Optimizing Visual Feature Perception for an Automatic Wearable Speech Supplement in Face-to-Face Communication and Classroom Situations

Dominic W. Massaro
University of California, Santa Cruz, Santa Cruz, CA
massaro@ucsc.edu

Miguel Á. Carreira-Perpiñán
University of California, Merced, Merced CA
mcarreira-perpinan@ucmerced.edu

David J. Merrill
MIT Media Lab
Cambridge, MA
dmerrill@media.mit.edu

Abstract

Given the limitation of hearing and understanding speech for many individuals, we plan to supplement the sound of speech and speechreading with an additional informative visual input. Acoustic characteristics of the speech will be transformed into readily perceivable visual characteristics. The goal is to design a device seamlessly worn by the listener, which will perform continuous real-time acoustic analysis of his or her interlocutor's speech. This device would transform several continuous acoustic features of the talker's speech into continuous visual features, which will be simultaneously displayed on the speechreader's eyeglasses. The current research evaluates how easily a number of different visual configurations are learned and perceived. The goal is to optimize the visual feature presentation and implement it in the wearable computer system.

1. Introduction

Traditionally, speech has been 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 from the speech and language of the interlocutor. This ability to exploit multiple modalities and multiple sources of information is a godsend to almost all individuals at some time in their lives. In addition, this richness of potential inputs facilitates the creation of language supplements. Before addressing the needs for language supplements 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.

2. Multimodal Speech Perception

Speech science evolved as the study of a unimodal auditory channel of communication because speech was viewed as primarily auditory [1]. There is no doubt that the voice alone is usually adequate for understanding for many individuals and, given the popularity of mobile phones, might be the most frequent medium for today's communication. However, there are many deaf and hard-of-hearing individuals who must have other sources of language input. The face is valuable even for normal hearing individuals because many communication environments 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 are influenced by a speaker's face, as well as the actual sound of speech [2,3,4].

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 [5]. 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.

2.1. Information Value of Visible Speech

The value of visible speech is demonstrated by the results of a series of experiments in which 71 typical college students reported the words of sentences presented in noise [6]. On some trials, only the acoustic sentence was presented (unimodal condition). On some other trials, the acoustic

978-0-7695-3450-3/09 $25.00 © 2009 IEEE

sentence was appropriately aligned with a highly realistic computer-animated face known as "Baldi" (bimodal condition). Baldi's presence facilitated performance for everyone. Accurate performance was more than doubled for those participants performing particularly poorly when given acoustic speech alone. Although a unimodal visual condition was not included in the experiment, we know that participants would have performed much more poorly than the unimodal acoustic condition [3,5]. 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 using either modality alone.

Similar results are found when noise-free speech is presented to persons with limited hearing [7]. Adolescents and young adults who were either profoundly deaf or had severely-impaired hearing benefited from face-to-face speech relative to just acoustic speech. The severely impaired perceivers (having a hearing loss between 75 and 90 dB) experienced the largest performance gain with nearly perfect performance in the bimodal condition relative to either of the unimodal conditions [5, p. 159,8].

2.2. 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 [9]. Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, for example), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer [5,10,11]. 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 will be combining speechread information with additional visual cues.

2.3 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 [8]. 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 fact that two sources of information are complementary makes their combined use much more informative than would be the case if the two sources were redundant [5, Chapter 14, pp. 424-427). In our application for deaf and hard-of-hearing individuals, our goal is to make visible the linguistic information that is particularly difficult to see on the face.

2.4. Optimal Integration

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 [3,8,12]. 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 in such a fashion that both sources are used but that the least ambiguous source has the most influence. Perceivers integrate the information available from each modality extremely efficiently, a pattern described by the Fuzzy Logical Model of Perception (FLMP) [5]. The FLMP assumes that the visible and audible speech signals are each evaluated (independently of the other source) to determine how much that source supports each alternative. The integration process optimally combines these support values to determine how much their combination supports the various alternatives. The perceptual outcome for the perceiver will be a function of the relative degree of support among the competing alternatives. As demonstrated elsewhere, the FLMP is mathematically equivalent to Bayes' theorem [5, Chapter 4), which is an optimal method for combining two sources of evidence to test among hypotheses.

