CAN BACKWARDS SPEECH PRODUCE AUDIO-VISUAL FACILITATION?

Jeesun Kim & Chris Davis

Dept. of Psychology, The University of Melbourne

ABSTRACT: We report on an experiment that examines the basis for the facilitation of the detection of speech in noise afforded by seeing the synchronized moving face of the talker (an AV facilitation effect). This work follows up research by Grant & Seitz (2000) and Kim & Davis (in press). Kim and Davis demonstrated that this AV facilitation effect occurred regardless of whether participants knew the language of test. In the current experiment we test to see if the AV facilitation occurs 1. For a computer generated face with synthesised speech and 2. When the AV presentation is played backwards, e.g., (both speech and vision time-reversed). Our findings suggest that the AV facilitation effect only occurs for the time forward natural talker presentation and we discuss these results with respect to Audio-Visual cuing.

In recent years a number of experiments have investigated whether the presentation of the visual speech of a talker (the facial movements associated with articulation) will facilitate detection of their auditory speech when presented masked by noise (e.g., Grant & Seitz, 2000, Grant, 2001; Kim & Davis, in press). Typically, these experiments have employed a two-interval, two alternative forced-choice (2IFC) procedure to determine detection performance of spoken sentences in noise and have shown improved detection thresholds with Audio-Visual (AV) presentation. For example, Grant and Seitz found that average detection thresholds improved in the AV presentation condition by about 1.6 dB relative to an auditory only condition (we will call this an AV advantage). These results are interesting because they suggest that interactions between monosensorial processes can occur very early. That is, the extraction of auditory cues via exposure to the associated visual speech movements can be viewed as a species of early sensory fusion or organization (see Remez, 1996).

In discussing their own recent results demonstrating a similar multimodal enhancement effect, Schwartz, Berthommier and Savariaux (2002) posed the question of whether any kind of audio-visual “comodulation” reducing the spectro-temporal uncertainty would improve AV speech processing. Based on an experiment by Summerfield (1979), Shwartz et al (2002; p 1349) suggested that the “ecological “speech nature” of the visual input could be necessary” for the production of an AV advantage. The current experiment was designed to test whether this is the case by using different types of visual and auditory stimuli and determining if an AV advantage is obtained.

That is, the current experiment will use the critical materials of Grant and Seitz (2000); Grant (2001) but present them in four different conditions that will degrade the ‘naturalness’ of the auditory and visual comodulation. The first condition will consist of a straightforward replication of the Grant and Seitz experiment by using speech presented with either a synchronized moving or still face. The second condition will use the same stimuli only presented time-reversed. Time reversed speech preserves such acoustic information as fundamental frequency, frequency range and speaking rate. Furthermore, time reversed speech should preserve the general correlation between energy in the F2 region and the variation of inter-lip separation thought to be important by Grant and Seitz, (2000). On the other hand, time reversal severely distorts intelligibility and phonological cues (Ramus, Hauser, Miller, Morris & Mehler, 2000; Van Lancker, Kreiman, Emmorey, 1985). The third and fourth conditions follow those of the first two (time normal and time reversed) but use the auditory and visual speech of a virtual talker (Massaro and colleague’s Baldi, e.g., Cohen, Beskow, & Massaro, 1998) Use of simulated auditory and visual speech in which the quality of the natural speech cues are reduced will test whether the production of an AV advantage requires the richness of human speech cues.

In establishing detection performance, the current study will use the method of constant stimuli used by Kim and Davis (in press) rather than the adaptive staircase procedure of Grant and Seitz (2000). This is because an adaptive staircase procedure would have involved presenting the same stimulus multiple times at different signal to noise ratios (SNR) and this may have encouraged participants to learn which parts of an auditory signal were most likely to emerge from the masker (with the 3-up 1-down presentation contingencies acting as error feedback) In the method of constant stimuli all the experimental materials are presented at or near a previously

determined threshold level and an AV advantage will manifest in more accurate classification performance in that condition compared to a non audio-visual condition.

METHOD

Participants

Four participants were tested (one female, three males, mean age of 33 years; 23–42). All were native speakers of English. All participants had normal hearing and normal or corrected-to-normal vision.

Materials and design

The two sentences used by Grant (2001) were employed. These were phonetically balanced low-context sentences selected from IEEE/Harvard (1969) sentence lists: "'Both brothers wear the same size'' and ''Watch the log float in the wide river''.

