BRILL

Seeing and Perceiving 25 (2012) 87–106

SEEING AND PERCEIVING

brill.nl/sp

The Effect of a Concurrent Working Memory Task and Temporal Offsets on the Integration of Auditory and Visual Speech Information *

Julie N. Buchan [1,**] **and Kevin G. Munhall** [1,2]

[1] Department of Psychology, Queen's University, 62 Arch St., Kingston, Ontario, K7L 3N6 Canada
[2] Department of Otolaryngology, Queen's University, Kingston, Ontario, Canada

Received 14 January 2011; accepted 11 October 2011

Abstract
Audiovisual speech perception is an everyday occurrence of multisensory integration. Conflicting visual speech information can influence the perception of acoustic speech (namely the McGurk effect), and auditory and visual speech are integrated over a rather wide range of temporal offsets. This research examined whether the addition of a concurrent cognitive load task would affect the audiovisual integration in a McGurk speech task and whether the cognitive load task would cause more interference at increasing offsets. The amount of integration was measured by the proportion of responses in incongruent trials that did not correspond to the audio (McGurk response). An eye-tracker was also used to examine whether the amount of temporal offset and the presence of a concurrent cognitive load task would influence gaze behavior. Results from this experiment show a very modest but statistically significant decrease in the number of McGurk responses when subjects also perform a cognitive load task, and that this effect is relatively constant across the various temporal offsets. Participant's gaze behavior was also influenced by the addition of a cognitive load task. Gaze was less centralized on the face, less time was spent looking at the mouth and more time was spent looking at the eyes, when a concurrent cognitive load task was added to the speech task.

Keywords
Audiovisual speech, multisensory integration, McGurk effect, eye tracking, cognitive load, working memory

1. Introduction

Speech perception is an example of how we integrate information from different senses in our everyday lives. It has been known for some time that visual speech

* This article is part of the Multisensorial Perception collection, guest edited by S. Wuerger, D. Alais and M. Gondan.

** To whom correspondence should be addressed. E-mail: julie.buchan@queensu.ca

 DOI:10.1163/187847611X620937

information can influence the perception of acoustic speech. The presence of visual speech information in acoustically degraded conditions can increase the intelligibility of acoustic speech (Erber, 1969; Ross *et al*., 2007; Sumby and Pollack, 1954). Visual speech information can also be perceptually useful when the acoustic information has not been degraded. For example, visual speech information can increase the intelligibility of difficult to understand clear speech (Reisberg *et al*., 1987). Presenting conflicting visual information can influence the perception of auditory speech (i.e., the McGurk effect: McGurk and MacDonald, 1976; Summerfield and McGrath, 1984). This auditory and visual speech can be integrated, causing an illusory percept of a sound not present in the actual acoustic speech.

The McGurk effect has been used extensively in the literature to study the integration of auditory and visual speech information. Several lines of evidence suggest that this integration may be largely automatic. Young infants seem to be susceptible to the McGurk effect (Rosenblum *et al*., 1997), and informing participants about the mismatch between the auditory and visual stimuli does not seem to influence the McGurk effect (Liberman, 1982). Further, there does not seem to be a reaction time cost in terms of processing the illusory McGurk percept as compared to the actual acoustic speech token in a speeded classification task (Soto-Faraco *et al*., 2004). The actual speech token or McGurk percept can equally provide a benefit or interfere with a concurrent syllable categorization task (using a syllabic interference paradigm) depending on the perceived syllable and not the actual auditory syllable. On the other hand, attention may play a role in the integration of auditory and visual speech information. For example, paying attention to concurrent irrelevant visual, auditory (Alsius *et al*., 2005) and tactile (Alsius *et al*., 2007) stimuli has been shown to reduce the amount of audiovisual integration in a McGurk task. A reduction in audiovisual integration has also been shown by getting participants to direct visual attention to either a face or a concurrently presented leaf on a screen (Tiippana *et al*., 2004). Directing visual spatial attention to one of two simultaneously presented talkers (Andersen *et al*., 2009) has also been shown to influence the perception of the McGurk effect.

The above studies suggesting that attentional resources play a role in the integration of auditory and visual speech information all had concurrent competing perceptual information which had to be either monitored or ignored during the speech task. These studies suggest that challenging attentional resources to selectively attend or ignore competing perceptual information can influence how audiovisual speech information is perceived. However, what is not yet understood is the extent to which these attentional influences are due to interference with gathering the perceptual speech information, possibly due to attentional capture by information from the competing non-speech task.

The current paper seeks to extend the findings of the small literature on attentional influences on audiovisual speech perception by having participants perform a concurrent working memory task as a cognitive load task. The cognitive demands

of the task are overlapped, but the stimuli for each task are not presented at the same time; the perceptual information for the speech task is presented without other experimentally relevant competing perceptual information. A speech task was presented either alone (speech-only condition) or was placed within a cognitive load task (speech-numbers condition). In the speech-numbers condition the numbers to be memorized in the cognitive load task were presented before the speech task, and participants had to make their response to the cognitive load task after the speech task. Additionally, the cognitive load task was also presented alone for comparison (numbers-only condition).

The integration of auditory and visual speech information occurs not just for synchronous speech stimuli, but occurs over a range of temporal asynchronies. The tolerance for asynchrony, or 'synchrony window', tends to be asymmetric with a greater tolerance for visual stimuli leading the auditory stimuli, rather than *vice versa*, and the integration of auditory and visual information tends to decrease as the amount of asynchrony is increased (Conrey and Pisoni, 2006; Dixon and Spitz, 1980; Grant *et al*., 2004; Munhall *et al*., 1996; van Wassenhove *et al*., 2007). It is not yet known whether taxing cognitive resources will have an effect on the duration of the synchrony window for speech. It is possible that this synchrony window will become narrower as cognitive resources are taxed, causing less integration to be shown at more extreme offsets. Such a finding would imply that asynchronous integration was more costly in terms of cognitive resources. In this study, the audio and video were either synchronous (0 ms) or asynchronous. Two video-leading asynchronies were chosen where the video led the audio by 175 or 350 ms.

Finally, this research will examine whether the concurrent cognitive load task alters the gaze behavior used to gather the visual speech information. Previous research looking at gaze behavior during audiovisual speech perception found that gaze tended to become more centralized on the face, clustering around the nose, when moderate acoustic noise was added to a sentence comprehension task as compared to when participants heard the sentences without noise (Buchan *et al*., 2007, 2008). This occurred, despite the visual stimuli being the same when acoustic noise was present and when it was absent during the speech task. Since the visual stimuli were the same, changes in visual stimulus properties could not be responsible for this change in gaze behavior (Parkhurst and Neibur, 2003; Parkhurst *et al*., 2002). The reason for this change in gaze behavior with addition of acoustic noise is not understood. One possibility is that when acoustic noise is present the speech task becomes more cognitively demanding, and it is the cognitive demands of the task driving this gaze centralization on the face. The centralization of gaze behavior will be examined by looking at the average distance of the eye-tracker samples from the centre of the nose. Eye tracking data will be compared for the speech tasks with and without the concurrent cognitive load task. Eye tracking data will also be compared across the audiovisual offsets to examine whether the offset between the auditory and visual speech stimuli has an effect on gaze behavior.

