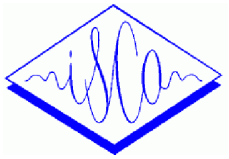
ISCA Archive
http://www.isca-speech.org/archive



Auditory-Visual Speech Processing
(AVSP) 2013
Annecy, France
August 29 - September 1, 2013

# How far out? The effect of peripheral visual speech on speech perception

*Jeesun Kim & Chris Davis*

The MARCS Institute, University of Western Sydney, Australia

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

## ABSTRACT

Seeing the talker's moving face (visual speech) can facilitate or distort auditory speech perception. Our previous study showed that these effects occur even when visual speech was presented in the periphery and participants performed a central visual task. The current study examined the extent to which these effects were modulated by the eccentricity of visual speech: Visual speech presented at a visual angle of 10.40 (Exp 1) and 23.60 (Exp 2). In both experiments spoken /aba/ stimuli were presented in noise (-6 dB) with congruent or incongruent visual speech in full-face or upper-face (baseline) conditions. Other AV vCv syllables were also presented as filler items. Participants were to identify what they heard while performing a central visual task with their eye-movements monitored. Congruent visual speech facilitated speech perception; incongruent interfered. The sizes of the visual speech effects were smaller for the more eccentric presentation but were still significant. We discuss these results in terms of the form and timing cues that visual speech provides for incoming auditory speech and the robustness of the speech processes that use these cues.

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

## 1. INTRODUCTION

The current study examined whether unattended and degraded visual speech information would affect speech perception in noise. This question stems from a consideration of multi-person interactions where it is reasonably common that one hears speech in background babble from an unattended speaker whose face is within the visual periphery.

It is well established that seeing a talker's face and head movements (visual speech) within central vision facilitates speech perception in noise [1;2]. The influence of visual speech even occurs when the visual (V) signal mismatches the auditory (A) one, e.g., the visual presentation of a spoken /ga/ paired with the sound /ba/ leads to the percept /da/ (the McGurk effect, [3]).

These AV speech effects appear remarkably robust, occurring even when the auditory and visual signals are temporally misaligned [4] or when explicit attention is not paid to the visual speech signals, e.g., due to the demands of performing a concurrent perceptual classification task [5]. The robustness of the effects of visual on auditory speech perception suggests that these would occur even when the source of visual speech was not in central vision. Indeed, there is evidence that visual speech in the periphery can affect speech perception [6, 7]. This finding is consistent with the view that visual speech can affect speech perception even if it has only low temporal/spatial resolution [8,9]. This phenomenon is clearly demonstrated by the finding that AV speech effects occur with degraded point-light displays (e.g., [10]).

The clearest evidence that unattended peripheral visual speech can affect speech perception comes from the findings of Kim and Davis [11]. In this study participants were required to identify a spoken syllable in babble noise while visual speech was presented in the periphery (with participants' eye movements tracked to ensure this); participants also had to attend to a secondary visual task in fovea. This secondary task consisted of the presentation of "+" and "x" symbols at the central fixation point that occurred only as the auditory speech was played and participants were instructed to monitor which symbol was presented and to respond to trials only when they saw the + symbol. The results showed a facilitation effect for AV congruent speech and an interference effect for AV incongruent speech compared to the control conditions (where the visual speech showed only the upper part of the talker's face). The results were similar regardless of whether participants performed the secondary visual task or not, suggesting the attention directed at the fixation point did not diminish the effect of visual speech.

The current study aimed to extend the findings of Kim and Davis [11] by adjusting two aspects of the study. First, we manipulated the eccentricity at which visual speech was presented. In a previous study [6], it was suggested that the influence of visual speech effect started to decline when visual speech was presented beyond an eccentricity of between 10º–20º (there was however still an effect up to 60º). Based on this, in the current study we examined two presentation eccentricities, 10.4º (used in the earlier study) and a more eccentric one of 23.6º. Given the results of [6], a visual speech effect is still expected at the further eccentricity but it may be reduced compared to the smaller one.

The second addition to the original study [11] was to make the secondary task more attention demanding. The original task (described above) was not a particularly demanding one as it only required that participants notice the difference between a + and an x symbol. This raises a question of whether the degree of attention required was sufficient to affect the resources allocated to AV processing. With visual speech presented in central vision it has been shown that a secondary task that exhausts participant's attentional resources can reduce the strength AV effects [5]. Given this, it is an empirical question whether with a more attention demanding task, visual speech presented in the periphery would still exert an effect.

