Visual Speech Speeds Up Auditory Identification Responses

Tim Paris, Jeesun Kim, Chris Davis
MARCS Auditory Laboratories, University of Western Sydney, Milperra, Australia
t.paris@uws.edu.au, j.kim@uws.edu.au, chris.davis@uws.edu.au

Abstract

Auditory speech perception is more accurate when combined with visual speech. Recent ERP studies suggest that visual speech helps 'predict' which phoneme will be heard via feedback from visual to auditory areas, with more visual salient articulations associated with greater facilitation. Two experiments tested this hypothesis with a speeded auditory identification measure. Stimuli consisted of ‘the sounds ’apa’, ’aka’ and ’ata’, with matched and mismatched videos that showed the talker’s whole face or upper face (control). The percentage of matched AV videos was set at 85% in Experiment 1 and 15% in Experiment 2. Results showed that responses to matched whole face stimuli were faster than both upper face and mismatched videos in both experiments. Furthermore, salient phonemes (aPa) showed a greater reduction in reaction times than ambiguous ones (aKa). The current study provides support for the proposal that visual speech speeds up processing of auditory speech.

1. Introduction

Seeing the talker’s articulating face (visual speech) facilitates the processing of auditory speech. Yet behavioural studies that have examined the effects of visual speech on speech perception have predominately focussed on speech identification (e.g. [1]) and have overlooked how the timing of processes might be affected. Determining whether visual speech facilitates the speed of speech processing could provide evidence as to how auditory and visual information combine to facilitate speech perception. Indeed, results from recent neurophysiological studies [2,3] indicate that visual speech can directly speed up the cortical responses associated with the processing of auditory speech. The current experiment was designed to determine if there is a behavioural correlate of this effect (i.e., whether visual speech will speed up speech identification responses). The following will set out the details and key results of the neurophysiological studies, and describe the derivation of the design for the current behavioural one.

Electrophysiological measures of cortical activity provide a direct method for probing the time course of neural processing. In this regard, the N100 waveform (an event related potential generated from auditory cortical neurons that occurs approximately 100 ms after an acoustic event) provides a useful index of the activation of the auditory cortex. Research has shown that the addition of visual speech not only reduces the size of the N100 [2,4] but it can also speed up its occurrence [3,5]. While the reduction in the size of the N100 occurs even with mismatched auditory and visual (AV) speech, the speeding up of neural response occurs only for matched AV speech [2]. Moreover, the degree to which the N100 is sped up appears to depend on how well the visual information predicts the acoustic signal. That is, in the study, highly salient lip movements (e.g. ‘pa’) lead to a larger temporal facilitation than less salient movements (e.g. ‘ka’). This effect may derive from the perceivers ability to exploit the temporal order of speech [6], as visual information can be available 150ms prior to any acoustic signal [7].

Given that visual speech speeds up the neural processing of the auditory stimuli, a clear prediction is that there should be concomitant speeding up of identification responses to AV speech stimuli (compared to auditory only ones).

2. Experiment 1

2.1. Method

2.1.1. Participants

Eight postgraduate students from the University of Western Sydney (UWS) voluntarily participated in the study (Mage = 25 years). All were fluent speakers of Australian English and had self reported normal or corrected-to-normal vision and no history of hearing loss.

2.1.2. Materials

The auditory stimuli used in the study consisted of the syllables ‘aKa’, ‘aPa’, and ‘aTa’. These were produced by a single native male speaker of Australian English (Age = 23 years) and were recorded in a well-lit, sound attenuated room using a Sony TRV19E digital video camera (25 fps), with audio captured using an externally connected lapel microphone (44.1 kHz, 16-bit mono). Videos were then processed using VirtualDub [8] to create matched and mismatched AV stimuli, as well as upper-face only control stimuli. For each video, the Acoustic Onset (AO) of the consonant (i.e., ‘p’, ‘k’ or ‘t’) occurred at 1000ms (see Figure 1). Each video was then clipped to 52 frames in length with the initial mouth opening occurring at the 6th frame (240ms). The reason for this was to control for the length of time visual information was available before the acoustic onset.

The three types of auditory stimuli, ‘aKa’, ‘aPa’ and ‘aTa’ (now to be referred to as A1, A2, A3) were the only acoustic stimuli used in the experiment. The matched videos included V1A1, V2A2 and V1A2 and the mismatched videos included V1A3, V2A3 and V2A3. To create the upper face control videos half of the face was removed (from the tip of the nose, down) at the 18th frame (280 ms before the onset of the acoustic consonant). This control was chosen so that participants would be presented with a visual stimulus similar to the AV full face condition but one that provided very little if any articulatory information.