2. Methods

All procedures were approved by Queen's University's General Research Ethics Board.

2.1. Participants

Participants were native English speakers and reported having normal or corrected to normal vision, and no speech or hearing difficulties. Written consent was obtained from each participant. There were 25 participants (22 females) in Experiment 1 with a mean age of 21.4 years (range 19–35), and 25 participants (20 females) in Experiment 2 with a mean age of 20.5 years (range 18–30).

2.2. Stimuli

For both experiments, a male volunteer was used as the talker, and was filmed in color saying vowel–consonant–vowel nonsense syllables. The video was edited into clips in Final Cut Pro.

2.2.1. Experiment 1

In Experiment 1 the syllables /aba/, /ada/, /atha/, /ava/, /ibi/, /idi/, /ithi/ and /ivi/ were used. The syllables /aba/ and /ibi/ were used for the congruent stimuli. The congruent stimuli had the auditory /aba/ paired with the visual /aba/, and the auditory /ibi/ paired with the visual /ibi/. The congruent stimuli were used to ensure that participants were performing the speech task correctly and to serve as a baseline to contrast with the McGurk effect. The incongruent stimuli were created to elicit the McGurk effect by dubbing the auditory syllable /aba/ onto the videos of the syllables /ada/, /atha/ and /ava/, and the auditory syllable /ibi/ onto the videos for /idi, /ithi/ and /ivi/ using custom MATLAB software. To maintain the timing with the original soundtrack, the approximate acoustic releases of the consonants in the dubbed syllables were aligned to the acoustic releases of the consonants in the original acoustic syllable.

2.2.2. Experiment 2

In Experiment 2 the syllables /aba/, /ava/, /ibi/ and /ivi/ were used. For congruent stimuli the auditory /aba/, /ava/, /ibi/ and /ivi/ were paired with the visual /aba/, /ava/, /ibi/ and /ivi/, respectively. For incongruent stimuli, an acoustic syllable with the same vowel as the one articulated in the video but different auditory consonant was dubbed onto the video using custom MATLAB software. That is, the /aba/ and /ava/ syllables were paired with one another, and could each be either the visual or the auditory token. The /ibi/ and /ivi/ syllables were paired with one another, and could each be either the visual or the auditory token. That is, an auditory /aba/ was paired with a visual /ava/, an auditory /ava/ was paired with a visual /aba/, an auditory /ibi/ was paired with a visual /ivi/, and an auditory /ivi/ was paired with a visual /ibi/. The acoustic syllable of each member of the pair was dubbed onto the video of the other member of the pair to create incongruent stimuli. As in Experiment 1, to maintain the timing with the original soundtrack, the approximate

acoustic releases of the consonants of the dubbed syllables were aligned to the acoustic releases of the consonants in the original acoustic syllable.

2.2.3. Temporal Offsets

In both experiments, the strength of the McGurk effect (as measured by the proportion of responses that do not correspond to the auditory token) was manipulated by varying the temporal offsets of the auditory and visual streams. Video-leading asynchronies were chosen since they tend to be more naturalistic, and show a greater asynchrony tolerance for video-leading speech stimuli than for auditory-leading speech. While the influence of the visual information at a 350 ms offset tends not to be as strong as when the audio and video are synchronous, the influence of the video on the auditory token in a McGurk task has been shown in several studies to extend out to rather large video-leading offsets: around 30% of responses still do not correspond to the auditory token. For example, Jones and Jarrick (2006) have shown that a 360 ms offset still produced about 45% non-auditory token responses, Munhall *et al*. (1996) have shown that a 360 ms offset produced about 30–40% non-auditory token responses. Grant *et al*. (2004) and van Wassenhove *et al*. (2007) have also found that there is still a noticeable influence of the visual token on response between 333 and 467 ms, with about 30–40% of the responses corresponding to non-auditory token responses. The offsets were created using custom MATLAB software. To create the 175 and 350 ms offsets, the onset of the syllable was selected, and then offset so that the audio trailed the video by either 175 or 350 ms. The beginning of the audio track was zero padded to make the audio and video of equal duration.

2.3. Experimental Task

The experiments were both carried out as a within-subjects design. There were two tasks, a speech task and a numbers task.

2.3.1. Speech Task

The speech task involved watching and listening to the talker say a syllable and choose which consonant sound they heard. In Experiment 1 the response choices were ‘B’, ‘D’, ‘TH’, ‘V’ and ‘other’. In Experiment 2 the response choices were ‘B’ and ‘V’. Participants were informed that they had to wait until the video was finished to respond, and that their key press would not be recorded until after the video was finished playing.

2.3.2. Numbers Task

For the numbers task participants were presented with a random set of eight digits from the digits 0–9 at the beginning of each trial. The numbers were randomized without replacement. A new set of eight digits was generated every trial. The numbers were presented sequentially. Each number was on the screen for 550 ms. After the last digit, two masker screens with greyscale ‘noise’ were presented for 550 ms each to reduce an afterimage of the last digit during the presentation of the video. Participants were asked to remember the order of the digits, and at the end of the

trial were presented with one digit from the set, and asked to respond on the keyboard which digit came after that digit in the series. If the digit happened to be at the end of the series, then participants were to report the number that came at the beginning of the series. The digit remained on the screen until participants made their response. Participants were asked to try and be as accurate as they possibly could be with the numbers task, and they were instructed that it might be helpful if they rehearsed the numbers silently to themselves.

2.4. Experimental Conditions

These two tasks were used to create three experimental conditions: (1) the speech-only condition where participants were just given the speech task, (2) the numbers-only condition where participants were just given the numbers task and (3) the speech-numbers condition where participant were given the speech task sandwiched between the numbers task. In the speech-numbers condition participants were first presented with the digit series from the numbers task, then the speech stimulus was presented. After the speech stimulus was presented, participants responded to the speech task, then were presented with a digit from the numbers task and made their response to the numbers task.

2.5. Experimental Equipment

The experiments took place in a single walled sound booth and participants were seated approximately 57 cm away from a 22-inch flat CRT computer monitor (ViewSonic P220f). Participants' heads were stabilized with a chin rest. The audio signal was played from speakers (Paradigm Reference Studio/20) positioned on either side of the monitor. Eye position was monitored with an Eyelink II eye tracking system (SR Research, Osgoode, Canada) (see Eyetracking analysis for further details).

2.6. Speech Task Analysis

Congruent and incongruent trials were analyzed separately. Congruent trials were measured using proportion of trials correct. Incongruent trials were measured using proportion of trials showing the McGurk effect. Trials in which participants reported hearing a consonant sound other than the one present in the audio file were considered to show the McGurk effect. Participant responses to the speech task were analyzed using a 2×3 (task conditions containing a speech task $\times$ temporal offsets) repeated measures ANOVA. In instances where there was a violation of sphericity, a Greenhouse–Geisser correction was used. Participant responses to the numbers task were analyzed with a repeated measures ANOVA. Pairwise comparisons were done with paired samples t-tests with Bonferroni corrections used for multiple comparisons.