Two experiments were conducted to test these extensions to Kim and Davis [11]; these used the same basic paradigm but modified the relevant display properties. Thus in Experiment 1, a more attention demanded secondary task was implemented and in Experiment 2 (that used the same attention task) the eccentricity of visual speech was increased.

Both congruent and incongruent AV effects were tested by presenting auditory /aba/ with a talker's face uttering "aba" or

"aga" with the effect of this examined by determining how the concurrently presented sound was perceived. In addition to these two key stimulus conditions, a set of AV congruent filler items (consisting of other consonants in an /a/ vowel context, i.e., vCv syllables) was also included. This was done to encourage participants to more fully analyze the target by increasing the range of options in the identification task and so that the majority of stimuli presented were in the customary AV congruent configuration.

## 2. Experiment 1

The participants' task was speech identification in noise and although visual speech was presented in the periphery, no comments were made about the presence of these stimuli. In addition to the speech identification task, on each trial participants also were required to monitor a sequence of geometric figures presented at the central fixation point and auditory speech identification responses were not to be made if participants saw a particular combination of shape and colour (e.g., a green triangle). This manipulation was employed so that participants needed to pay attention to the central display (away from the talking face in periphery).

This task is more attention demanding than the task used in Kim and Davis [11] as detecting the target (e.g., a green triangle) requires combining colour and shape; there were non-green triangles and green non-triangles. So the research question is whether increasing the attention manipulation would prevent AV speech processing.

### 2.1. Method

#### *2.1.1. Participants*

Fifteen 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/, /z/) presented in a vCv syllablic 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 consisted 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 [13]). First, the movies were rendered to grey-scale 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 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).

Six slightly smaller different static faces (without hairlines) were positioned around the central talker using tailored VirtualDub scripts (see Figure 1). These still faces were to added to produce visual crowding, a phenomenon that has the effect of regularizing the appearance of the array in the visual periphery by making the appearance of adjacent objects more consistent [12]. Crowding created an overall percept of face-like objects in the visual periphery while avoiding the attention capture that an isolated peripheral face might induce. Note that hairlines of the additional still faces were removed to increase the face crowding effect.

The peak intensity of the auditory speech stimuli were normalized and combined with babble speech at a SNR of -6dB 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 88 stimuli in total (including the incongruent stimuli): these consisted of 72 congruent AV stimuli (9 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.

**Figure 1:** Examples of the visual stimuli used. The left panel shows a full-face stimulus; the right panel the upper-face control. Note that the target talker's face in the centre was moving whereas the other faces were static.

In addition, 35 visual icons (5 geometric figures, i.e., circle, triangle, square, star, and pentagon in 7 different colours) were prepared for the secondary task. Using these icons, 20 visual stimuli were constructed. Each visual stimulus presented 7 different visual icons in sequence for 1750 ms (with each icon lasting for 200 ms with 50 ms inter-stimulus-interval).

#### *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 a fixation point that was displayed on the monitor (See Figure 2); in each trial after they pressed

spacebar, a series of visual symbols appeared (at the same time as the participant heard the spoken target) and if one of the symbols was a green triangle, participants were told that they should not make a response to the phoneme. There were 48 no-go trials and these were presented intermixed with the 264 go trials. There were 14 practice trials.

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/, /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 256 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 88 stimuli were repeated three times (264 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.

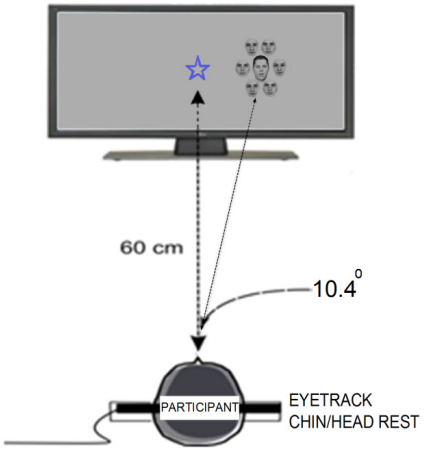



**Figure 2:** The participant rested his/her chin on a chin rest and forehead against a fitted constraint. For the trial to be valid, participants had to maintain their gaze at the fixation point.

## 2.2. Results & Discussion

The summary of participant's phoneme identification performance is presented in Figure 3. Note that in this and the following experiment, data of the participants who made responses on any of the no-go trials were not included in the analysis (i.e., none of the 15 participants whose data is reported here responded to a no-go trial).

