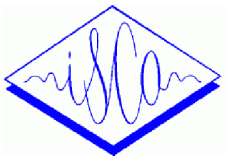
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
(AVSP) 2011
Volterra, Italy
September 1-2, 2011

Audiovisual speech processing in visual speech noise

Jeesun Kim & Chris Davis

MARCS Auditory Laboratories, University of Western Sydney
j.kim@uws.edu.au, chris.davis@uws.edu.au

ABSTRACT

When the talker's face (visual speech) can be seen, speech perception is both facilitated (for congruent visual speech) and interfered with (for incongruent visual speech). The current study investigated whether the degree of these visual speech effects was affected by the presence of an additional irrelevant talking face. In the experiment, auditory speech targets (vCv syllables) were presented in noise for subsequent speech identification. Participants were presented with the full display or upper-half (control) display of a talker's face uttering single syllables either in central vision (Exp 1) or in the visual periphery (Exp 2). In addition, another talker was presented (silently uttering a sentence) either in the periphery (Exp 1) or in central vision (Exp 2). Participants' eye-movements were monitored to ensure that participants always fixated centrally. Congruent AV speech facilitation and incongruent McGurk effects were tested by comparing percent correct syllable identification for full face visual speech stimuli compared to upper-face only conditions. The results showed more accurate identification for congruent stimuli and less accurate responses for incongruent ones (full face condition vs. the upper-half face control). The magnitude of the McGurk effect was greater when the face articulating the syllable was presented in central vision (with visual speech noise in the periphery) than when it was presented in the periphery (with central visual speech noise). The size of the congruent AV speech effect, however, did not differ as a function of central or peripheral presentation.

Keywords: Visual speech; AV congruency; Peripheral visual speech

1. Introduction

Speech communication occurs both with and without interlocutors being able to see each other and also with and without irrelevant auditory and/or visual speech occurring in the background (from here on: auditory/visual noise). When the interlocutor's face (visual speech) can be seen, speech perception occurs based on visual as well as auditory information integration. Auditory-visual speech processing has been shown to lead to better perception of speech particularly when there is auditory noise in the background [1]. The integration of auditory (A) and visual (V) speech information appears to be very robust as it occurs even when auditory and visual speech information is incongruent as in the McGurk effect [2], when the talker's face is not in central vision, [3] and even when a perceiver does not pay direct attention to the AV signals [4].

The current study was conducted to determine whether AV speech integration is affected by the presence of visual speech noise, i.e., the presence of another talker's face. The rationale for this investigation comes from a straightforward comparison with the effects of auditory noise, i.e., since auditory speech perception is degraded when auditory noise is present, it might be that the perception of visual speech will be degraded when visual speech noise is present (although auditory and visual signals propagate differently). Here, we will index any effects of the degradation of the visual speech signal by comparing the size of AV facilitation and interference effects when competing visual speech (visual noise) is and is not present.

The effect of visual speech on auditory speech perception has almost exclusively involved presenting a single target talker's face without visual speech noise. However, there have been a few studies that have examined elements of the current issue concerning the effects of visual speech noise on AV speech perception. For example, as mentioned, AV effects on speech identification still occur even when the talkers face is not in central vision [4]. Further, it has been shown that AV effects can be produced by selectively attending to one of two articulating talkers [5]. However, to date, no experiments have directly examined congruent and incongruent (McGurk) AV effects with a competing talker presented either in central or in the visual periphery.

In this series of experiments we tested if the size of congruent and incongruent visual speech effects varied as a function of where the congruent or incongruent visual speech (the "target talking face") was presented, in addition to where an irrelevant talking face was presented. That is, we examined AV effects by presenting the target talking face either in central vision and the competing talker in the periphery (Experiment 1) or vice versa (Experiment 2). In both cases, we used eye tracking to ensure that participants always fixated centrally. We tested for a congruent AV speech facilitation effect and also for an incongruent AV McGurk effect (visual 'aga', auditory /aba/).

2. Experiment 1

To begin with, we examined whether AV integration effects would be observed when the talking face uttering syllables (which would match or mismatch with the spoken target) was presented in central vision concurrently with other peripheral faces, one of which was also articulating.