2.7. Numbers Task Analysis

Participant responses to the numbers task was analyzed using an ANOVA. Performance on the numbers task in the numbers-only condition was compared directly

with performance on the numbers task for each of the audiovisual offsets in the speech task in the speech-numbers task. Pairwise comparisons were done with paired samples t-tests with Bonferroni corrections used for multiple comparisons.

2.8. Eyetracking Analysis

Eye tracking data was analyzed for the two conditions with speech tasks (the speech-only and the speech-numbers conditions). Eye tracking data from one participant in Experiment 1 was not collected due to equipment problems. Eye position was monitored using an Eyelink II eye tracking system (SR Research, Osgoode, Canada) using dark pupil tracking with a sampling rate of 500 Hz. Each sample contains an x and y coordinate which corresponds to the location of gaze on the screen. A nine-point calibration and validation procedure was used. The maximum average error was 1.0 visual degree, and maximum error on a single point was 1.2 visual degrees with the exception of the central point which was always less than 1.0°. A drift correction was performed before each trial.

Four of the videos (/aba/, /ada/, /atha/ and /ava/) had been used in a previous experiment where the positions of eyes, nose and mouth in each frame had been coded. The videos for /ibi/, /idi/, /ithi/ and /ivi/ show very similar head position and movement but had not been coded. For each experiment, the position of the eyes, nose and mouth in each trial were estimated based on average position of the eyes, nose and mouth the /a*a/ syllables. To further describe the eyetracking data, and allow for comparison with other experiments in the literature, the overall proportion of samples in each experiment falling within both 4 and 10° of visual angle of the nose are reported.

Based on the coding of eyes, nose and mouth, three analyses of the eyetracking data were performed. The first gaze analysis was a gaze centralization analysis where the average distance of the eyetracking samples from the centre of the nose was calculated for each trial (containing a video of the talker) for each participant, for each task condition and offset. Paré *et al*. (2003) showed similar gaze patterns for congruent and incongruent trials, so gaze for congruent and incongruent trials were pooled together. The average distance of the eye-tracking samples from the centre of the nose was analyzed using a 2×3 (task conditions containing a speech task $\times$ temporal offsets) repeated measures ANOVA. In instances where there was a violation of sphericity, a Greenhouse–Geisser correction was used. The second and third gaze analyses looked at the proportion of each trial that participants spent looking at the eyes and the mouth, respectively. Based on previous coding of the videos mentioned in the preceding paragraph, boxes 3.1° of visual angle on the x axis by 2.5° of visual angle on the y axis were centered around the average position of each eye. A box 5° of visual angle on the x axis, and 3.1° of visual angle on the y axis was positioned around the centre of the leftmost, rightmost, topmost and bottommost boundaries of the mouth. The box around the mouth was large enough to contain the maximal mouth movements in the coded videos (see Fig. 1). For each of the proportion of the trial spent looking at the eyes, and the proportion of

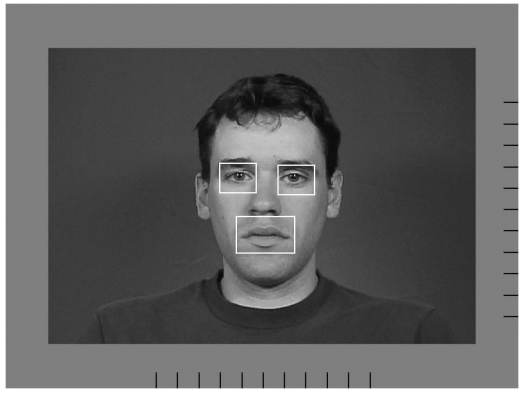


Figure 1. The white boxes illustrate the regions used for the proportion of the trial spent looking at the eyes and the mouth. The space between each black line around the edge is approximately 2° of visual angle. The screen subtended approximately 45° of visual angle along the horizontal, and the video of the talker subtended approximately 40° of visual angle.

the trial spent looking at the mouth, a 2 × 3 (task conditions containing a speech task × temporal offsets) repeated measures ANOVA was performed. In instances where there was a violation of sphericity, a Greenhouse–Geisser correction was used. Pairwise comparisons were done with paired samples t-tests with Bonferroni corrections used for multiple comparisons.

3. Results

3.1. Behavioral Data

3.1.1. Experiment 1

Performance on the speech task was compared between the speech-only condition and the speech-numbers task. Performance was very high, with a proportion of at least 0.89 correct, for the congruent trials in the speech task in both the speech-only and speech-numbers task (see Fig. 2A). The proportion of correct responses in the congruent trials was not affected by the concurrent cognitive load task ($p > 0.05$), although performance in the congruent trials was somewhat affected by offset ($F(1.56, 37.40) = 4.54,\ p = 0.025$). While this influence was statistically significant, the extent of the influence was rather subtle.

In the incongruent trials in the speech task, the proportion of trials showing the McGurk effect was lower in the speech-numbers task than it was in the speech-only task ($F(1, 24) = 14.07,\ p = 0.001$). The difference in proportions of McGurk responses between the speech-only task and the speech-numbers task was rather modest, ranging from 0.035–0.067 (see Fig. 2B). As expected, the proportion of trials showing the McGurk effect was affected by offset ($F(1.13, 27.21) = 13.67$, $p = 0.001$), with fewer McGurk responses as the offset between the audio and video was increased. There was no significant interaction between the task condition and

