AUDIOVISUAL SPEECH GAZE STRATEGIES

Willy Wong, Astrid Yi and Moshe Eizenman
Dept. of Electrical and Computer Engineering and
Institute of Biomaterials and Biomedical Engineering,
University of Toronto, Toronto, Canada
willy.wong@utoronto.ca

Abstract

The goal of this study was to examine the role of gaze in speech perception and to investigate gaze strategies for listening to speech in noise. Eye tracking was conducted on subjects engaged in a noisy audiovisual speech paradigm. Speech intelligibility was measured for eleven subjects listening to low-context sentences while viewing the talking face on a computer monitor. We found that speech intelligibility was similar for all fixations within 10° of the mouth area. However gaze strategy changed with speech signal-to-noise-ratio. When signal-to-noise-ratio was decreased, the number of gaze fixations to mouth region increased as expected. Other experiments were performed whereby gaze was fixed at different eccentricities and speech intelligibility was measured. These results were compared to results which would be obtained by mapping reduced acuity in the peripheral region to various levels of spatial degradation. Our findings suggest that the visual enhancement of speech occurs when subjects are able to see spatial frequencies of 6 cycles/degree or higher.

In this study we combined eye tracking with audiovisual speech to ask what salient visual information is required for speech perception in noise. Classic studies have shown that a talking face aids in speech intelligibility in noisy environments (Sumby and Pollack, 1954; Macleod and Summerfield, 1987). In our own studies, we have shown that auditory thresholds improves with co-modulated, synchronized visual signals (Luu, 2008; Qian. 2009). More generally, we know that the deaf are able to understand speech at a high level of proficiency on the basis of lipreading. An improved understanding of the role of vision in face-to-face communication can not only enhance our knowledge of how the brain processes speech, but also allows us to better find ways to develop engineering aids to help those with hearing impairment.

Despite the number of studies conducted on audiovisual speech, little is known about the role of vision in speech intelligibility. A study by Grant and Walden (1996) involving filtered speech showed that intelligibility was not affected when subjects were able to view the speaker's face. A similar study by Boothroyd et al (1998) showed that subjects had improved recognition performance in audiovisual speech when listening to speech encoded with only the fundamental frequency. Finally, the McGurk effect has spawned an entire area of study exploring how visual information (or visual information *mismatch*) plays a role in speech comprehension (McGurk and MacDonald, 1976).

Few theories have been proposed to explain how audiovisual enhancement of speech occurs. Most notable is the 'common format' theory which suggest that auditory and visual information transform to a common metric and are processed by cortical neurons responding to both audio and visual speech stimulations. (Calvert et al, 2000) Other studies have shown the co-dependence of the visual and auditory cortices. For example Pekkola et al has demonstrated that speechreading was found to activate the auditory cortex.

Our interest is in uncovering the patterns of gaze during speech perception in noise. Does there exist optimal gaze strategies and how does these strategies change with differing levels of noise? What happens to speech intelligibility when gaze is fixed at increasing viewing angles away from the face? Moreover, what is the role of peripheral vision in processing visual speech information and do experiments on spatial degradation reveal more about mechanism of audiovisual speech perception?

A number of studies on gaze behaviour and audiovisual speech perception have reported varying findings. One study showed that as auditory noise is increased, the number of fixations on the mouth increased. Other studies (Buchan et al., 2007), (Lansing & McConkie, 2003) have shown that a greater number of fixations were made on the nose and the mouth. Nevertheless, the general tendency of these studies suggest that the areas of primary fixation were on the eyes, nose and mouth. However, there hasn't been to our knowledge studies that examine the general role of gaze strategy within audiovisual speech perception.

