

# Visual Speech Influences Speech Perception Immediately but not Automatically

Holger Mitterer

*Department of Cognitive Science  
University of Malta*

Eva Reinisch

*Institute of Phonetics and Speech Processing  
Ludwig Maximilian University Munich*

Two experiments examined the time course of the use of auditory and visual speech cues to spoken word recognition using an eye-tracking paradigm. Results of a first experiment showed that the use of visual speech cues from lip-reading is reduced if concurrently presented pictures require a division of attentional resources. This reduction was evident even when listeners' eye gaze was on the speaker rather than the (static) pictures. Experiment 2 used a deictic hand gesture to foster attention to the speaker. At the same time the visual processing load was reduced by keeping the visual display constant over a fixed number of successive trials. Under these conditions, the visual speech cues from lip-reading were used. Moreover, the eye-tracking data indicated that visual information was used immediately and even earlier than auditory information. In combination, these data indicate that visual speech cues are not used automatically, but if they are used, they are used immediately.

In spoken conversation, listeners can make use of the acoustic/auditory signal as well as visual information from the speaker's lip movements to interpret what is being said. While the auditory signal carries much more, and more fine-grained information about the phonetic content of speech, there is a consistent comprehension benefit for audiovisual speech over audio-only when speech is perceived in noise (e.g., Macleod & Summerfield, 1987). The McGurk effect (McGurk & MacDonald, 1976) shows that visual speech is even used with clearly audible stimuli: If perceivers see the video of a speaker saying *go* while they hear *bow*, they often report hearing *dough*. However, *how* the two streams of auditory and visual information are integrated has been hotly debated. Some claim that visual information influences speech perception at very early stages of processing (van Wassenhove, Grant, & Poeppel, 2005) while others argue that audiovisual integration occurs late (Colin et al., 2002; Stekelenburg & Vroomen, 2007;

---

Holger Mitterer, Department of Cognitive Science, Faculty of Media and Knowledge Sciences, University of Malta, Msida MSD 2080, Malta. Eva Reinisch, Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich, Schellingstr. 3, 80799 Munich, Germany.

This work was supported by a University of Malta Research Grant to the first author. The second author was supported by an Emmy-Noether Fellowship by the German Research Council (DFG, grant nr. RE 3047/1-1).

Correspondence concerning this article should be sent to: Holger Mitterer, Department of Cognitive Science, Faculty of Media and Knowledge Sciences, University of Malta, Msida MSD 2080, Malta or electronically to holger.mitterer@um.edu.mt

for details see below). On a theoretical level, these claims are based on motor-based theories of speech perception (see, e.g., Fowler & Dekle, 1991; Rosenblum, 2008) and the fuzzy-logic model of perception (FLMP, Massaro, 1998), respectively.

A first line of research trying to estimate the locus of audio-visual integration used lexical effects to anchor audio-visual integration. It was tested whether the McGurk effect is more likely to occur if audio-visual integration leads participants to report a word rather than a nonword (Brancazio, 2004) or when the combined audio-visual percept better fits a given sentence (Windmann, 2004). The effects tend to be variable, with some studies finding a stronger visual influence when it leads to a response that is a real word or semantically fits the context (Barutchu, Crewther, Kiely, Murphy, & Crewther, 2008; Brancazio, 2004; Windmann, 2004), while others do not (Sams, Manninen, Surakka, Helin, & Kättö, 1998). At first sight, it would seem that findings of an influence of lexical factors on audiovisual speech perception would be more in line with the view that the visual signal is used in lexical access, and hence at an early stage. However, as discussed by Windmann (2004), these findings are also consistent with late-integration models such as FLMP.