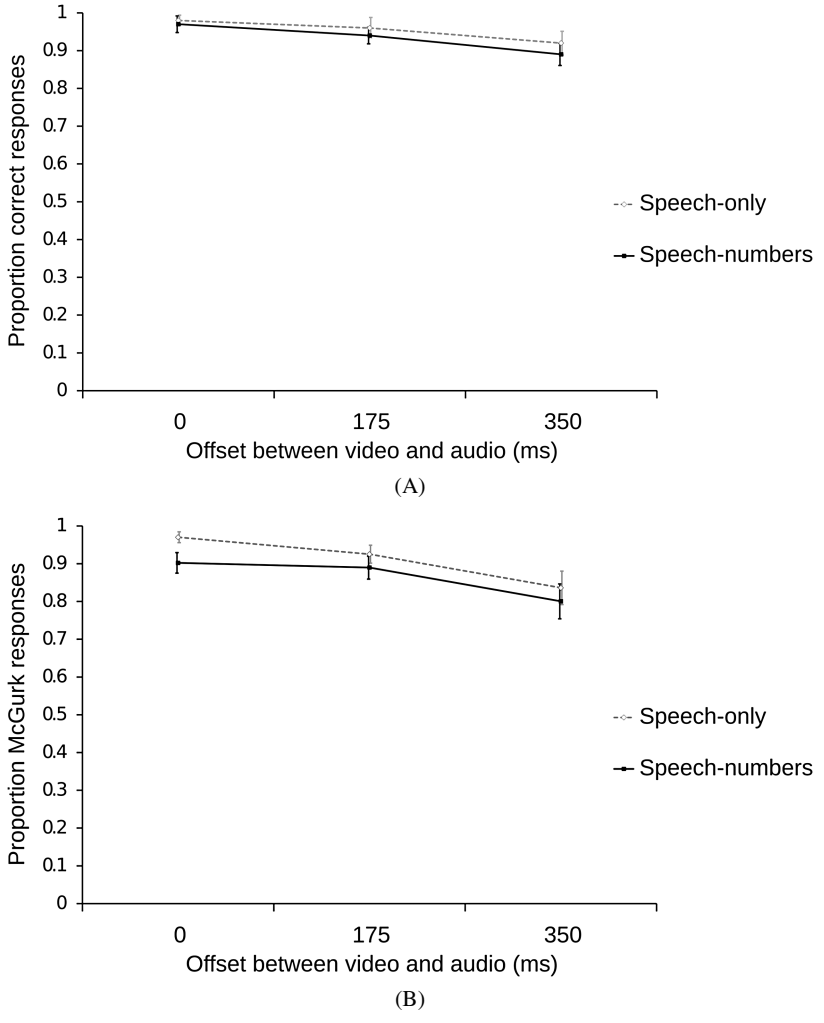


Figure 2. Responses to the speech task in Experiment 1. (A) shows the proportion of correct responses for the congruent trials, and (B) shows the proportion of McGurk responses for the incongruent trials. The error bars indicate standard errors of the mean.

offset ($p > 0.05$). The presence of a concurrent speech task had an effect on performance on the numbers task. Not surprisingly, performance on the numbers task was a bit higher when participants did not have to do the concurrent speech task (see Fig. 3A). Performance on the numbers task was higher in the numbers-only condition than in 0 ms speech offset ($t(24) = 4.33,\ p = 0.001$), 175 ms offset ($t(24) = 4.84,\ p < 0.001$) and 350 ms offset ($t(24) = 8.96,\ p < 0.001$) in the speech-numbers task. On average, the proportion of correct responses was about 0.189 higher when participants only had to do the numbers task. There was no difference in performance on the numbers task across the speech offsets in the speech-numbers condition ($p > 0.05$).

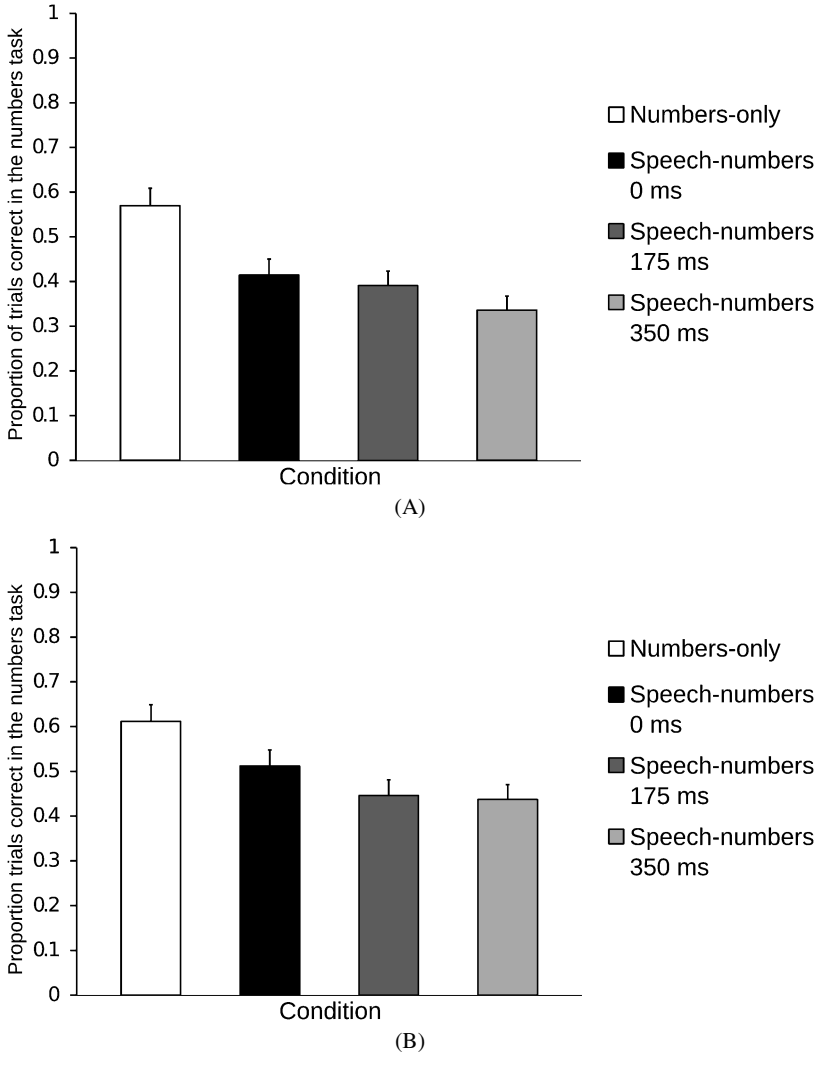


Figure 3. Performance on the concurrent working memory task (the numbers task) by condition. (A) shows the results for Experiment 1 and (B) shows the results for Experiment 2. The error bars indicate standard errors of the mean.

The concurrent numbers task did have an effect on the proportion of trials showing the McGurk effect in the speech task, but this effect was very modest. The auditory and visual speech stimuli used in this task were very conducive to integration, as evidenced by the high proportion of trials showing McGurk responses. The modest effect of the concurrent cognitive load task on the speech task could be due to the fact that the speech stimuli seem to elicit a strong McGurk effect. While the increasing offset did weaken the McGurk effect, it was still quite strong at 350 ms. The influence of the visual information at 350 ms was more pronounced than in

some other studies (e.g., Conrey and Pisoni, 2006; Grant *et al*., 2004; Jones and Jarrick, 2006). This could be due to the particular talker used in the experiment, who in previous experiments with no offsets has consistently produced very strong McGurk effects (Buchan *et al*., 2005; Wilson *et al*., 2008), and the particular speech tokens used. For instance, both talker and speech token can influence the number of McGurk responses (Paré *et al*., 2003). Would stimuli and response combinations that produced a weaker McGurk effect as the offset is increased show more interference from the cognitive load task? In Experiment 2, stimuli were used from another experiment that had been shown to produce a strong McGurk effect when the audio and video were aligned, but become much weaker at the 350 ms.