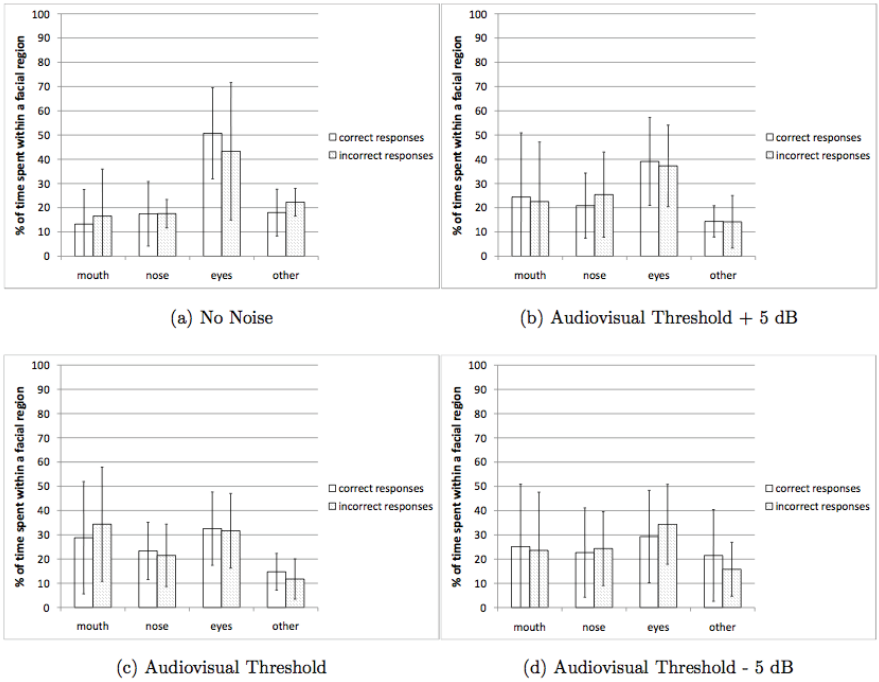
Method

Eleven naive adult subjects (1 female and 10 males) between the ages of 20 to 25 years old participated in the study consisting of 2 experimental sessions. All subjects were fluent in the English language with self-reported normal hearing and normal/corrected to normal visual acuity. The studies were approved by the Office of Research Ethics at the University of Toronto, and all subjects read and signed a consent form prior to the commencement of the research.

Low-context SPIN (Speech Perception in Noise) sentences were chosen from (Kalikow & Stevens, 1977) to minimize the probability that participants can determine the target word (last word) based on the initial words of the sentence. Sentences were spoken by a male talker fluent in the English language and recorded individually using a video camera. The audio stream (the speech signal) was normalized and combined with white noise to produce a new audio stream. The desired auditory SNR level was achieved by varying the amplitude of the speech signal level while maintaining the noise level.

The experimental stimuli were presented on a 19" LCD monitor with audio output through headphones (AKG K301xtra). An advanced remote, non-contact point-of-gaze estimation system requiring only a single point subject calibration routine and consisting of 2 cameras and 4 infrared light sources (Guestrin & Eizenman, 2006) was used to monitor subjects' eye movements. The system provided a point-of-gaze estimation with accuracy of better than 1 degree at a sampling rate of 30 Hz. The experiments were carried out in a quiet laboratory room.

The experiments began with an estimate of the subject's audiovisual speech reception threshold. Within an experimental session, the presentation order of sentences and test conditions were randomized and no sentence was repeated. For the second experimental sessions, subjects were asked to maintain their gaze on a fixation cross, which was placed either at 0˚, 2.5˚, 5˚, 10˚, or 15˚ from the center of the mouth of the talker. The first 3 fixation points (0˚, 2.5˚, 5˚) were chosen to correspond to the primary fixation regions that were found in previous studies (Vatikiotis-Bateson et al., 1998; Buchan et al., 2007; Lansing & McConkie, 2003). These points were mapped to the mouth, nose, and eyes of the talker, respectively. The remaining fixation points (10˚, 15˚) were chosen such that subjects' speech intelligibility could be tested beyond the primary fixation regions. These points were mapped to the top of the hair of the talker and to the top of the computer monitor. All fixation points were vertically aligned with the center of the mouth of the talker.



(a) No Noise (b) Audiovisual Threshold + 5 dB

(c) Audiovisual Threshold (d) Audiovisual Threshold - 5 dB

Fig. 1. Average audiovisual speech intelligibility scores as a function of auditory SNR when subjects gazed naturally at talker.

Subjects were then presented with 4 sets of SPIN sentences which tested their speech intelligibility under 4 different audiovisual conditions: at threshold, threshold – 5 dB, threshold + 5 dB, and no noise. For each of these conditions, subjects were free to look anywhere on the computer monitor while reporting the last word heard after each stimulus presentation. In the second session, subjects completed 5 sets of trials while fixating on a cross that was placed at a specific distance from the center of the mouth (0˚, 2.5˚, 5˚, 10˚, and 15˚).