2.1. Method

2.1.1. Participants

Eight undergraduate students at the University of Western Sydney took part in the experiment for course credit. They were native speakers of Australian English and all reported normal hearing and normal or corrected-to-normal vision.

2.1.2. Stimuli

The speech materials consisted of a sentence ("The green light in the brown box flickers") articulated by one person (which would be the irrelevant visual speech stimuli in the experiment) and 10 phonemes (/b/, /d/, /f/, /g/, /k/, /l/, /m/, /n/, /p/, /z/) presented in a vCv syllabic context (e.g., /aba/, /ada/,

etc) spoken by another (here visual speech would match the target syllable for congruent trials). The auditory and visual speech material was produced by two native talkers of Australian English who were recorded in a well-lit, sound attenuated room using a Sony TRV 19E digital video camera (25 fps), with audio recorded at 44.1 kHz, 16-bit mono with an externally connected Sony lapel microphone. Multiple repetitions were recorded. For each talker, two tokens of each phoneme were selected so that the durations were similar across phonemes and talkers.

Auditory and visual speech syllable stimuli were selected in order to construct two AV speech conditions: a set of full-face experimental stimuli and a set of upper-half control stimuli. The upper-half face stimuli were used as controls because they presented only limited articulatory speech information but still presented a visual stimulus similar to the AV experimental condition. Stimuli for these AV conditions were generated in the following fashion (video manipulation was done using VirtualDub [6]). First, the movies were rendered to black and white to allow for simpler control of intensity values (these were normalized for all the faces). Next, the auditory and visual streams of videos were separated. Each video stimulus consisted of a "target" talker's moving face (including hairline) that (under experimental viewing conditions) subtended a visual angle of 5.2^0 (width) by 5.7^0 (height). In addition to these experimental stimuli, the videos were also cropped to generate a set of the upper-half face control videos (these showed the talker from above the tip of the nose only). The background colour of all videos was rendered to grey and 6 slightly smaller different static faces (without hairlines) were positioned around the central talker using tailored VirtualDub scripts (see Figure 1).

Figure 1: Examples of the visual stimuli used. The left panel shows a full-face stimulus that pronounced the target syllable; the right panel shows the upper-half face control. Note that the target talker's face in the centre was moving whereas the other 6 surrounding faces were static. In both panels, the single face of the left-side is the irrelevant talking one.

The peak intensity of the auditory speech stimuli were normalized and combined with babble speech at a SNR of -8dB and the resultant auditory stimuli were recombined with the matching visual speech ones (both whole face and control) to form the congruent AV stimuli. In addition, visual /aga/ was combined with auditory /aba/ to create incongruent (McGurk) stimuli. The auditory and visual speech was aligned to maximize the /ada/ percept for the combined token and from this both full- and upper-face stimuli were constructed. (This means that there are two different types of the upper-face /aba/ stimuli: AV congruent and AV incongruent.) There were 92 stimuli in total (including the incongruent stimuli): these consisted of 80 congruent AV stimuli (10 syllables x 2 tokens x 2 talkers x 2 face conditions) and 16 incongruent AV stimuli (2 syllables x 2 tokens x 2 talkers x 2 face conditions). The approximate duration of each stimulus was 1700 ms.

The sentence stimuli were processed to be used as the task irrelevant talking faces (i.e., background visual speech noise). For this, the sound tracks were removed; the movies were rendered to black and white and the intensity values were slightly reduced compared to the face uttering the syllable stimuli. This was to make the visual noise analogous to auditory noise with respect to differences in target/noise SNR (see Figure 1).

2.1.3. Procedure

Participants were tested individually in a quiet room. They were seated with their heads positioned and stabilized by a chin and forehead rest.