3.1.2. Experiment 2

In Experiment 2 participants were very good at discriminating the consonant sound in the congruent trials. Performance on the congruent trials was at ceiling in all conditions (see Fig. 4A). Unlike Experiment 1 there was no influence of offset on performance in the congruent trials ($p > 0.05$). For the incongruent trials, there was an obvious influence of offset ($F(1.37, 32.97) = 47.12,\ p < 0.001$), with the proportion of McGurk responses dropping from an average of 0.89 at 0 ms, to 0.80 at 175 ms, to 0.56 at 350 ms (see Fig. 4B). This is not surprising since the particular stimuli were chosen because they show considerably less audiovisual integration as offset is increased. As in Experiment 1, there was an effect of the cognitive load task ($F(1, 24) = 8.05,\ p = 0.009$), with slightly more McGurk responses in the speech-only task compared with the speech-numbers task. The difference in proportion of McGurk responses between the speech-only task and speech-numbers task ranged between 0.040 and 0.081 across offsets. Like Experiment 1, there was no significant interaction between the task condition and audiovisual offset ($p > 0.05$). While the overall pattern of performance on the numbers task in Experiment 2 was similar to Experiment 1 (see Fig. 3B), the difference at the 0 ms speech offset between the numbers-only task and the speech-numbers task was not significant ($p > 0.05$). Performance on the numbers task was significantly different between the numbers-only condition and the 175 ms and 350 ms speech offsets in the speech-numbers task ($t(24) = 3.53,\ p = 0.011$) and ($t(24) = 4.86,\ p < 0.001$), respectively. The average difference in the proportion of correct responses on the numbers task between numbers-only condition and the 175 and 350 ms offsets was 0.170. The 175 ms speech offset and the 350 ms speech offset in the speech-numbers were not significantly different from one another ($p > 0.05$). The proportion of correct trials for the numbers task at the 0 ms offset was not significantly different from the 175 or 350 ms offset ($p > 0.05$).

3.2. Gaze Behavior

3.2.1. Experiment 1

Overall, eye-tracking samples tended to fall quite close to the centre of the nose. Approximately 0.68 of all eye-tracker samples in Experiment 1 fell within 4° of visual angle from the centre of the nose. Most samples fell either on the face, or

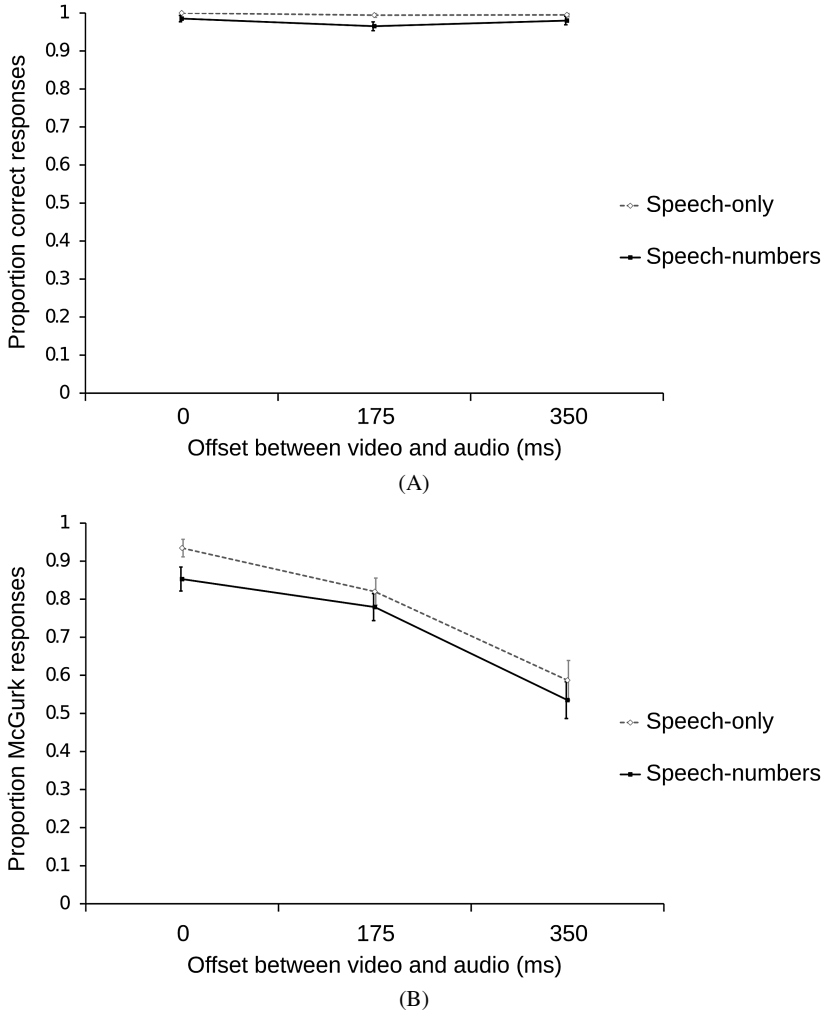


Figure 4. Responses to the speech task in Experiment 2. (A) shows the proportion of correct responses for the congruent trials, and (B) shows the proportion of McGurk responses for the incongruent trials. The error bars indicate standard errors of the mean.

very close to the face. Approximately 0.90 of eye-tracker samples in Experiment 1 fell within 10° of visual angle from the centre of the nose.

The presence of a concurrent cognitive load task in the speech-numbers condition did influence centralization of gaze compared to the speech-only condition (see Fig. 5A), although this influence was fairly subtle ($F(1, 23) = 5.18,\ p = 0.033$). Surprisingly, gaze was more centralized in the speech-only condition than in the speech-numbers condition. The offset in the speech task also had an effect on gaze centralization ($F(1.36, 31.28) = 15.60,\ p < 0.001$), with gaze showing a general tendency to become more centralized as the offset was increased. There was no significant interaction between task condition and offset ($p > 0.05$).

