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

Dominic W. Massaro[1], Miguel Á. Carreira-Perpiñán[2], David J. Merrill[3], Cass Sterling[1,]
Stephanie Bigler[1], Elise Piazza[1], Marcus Perlman[1]
[1]Psychology, University of California, Santa Cruz, Santa Cruz, CA 95064 USA
[2]Electrical Engineering & Computer Science, University of California, Merced
[3]Media Lab, Massachusetts Institute of Technology
1 831 459 2330
massaro@ucsc.edu

ABSTRACT

The need for language aids is pervasive in today's world. There are millions of individuals who have language and speech challenges, and these individuals require additional support for communication and language learning. We demonstrate technology to supplement common face-to-face language interaction to enhance intelligibility, understanding, and communication, particularly for those with hearing impairments. Our research is investigating how to automatically supplement talking faces with information that is ordinarily conveyed by auditory means. This research consists of two areas of inquiry: 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.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences – *Psychology.*

General Terms

Experimentation

Keywords

Automatic speech supplement; multimodal speech perception

1. Introduction

The Problem. The need for language aids is pervasive in today's world. There are millions of individuals who have language and speech challenges, and these individuals require additional support for communication and language learning. Currently, however, the needs of these persons, such as limited understanding in face-to-face communication, are not being met. In California alone, there are almost 200,000 deaf, hard of hearing, and speech-language impairment children enrolled in Special Education [1]. 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 [2, 3]. A similar situation exists for autistic children, who lag behind their typically developing cohort in language acquisition [4]. 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. This research develops technology to supplement the common face-to-face language interaction to enhance intelligibility, understanding, and communication.

The Solution. Given the limitation of hearing speech for many individuals, we supplement the sound of speech and speechreading with an additional informative visual input [5, 6]. Acoustic characteristics of the speech will be transformed into readily perceivable visual characteristics. The acoustic (not phonetic) features are voicing, nasality, and frication, which will be transformed into continuous visual features, which will be simultaneously displayed on the speechreader's eyeglasses. These acoustic features provide important linguistic information not directly observed on the face and are transformed into visual cues intended to enhance intelligibility and ease of comprehension. This system does not require any learning on the part of the talker and is perceptually and linguistically motivated because it is directly based on acoustic and phonetic properties of speech and gives continuous rather than only categorical information.

Background. Cued Speech Supplementing Visible Speech. Cued Speech has become an accepted form of communication for deaf and hard of hearing individuals [7,8]. 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/hard of hearing or family and friends of the deaf/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.

2. Research

Previous research and pilot research have demonstrated that neural networks can detect and track robust characteristics of speech, which include frication, voicing and nasality [9]. The proposed research extends this work and implements a complete system of transforming continuous acoustic features into continuous supplementary visible features displayed on eyeglasses

ICMI'08, October 20–22, 2008, Chania, Crete, Greece.
ACM 978-1-60558-198-9/08/10.

during face-to-face communication. Pilot research indicates that people can learn to combine these visual cues with the visual information from the face to enhance intelligibility and comprehension [9]. The research evaluates the learning of several potential visible features in real-world contexts. 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.

Research. We are carrying out research to investigate how to supplement talking faces with visual information corresponding to acoustic speech. This research is divided into two areas: 1) developing a neural network to perform real-time analysis of certain acoustic features for visual display, and 2) determining the ideal format for presentation of the visual cues, how quickly subjects can learn to use these selected cues, and how much they benefit from them when combined with speechreading.

Training and Testing Neural Networks. We will train and test neural networks on spoken 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 3 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 back propagation 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.

Perception of Visual Features. Our previous research indicated that perceivers are able to use supplementary visual features presented in the periphery to improve speech perception while still attending to the speaker's face and lips [9]. 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 are replicating 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 OLEDs in their peripheral vision. The supplementary feature displays will be computer-generated and their output will be displayed on the eyeglass-mounted OLEDs.

3. Demonstration

Demonstration. The demonstration will center around the use of a pair of functional iGlasses that can be used in face-to-face conversations. The iGlasses will consist of a microphone, processing module, battery, and visual display. There will be a demonstration of how these iGlasses are used to facilitate intelligibility and comprehension in face-to-face conversations.

4. Conclusion

Broader impacts. A successful outcome will benefit 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 automatic speech recognition, and will always be more accurate regardless of the advances or lack of advances in speech recognition technology. 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 all languages share the 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.

5. REFERENCES

[1] (http://www.cde.ca.gov/re/pn/sm/index.asp)

[2] 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.

[3] 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.

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

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

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

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

[8] 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.

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