The best evidence for the FLMP comes from an important experimental manipulation that systematically varies the ambiguity of each source of information [5]. We have also found that, like adults, typically developing children integrate information from both the face and the voice [13] as well as do deaf and hard-of-hearing children [8] and autistic children [14,15]. Critical for the requirements of our work is that this optimal integration occurs even if the auditory and visual speech are not perfectly synchronous (up to at least 100 ms). Finally, the pilot results described below indicate that individuals can easily learn to integrate facial information with supplementary visual features.

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

3. Language Supplements

There are millions of individuals who have language and speech challenges, and these individuals require additional support for language understanding and learning. In California alone, there are almost 200,000 deaf, hard-of-hearing, and speech-language impairment children enrolled in Special Education [16] (http://www.cde.ca.gov/re/pn/sm/index.asp). As an example of a specific need, it is well known that deaf and hard-of-hearing children have significant deficits in both spoken and written vocabulary knowledge [16]. A similar situation exists for autistic children, who lag behind their typically developing cohort in language acquisition [17]. Currently, however, these needs are not being met. One problem that the people with these disabilities face is that there are not enough skilled teachers, interpreters, and professionals to give them the one on one attention that they need.

In fact, humans can learn and use language successfully without adequate auditory input. Sign language parallels spoken language in acquisition, use, and communication. But even oral language can serve communication when the auditory input is degraded or even absent. Lipreading (called speechreading because it involves more than just the lips) allows these individuals to perceive and understand oral language and even to speak [18,19,20]. Speechreading seldom disambiguates all of the spoken input, however, and other techniques have been used to allow a richer input. Cued Speech, for example, is a deliberate solution to having a limited auditory input, and consists of hand gestures while speaking that provide the perceiver with information that potentially disambiguates what is seen on the face. However, very few people know Cued Speech or have the motivation to learn it, and therefore, individuals with limited auditory speech input are faced with insufficient input in many face-to-face and classroom-like environments.

Building on the innovative idea of Upton [21], [5, Chapter 14] proposed a device to perform acoustic analysis of speech and transform several acoustic features into visual features, which the speechreader would use in conjunction with watching the speaker's face. The acoustic features associated with important linguistic information not directly observed on the face will be transformed into visual cues intended to enhance intelligibility and ease of comprehension. A significant body of research supports the idea that people can easily learn to integrate such linguistic features with the incomplete visual information to achieve productive outcomes. Furthermore, similar to Cued Speech, the users of this device will have the advantage of gaining additional phonological awareness through the use of the linguistic features. We now discuss research that illustrates the value of providing additional visual cues to supplement the speech input.

3.1 Previous Supporting Research

Cued Speech has become an accepted form of communication for deaf and hard-of-hearing individuals (http://www.youtube.com/watch?v=plPw4H-ZsMg) [21]. Cued Speech was designed as a means for supplementing lipreading by providing manual cues to phoneme identity to replace information not normally seen on the talker's face. Properties of Cued Speech include: 1) its hand gestures can be learned, 2) it is based on the phonemes of the spoken language, and 3) it can used 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 or family and friends of the deaf or hard of hearing might be motivation enough to learn a system of cues, we cannot expect other individuals to be similarly motivated. Thus, a solution for supplementing communication that does not depend on any special skills of the talker would be ideal.

Another drawback is that Cued Speech implements an awkward mapping between gestures and phonemes that is arbitrary and has not changed since it was first proposed by Cornett [22]. His idea was based on the realization that speechreading does not provide sufficient detail to distinguish all of the phonemes but only different subsets of phonemes, such as /b, p, m/ versus /f,v/ in a language. Different Cued Speech hand gestures were therefore designed to denote different subsets of phonemes so that both subsets together would indicate just a single phoneme. For example, the hand gesture with the index finger extended would signal the subset of phonemes /d, p, zh/ which when combined with the speechread /b, p, m/ would denote /p/. However, there is no linguistic or psychophysical structure within a Cued Speech category, which necessarily makes learning and understanding of the categories difficult. Meaningful categories such as birds, fish,

and chairs share perceptual and conceptual properties [23]. The supplementary feature solution we propose is perceptually-based and conceptually-based and provides continuous information indicating the degree to which a feature is present.

