ISCA Archive
http://www.isca-speech.org/archive

AVSP 2003 – International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Testing the cuing hypothesis for the AV speech detection advantage

Jeesun Kim & Chris Davis

Department of Psychology, The University of Melbourne, Australia
jeesun@unimelb.edu.au; cwd@unimelb.edu.au

Abstract

Seeing the moving face of the talker permits better detection of speech in noise compared to not seeing their face. We report on an experiment that examined the basis of this AV facilitation effect. This work follows up research by [1] and [2] that developed a procedure for demonstrating an AV speech detection effect and [3] that showed that this facilitation occurred regardless of whether participants knew the language of test. In the current experiment we tested to see if AV facilitation occurred because participants were cued to when to pay attention by relatively simply properties of the visual speech of the talker (e.g., when the talker's mouth opened wide). This cuing idea was tested for two types of auditory and visual information that altered the naturalness of speech but maintained many simply cues. The first alteration was to present the AV stimuli backwards, e.g., (both speech and vision played time-reversed). The second used a computer-generated face (Baldi) with synthesised speech. We also tested with a human talker with time-forward presentation. Our findings indicated that AV facilitation only occurred for the time forward human talker presentation; we discuss these results with respect to different types of Audio-Visual cuing.

1. Introduction

A number of recent experiments have shown that the visual speech of a talker (the facial movements associated with articulation) will facilitate detection of their auditory speech when this is presented masked by noise (e.g., [1]; [2]; [3]). Typically, these experiments have employed a two-interval, two alternative forced-choice (2IFC) procedure to determine detection performance of spoken sentences in flat white noise. For example, [1] found that average detection thresholds improved in the AV presentation condition by about 1.6 dB relative to an auditory only condition (an AV advantage). These results are interesting because they suggest that interactions between modalities can occur very early in processing. However, the nature of this interaction is not clear.

One possibility is that any kind of audio-visual comodulation that reduced spectro-temporal uncertainty would improve AV speech processing. However, an experiment by [4] has suggested that the "ecological speech nature" of the visual input could be necessary for the production of an AV advantage (see [5] for a recent discussion of these issues). The current experiment was designed to investigate whether this is the case by using different types of visual and auditory stimuli and determining if an AV advantage is obtained.

The current experiment will use the critical materials of [1] and [2] 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 replication of [1] 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 [1; 2] for the generation of AV facilitation. On the other hand, time reversal severely distorts intelligibility and phonological cues [6; 7]. 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., [8]). 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 full richness of human speech.

In establishing detection performance, the current study will use the method of constant stimuli used by [3] rather than the adaptive staircase procedure of [1; 2]. 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. An AV advantage is demonstrated in terms of there being more accurate classification performance (less errors) in the AV condition compared to an auditory only condition.

2. Method

2.1. Participants

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

2.2. Materials and design

The two sentences used by [2] were employed. These were phonetically balanced low-context sentences selected from IEEE/Harvard [9] sentence lists. Following the procedure in [2] the degree of correspondence between area of mouth opening and rms envelope (for the region 800-2200 Hz) functions for both sentences was calculated. For one sentence “Both brothers wear the same size’’ there was a comparatively low correlation (Low correlation sentence). For the other sentence ‘‘Watch the log float in the wide river’’ the correlation was higher (High correlation) see [1].

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 talker (a native Australian English 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 [1].

To determine the SNR at which to present each test stimuli, 75% correct audio-only detection thresholds were calculated for each of the two 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 “Easy” trials and an additional 10 versions of each trial were also prepared with the noise being increased by 2 dB (Hard trials). These easy and hard versions were prepared 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 taken from the video of the appropriate phrase displayed throughout (this single frame was of the maximum mouth opening). In all there were 320 trials, 4 presentation conditions (Human face: Time Forward and Time Reversed; Virtual talker (Baldi): Time Forward and Time Reversed) of 10 moving face and 10 still face presentations and two signal-to-noise levels (Easy and Hard).

2.3. Procedure

The participants were tested individually in a sound attenuated chamber. Stimulus presentation and response collection were controlled by computer (PIII 1000 MHz) using the DMDX software program [9] 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 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 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.

3. Results

An AV advantage is determined by comparing the number of errors made in the Moving compared to the Still Face conditions. An ANOVA was conducted that examined whether there was an interaction between the AV effect and the different stimulus variables (Human versus Baldi; Time Forward versus Time Reversed speech; High correlation versus Low correlation sentences and the Easy versus Hard noise levels).

There was an interaction between the sizes of the AV speech with Human vs Baldi stimuli, with the AV effect for the human face being significant larger than that obtained from Baldi (F_i (1,144) = 4.21, $p < 0.05$). It was therefore decided to analyse the data from the human and baldi stimuli separately.

For the human face and voice stimuli, the percentage errors for both stimuli as a function of the Moving and Still Face conditions for the Time Forward and Time Reversed conditions are shown in Figure 1.

There was an overall effect of AV speech (F_i (1,72) = 5.27, $p < 0.05$). There was also a strong trend for the high correlation items to produce a larger AV effect than the low correlation ones (F_i (1,72) = 3.66, p = 0.06). Therefore it was decided to analyse the high and low correlation stimuli separately.

For the high correlation items there was a significant AV effect (F_i (1,36) = 8.22, $p < 0.05$). There was also a strong trend for an interaction between the AV effect and the time forward/time backward presentation (F_i (1,36) = 3.33, p = 0.07). The AV effect did not interact with the hard/easy variable (F <1).

