adjusted to minimize the differences between the actual and predicted features. Training gave a .057 root mean square deviation (RMSD) between the actual and predicted feature values on a 0-1 scale. To summarize, the neural net model was successfully trained to provide moment-by-moment outputs for the three features on the basis of acoustic input.

Thus, in principle, we have learned that we can use a network to transform the Bark scale energies from each speech frame into continuous visual features for presentation. Extensions of this work to a functional, effective, real-time system are described later in this chapter.

/H2/ Visual Feature Perception.

To provide a direct test of the perception of supplementary visual feature information, we used simulated rather than real-time analysis of acoustic features. We wished to see how difficult it would be for participants to learn to effectively use the visual features we had selected to supplement speechreading. A printed table mapping the relationships between phonemes and their corresponding visual features as well as phonetic and coarticulatory information was provided to participants. The following listing is an example of information provided in the table: vowels are voiced, fricatives have frication, frication can occur during the onset of stop consonants, and a nasal following a vowel can produce nasality during the vowel as well as during the nasal segment. In a five-day experiment, participants speechread 318 one-syllable words from the Bernstein & Eberhardt (1986) corpus that were presented visually. Visual speech was conveyed via the face of a humanoid speaker, presenting an image 13.7 deg horizontal and 20.4 deg vertical on a 30.5 cm diagonal screen 50 cm from the viewer. One group of 4 participants was presented with feature information along with this silent talking face, whereas a control group of 3 participants received only the silent talking face and no feature information.

For the feature group, three phonetic features (nasal, voiced, and fricative) were presented at the left side of the screen (centered 10.2 deg from face midline) in the form of intensity (saturation) of colored bars (5.1 deg horizontal by 2.0 deg vertical in size, spaced 2.9 deg apart vertically). Figure 1 gives an example of the display with the features. A series of trials is given on Band 14.8 in Massaro (1998), and is available online at http://mambo.ucsc.edu/psl/mmc/14_8.mov. It shows the

Insert Figure 1 about here

continuous nature of the colored features during the speech input. The top bar indicated the nasal sounds by lighting up orange during the period they occurred. The middle bar indicated voiced sounds by lighting up white when they occurred and remaining off when they did not. Two bars could light up at the same time, as during a voiced fricative, for example. Silence would be indicated when all three features are dark. The bottom bar, which corresponded to frication, lit up during the frication in fricatives and the burst/aspiration period in stop consonants. In all cases, the intensity of each of the three cues corresponded to the degree to which the corresponding acoustic feature was present in the speech signal. Speech cues were generated based on the phonetic labels of the acoustic speech as determined by Viterbi alignment (when knowledge of the

words was provided). This process will be described in more detail below. Participants responded by typing a word on a keyboard, which was followed by feedback during which the word was said again (with features for the feature group) with the sound on, and the word shown in print on the left side of the screen.

Several analyses were carried out, including accuracy of word identification; accuracy in identifying initial consonants, vowels, and final consonants; consonant and vowel confusions; and accuracy of feature identification for initial and final consonants. The left panel of Figure 2 shows the proportion of words correctly identified as a function of the five successive experimental blocks. Both groups improved with experience, but the feature group was significantly more accurate overall and improved faster. The center and right panels of Figure 2 show a d-prime measure of accuracy for identification of initial voicing and nasality for the two groups, respectively. (The d-prime measure is bias-free and is measured in z-scores. Note that these two panels have different scales based on the ranges of performance.) Relative to the control group, the feature group improved quickly by utilizing the supplementary visual feature information. It should be noted that word accuracy was still below perfect performance. This could mean either that the speechreading and features together were still insufficient to disambiguate the words or that the participants had not yet learned to use the information to achieve perfect word recognition.

Insert Figure 2 about here

Analysis of consonant confusions for the control and feature groups indicated that the feature group was able to make discriminations that the control group was not able to make. For example, within the category of labial stops (/b/, /m/, /p/), the feature group could discriminate between the three phonemes whereas the control group split their responses equally among the three alternatives. This experiment demonstrates that speechreading using visual features is learnable and greatly improves speechreading accuracy. However, it is necessary to determine if this positive outcome will also occur in more challenging situations, including scenarios with conversational speech and multiple speakers. In the next two sections, we outline plans for our ongoing and future work towards the goals of determining suitable acoustic features to extract from the speech, transforming them for presentation via the wearable supplement, and evaluating the prototype system.

/H2/ Acoustic Feature Analysis

/H3/ Evaluating Different Approaches to the Extraction of Features.

The proposed system is critically dependent on successfully extracting informative acoustic features from the speech signal, which several recent investigations have shown to be feasible. For example, the extraction of abstract phonetic features is apparently advantageous as a preliminary stage of automatic speech recognition (ASR) (King & Taylor, 2000). Typically, investigators have used TIMIT sentences with a preliminary Mel frequency cepstral coefficient (MFCC) analysis with a 25.6 ms window moving in 10-ms steps. This representation, sometimes with additional sets of first- and second-time derivatives (delta and double delta features), is then analyzed using neural nets (NNs), hidden Markov models (HMMs), support vector machines (SVMs), or decision trees. Chang, Greenberg, and Wester (2001), for instance, used a NN solution with separate networks for each phonetic feature with a mechanism for using only high-confidence results. Eide (2001) instead used a Gaussian-mixture (GM) approach. Frankel, Wester, and King (2004) compared dynamic GM models with NNs but found no advantage for either. Abu-Amer & Carson-Berndsen (2003) used independent HMMs. We now review the literature on the feasibility of extracting specific phonetic features.

/H3/ Extracting Voicing

Aioanei, Carson-Berndsen, and Kanokphara (2006) developed two phonetic feature-extraction engines for voicing and frication that were similar to two of the four acoustic features we used in our initial research. The output of their engine was the presence or absence of a feature. “The TIMIT training set was parameterized into 13 dimensional MFCCs, energy and their delta and

acceleration (39 length front-end parameters)." The feature models were trained from these parameters. Performance for voicing was very good, averaging about 90% accuracy.

/H3/ Extracting Frication

In contrast to the relatively accurate performance for voicing, the overall accuracy for frication was only about 37% correct. Lower accuracy was observed for frication because, although fricatives were usually recognized, the burst and aspiration phases of stops were also erroneously recognized as fricatives. Although this type of false alarm is problematic for categorizing phonetic features, it would not occur in our planned system, which extracts the *acoustic* characteristic of frication, not the *phonetic* feature of frication. Acoustic frication occurs in sub-segments of both fricatives and stops. Our supplementary visual feature for frication conveys the acoustic feature frication, not the phoneme fricative. Participants in our research learn that acoustic frication characterizes not only fricatives but also stop bursts and voiceless transitions. We found in our study of visual feature perception that perceivers benefited from this visual cue in identifying both stops and fricatives. That is, when perceivers learn the appropriate mapping, activation of acoustic feature frication during stop onsets is informative, as it is during fricatives. Supporting this analysis, when accuracy was recomputed for frication occurrence for both stops and fricatives in the Aioanei et al. (2006) study, feature extraction performance improved to over 80%.