Results and Discussion

Natural Gaze as a Function of Noise and Time

In this experiment, we measured where a subject looked during a noisy speech perception task. Subjects were free to gaze anywhere during the task and the eye tracker recorded where the person looked. Fig. 1 shows data pooled across different subjects. No discernible pattern is observed for where a person looked as function SNR. The data for both correct and incorrect responses are shown. The tendency for the correct responses to mirror the incorrect responses demonstrates that an incorrect response is not likely to be due to the fact that the person was looking at the wrong region. In terms of the temporal variation of gaze (i.e. duration of stimulus), we reanalyzed the data to determine average gaze (as measured by Euclidean distance to mouth) as a function of time. The results, not shown here, show a clear monotonic convergence of gaze towards the mouth by the end of the sentence for all conditions, although convergence is stronger for noisier conditions (lower SNR).

Natural Gaze Strategies

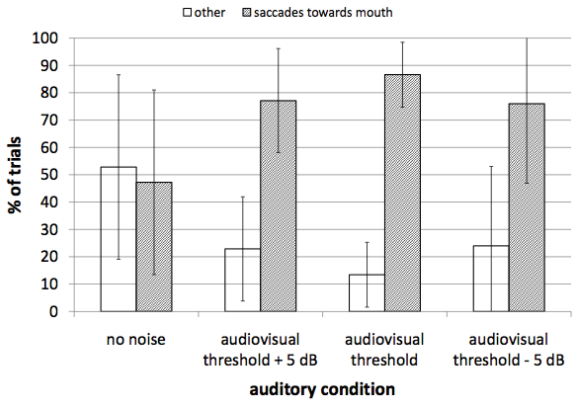


Fig. 2. Gaze strategies comparing the percentage of trials where there were saccades towards the mouth versus no saccades towards the mouth for different SNR conditions.

We also attempted to parse out different gaze strategies used by the subjects. We identified two major strategies. The first strategy entitled "saccades towards mouth" describes the strategy where subjects shifted their gaze from an initial starting point to a region within 2.5˚ of the mouth. 2.5˚ was selected because this encompassed 98% of our data for cases where subjects shifted their gaze to the mouth. The second strategy describes all other situations where the person did not fixate on a region within 2.5˚ of the mouth. The results we obtained (figure 2) show that the tendency is for subjects to gaze at the mouth when speech comprehension becomes difficult. While we would expect monotonic behaviour in terms of increased number of looks towards the mouth as SNR decreased, we could not draw such a conclusion from our data.

Audiovisual Performance Under Constrained Gaze

Very few experiments have been conducted to explore performance of audiovisual speech perception with respect to the proximity of fixation to visual cues. In this experiment, we carried out the same procedure but this time the subjects were asked to fixate on a cross placed either at the center of the mouth (0˚) or at 2.5˚, 5˚, 10˚, or 15˚ relative to the center of the mouth. The results in figure 3 illustrate that audiovisual performance is unchanged when the gaze is within 10˚ of the mouth. This is surprising in that this would indicate that much of the relevant visual information can be obtained by looking at just about anywhere on the face. The result also supports some very basic notions that we have about face-to-face communication -- that most people tend to look at the eyes when communicating with another person. According to these findings, however, looking at the eyes does not imply that the person will then "miss out" on the visual cues of audiovisual speech. They can still get this information provided that they are looking within 10˚ of the mouth.

The Role of Peripheral Vision in Audiovisual Speech

In the periphery, the ability to resolve fine spatial details is limited. Past studies have suggested that peripheral vision may be sufficient for audiovisual speech perception (e.g. Munhall et al, 2004). However, no study has explicitly investigated the role of peripheral vision in audiovisual speech perception, nor has there been a study which compared audiovisual speech perception between the conditions of viewing spatially filtered images with *foveal/parafoveal vision* and viewing unfiltered information with *peripheral vision*.

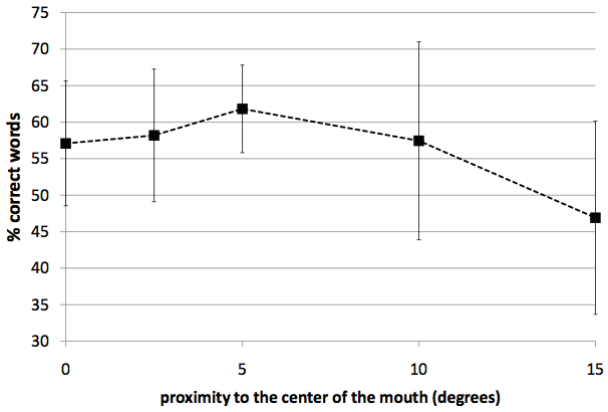


Fig. 3. Recognition performance as a function of fixed proximity of gaze from mouth.