Planned contrasts showed that there was an AV effect for the Time Forward condition (F_i (1,18) = 12.36, $p < 0.05$). The planned contrast of the Time Reversed condition revealed that there was no AV effect (F < 1).

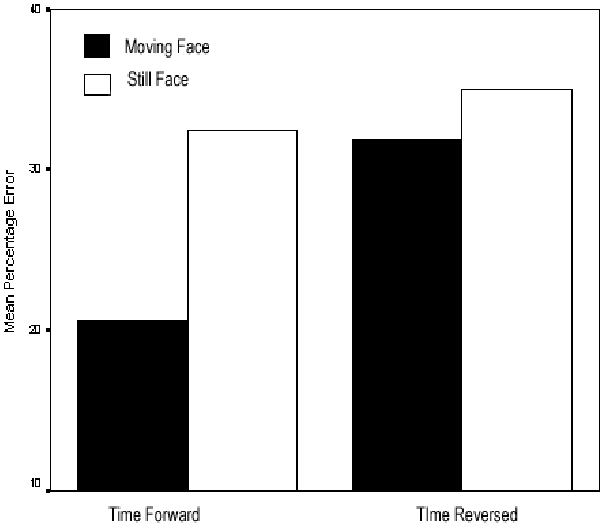

Figure 1. Mean percent 2IFC detection errors for the human talker as a function of the Moving and Still face presentation conditions and Time Forward versus Time Reversed.

There were no significant effects for the low correlation items (F <1).

The same sets of analyses were conducted for the stimuli displayed with Baldi. The percentage errors for both stimuli as a function of the Moving and Still Face conditions for the Time Forward and Time Reversed conditions are shown in Figure 2.

There was no overall AV effect (F < 1). There was no interaction between the AV effect and the high and low correlation items (F < 1).

For consistency with the human analyses, the high and low correlation items were analysed separately. There was no AV effect for the high correlation items (F < 1).

Likewise there was no AV effect for the low correlation condition (F < 1).

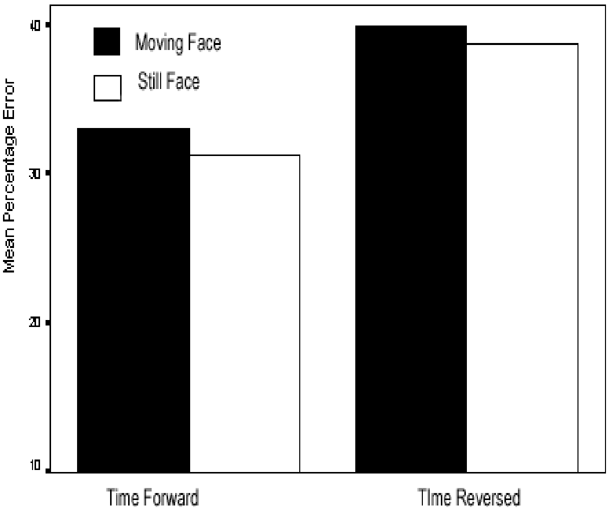

Figure 2. Mean percent 2IFC detection errors for the Baldi as a function of the Moving and Still face presentation conditions and Time Forward versus Time Reversed.

4. Discussion

Four different AV presentation conditions were tested (human face/voice: Time Forward and Time Reversed; Baldi face/Festival speech synthesis voice: Time Forward and Time Reversed). The time reversed and Baldi conditions reduced the naturalness of the visual and auditory signals. However, in itself, time reversal would not have interfered with a simple correlation between energy in the F2 region and variation of mouth area that has been suggested by [1; 2; 3] to be important in generation of the AV effect. Indeed, if the AV detection advantage arose because visual speech simply indicated when increased attention should be paid to the task, then any visual stimulus that had this property should have produced an AV advantage.

The results confirmed an AV advantage for a normally presented talker (at least for the high correlation sentence). There was, however, no AV facilitation effect for time-reversed presentation or for the synthetic visual and auditory speech. This pattern of results suggests that the AV advantage was not due to a simple correlation between the auditory and visual signals that might indicate when maximum attention should be allocated to the detection task. It is, of course, extremely difficult to determine the content of time-reversed visual speech (by speech reading) and one might argue that knowledge of the spoken phrase is needed to produce AV facilitation. However, this does not appear to be the case as [3] showed that robust AV facilitation occurs for phrases presented in an unfamiliar language. Therefore, it is possible that the failure to show an AV advantage for backwards speech reflects the importance of the fine temporal relationships that occur as the speech articulators change rapidly from one gesture to the next.

The current results do not rule out all visual-auditory cuing explanations, for it may be that standard visual presentation (time forward) provides a complex of movement cues that point to detection relevant auditory properties. In this regard, the results complement those of [10] that demonstrated that the accuracy of judgments of the sex of the talker based on non-rigid facial movements was significantly reduced in time-reversed presentation. The precise cues that may be affected by time reversal were not identified; likewise, in the current study, the key properties of normal speech that determine AV facilitation remain to be established. However, it should be noted that the more complex these cues turn out to be, the more they will be specific to speech.

5. References

[1] 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.

[2] 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.

[3] Kim, J. & Davis, C. (2003). Hearing foreign voices: does knowing what is said affect masked visual speech detection? *Perception*, **32**, 111-120.

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

[5] 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. *Proceedings of the 7th ICSLP*, vol. 3, pp. 1937-1940. Denver, Colorado (USA).

[6] 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.

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

[8] 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.

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

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