Auditory speech processing is affected by visual speech in the periphery

Jeesun Kim & Chris Davis

MARCS Auditory Laboratories, University of Western Sydney, Australia

j.kim@uws.edu.au; chris.davis@uws.edu.au

ABSTRACT

Two experiments were conducted to determine whether visual speech presented in the visual periphery affects the perceived identity of speech sounds. Auditory speech targets (vCV syllables) were presented in noise (-8 dB) with congruent or incongruent visual speech presented in full-face or upper-half face conditions. Participants’ eye-movements were monitored to assure that visual speech input occurred only from the periphery. In Experiment 1 participants had only to identify what they heard. The results showed that peripherally presented visual speech (full-face) facilitated identification of AV congruent stimuli compared to the upper-face control. Likewise, visual speech reduced correct identification for the incongruent stimuli. Experiment 2 was the same as the first except that in addition participants performed a central visual task. Once again significant effects of visual speech were found. These results show that peripheral visual speech affects speech recognition.

Index Terms: Visual speech; Auditory-visual speech; Auditory-visual congruency; Visual periphery; Attention

1. INTRODUCTION

When an interlocutor can be seen, speech perception typically involves the integration of both auditory and visual speech information. Determining how signals from different sensory modalities interact in perception remains a central topic in the study of human perception. One method for examining how different signals interact has been to manipulate the circumstances in which such are presented and then determine whether the presentation of one still affects the perception of the other. In the case of speech, the interaction between the auditory (A) and visual (V) signals appears remarkably robust. For instance, the McGurk effect [1] (an illusion in which the visual presentation of a spoken /ga/ with the sound /ba/ leads to the percept /da/) occurs even when the auditory and visual signals are temporarily misaligned (e.g., depending on the task, the visual signal can lead by 250 ms or more [2]). Also the McGurk effect does not require that explicit attention be paid to the visual speech signals and it still takes place even when its occurrence interferes with performance on a concurrent perceptual classification task [3].

The effect of visual speech on auditory speech perception has almost exclusively involved presenting the visual speech signal (the articulating speaker) in central (foveal) vision. However, it is reasonably common for a person to be in a situation in which a talker’s mouth area is not in central vision but where visual speech information would nevertheless be helpful (e.g., in a noisy party). Yet, the extent to which visual speech presented in the periphery will help auditory speech perception is not clear. On the one hand, peripheral vision is less acute than central vision; with the acuity of spatial processing decreasing with increasing visual eccentricity. Reduced acuity functions like a low-pass filter and has the consequence that fine, high spatial frequency information is not encoded. On the other hand, the perception of motion in the periphery is relatively unaffected [4]. This means that visual speech motion information presented in the periphery may be relatively intact even though structural information about the talker’s face may be impaired.

In this study of experiments we examined the robustness of AV speech integration by presenting visual speech in peripheral vision (here we follow [5] in the use of the terms fovea and periphery: foveal vision refers to the central 2 degrees of visual angle and visual periphery refers to regions outside the central 10 degrees). We note that a previous study has examined whether AV speech interactions occur when the talker’s face was presented in peripheral vision [6]. However, in the relevant experiment (where participants looked away from the talker’s face) eye-tracking was not used, so it cannot be assured that participants did not occasionally glimpse the talker in central vision (indeed, faces are much more likely to involuntarily attract gaze than other types of objects [7]).

In the current procedure visual speech was presented in the visual periphery and foveal viewing was prevented by monitoring eye gaze and aborting trials in which participants looked away from a central fixation point. We also reduced the visual salience of the form of the peripherally presented talker by using visual crowding (surrounding it by flanking faces). Crowding has the effect of impairing identification, although crowded objects do not simply disappear but their perceived form becomes grouped and simplified with the flankers [8]. Using other faces as flankers increases the crowding effect [9] and in part simulates a multiperson environment.

AV integration was tested by examining two types of AV effect (facilitation and interference) using a speech perception in noise task. Noise was added to the speech signal in order to be able to test for a congruent AV speech facilitation effect (i.e., the performance level was moved away from ceiling).

2. Experiment 1

2.1. Method

2.1.1. Participants

Eight participants (all native speakers of Australian English) took part in the experiment for course credit at the University of Western Sydney. All reported normal hearing and normal or corrected-to-normal vision.

2.1.2. Stimuli

The speech materials consisted of 10 phonemes (/b/, /d/, /f/, /g/, /k/, /l/, /m/, /n/, /p/, /z/) presented in a vCV syllabic
context (e.g., /aba/, /ada/, etc.). Auditory and visual speech of two native speakers of Australian English 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 speaker, two tokens of each phoneme were selected so that the durations were similar across phonemes and talkers.

Auditory and visual speech 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 [10]). 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° (width) by 5.7° (height). In addition to the experimental stimuli, the videos were also cropped to generate a set of the upper-half face control videos (these showed the speaker from above the tip of the nose only). The background colour of all videos was changed to grey and 6 slightly smaller different static faces (without hairlines) were positioned around the central talker using tailored VirtualDub scripts (see Figure 1).

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. 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.