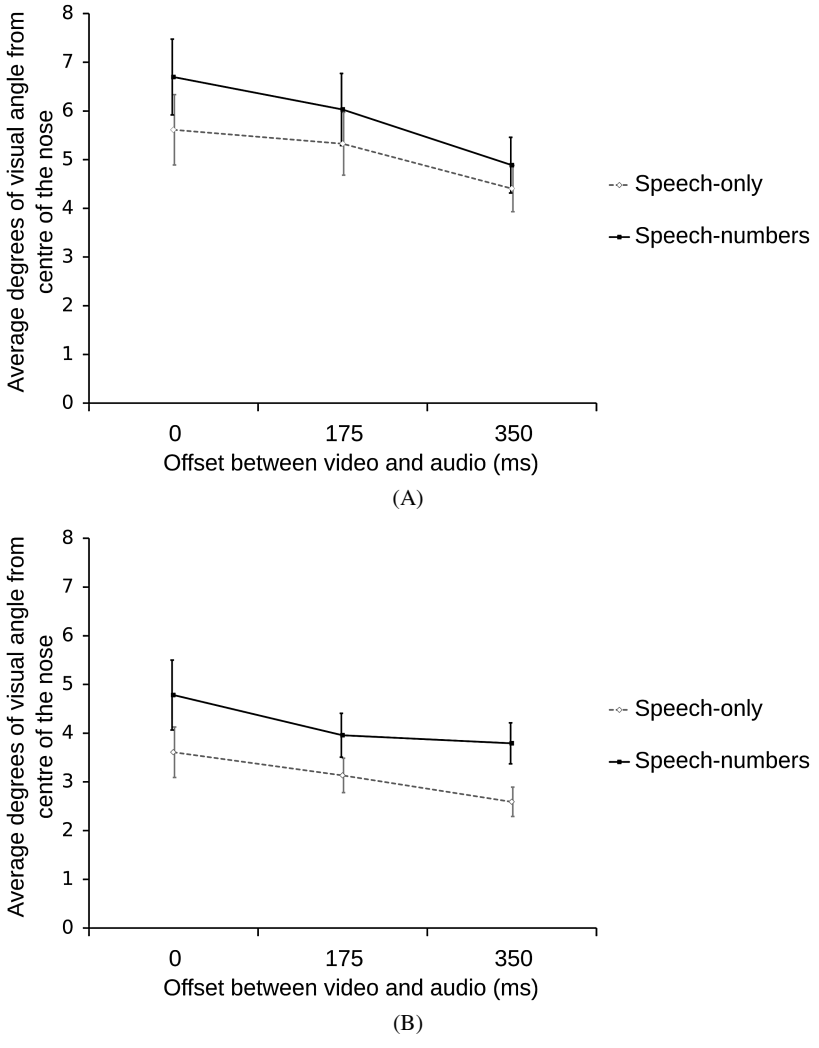
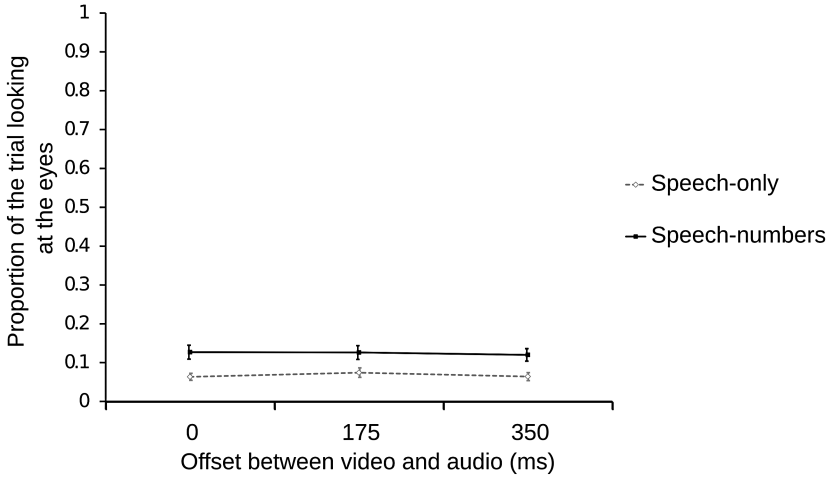
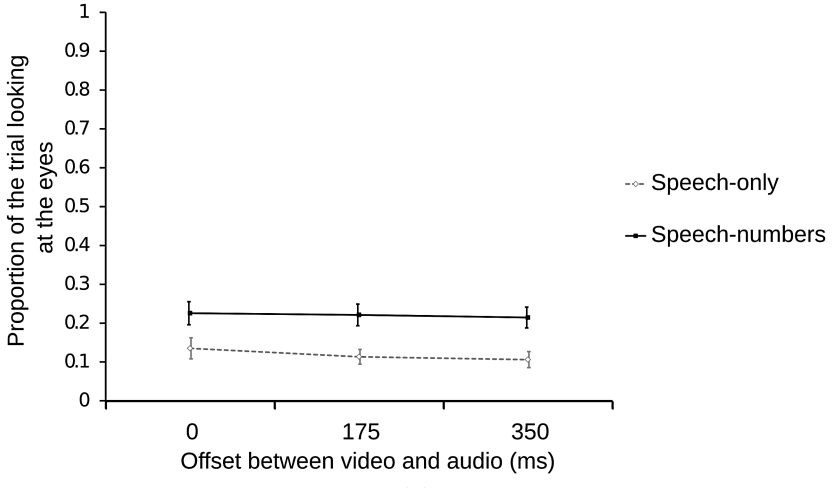


Figure 5. These figures show the average distance of the eye tracking samples from the centre of the nose in degrees of visual angle showing the amount of gaze centralization on the face. (A) shows the average distance from the centre of the nose for Experiment 1, and (B) shows the average distance from the centre of the nose for Experiment 2. The error bars indicate standard errors of the mean.

The presence of a cognitive load task did influence the amount of time spent looking at the eyes and mouth in Experiment 1. The addition of a cognitive load task to the speech task is accompanied by a shift of gaze away from the mouth, and towards the eyes. A larger proportion of the trial was spent looking at the eyes in the speech-numbers condition compared with the speech-only condition ($F(1, 23) = 21.52$, $p < 0.001$) (see Fig. 6A). There was no significant effect of offset, nor a significant interaction between task condition and offset ($p > 0.05$). A larger proportion of the trial was spent looking at the mouth in the speech-only condition compared



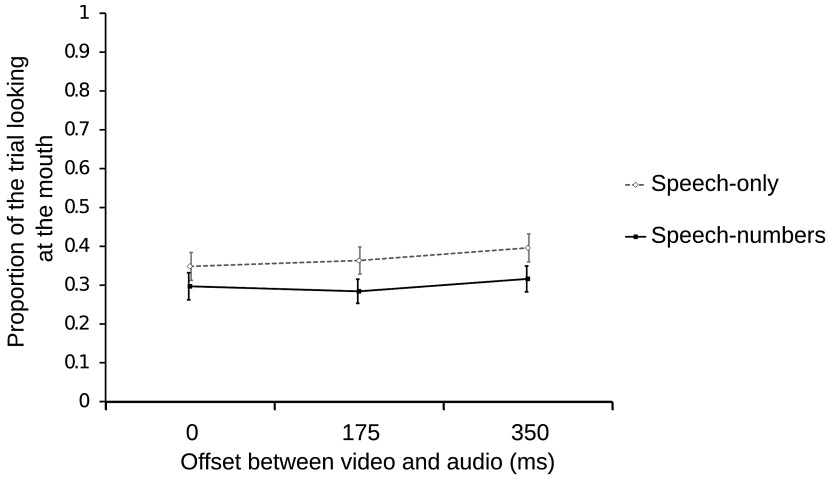
(A)



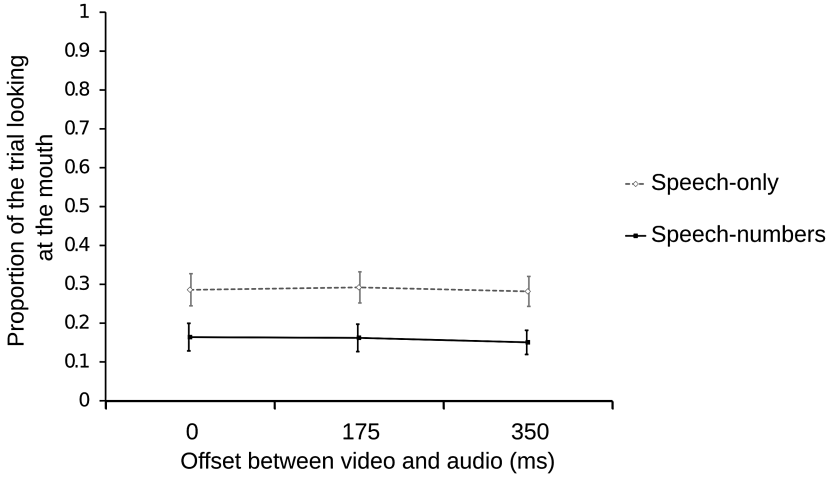
(B)

Figure 6. (A) shows the average proportion of each the trial spent looking at the eyes in Experiment 1 and (B) shows the average proportion of each the trial spent looking at the eyes in Experiment 2. The error bars indicate standard errors of the mean.

with the speech-numbers condition ($F(1, 23) = 13.28,\ p = 0.001$) (see Fig. 7A). A slightly, though significantly, larger proportion of the trial was also spent looking at the mouth with increasing offset ($F(1.40, 32.11) = 4.52,\ p = 0.030$). There were significant differences between the 0 ms offset and the 350 ms offset ($t(23) = -2.75,\ p = 0.034$), and the 175 ms offset and the 350 ms offset ($t(23) = -3.55,\ p = 0.003$). There was no significant interaction between task condition and offset ($p > 0.05$).