Video and audio were captured using a Sony TRV 900E digital camera, video at 25 fps and audio at 48000 HZ, 16-bit stereo. The male speaker was positioned 1.5 metres from the camera and recorded against a blank background. Only the lower region of the face (from the bottom of the eyes down) was recorded. The acoustic energy of the phrases was measured for the original unfiltered utterances and also for three spectral regions that correspond broadly to the F1 (100–800 Hz), F2 (800–2000 Hz), and F3 (2200–6500 Hz) formant regions. The rms output from the filtered waveforms was computed in 40 ms intervals to accord with the sampling window of video data. These data were then time aligned with the measures of mouth area obtained by measuring each frame of the video (using Sigmascan software). These data had the same characteristics as those reported by Grant and Seitz (2000).

To determine the SNR at which to present each test stimuli, 75% correct audio-only detection thresholds were calculated for each of the eight selected phrases by adjusting the intensity of the white noise masker. Thresholds were estimated using a 2IFC procedure by an adaptive tracking procedure for two participants. The initial step size in masking noise intensity (digitized 48 kHz, 16-bit white noise) was 3 dB and the final step size 1 dB. Thresholds were calculated as the geometric average of the last 8 of 10 reversals. Final threshold values averaged three separate threshold estimates.

Once the thresholds were determined, ten versions of each phrase (signal-plus-noise) were constructed by dubbing the signal-plus-noise sound track onto the video track using Adobe Premier 6. These trials were the "hard" trials and an additional 10 versions of each trial (easy) were also prepared with the signal being increase by 2 dB as previous experiments suggested that thresholds obtained using an adaptive staircase were lower than those estimated with constant stimuli. A new sample of white noise was generated for every stimulus phrase. The duration of the white noise masker on each trial was the same as the duration of the target phrase plus a random amount that varied between approximately 100 to 200 ms added equally to both the beginning and end of the target phrase. Each experimental item consisted of two intervals, signal-plus-noise and noise-alone or vice versa. For any given item, the same sample of white noise was used for the signal-plus-noise and noise-alone stimuli. In the experiment, all items were presented with both a moving and a still face. The same video files were used for the moving and still face stimuli except that for the still face condition the video was displayed only as a single pixel and a single frame of maximum mouth opening taken from the video of the appropriate phrase displayed at the same time. In all there were 320 trials, 4 presentation conditions (human face: time normal and reversed; virtual taker: time normal and reversed) of 10 moving face and 10 still face presentations and two signal-to-noise levels (easy and hard).

Procedure

The participants were tested individually in a sound attenuated chamber. Stimulus presentation and response collection was controlled by computer (PIII 1000 MHz) using the DMDX software program (Forster & Forster, 1999) that can display synchronized audio and video sequences. The computer was positioned outside the experimental chamber to reduce extraneous noise. Stimuli were presented on a Sony 18" flat screen monitor with the video or still face (a single video frame with maximally lip-opening) subtending approximately 10

degrees of visual angle. The auditory component of the stimuli was presented binaurally over headphones (Sennheiser HD 400) at 60 dBA.

For each trial, first the word "ready" was presented for 800 ms then a signal-plus-noise or noise-alone stimulus followed by an 800 ms gap then the complementary noise-alone or signal-plus-noise stimulus. After this the word "respond" appeared and the participants had to identify the interval containing the target phrase (signal-plus-noise) by pressing one of two numbered buttons. Half the trials began with a signal-plus-noise stimulus and the other half with a noise-only stimulus. For both intervals, half the trials showed a synchronized moving face and the other half a still face. The presentation of items was blocked into stimulus sets the sets of 20 trials of each phrase; within which the presentation of the Moving- and Still-Face trials was at random, as was the order of the signal-plus-noise or noise-alone intervals. The human face trials preceded the virtual talker ones and the easy signal to noise trials preceded the hard ones. Testing lasted approximately 80 minutes and several breaks were included.

RESULTS

As suspected, there was a considerable number of errors made for the hard signal to noise ratio trials, as such, the data presented will only be for the easy trials Figure 1 presents the percentage of detection errors made in the trials with a human talker.