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 aCa disyllables (e.g., /aba/) in babble noise through two loudspeakers (positioned out of sight behind the monitor); that they were required 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 a fixation point that was 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 in peripheral vision on a Dell 23" LCD monitor. The videos were displayed at a size of 236 x 304 pixels (14 cm horizontal x 16.7 cm vertical). The distance from the participant headrest (eyes) to the monitor was 60 cm. Distance between the centre (visual fixation point) and the centre of the talker's face was 11 cm and this resulted in visual angle of 10.4 degrees (see Figure 2).

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 0.25° to 0.5°) 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° 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 71% of all syllables in the AV congruent full-face condition and 60% in the upper face control condition. Figure 3 presents the size of the visual speech benefit (correct identification in the full-face minus the upper-face conditions in percent) and the standard error (SE) for all the AV congruent syllables; just the AV congruent /aba/ syllables and the incongruent McGurk “aga” (visual) /aba/ (auditory syllables).

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) = 5.96, p = 0.001. This demonstrates that the peripheral presentation of a talker’s articulation of a syllable can influence auditory identification. Consistent with this finding, the AV congruent /aba/ stimuli were also better identified in the full-face than in the upper-face condition, t(7) = 7.03, p = 0.000. For the AV incongruent condition (McGurk), where the visual speech was not congruent with the auditory speech, its provision had a detrimental effect on correct phoneme identification with the number of correctly identified /aba/ stimuli significantly less in the full-face condition compared to the upper-face condition, t(7) = 5.75, p = .001.

In sum, the results showed peripherally induced visual-on-auditory speech effects for both congruent and incongruent stimuli. In Experiment 1 participants were free to direct their attention to the peripheral input, the following experiment was conducted in order to determine whether constraining participant’s attention to the central fixation point would reduce these effects. One possibility is that such a manipulation may reduce the potency of the incongruent (McGurk) effect more than the congruent facilitation effect since it has been suggested that more visual speech information may be required to affect the identification of incongruent auditory speech than to improve the identification of congruent auditory speech [11].

3. Experiment 2

As in Experiment 1, the production of AV speech effects was investigated for visual speech presented in the visual periphery. In this experiment, participants were required to identify the style of the central fixation marker and to only respond when the fixation symbol was of a particular type. This manipulation was employed to encourage participants to pay attention centrally, away from the talking face in periphery. This manipulation may differentially affect the production of AV effects for congruent and incongruent stimuli.

3.1. Methods

3.1.1. 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.

3.1.2. Materials

The same speech materials were used as in Experiment 1.

3.1.3. 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 /Ca syllables. In addition, participants were told that in each trial after they pressed the spacebar, a ‘+’ or ‘x’ would appear in the centre of the monitor (with the target speech played at the same time) and they were instructed that they should respond only to speech when the ‘+’ sign was shown but not if an ‘x’ sign was displayed. In addition to the material outlined in Experiment 1, there were 48 no-go trials (12 AV speech x 2 talkers x 2 face conditions) and these were presented intermixed with the 276 go trials. There were 14 practice trials.

3.2. Results & Discussion

All the participants, except one, adhered to the go/no-go signals with accuracy ranging from 85.4 to 100% (M = 94%). The one participant who did not, made responses on most of the no-go trials. This person’s data was excluded from the analysis; hence the analyses were conducted on the data from 13 participants. The summary of their performance in phoneme identification in noise was presented in Figure 4. The pattern of the data and the AV effect sizes were very similar to the results of Experiment 1. Participants correctly identified 71% of all syllables in the AV congruent full-face condition and 60% in the upper face control. For all the AV congruent speech stimuli considered together, there was a significant difference between the visual speech advantage (percent correct identified in the full-face minus the upper-face condition), t(7) = 7.15, p = .000. Also, the identification of AV congruent /aba/ speech was better in the full-face condition compared to the upper-half one, t(7) = 3.79, p = .007. AV incongruent speech (McGurk) in the full-face condition was misidentified significantly more often than in the upper-face condition, t(7) = 3.12, p = .017.
The results showed that the current secondary task did not affect the production of either the AV congruent speech facilitation effect or AV incongruent speech inhibition effect. One issue with the setup of Experiment 2 is whether the degree of attention manipulation was sufficient to affect the resources allocated to AV processing. That is, although the current secondary task required that some attention be directed at the fixation point, the task was not a particular demanding one. Given the results of Alsius and colleagues [12], the expectation would be that with a secondary task that exhausts participant’s attentional resources AV effects should be reduced and in such a case possibly the production of AV congruent and incongruent effects might have differed. The current results are nevertheless noteworthy in demonstrating that the perception of an auditory presented syllable, to participants who are fixating and identifying a centrally directed at the fixation point, the task was not a particular demanding one. Given the results of Alsius and colleagues [12], the expectation would be that with a secondary task that exhausts participant’s attentional resources AV effects should be reduced and in such a case possibly the production of AV congruent and incongruent effects might have differed. The current results are nevertheless noteworthy in demonstrating that the perception of an auditory presented syllable, to participants who are fixating and identifying a centrally display visual symbol, is influenced by a talker’s visual speech displayed in the periphery.

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

5. Acknowledgements

6. REFERENCES