Auditory-Visual Speech Processing
2005 (AVSP'05)
British Columbia, Canada
July 24-27, 2005

Visual Contribution to Speech Perception: Measuring the Intelligibility of Talking Heads

Slim Ouni, Michael M. Cohen, Hope Ishak, and Dominic W. Massaro

Perceptual Science Laboratory - University of California at Santa Cruz, CA, USA

ABSTRACT

Animated agents are becoming increasingly frequent in research and applications in speech science. An important challenge is to evaluate the effectiveness of the agent in terms of the intelligibility of its visible speech. Sumby and Pollack (1954) proposed a metric to describe the benefit provided by the face relative to the auditory speech presented alone. We extend this metric to describe the benefit provided by a synthetic animated face relative to the benefit provided by a natural face. The validity of the metric is tested in a new experiment in which auditory speech is presented under 5 different noise levels and is paired with either our synthetic talker Baldi®[1] or a natural talker (the standard). A valid metric would allow direct comparisons across different experiments and would give measures of the benefit of a synthetic animated face relative to a natural face and how this benefit varies as a function of the type of synthetic face, the test items (e.g., syllables versus sentences, viseme class), different individuals, and applications.

1. INTRODUCTION

It is not surprising that face-to-face communication is more effective than situations involving just the voice. One reason is that the face improves intelligibility, particularly when the auditory signal is degraded (Sumby & Pollack, 1954; Summerfield, 1979; Jesse et al., 2000). Given this observation, there is value in developing virtual 3-D animated talking heads that are aligned with the auditory speech (Massaro, 1998). Given that the effectiveness of animated agents is critically dependent on the quality of their visible speech and emotion, it is important to assess their accuracy. A natural standard for measuring this accuracy is to compare the effectiveness of an animated agent to that of a natural talker. We know that a natural face improves the intelligibility of auditory speech in noise and we can evaluate an animated agent relative to this standard. Following this logic, a defining characteristic of our research has been the empirical evaluation of the intelligibility of our visible speech synthesis relative to that given by a human talker. The goal of the evaluation process is to determine how the synthetic visual talker falls short of a natural talker and to modify the synthesis accordingly. It is also valuable to be able to contrast the effectiveness of different animated agents.

2. PREVIOUS WORK

The purpose of this research is to facilitate the evaluation of the effectiveness of the agent in terms of the intelligibility of its visible speech. Sumby and Pollack (1954) proposed a metric to describe the benefit provided by the face relative to the auditory speech presented alone. Our goal is to extend this metric to describe the benefit provided by a synthetic animated face relative to the benefit provided by a natural face. The validity of the metric is tested in a new experiment in which natural auditory speech is presented under different noise levels and is paired with either our synthetic talker Baldi or the natural talker (the standard). We can expect the overall noise level to greatly impact performance accuracy but a valid metric would remain constant across noise levels.

Sumby and Pollack's metric measures the contribution of a single talking head. In our assessment of animated agents, the evaluation is made with respect to a natural talking head. A metric indicating the quality of an animated agent should be made relative to this standard of a natural talking head. A completely ineffective agent would give performance equal to the unimodal auditory condition and complete success would be the case in which the effectiveness of the animated agent would be equal to (or possibly better than) the standard. We describe and test a modification of Sumby and Pollack's formula, which provides a direct measure of the effectiveness of an animated agent relative to that of a natural talker.

[1] Baldi is a registered trademark of Dominic W. Massaro.

3. RESEARCH

Given the potential value of this metric, it is important to demonstrate its validity. A critical assumption underlying the metric is that it remains invariant with differences in unimodal auditory performance (of course, when all other experimental conditions are constant). To test this assumption, we carried out an experiment with different noise levels to modulate baseline performance.

We carried out an expanded factorial experiment with the conditions: (a) unimodal auditory; (b) unimodal visual; (c) bimodal synthetic talker Baldi and (d) bimodal natural talker (the reference). The auditory input was the same natural speech presented in the different conditions. For the synthetic face, the visual phonemes were viterbi aligned and then manually adjusted and corrected. The auditory speech was paired with 5 different noise levels chosen to give 5 difficulty conditions. Participants were asked to recognize 27 consonant-vowel syllables (9 consonant viseme categories times 3 vowels).

Performance improved dramatically with decreases in noise level. As expected, the natural talker gave the best performance, Baldi the second best, and the auditory alone the poorest. The results were used to test Sumby and Pollack's (1954) metric and our new metric for the relative visual contribution. We found that the metrics did not appear to provide sufficient robust measures that were constant across noise level (see also Grant & Walden, 1996).

4. NEW DEVELOPMENTS

One potential limitation of these two metrics is that they do not consider performance based on just the visual information. This is not unreasonable because visual alone trials are not usually tested in experiments of this kind. However, we propose that any robust measure of the benefit of visible speech must include visual only trials. Given that this type of trial was actually included in the present experiment, we can test the value of including the performance on visual alone trials in the evaluation analysis.

The fuzzy logical model of perception (FLMP) assumes that the various speech signals specifying a single event are continuously integrated during categorization, leading to perceptual experience and action. Before integration, however, each source is evaluated (independently of the other source) to determine how much that source supports various alternatives. The integration process 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.

Given this framework for speech perception, the FLMP can be used to assess the visual contribution to speech perception and therefore provide a measure of the relative visual contribution of the synthetic face relative to the natural (Massaro, 1998). If the FLMP gives a good description of the observed results, its parameter values can be used to provide an index of the relative visual contribution. Previous tests of the FLMP did not include both a synthetic and a natural talker, and previous tests of intelligibility as a function of noise level did not include a measure of the intelligibility of visible speech. This study includes these additional conditions, and allows a test of whether its parameter values can be used to provide an index of the relative visual contribution. These metrics and their measures will be presented and discussed.

5. REFERENCES

[1]. Grant, K.W., and Walden, B.E. (1996). Evaluating the articulation index for auditory-visual consonant recognition, Journal of the Acoustical Society of America, 100, 2415-2424.

[2]. Jesse, A., Vrignaud, N., Cohen, M. M., & Massaro, D. W. (2000). The processing of information from multiple sources in simultaneous interpreting. Interpreting, 5(2), 95-115.

[3]. Massaro, D. W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press: Cambridge, MA.

[4]. Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. Journal of Acoustic Society of America, 26, 212-215.

[5]. Summerfield, A. Q. (1979). Use of visual information in phonetic perception. Phonetica, 36, 314-331.

6. ACKNOWLEDGMENT

The research and writing of the paper were supported by the National Science Foundation (Grant No. CDA-9726363, Grant No. BCS-9905176, Grant No. IIS-0086107), Public Health Service (Grant No. PHS R01 DC00236), the Cure Autism Now Foundation , and the University of California, Santa Cruz. Electronic mail may be sent to Slim.Ouni@loria.fr, mmcohen@ranx.ucsc..edu, rishak@ucsc.edu, or massaro@fuzzy.ucsc.edu.