Using a grating visual acuity curve, eccentricities were mapped to levels of spatial degradation. We then compared speech intelligibility performance by using this mapping. The performance was found to be identical (see figure 4). Our findings suggest that when subjects viewed low-pass filtered video recordings, their speech intelligibility was optimal (in terms of visual enhancement of speech perception) so long as they were able to see spatial frequencies below 6 cycles/degree.

Acknowledgements

This research was supported by the National Sciences and Engineering Research Council of Canada. We would also like to thank Eric Dacquay for his assistance with the experiments and with various technical matters in the project.

References

Boothroyd, A., Hnath-Chisolm, T., Hanin, L., & Kishon-Rabin, L. (1988, Dec.). Voice fundamental frequency as an auditory supplement to the speechreading of sentences. Ear and Hearing, 9(6), 316-312.

Buchan, J. N., Par´e, M., & Munhall, K. G. (2007, Mar.). Spatial statistics of gaze fixations during dynamic face processing. Social Neuroscience, 2(1), 1-13.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000, Sept.). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Current Biology, 10, 649-657.

Grant, K. W., & Walden, B. E. (1996, Oct.). Evaluating the articulation index for auditory-visual consonant recognition. Journal of the Acoustical Society of America, 100(4), 2415-2424.

Guestrin, E. D., & Eizenman, M. (2006, Jun.). General theory of remote gaze estimation using the pupil center and corneal reflections. IEEE Transactions on Biomedical Engineering, 53(6), 1124-1133.

Lansing, C. R., & McConkie, G. W. (2003, May). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. Perception & Psychophysics, 65(4), 536-552.

Luu, S. (2008). Visual enhancement of auditory detection in the quiet. Unpublished master's thesis, University of Toronto.

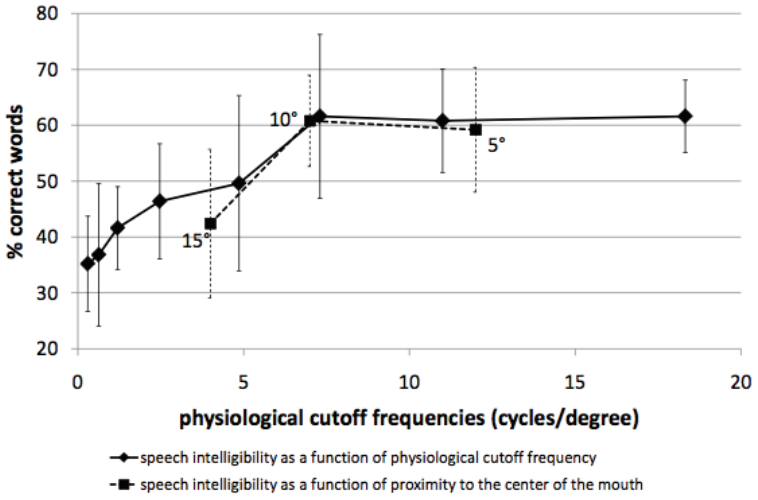


Fig. 4. Comparison of results between spatially-filtered foveal vision versus unfiltered peripheral vision.

McGurk, H., & MacDonald, J. (1976, Dec). Hearing lips and seeing voices. Nature, 264, 746-748.

MacLeod, A., & Summerfield, Q. (1987, May). Quantifying the contribution of vision to speech perception in noise. Journal of Audiology, 21(2), 131-141.

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004, May). Spatial frequency requirements for audiovisual speech perception. Perception & Psychophysics, 66(4), 574-583.

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I. P., Mottonen, R., Tarkiainen, A., et al. (2005, Jan.). Primary auditory cortex activation by visual speech: an fMRI study at 3T. Neuroreport, 16(2), 125-128.

Qian, C. L. (2009). Crossmodal modulation as a basis for visual enhancement of auditory performance. Unpublished master's thesis, University of Toronto.

Sumby, W. H., & Pollack, I. (1954, Mar.). Visual contribution to speech intelligibility in noise. Journal of Acoustical Society of America, 26(2), 212-215.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998, Aug.). Eye movement of perceivers during audiovisual speech perception. Perception & Psychophysics, 60(6), 926-940.