/H3/ Extracting Nasality

Frankel et al. (2007) trained multi-layer perceptrons (MLPs) for feature classification on nearly 2,000 hours of telephone speech. A context window of 9 10-ms frames (central frame plus 4 frames each of left and right context) was used on the input layer for all MLPs. Given the 39 dimensional input feature, this amounts to 351 input units. The number of units on the output layer of each MLP corresponded to the number of levels of the feature group. For the nasality feature, the system achieved an accuracy of 90% correct. This study provides initial support for the feasibility of accurately tracking the acoustic feature nasality in our proposed system.

/H3/ Extracting Sonorant

Although we have not evaluated the sonorant feature, it could serve as a robust and valuable cue. Sonorant categorizes "+ sonorant" for vowels, semi-vowels, and nasals, as well as "- sonorant" for fricatives, stops, and non-speech. Shutte and Glass (2005) implemented a support vector machine to detect sonorant features in TIMIT sentences. For each utterance, 14 Mel-frequency cepstral coefficients (MFCCs) were computed every 10 ms over a 25.6 ms Hamming window, with cepstral-mean subtraction performed over each 500-ms window. After training, fairly accurate performance was obtained even when a moderate amount of noise was added to the speech. These representative studies add credibility to our proposed system, which would require robust extraction of three or four acoustic (rather than phonetic) features: voicing, frication, nasality, and sonorant.

/H3/ Discrete Features with Continuous Outputs

Our planned system would display an analog or continuous measure of each acoustic feature. This might seem problematic because our feature labeling of the acoustic input would be discrete (a feature would be either present or absent in a sub-segment). King and Taylor (2000), however, pointed out that while feedback data used to train the feature-detection mechanism (e.g. an NN) might be discrete, output features of the NN may be continuous in value. In our prior use of visual features for supplementing speechreading this type of continuous output is explicitly the case. Figure 3, a screen shot from our development system, shows an illustration of the mapping of the acoustic signal to the fricative indicator for the phrase "he then sat."

Insert Figure 3 about here

We generated the display intensity in the same way as other synthesis control parameters for a computer-animated talking face (Cohen & Massaro, 1993; Massaro, 1998, Chapter 12) using coarticulation blending functions with specific target values and co-articulation dominance functions for each phoneme. Although the occurrence of frication at each moment is coded as

either present or absent, the coarticulation blending function produces continuous outputs. For example, in Figure 3, the output of frication for /h/ is computed as having weaker frication than the output for /s/. In addition, the amount of frication changes dynamically at transitions between successive phonemes. Figure 3 also shows that short frication is displayed during the consonant release burst of /t/.

In evaluating competing approaches to the signal analysis, real-time performance is a critical factor because there should not be a substantial time lag between visual facial information and feature displays. Research in auditory/visual speech perception suggests that a short delay (e.g. up to 50-100 ms) would nonetheless be acceptable (Grant, 2002; Grant et al., 2004; Grant & Greenberg, 2001; Massaro, 1998, Chapter 3).

/H3/ Other Potentially Informative Features

The features voicing, frication, nasality, and sonorancy are derived from prior knowledge of acoustic phonetics and as such are likely to be highly effective for determining phoneme identity when combined with a facial display. However, these features may not necessarily be the most distinguishable of all possible features, and they may not be completely independent of one another and may be jointly redundant with the facial display. Furthermore, the number of visual features of interest should be kept to a minimum because of a limit in the number of visual cues that the glasses can accommodate (3), and because of a limit in the number of visual cues that participants may reasonably be able to learn to use. These criteria suggest the use of machine-learning techniques to determine optimal speech features to display visually, where we roughly define "optimal speech features" as the 3 functions of the acoustic vector that, when augmented with the face, most accurately classify acoustics into phonemes. While these features will likely lack a straightforward interpretation such as, for instance, acoustic frication, this lack of pre-existing categorical labels may not necessarily result in increased learning difficulty for participants.

There are specific methods to determine optimal visual speech features. One possibility is to extract informative features from each acoustic frame (13-D MFCCs) using dimensionality reduction methods. These methods seek, in an unsupervised way, a small number of continuous features that preserve information about the acoustic vector. Many methods exist in the machine learning and statistics literature, including linear methods (such as PCA, factor analysis and independent component analysis) and nonlinear methods (such as autoencoders, nonlinear latent variable models, Isomap, LLE, Laplacian eigenmaps and others) (Carreira-Perpiñán, 2001; Saul et al., 2007). For our task, which requires us to extract 3 latent features and to extract features from test data unseen during training, we plan to use a recently-introduced nonlinear method, the Laplacian Eigenmaps Latent Variable Model (LELVM; Carreira-Perpiñán & Lu, 2007). This method has the advantages of scaling well to latent spaces of three or more dimensions, having no local optima, and providing both continuous mappings (for dimensionality reduction and reconstruction) and probability densities (for the data and latent features); no other nonlinear method that we know of possesses all of thesebenefits.

A more direct approach that actively seeks features that maximize discriminant ability jointly with the face display is described as follows. Consider, for simplicity, visual features **W*x** that are a linear projection of the 13-D MFCCs **x** (where **W** is a matrix of weights) and assume we can extract a face feature vector **y** from the face image, for example, from the lip and tongue contours (note that these face features are used exclusively in an offline analysis to determine the most informative visual features, not in real-time use with the glasses). Now, we can use the joint vector (**W*x**,**y**) of linear visual features and face features as inputs to a neural-net classifier that maps its input to the corresponding phoneme. By minimizing the classification error over the neural-net weights and the linear projection weights **W**, we can find the best linear projection. In this case, "best" means optimal for classification, and the projection may differ from the PCA projection that we would obtain in an unsupervised training. Thus, our projection would be a form

of supervised dimensionality reduction. Clearly, we can also use a nonlinear projection (e.g., another neural net) of the acoustic vector **x** instead of a linear projection **W*x**. The classification error would also allow a comparison with the phonetic features of voicing, frication, nasality, and sonorant.

/H3/ Labeling New Data Sets

The use of acoustic rather than phonetic features requires new labeling of existing speech corpora, which are phontically but not acoustically labeled. Two feasible candidates are the well-known TIMIT Acoustic-Phonetic Continuous Speech Corpus, which contains a total of 6,300 sentences, consisting of 10 sentences spoken by 630 speakers from 8 major dialect regions of the United States (DARPA TIMIT), and a Buckeye database of face-to-face spontaneous speech (Pitt et al., 2007). Forty speakers, all natives of Central Ohio, contributed about 300,000 words of spontaneous speech in interviews in which they expressed their opinions in conversation. The speech of central Ohioans is fairly representative of American speech without the accents of the northeast or southeast.

Both of these databases have been orthographically transcribed and phonetically aligned, but this markup is not adequate for the training of acoustic features. Given their large size, only a subset of each database is necessary to mark up the presence or absence of acoustic features. The existing databases contain waveforms and spectra, phoneme labels and boundaries, and written transcriptions. Each frame of the waveform can be labeled as plus (1) or minus (0) on, for example, the four acoustic features of interest: voicing, frication, nasality, and sonorant. This markup can be aided by both the existing phoneme markup and acoustic properties, which will be evaluated by both listening and visual inspection. These analyses can be carried out using readily-available applications such as Wavesurfer (Wavesurfer) or Pratt (Pratt). To obtain a sufficiently thorough database for neural network training and testing, it will probably be necessary to mark up about 24 sentences sampled from 12 speakers for a total of 288 sentences from the TIMIT Corpus and about 2 minutes sampled from 12 speakers for a total of 24 minutes of spontaneous speech in the Buckeye corpus.