(A)



(B)

Figure 7. (A) shows the average proportion of each the trial spent looking at the eyes in Experiment 1 and (B) shows the average proportion of each the trial spent looking at the eyes in Experiment 2. The error bars indicate standard errors of the mean.

3.2.2. Experiment 2

Like Experiment 1, the overall eye-tracking samples tended to fall quite close to the centre of the nose. Approximately 0.54 of eye-tracker samples in Experiment 2 fell within 4° of visual angle from the centre of the nose. Most samples fell either on the face, or very close to the face. As in Experiment 1, approximately 0.90 of samples in Experiment 2 fell within 10° of visual angle.

Gaze behaviour in Experiment 2 was very similar to that seen in Experiment 1, although overall gaze seemed to be more centralized in Experiment 2 (see Fig. 5B). The addition of the cognitive load task causes gaze to become less centralized. Gaze

was more centralized in the speech-only condition than in the speech-numbers condition ($F(1, 24) = 19.36$, $p < 0.001$). Gaze also tended to become more centralized as offset was increased ($F(1.15, 27.59) = 8.94$, $p = 0.004$). There was also no significant interaction between task condition and offset ($p > 0.05$).

As in Experiment 1, the addition of a cognitive load task in Experiment 2 also caused a slight shift of gaze away from the mouth, and towards the eyes. A larger proportion of the trial was spent looking at the eyes in the speech-numbers condition compared with the speech-only condition ($F(1, 24) = 21.51$, $p < 0.001$) (see Fig. 6B). There was no significant effect of offset, nor a significant interaction between task condition and offset ($p > 0.05$) to the proportion of the trial looking at the eyes. A larger proportion of the trial was spent looking at the mouth in the speech-only condition compared with the speech-numbers condition ($F(1, 24) = 31.59$, $p = 0.001$) (see Fig. 7B). There was no significant effect of offset, nor a significant interaction between task condition and offset ($p > 0.05$) to the proportion of the trial looking at the mouth.

4. Discussion

The results of both experiments show an effect of a concurrent cognitive load task on the integration of audiovisual speech. Slightly less audiovisual integration was observed when participants had to perform the secondary cognitive load task. Although the reduction in reported audiovisual integration was quite modest, this effect was replicated across both experiments. As expected, increasing the temporal offset between the auditory and visual speech information decreased the observed integration in both experiments, although the decrease was much more pronounced in Experiment 2.

The overall influence of the secondary task on the integration of auditory and visual speech information in the current experiments was quite modest. The very modest nature of the interference is interesting since the cognitive load task (numbers task) in this experiment was reasonably difficult. For example, a similar task to the one used in the current experiment was used by de Fockert *et al*. (2001), and was shown to influence the processing of distractors in a visual attention task. The task used in the current study was likely more difficult than the task used in de Fockert *et al*. (2001), since in the current experiment participants had to remember the order of eight digits rather than the five digits required of de Folkert *et al*.'s participants. However, the influence of visual speech information on the perception of speech has previously been shown not to be influenced by a concurrent cognitive load task. Baart and Vroomen (2010) found no influence of either a secondary visuospatial or verbal cognitive load task on phonetic recalibration, a phenomenon where lipread information can adjust the phonetic category between two speech categories. The verbal cognitive load task that was used by Baart and Vroomen (2010) appeared to be somewhat easier than the task used in the current paper. Baart and Vroomen (2010) used either three, five or seven memory items, while the current paper used

eight. In addition, their task was a recognition task, whereas participants were required to memorize the order of the digits in the current task. It is possible that in both the current paper and the Baart and Vroomen (2010) paper that a more difficult cognitive load task could have a greater influence on audiovisual integration, although at some point it is possible that participants would start to give up on the task.

The fact that a concurrent cognitive load task had a slight effect on audiovisual integration shows a reduction of audiovisual integration can be achieved in the absence of competing perceptual information. The exact nature of the interference remains unclear. A general issue in dual task paradigms is that participants adopt strategies in order to perform both tasks. The interference could possibly be due to participants silently repeating the items from the cognitive load task to themselves during the speech task. However, as Baart and Vroomen (2010) have pointed out, there is no guarantee that participants were silently repeating the items during the speech task. While participants may have used a verbal strategy to help them with the cognitive load task, high performance on the congruent trials during the speech task would suggest that the strategy used by participants during the cognitive load task was not particular detrimental to the speech task.

It is certainly possible that a different secondary task could show greater deficits in the integration of auditory and visual speech information. However, the lack of interaction between the effect of the cognitive load task and the effect of offset on the McGurk responses suggest that the large size of the synchrony window is relatively independent of cognitive resources that overlap with the cognitive load task. The window for synchrony perception for audiovisual speech stimuli tends to be more generous than that observed for non-speech audiovisual speech stimuli (see Conrey and Pisoni, 2006; Dixon and Spitz, 1980; Lewkowicz, 1996). It seems likely that the rather large size of the synchrony window observed in audiovisual speech is determined by the inherent dynamic properties of the stimuli, with speech generally having a richer time series in both the visual and auditory modalities than the stimuli tested in the non-speech tasks. For example, Arrighi *et al*. (2006) showed that video sequences of conga drumming with natural biological speed variations showed generally greater temporal delays for perceptual synchrony than for artificial stimuli based on the videos that moved at a constant speed. The non-speech stimuli used in studies on perceptual synchrony (e.g., Dixon and Spitz, 1980; Lewkowicz, 1996) have less dynamic variation than the speech stimuli.

Both the concurrent cognitive load task and offset did have an influence on gaze behavior. Gaze was somewhat less centralized on the video of the talker during the concurrent cognitive load task than when no concurrent cognitive load task was present. Increasing the temporal offset generally showed an increase in the amount of gaze centralization on the face. Increasing the temporal offset could have increased the difficulty of the speech task. However, in the behavioral data in the incongruent trials, there was no interaction between the effect of the cognitive load task and the effect of offset. This increase in gaze centralization with the addition of

a cognitive load task was also accompanied by a decrease in the proportion of the trial spent looking at the mouth, and an increase in the proportion of the trial spent looking at the eyes. The general proportions of the trial spent looking at the eyes and mouth are in line with Paré *et al*. (2003) and Buchan *et al*. (2005) who also looked at gaze using a McGurk task, although Paré *et al*. (2003) did show more fixations on the mouth. It is interesting that experiments using longer stimuli such as sentences (Buchan *et al*., 2007), and extended monologues (Vatikiotis-Bateson *et al*., 1998) do show far more time spent looking at the eyes than both the current study, and Paré *et al*. (2003), that both used short vowel–consonant–vowel stimuli.