Another line of research used event related potentials (ERP) to investigate the time course of audiovisual integration. ERPs are well suited to investigate early perceptual processes and make it easier to exclude post-perceptual explanations. Van Wassenhove et al. (2005) present data that indicated an early interaction of visual and auditory channels, based on an earlier and smaller N1 component to audiovisual speech compared to audio-only speech. The N1 is a component occurring about 100 ms after stimulus onset and is considered to reflect early auditory processing. The authors argued that their results show that the auditory and visual channels are integrated early, especially because the N1 reduction was largest for highly salient visual speech stimuli (i.e., a labial stop /p/). These results, however, have not stood the test of time. Stekelenburg and Vroomen (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010) showed that these effects can be understood as a reduction of the N1 signal independent of the phonetic content of the speech signal. They showed that the visual signal acts as a warning signal that an auditory signal is going to arrive. Accordingly, the reduction of the N1 amplitude and its consequent earlier peak (see Luck, 2014, for the interrelatedness of amplitude and latency of ERP components) occur with stimuli in which visual motion precedes the sound (e.g., one-syllable utterances and hand clapping) but not with those in which the visual signal does not lead the auditory signal (e.g., tearing a piece of paper). Later research showed that the reduction of the N1 component may in fact be proportional to the salience of the visual signal for predicting the onset of the auditory signal (Brunellière, Sánchez-García, Ikumi, & Soto-Faraco, 2013).

In line with these findings, Meyer, Harrison, and Wuerger (2013) presented evidence that audiovisual congruency does not influence the ERP before 200ms. Additionally, Colin et al. (2002) report that it is possible to evoke a Mismatch Negativity (MMN) purely by changing the visual signal; however, this MMN did not lead to a polarity reversal at the mastoid electrodes. Since the latter effect is usually associated with early auditory processes, this finding aligns with the other ERP evidence to suggest that the integration of auditory and visual *speech* information occurs late as proposed in the FLMP (Massaro, 1998). First the auditory and visual signals are evaluated separately and the percept is decided upon by a combination of both signals.

What these findings also highlight is that the interpretation of the ERP signal is often not that straightforward. For instance, the finding that labial stops elicit the strongest N1 reduction may simply reflect that a lip closure allows a better prediction of when a sound is going to arrive than other (visible) articulatory gestures. The phonetic content of the visual signal, that is, the phonetic information that it is a labial sound, may not matter for the early processing of the auditory signal. It may hence be useful to have a dependent variable that reflects how the visual speech information influences lexical access, but ERP components are not always straightforwardly linked to cognitive processes.

Ostrand, Blumstein, Ferreira, and Morgan (2016) used a measure that can be more directly interpreted as reflecting lexical processing: associative priming. Ostrand et al. (2016) made use of minimal pairs such as *bait-date*, which were recorded audiovisually and then presented in congruent and incongruent conditions (i.e., McGurk stimuli such as auditory *bait* and visual *date*). In a pretest, participants were asked to judge how well these items sound as *date* or *bait*. This gave rise to a McGurk like effect so that visual *bait* coupled with auditory *date* was deemed as a good example of *bait* but a bad example of *date*. In their main experiment, the same stimuli were used in an associative priming task. Auditory target words related to both members of the minimal pair (*bait* → *fish* and *date* → *time*) were presented with a 50 ms inter-stimulus interval. In the congruent condition, the expected effect of associative priming was found so that *bait* primed *fish* and *date* primed *time*. In the incongruent McGurk condition, however, the priming data indicated that only the auditory signal seemed to matter. That is, visual *date* combined with auditory *bait* primed only *fish* but not *time*. This indicates that early lexical activation may be independent of the visual signal when the speech signal is acoustically unambiguous.

However, there is a possible alternative interpretation for these results. While it is often assumed that audiovisual integration is automatic (e.g., Rosenblum, 2008), there is now a substantial body of evidence that audiovisual integration in speech perception is hindered if there is an additional cognitive load (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Navarra, Alsius, Soto-Faraco, & Spence, 2010;

Tiippana, Andersen, & Sams, 2004). The priming method used by Ostrand et al. (2016) included performing a secondary task—performing a lexical decision on the potentially primed target word—which is a form of cognitive load. It may be that the lack of an effect of the visual stimulus was due to this secondary task.