Our approach bypasses full-blown speech recognition because accurate automatic speech recognition (ASR) is optimized for recognizing words, not acoustic features, and requires huge computational resources and is limited to less than real-time performance. (Best performance occurs with at least a 3 GHz processor when a complete sentence is available. Successful systems carry out something like a cepstral analysis with about 60-90 spectral features—with little relationship 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, close to real-time performance, and a light-weight portable device with limited computing power. Our proposed alternative is to simply detect a few robust acoustic features that can be mapped into visual cues simultaneously with their detection.

To compensate for the delay required for full-blown speech recognition, Duchnowski et al. [24,25] (unpublished), recorded a video of the talking face and replayed this video to the listener simultaneously with the Cued Speech with a 2-second delay. Although this may be feasible in televised broadcast or perhaps even in a classroom on a video iPod or other handheld device, it would be highly disruptive in any face-to-face encounter. Our envisioned system, on the other hand, would be highly functional in all foreseeable applications.

In summary, the widespread use of Cued Speech and the research with visual cueing systems show that automatically supplementing speech with visual features is a worthwhile research objective. Our future research will test improvements in such a manner that will lead to a successful system. The requirements of a successful system include a light footprint for a wearable device, operation in near real time, accurate tracking of acoustic features, learnable visual features, and integration of these features with auditory and visual speech.

3.2 Pilot Research

We have carried out pilot research to investigate how to supplement talking faces with information that is ordinarily conveyed by auditory means. We now describe our initial work on this problem of supplementing visual speech. We have separated this research into two areas, which will be discussed in the next two sections: 1) developing a neural network to perform real-time analysis of certain acoustic features for visual display, and 2) determining how quickly subjects can learn to use these selected cues and how much they benefit from them when combined with speechreading.

3.2.1 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. In pilot research, we developed and trained a neural network to recognize three auditory speech characteristics: nasality, voicing, and frication. Training gave a .057 root mean square deviation (RMSD) between the actual and predicted feature values on a 0-1 scale. Thus, the neural net model was successfully trained to provide moment-by-moment outputs for the three features on the basis of acoustic input. 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. The focus of the current research is to optimize the visual feature presentation.

3.2.2 Visual Feature Perception. Our studies of the perception of supplementary visual feature information were done using simulated rather than real-time analysis of acoustic features. We wished to see how difficult it would be for subjects to learn to effectively use the visual features we had selected to supplement speechreading. A table giving the mapping between the phonemes and the visual features, as well as phonetic and coarticulatory information, was provided in written form to the subjects. For example, vowels are voiced, fricatives have frication, frication can occur during the onset of stop consonants, and the nasal following a vowel can produce nasality during the vowel as well as during the nasal segment. In a five day experiment, subjects speechread 318 one-syllable words from the Bernstein & Eberhardt corpus [26] presented visually. The visual speech was presented by a human speaker whose facial image was 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 subjects was presented with feature information along with this silent talking head, whereas a control group of 3 subjects received only the silent talking face and no feature information.

Figure 1. An example of the video display with the visual features. The top nasal bar indicates the nasals by lighting up orange during the period they occur. The middle voicing bar indicates voiced sounds by lighting up white and the bottom frication bar lights up when there is frication noise. The intensity of each cue corresponded to the degree to which the corresponding acoustic feature was present in the speech signal.

For the feature group, the 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 the CD Band 14.8 in [5], and is available online at http://mambo.ucsc.edu/psl/mmc/14_8.mov. It shows the 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 corresponded to frication, which 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. The 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 will be described in more detail below. Subjects made their responses by typing a word on a keyboard, which was followed by feedback during which presenting the word (with features for the feature group) was presented again, 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 better 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 respectively for the two groups. (The accuracy score is transformed into a d-prime measure from signal detection theory, which is bias-free and measured in z-scores. For voicing, for example, correctly identifying a voiced test item is the hit rate and erroneously calling a non-voiced test item voiced is the false alarm rate. The d-prime value is computed from the z-score transformation of these probabilities. Therefore, the range of performance from chance to perfect is measured between 0 and 3 or so. Note also that these two panels also have different scales, given the different ranges of performance.) Relative to the control group, the feature group was able to improve 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 the subjects had not yet learned to use the information to achieve perfect word recognition.