Participants were informed that they would be presented a series of spoken disyllables (e.g., /aba/) in babble noise through two loudspeakers (positioned out of sight behind the monitor); that they were required to identify the central consonant in each of the disyllables. They were told that at all times during the stimulus presentation part of a trial they were required to look at the stimuli displayed centrally on the monitor (See Figure 2). After the stimulus presentation, a set of response options appeared (displayed as a column of labeled virtual buttons in central vision) and participants selected a response by clicking one of the buttons using the mouse. Response options consisted of /b/, /d/, /f/, /g/, /k/, /l/, /m/, /n/, /p/, /z/. The experiment was self-paced so that participants were required to press a spacebar to begin each trial (this would start the auditory and visual speech presentation after a gap of approximately 400 ms). Seven practice trials were given.

Visual speech was presented from video clips on a 23″ LCD monitor. The distance from the participant headrest (eyes) to the monitor was 60 cm (see Figure 2). The distance between the centre of the target talker's face (visual fixation point) and the centre of the competing talker's face was 11 cm. This resulted in visual angle of 10.4 degrees (see Figure 2).

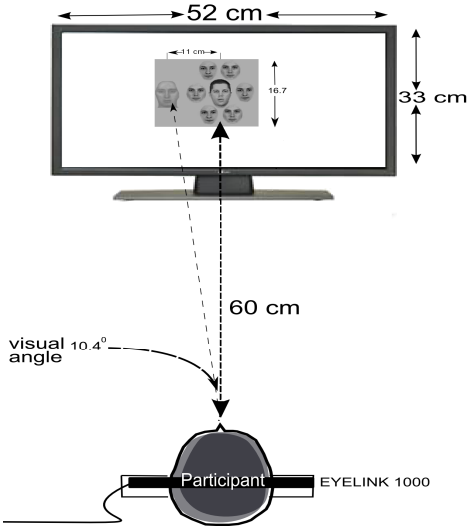


Figure 2: The participant rested his/her chin on a chin rest and forehead against a fitted constraint. In this experiment the face uttering the silent syllables was presented in central vision. For the trial to be valid,

participants had to maintain their gaze within a 3^0 radius of the fixation point.

In the experiment, the 92 stimuli were repeated three times (276 stimuli in total) and the presentation order within each repetition was randomized. An Eyelink 1000 (SR Research, Kanata, Ontario, Canada) was used for eye-tracking (monocularly, right eye, at a sampling rate of 1000 Hz and an average accuracy between 0.25º to 0.5º) and experiment builder for stimuli presentation and data collection. When there was a violation in eye-tracking (when the participant's gaze ventured outside of a virtual circle that was 6^0 visual angle in diameter), the trial was immediately terminated (disappeared) and a large red "X" was presented on the left side of the monitor to alert the participant to what had happened. The terminated trial was added to the rest of the trials in which it was randomly ordered.

2.2. Results & Discussion

Participants correctly identified 66% of all syllables in the AV congruent full-face condition and 53% in the upper face control condition. Figure 3 shows percentage correct identification and (SE) for all the AV congruent syllables; only the AV congruent /aba/ syllables and the incongruent McGurk "aga" (visual) /aba/ (auditory syllables).

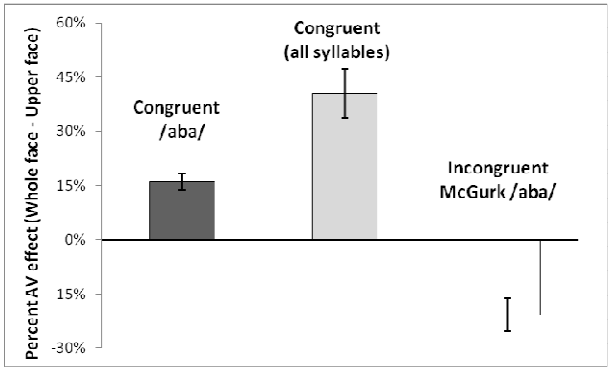


Figure 3: Mean percent AV effect (whole face correct minus upper face correct) for all AV congruent syllable (middle panel), AV congruent /aba/ (right panel) and AV incongruent speech (McGurk) syllables (left panel) in noise (the whiskers show the Standard Error, SE).