/H3/ Training and Testing Neural Networks

We will train and test neural networks on the corpora in order to determine whether the acoustic features of interest can be accurately tracked by a neural network and to determine which acoustic features give the most accurate performance. We will use feed-forward neural networks with a single layer of hidden units, which can approximate most useful functions to a high degree of precision when a sufficient number of hidden units are used. Several different configurations of the acoustic input and the number of hidden units will be used to converge on a successful representation. As an example, there would be 9 input frames, consisting of a center frame and four frames preceding and following the center frame, each corresponding to 10 ms of speech. For each input frame, the neural net would have 22 input units, 8 hidden-layer units, and 4 output units. The amount of energy in each of 20 Bark frequency bands combined with overall amplitude and number of zero-crossings would yield a 22-valued input vector. The target value for the four output nodes would be the subphonemic value computed in the proposed alignment process. The networks will be trained using backpropagation to minimize prediction error on the training set and weight decay to improve generalization. The best network architecture (i.e., the number of hidden units and the number of frames in the input window) will be determined by cross-validation. Training and test data will come, for example, from the TIMIT database (e.g. 12 sentences sampled from 12 speakers for a total of 144 sentences). Analogous training regimes will be employed for the conversation database.

/H3/ Remote Recordings of the Speech

Most ASR systems permit the talker to speak into a lapel microphone or a telephone during recording. To succeed, our envisioned system requires the recording of acoustic speech in face-to-face conversations by a microphone worn by the interlocutor. The biggest sources of potential

error using a remote microphone located on the listener include background noise and room reverberation during the conversation. Optimum training performance should occur when the system is trained on remote recordings of acoustic speech. In these sessions, signal processing will be based on live presentations of the acoustic speech at distances resembling those in face-to-face conversation rather than the original digital representation of the speech database..

/H3/ Goal of Neural Network Training

Our goal is to test the feasibility of our proposed feature analysis. Given the previous successes in phonetic feature analysis reviewed above and in our research, we expect this aspect of our work to be successful. Apart from the trained neural networks themselves (needed for real-time computation of the features), another outcome will be a direct measure of how accurately each of the four acoustic features of interest can be tracked by a neural network. This information will (in part) partially indicate which visual features would most likely be effective in representing the input in face-to-face communication.

/H2/ Visual Feature Perception

/H3/ Participant Population.

Our research concerning the development of a speech facilitation device would benefit greatly from a design in which participants recruited for testing have a vested interest in the success of the project. Thus, participants, including both hearing and hard-of-hearing persons, should have the goal of enhancing their communication interactions. Similarly, persons skilled in Cued Speech would likely be interested in the project, and their performance will illuminate whether proficiency in Cued Speech facilitates or inhibits speech perception with supplementary visual features.

/H3/ Use Neural Network Outcome to Choose Visual Features.

The outcome of the neural network experiments will provide direct measures of how accurately each of the four acoustic features of interest can be tracked by a neural network. We will use this information to choose the three visual features to be used in these experiments. Ideally, all four visual features might be tested and compared with combinations of three features. Given that each experiment requires a significant amount of learning, however, four independent experiments (i.e., the 4 combinations ABC, ABD, ACD, BCD of 4 acoustic features A, B, C, D) would be too time-consuming. Furthermore, if one of the acoustic features proves to be too difficult to track accurately, it would not be functional and could even be disruptive for performance. For these reasons, the design of the experiments on visual feature processing will be contingent on the outcome of the neural network experiments.

/H3/ Extend Presentation of the Visual Features to LED Eyeglasses

Our research indicates that perceivers were able to use supplementary visual features presented in the periphery to improve speech perception while still attending to the speaker's lips. Performance improved significantly with about 40 minutes of practice per day across 5 days of training. This improvement might have been due to the presentation of facial as well as visual feature information on the same computer display monitor. It is therefore important to replicate this study in a situation that more accurately approximates our envisioned real-world application. In this study, we will replicate our initial experiment so that the participants look through the instrumented eyeglasses to see the talking face on a computer monitor, with visual features displayed on LED's in their peripheral vision. The supplementary feature displays will be computer-generated and their output will be displayed on the eyeglass-mounted LED's.

It will be necessary to build and configure eyeglass frames to hold the LED display. Figure 4 shows a picture of our first mockup of the specialized eyeglasses. The display is mounted at the periphery of each lens. We will evaluate whether the LED display can be processed adequately or whether it must be moved to a more central position. It may also be necessary to adjust the forward-backward location of the display for some users in order to maximize comfort and view-ability. We will

Insert Figure 4 about here

assess whether it is valuable to mount LED displays on both sides or whether a one-sided LED display is sufficient. Eyeglasses with LED's on one side only would significantly streamline the design of the device. If a single-sided display is sufficient, the side of the eyeglasses that holds the display may be an important variable because it will determine whether the visual features will be seen to the right or left of the talking face. There is a substantial amount of literature on hemifield effects in visual perception and language processing (Smeele et al., 1998), and it would be advantageous to choose the side that leads to most accurate performance on which to mount the LED display. Thus, we will systematically vary whether the LED display is shown on the left side of the left lens or on the right side of the right lens. We expect that visual cues presented via the LED display will improve speech perception, as it did in our previous research. If the LED display does not lead to improved speech perception, we will explore the differences between these two situations in order to better design an effective wearable display.

Figure 4 shows that the microphone, the processor, the battery, and the LED display can be placed together on one side of the eyeglasses. We have also mounted a small vibrator because we will also consider and evaluate the possibility of transforming one of the acoustic features, such as sonorant, into a slight vibration of the frame of the eyeglasses.

/H3/ Evaluate Speechreading and Feature Perception with Sentences and Conversational Speech

Our research has indicated that supplementary visual features positively influenced the perception and recognition of words presented in isolation. Natural dialog provides continuous speech; therefore, we will test the functionality of our system with sentences. This extension is important because with isolated words, participants may be able to employ perceptual strategies with words presented in isolation that are difficult with longer verbal stimuli, such as sentences. Martony (1974) reports a supplementary feature advantage for participants' recognition of whole sentences that were previously trained with a closed response set method. For sentences not previously trained, however, they showed only a small non-significant advantage over unaided speechreading. Therefore, it is important to determine whether supplementary visual features contribute positively to speech perception for sentences from multiple speakers in our application.

The test materials will consist of 144 sentences from the TIMIT database. These sentences will be presented both with and without supplementary features. To measure learning, participants will be asked to type as much of each test sentence as they can. The results will be analyzed in terms of the number of words correctly reported under the two conditions. With the exception of continuous sentences as opposed to isolated words, all other procedural details and data analyses will be similar to our initial pilot study described above. The same evaluation task will be extended to include perception of speech from the corpus of conversational speech.

An important aspect of these studies involves learning. Our research indicates that participants learn to take advantage of supplementary visual features. Even though learning occurred, there are alternative learning regimens that may potentially increase the rate and asymptote of learning. One possible training situation is to practice a single supplementary visual feature at a given time. After one feature has been successfully learned, the presentation could be made more challenging by adding a second feature and then a third, in like manner. Another technique to facilitate learning would be to present a practice period on each feature directly rather than in the context of words and sentences. In this case, the test materials would consist of simple consonant-vowel syllables so that participants will be able to focus on the visual features for a single consonant. Another possibility is to train participants on the visual features representing speech sounds without the face present so that they can directly learn the supplementary features of interest without interference. If we observe that the benefit from the supplementary features is acquired slowly in our standard testing paradigm, we will experiment with optimizing the learning process by instantiating some or all of these potential learning aids.