Analyzing the consonant confusions for the control and feature groups indicated that the feature group was able to make discriminations that were not possible for the controls. For example, within the labial stops /b,m,p/, the feature group could discriminate between the three members of the class, while the control group split their responses equally among the three alternatives. This experiment demonstrates that speechreading with these visual features is learnable and greatly improves speechreading accuracy. However, it is necessary to extend the research to 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 to present on the wearable supplement, and evaluatiing the prototype system.

3.3 New Research

3.3.1 Participant Population. Participants will include normal-hearing as well as deaf and hard-of-hearing persons, who have a vested interest in

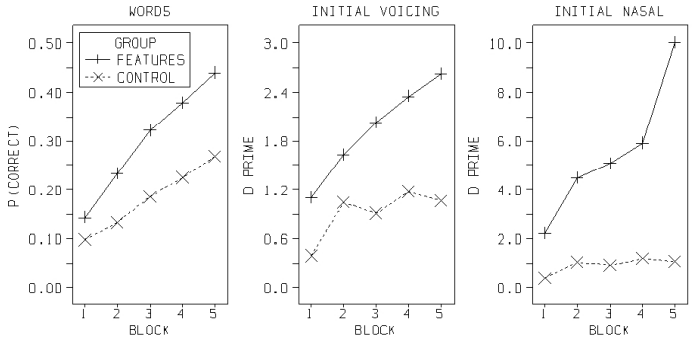

Figure 2. Proportion correct word identification (left panel), identification (d prime) of initial voicing (center panel), and identification (d prime) of initial nasality (right panel) as a function of experimental block, for feature and control groups [5, Chapter 14].

enhancing their communication interactions. We will also attempt to recruit persons who are skilled in Cued Speech to assess to what extent this experience and skill facilitates or inhibits performance with the supplementary visual features.

3.3.2 Implementing Eyeglasses Appliance. We have built and configured eyeglass frames to hold the LED display. Figure 5xxx shows a mockup of the glasses. The display is mounted beside one of the lenses, in a location that is visible in the person's periphery. It may be necessary to adjust the forward-back location of the display for some users, to maximize comfort and view-ability. The side of the face that is closest to the display may be an important variable, since this determines whether the visual features well be seen to the right or left of the talking face. There is a large literature on hemifield effects in visual perception and language processing and, although the research is not conclusive (Smeele et al., 1998), it would be advantageous to choose the side that leads to best performance. 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 performance with the LED display will improve speech perception, as it did in [5]. If it does not, we will explore the differences between these two situations in order to better design an effective wearable display.

3.3.3 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 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 all combinations of three features. Given that each experiment requires some significant amount of learning, however, these 4 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, then it would not be functional and, in fact, could 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.

An important aspect of these studies involves learning. Our pilot study indicated that participants do indeed learn to take advantage of the supplementary visual features. Even though learning occurred, there are potentially alternative learning regimens that can increase the rate and asymptote of learning. One possible training situation is to practice just a single supplementary visual feature at a given time. After achieving good learning on one feature, 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 simply consonant-vowel syllables so participants will be able to focus on the visual features for just a single consonant. Another possibility is to practice the participants on the features without the face present so they can directly learn the supplementary features. 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.

4. Significance and Advantages

The technology we are developing would be ideally designed for wearable computing so a person could have a face-to-face conversation while wearing a pair of simple glasses, which could also be fitted with the person's normal eye prescription. The wearable

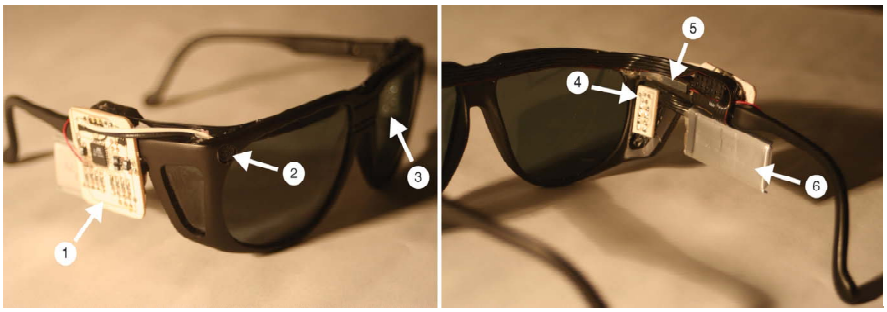

Figure 3. Physical system prototype. Identified parts are (1) circuit board with processor, (2) microphone, (3) secondary LED display (optional), (4) primary LED display, (5) vibrating actuator, and (6) rechargeable battery.

product would process primitive characteristics of the speech signal such as voicing (the presence of energy at the fundamental frequency such as heard in vowel sounds); frication (high frequency noise like energy characteristic of various consonants such as [s], [z], and [sh]; and nasality (which is 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 glasses [27].