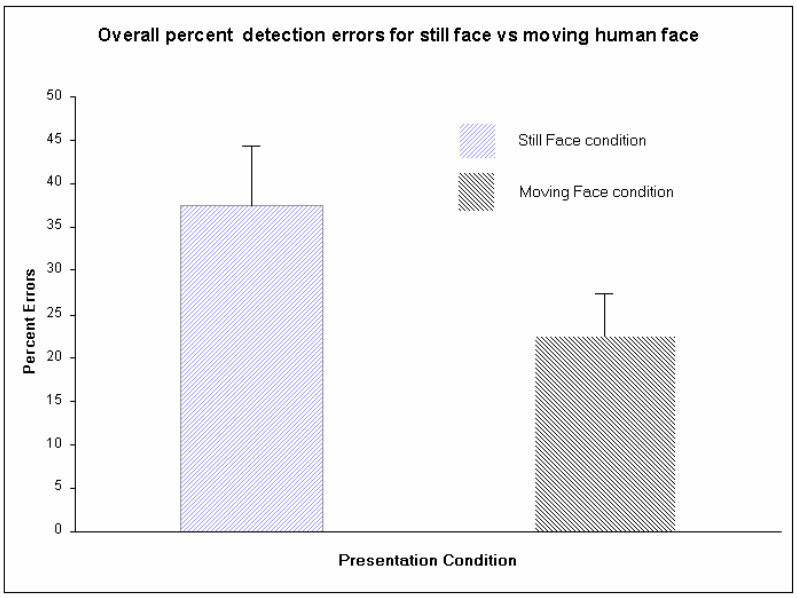

Figure 1. Mean percent 2IFC detection errors for the human talker as a function of the Moving and Still face presentation conditions.

As can be seen, there were fewer detection errors made in the Moving face compared with the Still face condition. This difference was significant, $F(1,18)= 4.836$, $p < 0.05$ and indicates that as expected, there was an AV facilitation effect.

The percentage of detection errors made in the trials with a time reversed human talker is presented in Figure 2. As can be seen from the figure (and consistent with the time normal presentation), there were fewer errors made in the Moving face condition. However, this difference was not significant, $F(1,18)= 0.966$, $p > 0.05$.

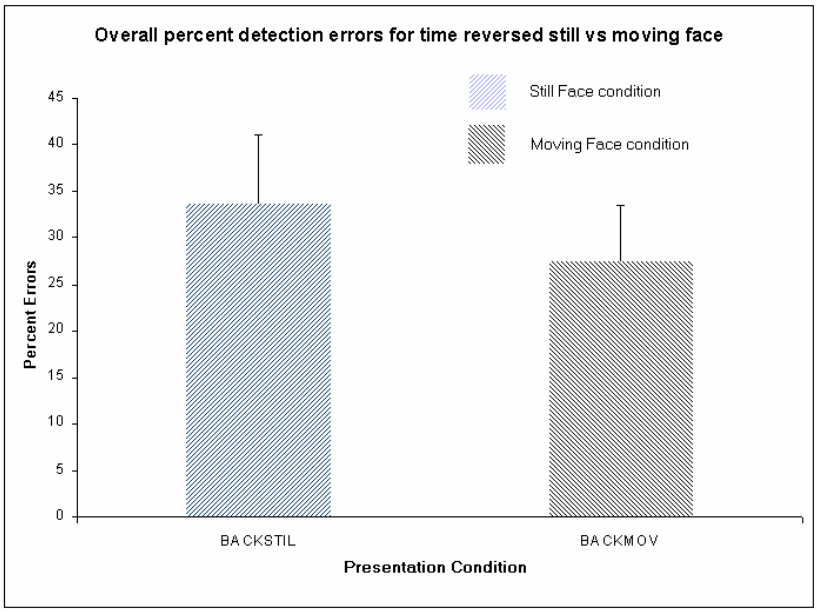

Figure 2. Mean percent 2IFC detection errors for the human talker as a function of the Moving and Still face presentation conditions with time reversed presentation.

The percentage of detection errors made in the trials with the virtual talker is presented in Figure 3. Unlike presentations with the human talker, presentation of the moving face of the virtual talker produced more detection errors than the Still face presentation condition. However, this difference was not significant, $F(1,18) = 0.284$, $p > 0.05$.