/H3/ Combine Visual Feature Perception and Acoustic Feature Analysis in Real Time

After the research on visual feature perception and acoustic feature analysis in real time has been completed, an experiment will be carried out to test the two paradigms in combination. In this experiment, acoustic feature analysis and visual feature presentation will be combined during evaluation testing. To implement the experiment, it will be necessary to add a microphone and a wearable processor to the eyeglasses with the LED display. The microphone will capture the speaker's conversational surroundings and an analog audio circuit will transmit the acoustic signal to the processor where it will be digitized. Given that the neural networks have already been trained off-line, it is only necessary to program the processor with any necessary signal processing for the sampled digital signal, the feed-forward network with the learned weights from the learning phase, and the output generation algorithm to drive the LED display. Running on the processor, the trained neural networks will process the digitized speech sounds in real time and directly control the LED display on the eyeglasses. All other procedural details will follow that given in the previous tests with sentences and conversational speech.

/H3/ Evaluation Summary

Normally, there are three sequential activities that can evaluate the quality of pedagogy and technology: 1) exploratory research, 2) formative evaluation, and 3) summative evaluation. Our exploratory research has already been carried out along with specified procedures for formative evaluation: accuracy of acoustic feature analysis and speech perception benefit of supplementary visual features. The summative evaluation will assess the effectiveness of these two components in the context of a complete system of supplementing talking faces with visual features.

/H3/ Extensions to Real World Performance

Given a successful outcome, selected participants will wear the eyeglasses throughout a typical day. Participants will include deaf and hard-of-hearing persons, who have a vested interest in enhancing their communication interactions. These participants will be monitored to determine how functional the system is in typical interlocutor situations. Based on these observations and participant reports, we can determine the positive and negative aspects of the system. Modifications can be made, if necessary, to improve the system and the resulting quality of the face-to-face conversations.

/H1/ Significance and Concomitant Advantages of the Proposed System

The technology we are developing would be ideally designed for wearable computing, so a person could have face-to-face conversations while wearing a pair of glasses that could also be fitted with the wearer's normal eye prescription. The wearable product would process primitive characteristics of the speech signal, such as voicing (the presence of energy at the fundamental frequency, such as in vowel sounds); frication (high-frequency noise similar to energy that is characteristic of various consonants such as [s], [z], and [sh]); and nasality (a unique resonance characteristic, as in [m], [n], and [ng]). These characteristics would be tracked in near-real time, and the output displayed on the LED display of the glasses (Costanza et al., 2006).

Our envisioned system holds promise because it does not replace auditory information with supplementary cues, but rather supplements auditory speech that is normally available to the listener. People naturally integrate auditory and visual information, so they should benefit from the availability of both visible and audible speech. In addition, this strategy is particularly effective because of the complementary nature of auditory and visual speech. Acoustic speech that is robust in signal and fairly easy to recognize consists of speech cues that are not readily perceivable from the face in visual form. This disparity in quality enhances recognition of the speech signal through simultaneous presentation of cues in both visual and acoustic form so that cues that are ambiguous in one modality are complemented by robust cues in the other modality.

The proposed technology qualifies as a transparent information appliance that adds to the listener's perceptual and cognitive resources (Clark, 2003; Norman, 1999; Weiser, 1991). We have developed an analysis of requirements, a conceptual design, and possible physical designs for this device. It consists of an affordable, non-invasive device that is seamlessly integrated with normal

dress, adding only a pair of glasses (which might be necessary regardless). This qualifies as an augmented-reality device that is available for use 24 hours per day, 7 days per week and requires very little maintenance.

/H2/ Usability for All Individuals.

The system we propose would be available to all individuals who can wear a pair of eyeglasses. The device does not require that speakers be literate 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 throughout the life span. Young children are able to learn sign language and even finger spelling of the spoken language. Therefore, young children should also be able to use the proposed supplementary speech cues. The phonetic basis for speech-based cues should also reinforce an understanding of the phonology of the language (Morais & Kolinsky, 1994). Studies have shown that deaf and hard-of-hearing children who have mastered Cued Speech have internalized much of the phonology of their language and learn to read naturally. Thus, with our system, we expect that children will learn vocabulary and grammar and will gain meta-awareness of the structure of the community's spoken language.

/H2/ Available to All Language Groups.

One of the major advantages of our envisioned system of communication is that it is language-independent because all languages share the same fundamental acoustic characteristics. Other non-automated systems such as Cued Speech and sign language are language-dependent. Thus, all language groups can use the proposed system without compromising their normal language processing in other domains, such as sign or Cued Speech conversations. The device would be optimally functional when the listener is faced with a person who is proficient in oral language but is not proficient in Cued Speech.

/H2/ Significant Help for People with Hearing Aids and Cochlear Implants.

There have been substantial improvements in the technology of hearing aids and cochlear implants, which now provide significant help for many individuals. However, these persons remain at a disadvantage in many natural environments, such as those with background noise and reverberation, and in many fast-paced and/or dense conversations. The technology we propose will provide an additional supplement to speechreading that will allow communication in these situations.

/H2/ Extended Reach of the Research.

The benefits of this research extend beyond the hearing-impaired community. There are many individuals, including autistic children and persons recovering from brain trauma, who have difficulty processing acoustic speech. Many of them successfully communicate by alternative communication methods. Our research will improve the state of the art in transforming acoustic speech into other forms on information, offering a greater number of potential communication methods for these individuals.

/H2/ Benefits to Pedagogy of Reading:

It is well-known that there are numerous irregularities that a number of children encounter in learning to read and spell. Children who have a substantially greater amount of difficulty in reading and spelling than would be expected based on their age and perceptual and cognitive abilities are labeled as dyslexic (Fleming, 1984; Willows, Kruk, & Crocos, 1993). Psychological science has established a close relationship between the mastery of written language and children's ability to process spoken language (de Gelder & Morais, 1995; Morais & Kolinsky, 1994; Taylor & Olson, 1994). That is, it appears that many dyslexic children also have deficits in spoken language perception. This difficulty with spoken language can be alleviated through improving children's perception of phonological distinctions in spoken language, which in turn improves their ability to read and spell (National Reading Panel, 2000). Experience with the wearable system could help these children gain insights into the spoken language and therefore improve their reading skill.

/H1/ Potential Limitations of the Proposed System

A potential limitation of our system is that some non-visible acoustic features of vowels cannot be mapped into visible features to help disambiguate the spoken message. Cuing vowels would obviously provide more potential information for the listener. As in all applications, however, there are trade-offs that must be considered. Cuing vowels could potentially entail a number of negative effects. First, recognition of vowels or vowel features from the waveform would be highly fallible relative to the features of consonants analyzed by our system. Second, there is a limit on the number of features that the listener can process in parallel with the audible and visible speech input. Adding several vowel features would probably exceed that limit. Third, vowels carry less a priori information than consonants in English. Thus, vowels are more predictable in word contexts than consonants. Fourth, for people with partial hearing due to hearing loss, noise, or cochlear implants, vowels appear to be less perceptually degraded and therefore more intelligible than consonants. In this case, the listener will probably benefit more from the acoustic signal for vowels than for consonants, making any visible feature for vowels less informative. Fifth, visible speech from the speaker is much more informative for vowels than it is for features of consonants such asvoicing, frication, and nasality, which are analyzed by the proposed system. For this reason, vowel information is perceived fairly accurately from the face alone. Montgomery and Jackson (1983) and Massaro (1998) found about 75% correct lipreading among 8 vowel categories. Our research will determine whether consonant cues are sufficient for accurate performance given a reasonable amount of training. If so, the case can be made that a robust system of augmented communication can be implemented even though no additional supplementary cues are provided for vowels.