As can be seen from the Figure 3, there was an AV facilitation effect from seeing a talker's whole face presented in the periphery: congruent /aba/ stimuli were better identified in the full-face than in the upper-face baseline condition, $t(14) = 3.51$, $p < 0.05$. There was also a McGurk effect for AV incongruent stimuli, i.e., there were significantly fewer correct /aba/ responses in the full-face condition compared to the upper-face condition, $t(14) = 3.12$, $p < 0.05$.

The research question tested in Experiment 1 was whether increasing the attention manipulation of Kim and Davis [11] would prevent AV speech processing. The results showed that the current secondary task did not prevent the production of either the AV congruent facilitation effect or the AV incongruent one. These significant AV speech effects were similar in the size to those found by Kim and Davis [11] a find that suggests that such AV processing makes little demand on attentional resources. The next question to be investigated was whether these effects would still occur when the visual speech was presented at a greater eccentricity. This was tested in Experiment 2.

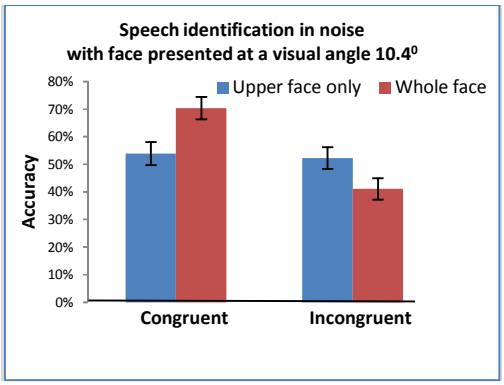



**Figure 3:** Mean percent AV effect (whole face correct minus upper face correct) for all AV congruent syllable (left panel), AV congruent /aba/ (centre panel) and AV incongruent (McGurk) syllables (right panel) in noise (the whiskers show one Standard Error, SE).

# 3. Experiment 2

As in Experiment 1, the production of AV speech effects was investigated with visual speech presented in the visual periphery while participants performed a secondary visual task. In this experiment, the eccentricity of the presented visual speech was increased from 10.4º to 23.6º.

## 3.1. Methods

### *3.1.1. Participants*

Fifteen 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 except that visual speech was presented at a visual angle 23.6º. As in Experiment 1, participants were told that their task was to identify spoken phonemes presented in aCa syllables while at the same time monitoring a stream of visual symbols presented in central vision. If a green triangle was presented participants were told to withhold their response to the spoken target. Once again, no comments were made about the display of visual speech. There were 14 practice trials.

## 3.2. Results & Discussion

The results were summarized in Figure 4. Overall, the pattern of the results was very similar to but the AV effect sizes tended to be smaller than those of Experiment 1.

More specifically, there was a significant AV facilitation effect for congruent /aba/ stimuli, with better identification in the full-face than in the upper-face baseline condition, $F(1,14) = 6.4$; $p < .05$; $\eta p^2 = .31$. There was also a McGurk effect for AV incongruent stimuli, with significantly fewer correct /aba/ responses in the full-face condition compared to the upper-face condition, $F(1,14) = 5.5$; $p < .05$; $\eta p^2 = .28$. These results clearly demonstrated that visual speech (that was not a focus of attention) presented even at this increased eccentricity still produce effects on speech perception.

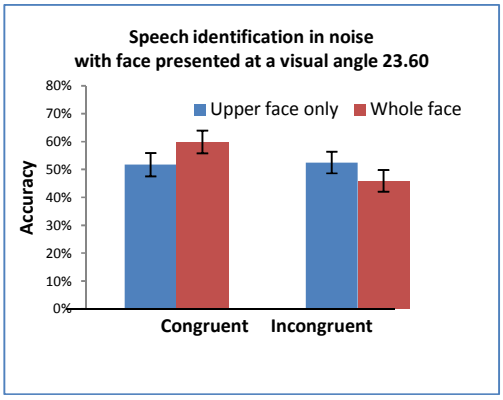



**Figure 4:** Mean percent AV effect (whole face correct minus upper face correct) and (SE) for all AV congruent syllable, AV congruent /aba/ and AV incongruent speech (McGurk) syllables in noise.