2.1.3. Procedure

Participants were tested individually in a sound-attenuated booth, with stimuli presented on a 17” CRT computer monitor. The experiment was broken into two blocks that were completed on different days to reduce participant fatigue. The order of presentation of the videos for the experiment were pseudo-randomized so that the ratio of matched to mismatched stimuli was constant throughout the experiment. In this experiment, 85% of AV videos had matched AV speech and...
15% had mismatched AV speech. Note that the ratio of matched versus mismatch videos do not include upper-face videos as these did not present any visual articulatory information. In each block a practice set of 15 items established the ratio of matched and mismatched videos before any test trials were recorded.

Each block lasted approximately 40 minutes and consisted of 372 trials; including 276 matched, 24 mismatched and 24 control videos. In each trial a fixation point was presented for 300 ms followed by the target video. Participants were told to listen to the audio and respond by quickly and accurately repeating what they heard. DMDX [9] was used to present videos and record responses. Responses were recorded for 2200 ms after the AO after which the next trial began.

2.2. Results and Discussion

In analysing the data, response times 2 SD above or below the participant’s mean RT for a given condition were eliminated (2 %). Note that response times were measured from the onset of the auditory signal (1000 ms). Incorrect responses1 (3%), including a visual response or a combination of auditory and visual responses were also removed from the analysis.

Mean response time in the baseline (upper face) condition was 211 ms. Responses did not significantly differ between baseline phonemes. On average, participants responded 12 ms faster to matched videos relative to the upper-face baseline, \( t(7) = 4.24, p < 0.05 \).

Figures 2a and 2b show the response times for matched and mismatched videos relative to upper-face baseline. This data was analysed using a repeated measures t-test. For matched \( V_pA_k \) and \( V_pA_t \) videos, responses were significantly faster than baseline (\( t(7) = 7.29, p < 0.05 \) and \( t(7) = 6.56, p < 0.05 \) respectively). Responses to \( V_pA_k \) videos, however were significantly slower than baseline (\( t(7) = 7.76, p < 0.05 \)).

In a test of interactions between the matched phonemes and the difference between AV and control videos, both \( V_pA_k \) (\( F(1,7) = 73.37, p < 0.05 \)) and \( V_pA_t \) (\( F(1,7) = 84.47, p < 0.05 \)) differed from \( V_kA_p \), so that for high salience (\( V_p \)) stimuli, response times relative to baseline slowed significantly more than for less salient (\( V_k \)) stimuli.

Responses to mismatched videos was on average 71 ms slower than baseline (\( t(7) = 5.39, p < 0.05 \)). For each mismatched combination; \( V_pA_k, V_pA_t \) and \( V_kA_p \), responses were significantly slower than baseline (\( t(7) = 7.00, p < 0.05; t(7) = 4.55, p < 0.05; t(7) = 3.35, p < 0.05 \) respectively).

In a test of interactions between the mismatched phonemes and the difference between AV and control videos, Both \( V_pA_k \) (\( F(1,7) = 34.98, p < 0.5 \)) and \( V_pA_t \) (\( F(1,7) = 7.56, p < 0.05 \)) differed from \( V_kA_p \), so that for high salience (\( V_p \)) stimuli, response times relative to baseline slowed significantly more than for less salient (\( V_k \)) stimuli.

In general, the results support the hypothesis that visual speech directly facilitates the speed of auditory processing and so should be apparent in faster responses to matched AV stimuli in a speech identification task. Furthermore, the pattern of temporal facilitation matched what was found in the ERP study [3] with a greater benefit found when the visual speech was salient compared to when it was not.

However, the pattern of results can also be explained in a way that does not require that visual speech directly speeds up auditory processing. That is, it could be that faster responses in the matched AV condition were due simply to participants using visual speech as a cue to what response to make. In other words, participants might have started their responses based on the identification of visual speech. To determine whether this was the case or not, we ran an experiment similar to the above, but this time, the validity of visual speech as a response cue was greatly diminished by increasing the number of mismatched AV stimuli to 85%. If visual speech simply provides a response cue, then the facilitation of AV matched

---

1 In the case of \( V_kA_p \) videos where responses could have been perceived as “fused” McGurk responses, a short pilot study with 4 of the 8 participants was conducted, however no one perceived any fused percepts (e.g. ‘ta’ or ‘kpa’).
responses would not be expected under such low cue validity conditions.