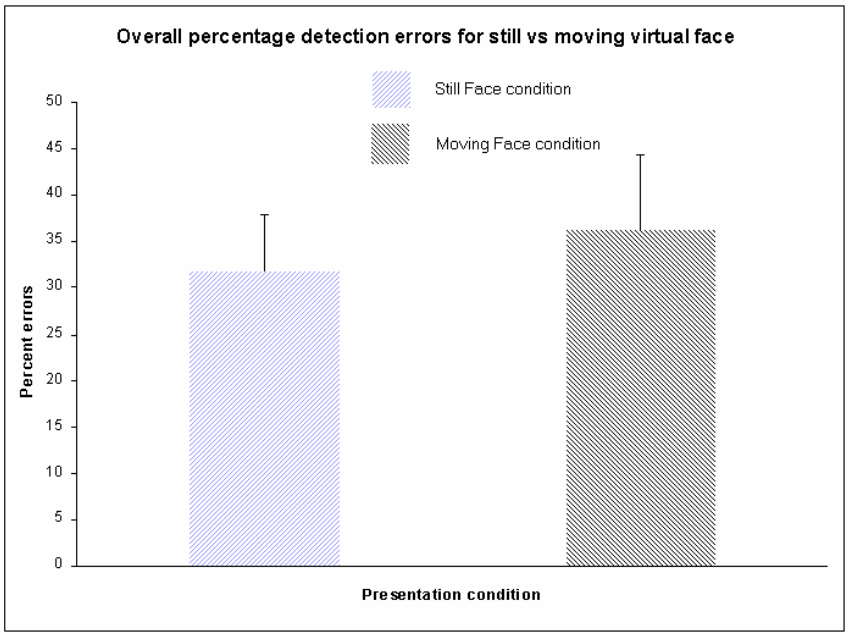

Figure 3. Mean percent 2IFC detection errors for the virtual talker as a function of the Moving and Still face presentation conditions.

Finally, the percentage of detection errors made in the trials with the virtual talker with time-reversed presentation is shown in Figure 4. Although there was a trend in the facilitation direction, the detection advantage for AV presentation was no significant, $F(1,18)= 2.025$, $p > 0.05$.

Figure 4. Mean percent 2IFC detection errors for the virtual talker as a function of the Moving and Still face presentation conditions with time reversed presentation.

DISCUSSION

The paper began by noting that a number of recent papers have reported that visual speech can assist in the detection of auditory speech in noise and posed the question as to the nature of this effect. To investigate this, four different AV presentation conditions were tested. These conditions varied the quality and nature of the cues available in the visual and auditory signals. On the one hand, it may be that an AV advantage requires some specific and precise correlation between auditory and visual speech only available from the presentation of standard speech. However, if the AV detection advantage arises because visual speech simply indicates when increased attention should be paid to the task, then any visual stimulus that did this should produced an AV advantage.

The results confirmed an AV advantage for a normally presented talker, however they also showed that this facilitation effect did not occur for time-reversed presentation or with synthetic visual and auditory speech. This pattern of results suggests that the AV advantage may be generated by a complex of specific speech features rather than simply from those visual cues (such as the overall amplitude of mouth opening) that might signal when maximum attention should be allocated to the detection task. It is, of course, the case that it is extremely difficult to determine the content of time-reversed visual speech (by speech reading) and it may be that knowledge of the spoken phrase is needed to produce AV facilitation. This does not appear to be the case as Kim and Davis (in press) showed that robust AV facilitation occurs for phrase of an unfamiliar language.

The current findings that standard visual presentation (time forward) provides cues to movement that may enable better performance compared to time reversed presentation complement those of Hill and Johnston (2001). Hill and Johnston demonstrated that the accuracy of sex judgments based on non-rigid facial

movements is significantly reduced in time-reversed presentation. Hill and Johnston did not speculate on the precise cues that may be affected by time reversal and likewise, the nature of the key properties of normal speech that determine AV facilitation remain to be established. Grant & Sietz (2000), Grant (2001) and Kim & Davis (in press) have all suggested that it may be the strength of the correlation between energy in the F2 region and variation of mouth area but this property should be the same whether the speech is presented time forward or reversed. Thus the details of the AV facilitation of speech detection remain still to be understood.

REFERENCES

Cohen, M. M., Beskow, J., & Massaro, D. W. (1998). Recent developments in facial animation: An inside view. Proceedings of Auditory Visual Speech Perception '98. (pp. 201-206). Terrigal-Sydney Australia, December, 1998.

Grant, K. W. (2001). The effect of speechreading on masked detection thresholds for filtered speech. *Journal of the Acoustical Society of America*, **109**, 2272-2275.

Grant, K. W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America,* **108**, 1197-1208.

Hill, H. & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology*, **11**, 880-885.

IEEE (**1969**). IEEE recommended practice for speech quality measurements. Institute of Electrical and Electronic Engineers, New York.

Kim, J. & Davis, C. (in press). Hearing foreign voices: does knowing what is said affect masked visual speech detection? *Perception*.

Ramus F, Hauser M. D, Miller C, Morris D, Mehler J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science,* **288**, 349–51.

Remez, R. E. (1996). Perceptual organization of speech in one and several modalities: Common functions, common resource. *Proc. ICSLP'96*, 1660-1663.

Schwartz, J-L., Berthommier, F. & Savariaux, C. (2002). Audio-visual scene analysis Audio-visual scene analysis. Evidence for a "very-early" integration process in audio-visual speech perception.

Summerfield, Q. (1979). Use of Visual Information for Phonetic Perception. *Phonetica*, **36**, 314-331.

Van Lancker D, Kreiman J, Emmorey K. (1985). Familiar voice recognition: patterns and parameters. Part I. Recognition of backward voices. *Journal of Phonetics* 1985, **13**, 19–38.