The results of the two experiments were contrasted by using an ANOVA with visual angle as non-repeated factor. The visual speech effect for congruent /aba/ stimuli was significant, $F(1,28) = 18.73$, $p < .001$, $\eta p^2 = .40$; the effect of visual angle was not significant, $F(1,28) = 1.57$, $p > .05$. The interaction did not reach significance, $F(1,28) = 2.15$, $p > .05$. For incongruent stimuli, similar results were found. That is, the visual speech effect was significant, $F(1,28) = 15.24$, $p < .001$, $\eta p^2 = .35$; the effect was not significant different across the visual angles, $F < 1$, or an interaction, $F(1,28) = 1.01$, $p > .05$.

These results suggested that the production of AV effects was not differentially affected by an increase in the eccentricity of the display from 10.4º to 23.6º. These findings are consistent with those of Paré et al [6].

# 4. General Discussion

The current results demonstrate the impressive range over which the processing of visible speech can influence speech processing. Not only did visual speech presented in the periphery both facilitate and interfere with auditory speech processing but it did so even while a participant was engaged in an attention demanding visual task.

We are not claiming that AV speech effects occur independently of attention, the study by Alsius and colleagues [5] indicates that they do not. What we suggest is that (at least for the current paradigm) visual and auditory speech processing are intimately connected and that this connection requires minimal attention resources. Further, we propose that this connection is based upon auditory and visual speech cues that are salient in the processing environment.

That is, the auditory target stimuli were presented in mild babble noise and under these circumstances the presentation of visual speech likely provides timing cues that can reduce any uncertainly as to when an utterance begins [14]. Such a timing cue might explain facilitatory effects of visual speech but seem unlikely to explain interference from visual speech (which was also observed). However, as MacDonald and colleagues [15] have shown, the McGurk effect can occur with coarse-spatial-scale visible speech information. Thus the same information that provides a timing cue (the visual bilabial clapper) may also be sufficient to bias auditory perception.

Of course, the current results may in part be determined by AV features being made salient by the experimental setup (a proposal similar to of Fujisaki and Nishida [16] in relation to the perception of AV synchrony). In this regard, the current paradigm had only one moving face presented and thus there was no ambiguity in the assignment of auditory and visual speech. It would be interesting to see what occurs when there are multiple visual speech signals present.

# 5. Acknowledgements

The authors would like to thank Leo Chong and Tim Paris for their assistance in preparing the stimuli and collecting the data, and acknowledge support from the Australian Research Council (DP130104447).

# 6. REFERENCES

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

[2] Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. The Quarterly Journal of Experimental Psychology, 57A, 1103-1121.

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

[4] Grant, K. W., Greenberg, S., Poeppel, D., & van Wassenhove, V. (2004). Effects of spectro-temporal

asynchrony in auditory and auditory-visual speech processing. Seminars in Hearing, 25, 241-255.

[5] Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch reduces audiovisual speech integration. Experimantal Brain Research, 183, 399-404.

[6] Paré, M, Richler, R. C., Ten Hove, M., & Munhall, K. G. (2003). Perception & Psychophysics, 65, 553-567.

[7] Krause, H., Schneider, T. R., Engel, A. K., & Senkowski, D. (2012). Capture of visual attention interferes with multisensory speech processing. Frontiers in Integrative Neuroscience, 6.

[8] Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirementsfor audiovisual speech perception. Perception & Psychophysics, 66, 574-583.

[9] Vitkovich, M., & Barber, P. (1996). Visible speech as a function of image quality: Effects of display parameters on lipreading ability. Applied Cognitive Psychology, 10, 121-140.

[10] Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Pointlight facial displays enhance comprehension of speech in noise. Journal of Speech & Hearing Research, 39, 1159-1170.

[11] Kim, J., & Davis, C. (2011). Auditory speech processing is affected by visual speech in the periphery. In P. Cosi, R. De Mori, G. Di Fabbrizio, R. Pieraccini (Eds.), Proceedings of Interspeech 2011 (pp. 2465-2468).

[12] Whitney, D. &, Levi D. M. (2011). Visual crowding: a fundamental limit on conscious perception and object recognition. Trends in Cognitive Sciences, 15, 160 - 168.

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

[14] Davis, C., & Kim, J. (2013). The effect of visual speech timing and form cues on the processing of speech and nonspeech. 14th InterSpeech conference, Lyon, France, 2013.

[15] MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: How much spatial degradation can be tolerated? Perception, 29, 1155-1168.

[16] Fujisaki, W., & Nishida, S. (2007). Feature-based processing of audio-visual synchrony perception revealed by random pulse trains. Vision research, 47, 1075-1093.