3. Experiment 2

3.1. Method

3.1.1. Participants

The participants were the same as in Experiment 1.

3.1.2. Materials

The same auditory stimulus as used in Experiment 1 was used in this experiment (i.e., aPa, aTa, aKa). However, the number of AV incongruent filler items was increased so that the proportion of the mismatched AV stimuli at any point of the experiment was 85%. This was done by including additional visual stimuli: ‘aHa’, ‘aKa’, ‘aNGa’, ‘aPa’, ‘aSHA’, ‘aTa’, ‘aTHa’, ‘aVa’ and ‘aWa’. These were produced by the same speaker as in Experiment 1. Thus the mismatched AV filler items consisted of V\text{h}Ap, V\text{h}At, V\text{k}At, V\text{ng}At, V\text{sh}Ap, V\text{t}Ak, V\text{th}Ak, V\text{th}Ap, V\text{v}Ak, V\text{w}Ap, V\text{w}At. These items were chosen to create the largest perceived mismatch between auditory and visual tokens.

3.1.3. Procedure

The procedure was the same as in Experiment 1, except that each block consisted of 414 trials; including 54 matched, 306 mismatched and 54 upper-face control videos.

3.2. Results and Discussion

Mean response time in the baseline (upper face) condition was 253 ms. Responses did not significantly differ between baseline phonemes. On average, participants responded 41 ms faster to matched videos relative to the upper-face baseline ($t(7) = 10.358, p < .05$).

Figure 3a and 3b shows the response times for matched and mismatched videos relative to the upper-face baseline. For matched V\text{p}Ap and V\text{t}At videos, responses were significantly faster than baseline ($t(7) = 12.69, p < 0.05$ and $t(7) = 7.94, p < 0.05$ respectively). There was no difference for V\text{k}Ak videos ($t(7) = 0.39, p > 0.05$).

4. General Discussion

The current study followed up the proposal that visual speech speeds up the neural processing of the auditory stimuli by seeking parallel behavioural evidence. The results clearly showed that visual information facilitated response times for matching auditory stimuli relative to controls. Furthermore, this facilitation effect appeared to be moderated by the visual salience of the stimulus as both V\text{p} and V\text{t} showed strong
facilitation effects, whereas \( V_k \) did not. The model proposed by van Wassenhove and colleagues [3] also made hypotheses about mismatched AV stimuli. They proposed that visual speech can also interfere with the processing of auditory speech when the visual prediction is wrong (i.e., mismatched stimuli) with the degree of this interference also moderated by visual salience. In both experiments, highly salient visual articulations (\( V_p, V_t \)) lead to greater interference than did ambiguous (\( V_k \)) stimuli. This suggests that for speeded behavioural responses to AV stimuli, more visual articulatory information leads to greater modulation in processing speeds.

Although visual information both facilitated and interfered with auditory processing, this varied depending on the context in which it was viewed. The facilitation effect for matched stimuli was more than twice as large in the incongruent context. On the other hand, the visual interference for mismatched stimuli decreased almost three fold in incongruent context. This greater change in mismatched interference highlights the possibility that other processes are involved in generating a naming response when a prior prediction from visual speech information is incorrect.

Finally, a possible reason for the discrepancy in mismatched responses between the two salient visemes (\( V_p \) and \( V_t \)) but not in matched conditions may be due to participants mimicry of lip articulation during the experiment. A number of participants self-reported that they moved their lips in conjunction with presented visual speech of the AV stimuli. As participants made lip movements prior to the acoustic onset they were in effect priming their articulation for a certain phoneme. When, however the phoneme did not match the acoustic one, they have to reconfigure their articulation in order to articulate the response. Of course this response articulation effect would not be the same for all phonemes as their place of articulation differed. For example the place of articulation for \( V_p \) is in the front of the mouth and consequently any prepared articulation of ‘pa’ would delay the articulation of ‘ka’ (back placing) relative to ‘ta’ (middle placing). This would explain the discrepancy between \( V_p \) and \( V_t \) responses in the incongruent context, however further studies will need to be done to measure the actual extent of this articulatory response effect on speeded response times.

5. Acknowledgements
The authors thank all the participants and acknowledge support from the Australian Research Council (TS0669874 and DP0666857).

6. References