The gaze centralization observed in previous speech-in-noise studies (Buchan *et al*., 2007, 2008) is not strictly due to the increased cognitive demands caused by the acoustic stimuli being harder to hear with the addition of acoustic noise. The addition of a cognitive load task actually caused a decrease in gaze centralization. This decrease in gaze centralization with the addition of a cognitive load task was also accompanied by a tendency to look slightly less at the mouth, and slightly more at the eyes.

It is possible that different gaze patterns seen in the speech task with the speech task alone compared to the speech task with the concurrent cognitive load task could be driving the decrease in integration seen with the addition of the cognitive load task. However, this seems unlikely as visual speech information can still be gathered without direct fixations on the mouth (Andersen *et al*., 2009; Paré *et al*., 2003). Also, highly detailed visual information is not necessary for the visual speech information to be acquired and integrated with auditory and visual speech information (MacDonald *et al*., 2000; Munhall *et al*., 2004). Paré *et al.* (2003) showed that fixating on either the mouth, eyes or hairline of a talking face seems to provide rather similar vantage points in terms of gathering visual information during audiovisual speech processing in a McGurk task. It is not until gaze is fixed more than 10–20° away from the mouth that the influence of the visual information on the McGurk effect is significantly lessened, and some visual speech information persists even at 40° of eccentricity.

In summary, the data presented here show a very small but statistically significant decrease in the number of McGurk responses when subjects perform a concurrent cognitive load task. This suggests a rather modest role for cognitive resources such as working memory in the integration of audiovisual speech information. While a distracting cognitive load task can slightly modulate the multisensory integration of auditory and visual speech information, it appears that integration of audiovisual speech occurs relatively independent of cognitive resources such as working memory and further suggests that this integration is primarily an automatic process.

Acknowledgements

This work was funded by the Natural Sciences and Engineering Research Council of Canada. J. B. held a Natural Sciences and Engineering Research Council of

Canada PSG D3 award and the Brian R. Shelton Graduate Fellowship for part of the duration of this work. Special thanks to Paul Plante and Fred Kroon for help creating the stimuli and analyzing the data.

References

Alsius, A., Navarra, J. and Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration, *Exper. Brain Res*. **183**, 399–404.

Alsius, A., Navarra, J., Campbell, R. and Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attentional demands, *Curr. Biol.* **15**, 839–843.

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I. and Sams, M. (2009). The role of visual spatial attention in audiovisual speech perception, *Speech Commun.* **51**, 184–193.

Arrighi, R., Alais, D. and Burr, D. (2006). Perceptual synchrony of audiovisual streams for natural and artificial motion sequences, *J. Vision* **6**, 260–268.

Baart, M. and Vroomen, J. (2010). Phonetic recalibration does not depend on working memory, *Exper. Brain Res*. **203**, 575–582.

Buchan, J. N., Paré, M. and Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing, *Soc. Neurosci.* **2**, 1–13.

Buchan, J. N., Paré, M. and Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception, *Brain Res.* **1242**, 162–171.

Buchan, J. N., Wilson, A., Paré, M. and Munhall, K. G. (2005). The effect of spatial frequency information in central and peripheral vision on natural gaze patterns and audiovisual speech perception, *J. Acoustic. Soc. Amer*. **117**, 2620.

Conrey, B. and Pisoni, D. B. (2006). Auditory–visual speech perception and synchrony detection for speech and nonspeech signals, *J. Acoustic. Soc. Amer*. **119**, 4065–4073.

de Fockert, J. W., Rees, G., Frith, C. D. and Lavie, N. (2001). The role of working memory in visual selective attention, *Science* **291**, 1803–1806.

Dixon, N. and Spitz, L. (1980). The detection of audiovisual desynchrony, *Perception* **9**, 719–721.

Doherty-Sneddon, G. and Phelps, F. G. (2005). Gaze aversion: a response to a cognitive or social difficulty? *Memory Cognit*. **33**, 727–733.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli, *J. Speech Hearing Res.* **12**, 423–425.

Grant, K. W., van Wassenhove, V. and Poeppel, D. (2004). Detection of auditory (cross-spectral) and auditory–visual (cross-modal) synchrony, *Speech Comm*. **44**, 43–53.

Jones, J. A. and Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time, *Exper. Brain Res*. **174**, 588–594.

Lewkowicz, D. J. (1996). Perception of auditory–visual temporal synchrony in human infants, *J. Exper. Psychol.: Human Percept. Perform*. **22**, 1094–1106.

Liberman, A. M. (1982). On the finding that speech is special, *Amer. Psychol*. **37**, 148–167.

MacDonald, J., Andersen, S. and Bachmann, T. (2000). Hearing by eye: how much spatial degradation can be tolerated? *Perception* **29**, 1155–1168.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices, *Nature* **264**, 746–748.

Munhall, K. G., Gribble, P., Sacco, L. and Ward, M. (1996). Temporal constraints on the McGurk effect, *Percept. Psychophys.* **58**, 351–362.

Munhall, K. G., Kroos, C., Jozan, G. and Vatikiotis-Bateson, E. (2004). Spatial frequency requirements of audiovisual speech perception, *Percept. Psychophys.* **66**, 574–583.

Paré, M., Richler, R. C., ten Hove, M. and Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect, *Percept. Psychophys.* **65**, 533–567.

Parkhurst, D. J. and Neibur, E. (2003). Scene content selected by active vision, *Spatial Vision* **16**, 125–154.

Parkhurst, D., Law, K. and Neibur, E. (2002). Modeling the role of salience in the allocation of overt visual attention, *Vision Res*. **42**, 107–123.

Reisberg, D., McLean, J. and Goldfield, A. (1987). Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli, in: *Hearing by Eye: The Psychology of Lip-Reading*, B. Dodd and R. Campbell (Eds). Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Rosenblum, L. D., Schmuckler, M. A. and Johnson, J. A. (1997). The McGurk effect in infants, *Percept. Psychophys.* **59**, 347–357.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C. and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments, *Cerebral Cortex* **17**, 1147–1153.

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

Sumby, W. H. and Pollack, I. (1954). Visual contributions to speech intelligibility in noise, *J. Acoust. Soc. Amer.* **26**, 212–215.

Summerfield, Q. and McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels, *Qtrly J. Exper. Psychol. A* **36**, 51–74.

Tiippana, K., Andersen, T. S. and Sams, M. (2004). Visual attention modulates audiovisual speech perception, *Eur. J. Cognit. Psychol.* **16**, 457–472.

van Wassenhove, V., Grant, K. and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception, *Neuropsychologia* **45**, 598–607.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S. and Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception, *Percept. Psychophys.* **60**, 926–940.

Wilson, A., Wilson, A., ten Hove, M., Paré, M. and Munhall, K. G. (2008). Loss of central vision and audiovisual speech perception, *Visual Impairment Res.* **10**, 23–34.