In the present study, we therefore tackled the issue of how the auditory and visual channels of speech are integrated with another method that has been successfully used to track the time course of the use of different cues in speech perception: visual-world eye-tracking (McMurray, Clayards, Tanenhaus, & Aslin, 2008; Mitterer & Reinisch, 2013; Reinisch & Sjerps, 2013). In this paradigm, participants are presented with a set of pictures on a screen and asked to click on the picture that matches what they perceive. Importantly, listeners have been shown to spontaneously fixate their eye gaze on visual referents that match the concurrent speech input in a closely time-locked fashion (Allopenna, Magnuson, & Tanenhaus, 1998; Cooper, 1974). For example, Allopenna et al. (1998) presented participants with the sentence *Click on the beaker* and tracked how much participants fixated on pictures of a beaker, a speaker, a beetle, and a phonetically unrelated distractor. They found that not only the eventual target picture of a beaker received more fixations than the unrelated distractor but also the pictures of both words that were phonologically related to the target (i.e., *speaker* and *beetle*). Since then, this method has been highly successful in showing that listeners are extremely quick to use fine-grained differences in the speech signal (for a review, see Huettig, Rommers, & Meyer, 2011). If audiovisual integration for lexical items is indeed late, we should find that the visual signal influences the eye gaze much later than the auditory signal. One advantage of the visual-world paradigm is that the eye-tracking record reflects online processing contingent with the unfolding speech signal—just as ERPs do, but the interpretation is more straightforward: If fixations on a picture referring to a word starting with a labial sound occurred earlier if there was visual evidence for a labial sound, this would be clear evidence that information about the place of articulation has been extracted from the visual signal and is used for spoken-word recognition (for a similar argument comparing ERPs and eye-tracking, see Mitterer, 2011).

One potential issue with using a visual-world paradigm with audiovisual stimuli might be that the referent pictures require participants to look away from the speaker towards the pictures. As a consequence, participants would perceive the talking face much blurrier due to the strong decrease of visual acuity outside the fovea. However, the McGurk effect has been shown to be surprisingly resilient to fixation position. Below 10 degrees of visual angle, there is no decrement of the effect of visual speech on speech perception and only above 20 degrees does the use of the visual speech signal deteriorate strongly (Paré, Richler, ten Hove, & Munhall, 2003). Hence, fixations on picture referents in the visual-

world paradigm should not prevent listeners from using visual speech cues from a speaker video that is centered on the screen, as this is possible within 20 degrees of visual angle. Another potential issue, however, is that the additional picture stimuli may lead to divided attention in processing the video versus picture stimuli. Divided attention in turn has been shown to lead to a decrement in the use of visual speech (Alsius et al., 2005). Therefore, we also included a control condition, with little visual load. That is, participants made repeated decisions to the same word pair with the pictures on the screen remaining in a fixed position, which is the standard format under which audio-visual integration is usually assessed (and unlike the random assignment of pictures to screen locations typical for the visual-world paradigm). This control condition will be referred to as minimal-variability condition. If visual speech does not influence speech perception in the visual-world condition, the minimal-variability condition will rule out the interpretation that our specific AV stimuli used in this experiment might be insufficient to influence speech perception. If an effect is found in the minimal-variability condition but not in the visual-world condition this may then be because the visual cues are not used under high task demands.

## Experiment 1

In this experiment, we aimed at testing how quickly eye movements reflect the use of auditory and visual information in speech perception. At this juncture it is useful to distinguish the use of the visual speech signal from audio-visual integration. If participants look more towards the picture of a labial referent when seeing a lip-closing speech gesture, this indicates that visual speech is used, but it does not necessarily mean that the visual and the auditory signal are integrated to a multi-sensory percept.

We presented participants with an audiovisual speech stimulus and asked them to click on the picture referred to by the speaker. By using congruent and incongruent audiovisual stimuli, we could estimate when the visual and the auditory signal were used for speech perception. Since a previous study indicated that initial lexical access might be strongly biased towards the auditory signal (Ostrand et al. 2016), we used slightly ambiguous auditory stimuli in an attempt to provide more leverage for the visual signal to influence perception. We decided to aim for stimuli with an odds ratio of 4 (80/20), which should still give rise to a clear effect of the auditory cues but at the same time allow for a visual influence.

### Method

#### Participants

Fourteen native speakers of German took part in the experiment. They all had normal or corrected-to-normal vision and no hearing impairment. They were paid for their participation.

## Materials and Procedure

We recorded a German native speaker producing eleven German minimal word pairs, of which ten were eventually used. The minimal pairs differed in the place of articulation of stop consonants, which were either labial or alveolar (e.g., *Panne-Tanne*, Engl. (*car*) *breakdown – fir*, Table A1 provides the list of minimal pairs used in this experiment). Half of the minimal pairs differed in word-initial position (as in *Panne-Tanne*), the other half differed in word-final position (e.g., *gelb-Geld*, Engl., *yellow-money*)<sup>1</sup>.