It could be argued that automatic speech recognition (ASR) by machine will improve sufficiently in the near future so that a full captioning of speech could be accurately rendered in real time. Although this significant breakthrough has always been possible, it seems unlikely that it will occur in the near future. ASR can be expected to function reasonably well as long as vocabulary and grammatical structure input is limited and the system is speaker dependent—that is, trained on a single speaker and/or used in a completely noise-free environment. We have also described how ASR functions with less than real-time performance and extensive computational resources. These constraints exist because most ASR systems do a poor job of recognizing phonemes (Greenberg, 2006; Greenberg & Chang, 2000), using sophisticated word models (usually bi- or tri-gram models) to deduce words from a flawed recognition of the phonemes. Our device, by contrast, will be functional in the natural setting of open dialogs and conversations from multiple speakers. Most importantly, however, our approach has five important advantages: 1) it supplements rather than replaces the acoustic signal, 2) it can be carried out in real time, 3) it requires relatively few computational resources, 4) it conveys a continuous analysis rather than a discrete categorization of the speech input, and 5) it is language independent because the acoustic features that will be analyzed should vary relatively little across languages.

Most ASR systems permit the talker to speak into a lapel microphone or a telephone during recording. In the present system, however, the microphone embedded in the eyeglasses necessitates a remote recording of acoustic speech. The most likely sources of potential error using a remote microphone on the listener include background noise and room reverberation in the location of verbal exchange and the speech of others who are not in the conversation. In addition to the distance of the microphone from the speaker, the speech signal would also be somewhat variable because the distances and directions of speech will vary in typical face-to-face conversations. This challenge is anticipated by training the system on remote recordings, which should also have less impact on the tracking of acoustic features relative to speech recognition. Techniques are available to adjust for these sources of degradation of the acoustic spectrum. By training our neural-net acoustic-feature recognition system on remote recordings, the number of possible sources of degradation will be reduced.

Regardless of the advances or lack of advances in speech-recognition technology, it will always be more accurate and effective to automatically detect speech features than phonemes. First, there are typically only two to five alternatives for features, as opposed to roughly 40 to 60 alternatives for phonemes. Second, features (voicing, frication, nasality, and sonorant) are relatively easy to recognize automatically. Our system does not attempt to analyze the most difficult acoustic feature, place of articulation, because it can be easily determined from visual cues that are readily perceivable from the face.

It might be argued that the tactile modality is more appropriate for presentation ofsupplementary speech features than the visual modality. For example, instead of providing three colored bars, the same information could be mapped into three vibratory transducers. There are well-known commonalities between the visual and tactile sensory systems (Freides, 1974; Hirsh & Sherrick, 1961; Lederman & Klatzky, 2004; Loomis & Lederman, 1986; Loveless et al., 1970; Sherrick & Cholewiak, 1986), and it may be that observers will be at a disadvantage dividing their attention between two visual sources of information relative to coordinating two sources of information from separate modalities. However, it is also known that the tactile modality has much poorer spatial and temporal resolution than vision. In an experiment with CV syllables, better recognition was achieved with visual than with tactile presentation of speech features (Martony, 1974). For voiced stops, there was no improvement with tactile feature presentation, but a significant improvement was observed with visual feature presentation. Martony suggests that this disparity was due to participants' difficulty in perceiving the exact temporal relation between the visual and tactile information. Thus, there may be an advantage to using two sources from the visual modality because of an enhanced ability to perceive the temporal relationship between speechreading and visual cues. Using two visual sources, listeners should be able to easily detect temporal relation cues such as voice onset time (a cue to voicing which, in this case, would be realized as a relation between a visible facial articulation and the activation of the supplementary voicing bar).

/H1/ Summary and Conclusion

As illustrated by the success of Cued Speech, the need for language aids is pervasive in today's world. Millions of individuals with language and speech impairments require additional support for language understanding and learning. Currently, however, these needs are frequently underserved because there are not enough qualified teachers, interpreters, and other professionals to provide the one-on-one attention that many of them need. Lipreading (also known as speechreading because it involves more than just the lips) allows deaf and hard of hearing individuals to perceive and understand oral language and even, in some cases, to speak. Speechreading seldom disambiguates all of the spoken input, however, and other techniques have been used to create a richer input. Cued Speech has been successful in providing additional supplementary linguistic cues, but very few people are proficient in Cued Speech or have the motivation to learn it.

To fill the need for rich language input, we are developing a real time system to automatically detect robust characteristics of auditory speech and transform these acoustic features into supplementary visible features representing aspects of speech. This information, combined with the perception of visual speech cues from the speaker's face, provides additional information so that people with limited hearing can perceive and understand oral speech. Our new technology, a wearable computing device, would recognize primitive characteristics of the speech signal in real time and display supplementary visual features via an LED display mounted on a pair of eyeglasses. This system may possess some advantages over Cued Speech because it is directly based on the acoustic and phonetic properties of speech and provides continuous rather than only categorical information. Our research has demonstrated that it is possible to recognize robust characteristics of isolated spoken words from a single speaker and to transform them into visual features in real time. The proposed research seeks to extend these findings to sentences

and conversation from multiple speakers and to determine the selection of the ideal set of features for display.

The successful outcome of this work would benefit society by providing an empirical and theoretical foundation for a system that would be available to most individuals at a very low cost. It does not require that users be literate because no written information is presented, as would be the case in a captioning system; it is age-independent in that it can be used throughout the life span; it is functional for all languages because it is language-independent given that all languages share the same phonetic features with highly similar corresponding acoustic characteristics; it would provide significant help for people with hearing aids and cochlear implants; and it would be beneficial for many individuals with language impairments and even for children learning to read. Finally, regardless of the advances or lack of advances in speech recognition technology, it will always be more accurate and effective to detect speech features than phones.

Baldi® is a trademark of Dominic W. Massaro. The author DWM acknowledges current support from the Special Hope Foundation and CITRIS (the Center for Information Technology Research in the Interest of Society) and previous support from the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), the Public Health Service (Grant No. PHS R01 DC00236), a Cure Autism Now Foundation Innovative Technology Award, the National Alliance for Autism Research, and the University of California, Santa Cruz. The author thanks Laura Morett for a careful reading and editing of this paper.

/H1/ References

Abu-Amer, T. & Carson- Berndsen, J. (2003, May 20-23). HARTFX: A Multi-Dimensional System of HMM Based Recognisers for Articulatory Features Extraction. ITRW on Non-Linear Speech Processing (NOLISP 03). Le Croisic, France.