A series of two-tailed paired samples t-tests were conducted to determine if the AV effects for all the congruent syllables, the congruent /aba/ syllables and the incongruent syllables were significant. For all the AV congruent stimuli, there was a significant visual speech effect, with phonemes being identified more accurately in full-face condition compared to the upper-face control, $t(7) = 7.17$, $p < 0.05$. If only the /aba/ stimuli are considered, it was found that there was the AV congruent /aba/ stimuli were also better identified in the full-face than in the upper-face condition, $t(7) = 6.00$, $p < 0.05$. For the AV incongruent condition (McGurk) in which the visual speech was not congruent with the auditory speech (/aba/), it was found that the full face lead to the identification of fewer correct phonemes than the upper-face condition, $t(7) = 4.54$, $p < 0.05$.

In sum, the results showed a typical visual-on-auditory speech effects for both congruent and incongruent stimuli in spite of the presence of the task-irrelevant talking face in the periphery. Given that the target visual speech was in central vision it was presumably fully attended and so a robust visual speech effect was expected [5]. The interesting question is whether this will be the case when the irrelevant talking face is in central vision (and presumably attracting attention) while the talking face uttering the matching target syllables is in periphery. This was tested in the following experiment.

2.3. Experiment 2

In this experiment, the target talking face was presented in the visual periphery in order to determine whether the task-irrelevant talking face in central vision would interfere with the AV effects.

Note that in our previous study [7] robust peripheral AV speech effects were found that were unaffected by participants being engaged in a secondary task (attending to the selective presentation of geometric shape that was of a particular colour). Given such a finding, it might be predicted that the presentation of an irrelevant talking face would also not reduce AV effects. On the other hand, the face is a stimulus that has been shown to be very effective in capturing attention [8] and as such an irrelevant talking face in central vision might be more effective in competing with the peripheral target talking face than the non-face stimuli used in [7]. If so, reduced AV effects should be observed.

2.4. Methods

2.4.1. Participants

Eight participants took part in the experiment for course credit at the University of Western Sydney. All were native speakers of Australian English, all reported normal hearing and normal or corrected-to-normal vision and none had participated in Experiment 1.

2.4.2. Materials

The same speech materials were used as in Experiment 1 except that here the face uttering the syllables was in the periphery and the competing talking face silently uttering a sentence was in central vision (see Figure 4).

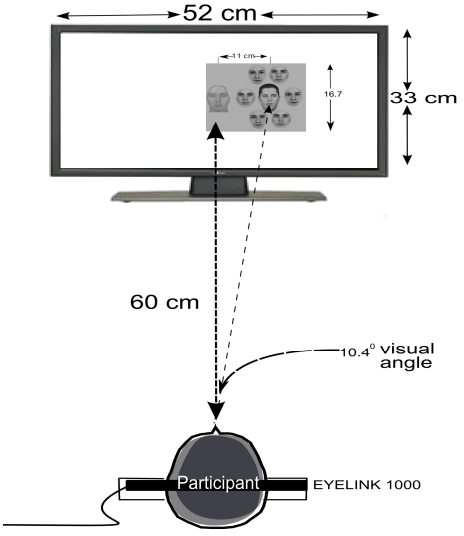


Figure 4. The participant rested his/her chin on a chin rest and forehead against a fitted constraint. In this version the participant viewed the talker silently uttering the sentence in central vision. If the participant glanced away, the trial was invalid.

2.4.3. Procedure

The basic procedure was the same as in Experiment 1 and participants were told that their task was to identify speech phonemes presented in aCa syllables. They were told that at all times during the stimulus presentation they were required to look at the talking face that was displayed centrally on the monitor.

The visual speech that would correspond to the auditory target (on AV congruent trials) was presented in peripheral vision. There were 7 practice trials.

2.5. Results & Discussion

The summary of the performance in phoneme identification in noise is shown in Figure 5. The pattern of the data and the AV effect sizes were similar to the results of Experiment 1. Participants correctly identified 71% of all syllables in the AV congruent full face condition and 65% in the upper face control. For the AV congruent speech overall, there was a significant difference between the whole face and upper face (control) conditions, $t(7) = 2.45$, $p < .05$. Also, the identification of AV congruent /aba/ speech was better in the full face condition compared to the upper-half one, $t(7) = 3.48$, $p < .05$. The AV incongruent speech in the full condition was misidentified significantly more than in the upper-face condition, $t(7) = 3.82$, $p < .05$.