Our envisioned system holds much promise because 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. People naturally integrate auditory and visual information so they should necessarily benefit from having both visible and audible speech. In addition, this strategy is particularly effective because of the complementarity of auditory and visual speech. The acoustic speech that is robust in the signal and fairly easy to automatically recognize is exactly that which is not visible on the face. This serendipitous occurrence makes it more likely to succeed at automatically recognizing the robust acoustic characteristics and simultaneously presenting them visually as supplementary cues.

The proposed technology qualifies as a transparent information appliance that adds perceptual and cognitive resources to the listener [29,30]. We have developed a requirements analysis, a conceptual design, and possible physical designs for this appliance. It consists of a very affordable noninvasive device that is seamlessly integrated with normal dress, adding a pair of glasses (which might be used regardless). This qualifies as an augmented-reality device, which is also available for use 24/7 and requires very little maintenance.

4.1 Usability for All Individuals

The system we propose is naturally available to all individuals who can wear a pair of eye glasses. 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. There is evidence that very young children can learn sign language and even finger spelling of the spoken language. The same should be true for the proposed supplementary cues. The phonetic basis for the speech driven cues should also reinforce an understanding of the phonology of the language [31]. 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.

4.2 Available to All Language Groups

One of the major advantages of our envisioned system 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 in signing or Cued Speech conversations. The device would be primarily functional in the frequent case when the listener is faced with oral language of a person who does not use Cued Speech.

4.3 Extended Reach of the Research

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 are still challenged in many natural environments such as those with background noise and reverberation, and in challenging conversations. The technology we propose will provide exactly the additional supplement to speechreading that will allow communication in these situations.

The benefits of this research extend beyond the deaf and hard-of-hearing 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, which will offer a larger number of potential communication methods for these individuals.

It is well-known that there are substantial out-of-the-ordinary problems that a number of children encounter in learning to read and spell. Children who have much more difficulty in reading and spelling than would be expected from their other perceptual and cognitive abilities are labeled as dyslexic [32]. Psychological science has established a tight relationship between the mastery of written language and the child's ability to process spoken language [31]. That is, it appears that many dyslexic children also have deficits in spoken language perception. The difficulty with spoken language can be alleviated through improving children's perception of phonological distinctions in their spoken language, which in turn improves their ability to read and spell [33]. Experience with the wearable system should help these children gain insights into the spoken language and therefore improve their reading skill.

5. Potential Limitations

A potential limitation of our system is that some non-visible acoustic features of consonants but not vowels are mapped into visible features to help disambiguate the spoken message. Cuing vowels would have a number of possible negative effects, however. First, recognizing vowels or vowel features from the waveform would be highly fallible relative to the other features being analyzed in 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. Fourth, vowels appear to be less perceptually degraded and therefore more intelligible than consonants. Fifth, the visible speech from the speaker is relatively informative for vowels, much more than it is for the voicing, frication, and nasality features currently in the proposed system. Our research will determine whether a robust system of augmented communication can be implemented even though no additional supplementary cues are provided for vowels.

It might be proposed that automatic speech recognition (ASR) by machine will improve sufficiently in the near future so that a full captioning of the speech being spoken can be accurately rendered. Although this significant breakthrough is always possible, it seems unlikely to occur in the near future. ASR recognition can be expected to be reasonably functional when there is a limited vocabulary and grammatical structure as input, if the system is speaker dependent—that is, trained on a single speaker, and/or used in a completely noise-free environment. Our device, on the other hand, will be functional in natural settings 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 different languages.