Aioanei, D., Carson-Berndsen, J., Kanokphara, S. (2006). "Diagnostic evaluation of phonetic feature extraction engines: A case study with the Time Map model." In Proceedings of the 19th International Conference on Industrial and Engineering Applications of Applied Intelligent Systems, pp. 691-700, LNAI 4031, Annecy, France, 2006.

Bernstein, L.E. (2005). Some Principles of the Speech Perceiving Brain. Handbook of Speech Perception. Blackwell. pp. 79-98

Bernstein, L.E. & Eberhardt, S.P. (1986). Johns Hopkins Lipreading Corpus Videodisk Set. The Johns Hopkins University: Baltimore, MD.

Bernstein, L. E., M. E. Demorest, & P. E. Tucker, (2000). Speech Perception Without Hearing. Perception & Psychophysics, 62, 233-252.

Blei, D., Ng, A., Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3,993-1022.

Breslaw, P. I., Griffiths, A. J., Wood, D. J., & Howarth,C. I. (1981). The Referential

Communication Skills of Deaf Children from Different Educational Environments. Journal of Child Psychology, 22, 269–282.

Carreira-Perpiñán, M. Á. and Lu, Z. (2007): "The Laplacian Eigenmaps Latent Variable Model". 11th Int. Workshop on Artificial Intelligence and Statistics (AISTATS'2007), pp. 59-66.

Carreira-Perpiñán, M. Á. (2001): Continuous latent variable models for dimensionality reduction and sequential data reconstruction. PhD thesis, University of Sheffield, UK.

Chang, S., Greenberg, S. & Wester, M. (2001). An Elitist Approach to Articulatory-Acoustic Feature Classification. Eurospeech 2001. Aalborg, Denmark.

Clark, A. (2003). Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence. Oxford University Press, USA.

Cooke, M., Barker, J., Cunningham, S. & Shao, X. (2006). An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition, J. Accoust. Soc. Am. 120, 5, November 2006. http://www.dcs.shef.ac.uk/spandh/gridcorpus/

Cornett, R.O., (1967). Cued Speech. American Annals of the Deaf, 112, 3-13.

Cornett, R.O., Beadles, R., and Wilson, B. (1977). Automatic Cued Speech. Processing Aids for the Deaf, pp 224-239.

Costanza E., Inverso S. A., Pavlov E., Allen R., Maes P. (2006). Eye-Q: Eyeglass PeripheralDisplay for Subtle Intimate Notifications. (Full paper) in Proc. of MobileHCI.

DARPA TIMIT (1993). http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html

Davis, H., & Silverman, S.R. (1970). Hearing and Deafness. New York: Holt, Rhinehart, and Winston.

de Gelder, B., & Morais, J. (Ed.) (1995).Speech and Reading: A Comparative Approach. Hove, England: Erlbaum (UK) Taylor & Francis, Publishers.

Denes, P. B., & Pinson, E. N. (1963). The Speech Chain. The physics and biology of spoken language. New York: Bell Telephone Laboratories.

Duchnowski, P., Lum, D.S, Krause, J., Sexton, M., Bratakos, M., & Braida, L. (2000). Development of Speechreading Supplements Based on Automatic Speech Recognition, IEEE Transactions on Biomedical Engineering, 47(4), 487-495.

DuMouchel, W. (1994). Hierarchical Bayes Linear Models for Meta-Analysis. Technical Report, 27, National Institute of Statistical Sciences.

Eide, E. (2001). Distinctive Features For Use in an Automatic Speech Recognition System. Eurospeech 200., Aalborg, Denmark.

Erber, N. P. (1972). Auditory, Visual, and Auditory-Visual Recognition of Consonants by Children with Normal and Impaired Hearing. Journal of Speech and Hearing Research, 15, 423-422.
Flaush, D., Stephens, M., & Pritchard, J.K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlate Allel Frequencies.
Fleming, E. (1984). Believe the heart. San Francisco: Strawberry Hill Press.
Frankel, J., Magimai-Doss, M., King, S., Livescu, K., & O¨zgu¨r, C. (2007). Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech. In Proc. Interspeech, Antwerp, Belgium, Aug 2007.
Frankel, J., Wester, M. & King, S. (2004). Articulatory Feature Recognition Using Dynamic Bayesian Networks. In Proc. ICLSP.
Freides, D. (1974). Human information processing and sensory modality: Cross-modal functions, information complexity, memory, and deficit. Psychological Bulletin, 81, 284-310.
Friedman, D., Massaro, D. W., Kitzis, S. N., & Cohen, M. M. (1995). A comparison of learning models. Journal of Mathematical Psychology, 39, 164-178.
Gengel, R. W. (1976). Upton's Wearable Eyeglass Speechreading Aid: History and Current Status. Hearing and Davis: Essays honoring Hallowell Davis. St. Louis: Washington University Press.
Grant, K.W. (2002). "Measures of auditory-visual integration for speech understanding: A theoretical perspective. Journal of the Acoustical Society of America, 112, 30-33.
Grant, K. W., & Greenberg, S. (2001): "Speech intelligibility derived from asynchronous processing of auditory-visual information", In AVSP-2001, 132-137.
Grant, K.W., Greenberg, S., Poeppel, D., & van Wassenhove, V. (2004). Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. Seminars in Hearing, 25, 241-255.
Greenberg, S. (2006). A Multi-tier Theoretical Framework for Understanding Spoken Language. In Listening to Speech: An Auditory Perspective. Mahwah, NJ: Lawrence Erlbaum Associates, pp.411-433.
Greenberg, S. & Chang, S. (2000). Linguistic Dissection of Switchboard-Corpus Automatic Speech Recognition System. Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, pp. 195-202.
Hage, C. & Leybaert, J. (2006). The effect of Cued Speech on the development of spoken language. In: P.E. Spencer & M. Marschark (Eds), Advances in the spoken language development of deaf and hard-of-hearing children. New York : Oxford University Press, pp, 193-211.
Hazen, T.J., Saenko, K., La, C. & Glass, J. (2004). A Segment Based Audio-Visual Speech Recognizer: Data Collection, Development, and Initial Experiments, ICMI 2004.
Hirsh, I. J., & Sherrick, C. E., Jr. 1961. Perceived Order in Different Sense Modalities. Journal of Experimental Psychology, 62, 423-432.
Holt, J. A., Traxler, C. B., & Allen, T. E. (1997). Interpreting the scores: A user's guide to the 9th Edition Stanford Achievement Test for educators of deaf and hard-of-hearing students. Washington, DC: Gallaudet Research Institute.
Jesse, A., Vrignaud, N., & Massaro, D. W. (2000/01). The processing of information from multiple sources in simultaneous interpreting. Interpreting, 5, 95-115.
King, S. & Taylor, P. (2000). Detection of Phonological Features in Continuous Speech Using Neural Networks. Computer Speech and Language, 14 (4), pp. 333-353.
Kisor, H. (1990). What's that pig outdoors? A memoir of deafness. New York: Hill and Wang.
Kitzis, S.N., Kelley, H., Berg, E., Massaro, D.W. & Friedman, D. (1998). Broadening the tests of learning models. Journal of Mathematical Psychology, 42, 327-355.
Lederman, S.J. & Klatzy, R. (2004). Multisensory Texture Perception. Handbook of Multisensory Processes, pp 107-122. Massachusetts: MIT Press.