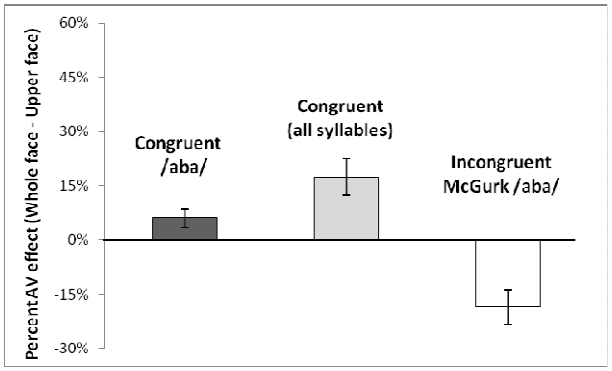


Figure 5: Mean percent AV effect (whole face correct minus upper face correct) for all AV congruent syllable (left panel), AV congruent /aba/ (centre panel) and AV incongruent speech (McGurk) syllables (right panel) in noise (the whiskers show the Standard Error, SE).

A comparison of Figures 3 and 5 shows that the size of the AV effects tended to be greater when the target visual speech was presented in central vision compared to in the periphery. An ANOVA showed that this was the case for the McGurk effect, as it was significantly greater when the video of talker uttering syllables was presented centrally compared to in the periphery, $F(1,22) = 9.87$, $p < .05$. The size of the AV effect for congruent speech did not differ as a function of where the target visual speech was presented, for all syllables, $F(1,22) = 2.22$, $p > .05$, and for the congruent /aba/ syllable, $F < 1$.

The difference in the sensitivity of congruent and incongruent AV speech effects to where the target and the distracting talkers were displayed is intriguing. That there were fewer McGurk responses with peripheral visual speech (and central distracting visual speech) appears to indicate that the perceptual system "recognizes" that the A and V signals are mismatched and under noisy visual speech conditions is less inclined to integrate these signals into a unitary percept. The finding that the congruent AV effect was not affected by the disposition of the target and distracting talkers suggests that when there is visual speech noise AV integration is modulated by an assessment of whether the A and V signals match.

One thing to note is that in the current experiment the position of the target talker and the distracting talker always changed together. This meant that the effects of presenting the target talker in the visual periphery cannot be assessed independently of the position of the distracting talker. In a recent study [7] we measured AV effects in central and peripheral vision without a distracting talker. What was found was that the size of both congruent and incongruent AV effects did not change as a function of the position in which visual speech was displayed. This finding bears on those found in Experiment 2 as it indicates that the McGurk effect itself is not affected by peripheral presentation per se and thus reinforces that idea that the key factor in the reduced McGurks in Experiment 2 was whether other interfering visual speech was present. To properly determine whether congruent and incongruent AV effects differ due the presence or absence of a competing talker requires a within-experiment contrast. This is currently a part of our on-going experimental program.

3. Acknowledgements

The authors wish to thank Tim Paris, Erin Cvejic & Michael Fitzpatrick for their assistance in conducting the experiments and acknowledge support from the Australian Research Council (DP0666857 & TS0669874).

4. References

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

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

[3] Paré, M, Richler, R. C., Ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Percpetion & Psychophysics*, 65, 553-567.

[4] Soto-Faraco, S., Navarra, J. & Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. Cognition, 92, B13–B23.

[5] Hirvenkari, L., Jousmäki, V., Lamminmäki, S., Saarinen, V-M, Sams, M. E., & Hari, R. (2010). Gaze-direction-based MEG averaging during audiovisual speech perception. *Frontiers in Human neuroscience*, 4, 1-7.

[6] Lee, A. (2001). VirtualDub home page. URL: www.virtualdub.org/index.

[7] Kim, J., & Davis, C. (submitted). Auditory speech processing is affected by visual speech in the periphery.

[8] Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, 9, 1-15.