Another potential limitation of our approach is the recording of the acoustic speech in face-to-face conversations. Most ASR systems have the luxury of having the talker speak into a lapel microphone or a telephone for the recording. In our system, the microphone will be worn by the listener. This challenge is anticipated in the present project by training the system on somewhat remote recordings. In addition, because it is simply necessary to transmit acoustic features, the challenge of remote recording is diminished significantly. 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 challenge of having the microphone distant from the speaker, it would also be somewhat variable because the distances and directions will vary in typical face-to-face conversations. 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 potential 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 pick off features than phonemes. First, there are typically only two to five alternatives for features, as opposed to roughly 40 to 60 phonemes. Second, the features (voicing, frication, nasality, and sonorant) are relatively straightforward to recognize automatically. We do not attempt to analyze the most difficult acoustic feature place of articulation, which is exactly the information that is so easily seen on the face.

It might be argued that the tactile modality might be more appropriate for the presentation of the supplementary features. 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 [34], 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 from separate modalities. However, it is also known that the tactile modality has much poorer spatial and temporal resolution than vision. 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. With two visual sources listeners should more 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 some visible facial articulation and the activation of the supplementary voicing bar).

6. Retrospective and Conclusion

Our research will advance engineering research and speech science by developing a real-time system to automatically detect robust characteristics of auditory speech and to transform these acoustic features into supplementary visible features. This information combined with watching the speaker's face provides enough information for a person with limited hearing to perceive and understand what is being said. This new technology will allow the application of a wearable computing device that would recognize primitive characteristics of the speech signal in real time, and to display the supplementary features on a pair of eyeglasses. This system improves on Cued Speech because it is directly based on acoustic and phonetic properties of speech and gives continuous rather than only categorical information.

Pilot research has demonstrated that it is possible to recognize robust characteristics of isolated auditory words and to transform them into visible features in real time. The proposed research extends this research to sentences from multiple speakers, along with tests of different feature detectors and automatic recognition models. The team has a synergy of expertise in psychology, speech science, machine learning and autonomous embedded system engineering. The proposed research will advance the state of the art in human machine interaction, speech, machine learning and assistive technologies.

The research benefits society by providing a research and theoretical foundation for a system that would be naturally available to almost all individuals at a very low cost. It does not require literate users because no written information is presented as would be the case in a captioning system; it is age-independent in that it might be used by toddlers, adolescents, and 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 challenges 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 pick off features than phones.

7. References

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

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

[3] Massaro, D. W. (1987). Speech perception by ear and eye: A Paradigm for psychological inquiry. Hillsdale, NJ: Erlbaum.

[4] Summerfield 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 Assoc's.

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

[6] Jesse, A., Vrignaud, N., & Massaro, D. W. (2000/01). The processing of information from multiple sources in simultaneous interpreting. Interpreting, 5, 95-115.

[7] 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.

[8] 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.

[9] 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.

[10] Munhall, K.G., Kroos, C., Jozan, G. & Vatikiotis-Bateson, E., (2004). Spatial frequency requirements for audiovisual speech perception. Perception and Psychophysics, 66, 574-583.

[11] 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.

[12] [12] Massaro, D.W., & Stork, D. G. (1998). Sensory integration and speechreading by humans and machines. American Scientist, 86, 236-244.

[13] Massaro, D. W. (1984). Children's perception of visual and auditory speech. Child Development, 55 , 1777-1788.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] Kisor, H. (1990). What's that pig outdoors? A memoir of deafness. New York: Hill and Wang.

[20] Mirrielees, D. I. (1947). Education of the Young Deaf Child: Special Subjects and Methods. University of Chicago Home Study Department.

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

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

[23] Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), Cognition and categorization. Hillsdale, NJ: Erlbaum.Bernstein, L.E. (2005). Some Principles of the Speech Perceiving Brain. Handbook of Speech Perception. Blackwell. pp. 79-98

[24] 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.

[25] Duchnowski, P., Lum, D.S, Krause, J., Sexton, M., Bratakos, M., & Braida, L. (in press). Development of Speechreading Supplements Based on Automatic Speech Recognition.

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

[27] 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.

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

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

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

[31] Morais, J., & Kolinsky, R. (1994).Perception and awareness in phonological processing: The case of the phoneme. Cognition, 50, 287-297.

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

[33] 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

[34] Lederman, S.j. & Klatzy, R. (2004). Multisensory Texture Perception. Handbook of Multisensory Processes, pp 107-122. Massachusetts: MIT Press.