Loomis, J. M., & Lederman, S. J. (1986). Tactual Perception. Handbook of perception and human performance, Vol. 2:Cognitive processes and performance pp. 31-1 to 31-41. New York: John Wiley & Sons.
Loveless, N. E., Brebner, J., & Hamilton, P. (1970). Bisensory Presentation of Information. Psychological Bulletin, 73, 161-199.
Martony, J. (1974). Some experiments with electronic speechreading aids. KTH Speech Transmission Laboratory: Quarterly Progress and Status Report, 2-3, 34-56. Stockholm: Royal Institute of Technology.
Massaro, D. W. (1984). Children's perception of visual and auditory speech. Child Development, 55 , 1777-1788.
Massaro, D. W. (1987). Speech perception by ear and eye: A Paradigm for psychological inquiry. Hillsdale, NJ: Erlbaum.
Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, Massachusetts: MIT Press.
Massaro, D.W. (1999). From theory to practice: Rewards and challenges. In Proceedings of the International Conference of Phonetic Sciences, San Francisco, CA, August.
Massaro, D. W. (2004). Symbiotic Value of an Embodied Agent in Language Learning.
Proceedings of 37th Annual Hawaii International Conference on System Sciences (CD/ROM), Computer Society Press, 2004. http://mambo.ucsc.edu/pdf/ETSIB10.pdf
Massaro, D.W. (2006). Embodied Agents in Language Learning for Children with Language Challenges. Proceedings of the 10th International Conference on Computers Helping People with Special Needs, ICCHP 2006 pp.809-816. University of Linz, Austria. Berlin, Germany: Springer.
Massaro, D.W. & Bosseler, A. (2003). Perceiving Speech by Ear and Eye: Multimodal Integration by Children with Autism. Journal of Developmental and Learning Disorders, 7, 111-144.
Massaro, D.W. & Cohen, M.M. (1993). Perceiving Asynchronous Bimodal Speech in Consonant-Vowel and Vowel Syllables. Speech Communication, 13, 127-134.
Massaro, D.W. & Cohen, M.M. (1999). Speech perception in hearing-impaired perceivers:
Synergy of multiple modalities. Journal of Speech, Language, and Hearing Science, 42, 21-41.
Massaro, D. W., Cohen, M. M., & Beskow, J. (2007). Visual display methods for in computer-animated speech production models. United States Patent 7,225,129, May 29, 2007
Massaro, D.W. & Light, J. (2004). Using visible speech for training perception and production of speech for hard-of-hearing individuals. Journal of Speech, Language, and Hearing Research, 47(2), 304-320.
Massaro, D.W., & Stork, D. G. (1998). Sensory integration and speechreading by humans and machines. American Scientist, 86, 236-244.
Miller, J. D., Engebretson, A. M., & DeFillipo, C. L. (1974). Preliminary research with a three
channel vibrotactile speech-reception aid for the deaf. Speech Communication, 4, Proceedings of the Speech Communication Seminar. New York: John Wiley & Sons.
Mirrielees, D. I. (1947). Education of the Young Deaf Child: Special Participants and Methods. University of Chicago Home Study Department.
Montgomery, A. A., & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. Journal of the Acoustical Society of America, 73, 2134-2144.
Morais, J., & Kolinsky, R. (1994).Perception and awareness in phonological processing: The case of the phoneme. Cognition, 50, 287-297.
Munhall, K.G., Kroos, C., Jozan, G. & Vatikiotis-Bateson, E., (2004). Spatial frequency requirements for audiovisual speech perception. Perception and Psychophysics, 66, 574-583.

Munhall, K., & Vatikiotis-Bateson, E. (2004). Spatial and Temporal Constraints on Audiovisual Speech Perception. Handbook of Multisensory Processe,s pp. 177-188. Cambridge, MA: MIT Press.

National Reading Panel, (2000). Teaching children to read. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Institute of Child Health and Human Development, NIH Pub. No. 00-4769

Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., & Robinson, A. (1995, September). Speaker Adaptation for Hybrid HMM-ANN Continuous Speech Recognition System. Eurospeech, pp 2171-2174.

Norman, D. (1988). The Psychology of Everyday Things. Basic Books.

Norman, D.A. (1999). The Invisible Computer. Cambridge, Massachusetts: MIT Press.

Pickett, J. M., Gengel, R. W., Quinn, R., & Upton, H. W. (1974). Research with the Upton eyeglass speechreader. Speech Communication, 4, Proceedings of the Speech Communication Seminar. New York: John Wiley & Sons

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

Pratt http://www.crlmb.ca/latlang/2008/02/pratt-acoustic.html

Saul, L. K., Weinberger, K. Q., Ham, J. H., Sha, F. & Lee, D. D. (2006). "Spectral methods for dimensionality reduction". In O. Chapelle, B. Schoelkopf, and A. Zien (eds.), Semisupervised Learning. MIT Press: Cambridge, MA.

Sherrick, C. E., & Cholewiak, R. W. (1976). Cutaneous sensitivity. Handbook of perception and human performance, 2, Cognitive processes and performance, pp. 12-1 to 12-58. New York: John Wiley & Sons.

Schutte, K., & Glass, J. (2005). Robust detection of sonorant landmarks. Interspeech, 105-1008.

Smeele, P.M.T., Massaro, D.W., Cohen, M.M., & Sittig, A.C. (1998). Laterality in visual speech perception. Journal of Experimental Psychology: Human Perception and Performance, 24, 1232-1242.

Summerfield, A. Q. (1987) Some preliminaries to a comprehensive account of A/V speech perception. In: Hearing by eye: The psychology of lip-reading (Dodd B, Campbell R, eds). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tager-Flusberg, H. (2000). Language development in children with autism. Methods For Studying Language Production, pp., 313-332. New Jersey: Mahwah.

Taylor, I., & Olson, D. R. (Eds.). (1995). Scripts and literacy: Reading and learning to read alphabets, syllabaries and characters. Dordrecht, Netherlands: Kluwer Academic Publishers.

Upton, H. W. (1968). Wearable eyeglass speechreading aid. American Annals of the Deaf, 113, 222-229.

Wandel, J.E. (1998) Use of Internal Speech in Reading by Hearing and Hearing-Impaired Students in Oral, Total Communication, and Cued Speech Programs, PhD Dissertation, Columbia University Teacher's College.

Wavesurfer (http://www.speech.kth.se/wavesurfer/)

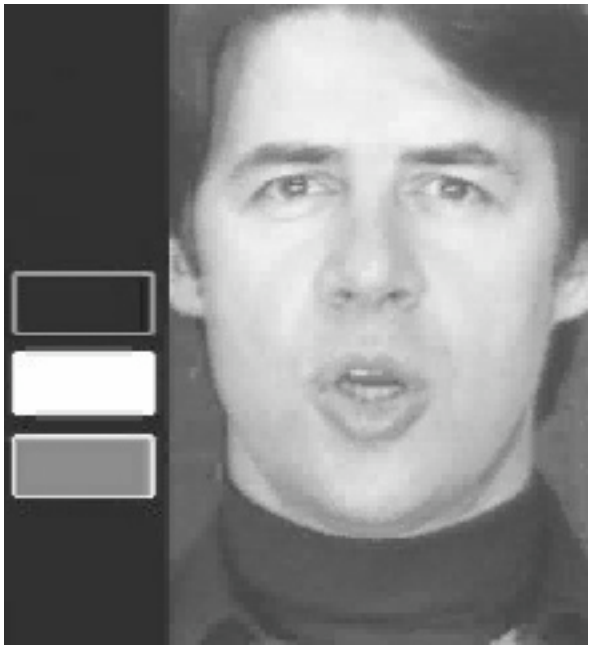
Weiser, M. (1991). The computer for the 21st century. Scientific American, 265 (3), pp.94-104.

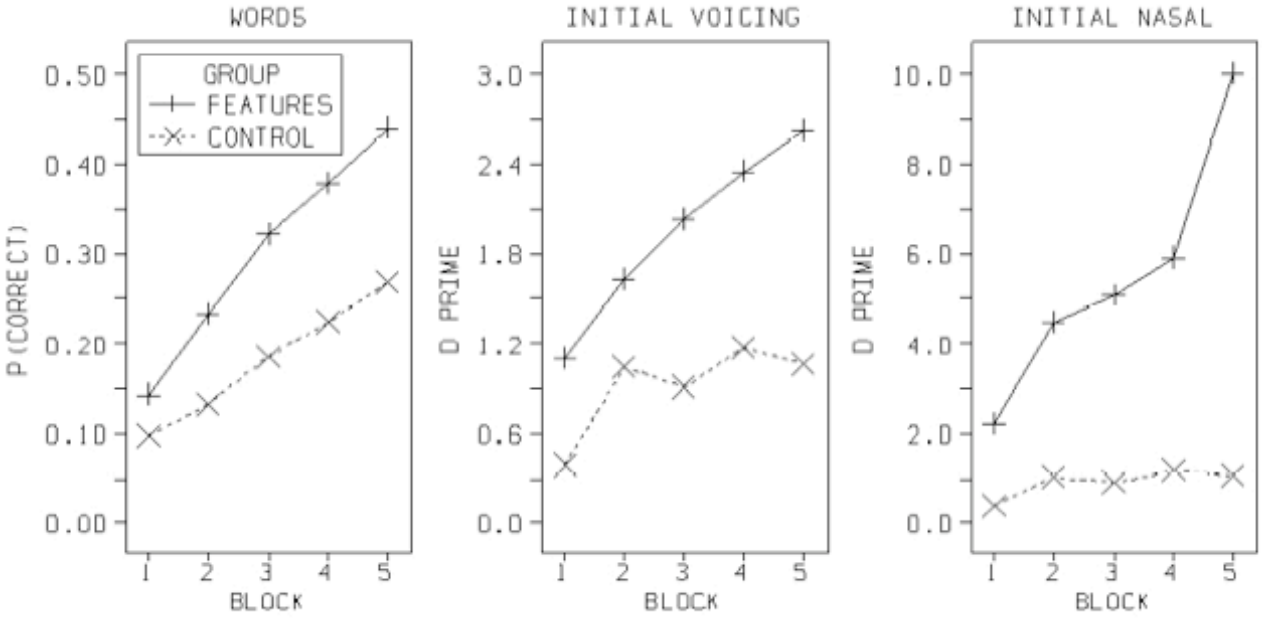
Williams, J.H.G., Massaro, D.W., Peel, N.J., Bosseler, A., & Suddendorf, T. (2004). Visual-Auditory integration during speech imitation in autism. Research in Developmental Disabilities, 25, 559-575.

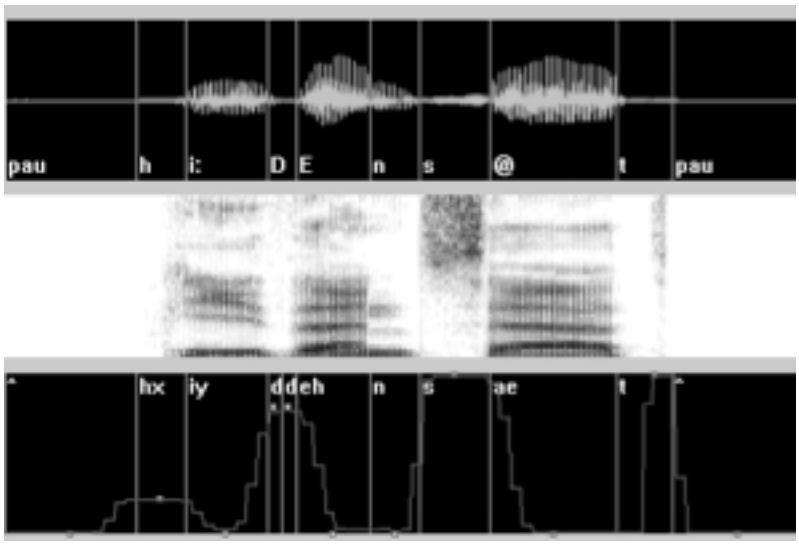
Willows, D. M., Kruk, R. S., & Corcos, E. (Eds.), Visual processes in reading and reading disabilities. Hillsdale, NJ: Lawrence Erlbaum

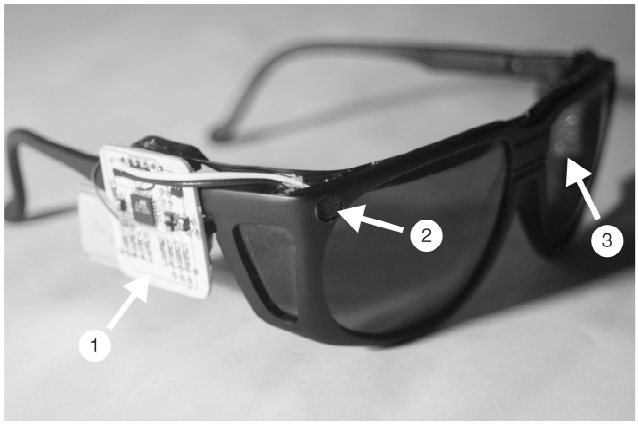
Zigoris, P., & Zhang, Y. (2006) Bayesian Adaptive User Profiling with Explicit & Implicit Feedback, Conference on Information Knowledge Management (CIKM).

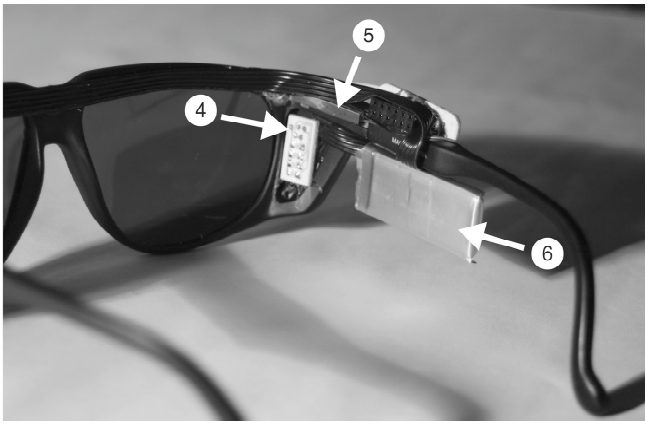
List of Figures




WORDS
INITIAL VOICING
INITIAL NASAL
GROUP
FEATURES
CONTROL
P(CORRECT)
D PRIME
D PRIME
BLOCK
BLOCK
BLOCK


pau h i: D E n s @ t pau
^ hx iy ddeh n s ae t ^


1
2
3


4
5
6