Two recordings were made for each of the minimal pairs, one audiovisual recording and one high-quality audio-only recording with the microphone close to the speaker's mouth (for details on the audio recordings see below). The video was recorded at 25 frames per second and focused on the speaker's head (720x576 pixels per frame). From this video recording, short clips of 1.2 s were extracted. The audio in these clips started between frame 12 and 13 (i.e., after 480-520 ms). The first 200 ms were overlaid with a fade-in from a black frame up to the still of the sixth frame and the last 200 ms were overlaid with a fade-out from a still of the twenty-fifth frame to a black frame. These transitions were added using Adobe Premiere (Adobe Systems Inc.). The videos were then further cropped to a size of 350 x 496 pixel using the VirtualDub software ([www.virtualdub.org](http://www.virtualdub.org)).

Multiple high-quality audio-only recordings were made for each word. The one matching closest with the audio from the video was then selected for the generation of audio continua. Discrepancies between the video and chosen high-quality audio recording were below 30ms<sup>2</sup>. To select audio tokens that were slightly ambiguous between the labial and alveolar endpoints of the minimal pairs, the selected audio recordings were morphed into eleven-step continua using the STRAIGHT audio morphing algorithms (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999).

The continua were pretested by asking nine native speakers of German to categorize 6 steps for each of the eleven minimal pairs (steps 1,3,4,5,6, and 8) five times. Based on this pre-test we selected audio-files that elicited roughly 20% and 80% of labial responses. Since all pairs gave rise to clear identification functions, we decided to exclude the pair *Korb – Cord* (Engl., *basket-cord*), because *Cord* is

---

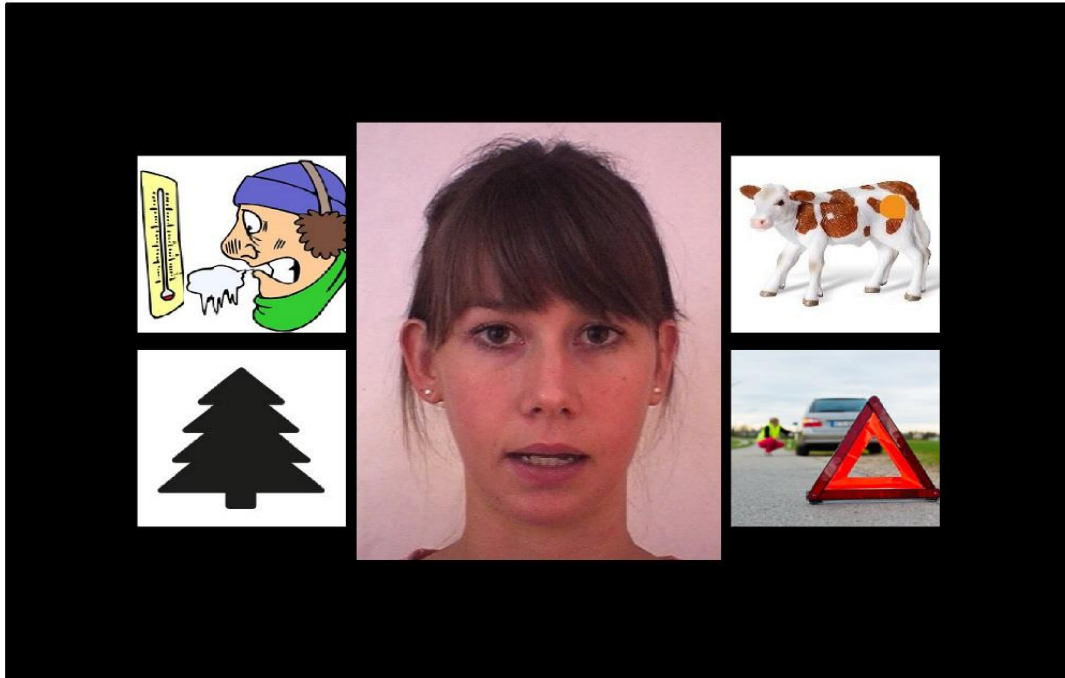
<sup>1</sup> Barutchu, Crewther, Kiely, Murphy, and Crewther (2008) reported that audiovisual integration differs over position in the word, being stronger in the onset than the offset position. However, their statistical analysis only takes into account variation over participants, even though their Appendix shows massive variation between items (with McGurk responses varying between 0 and 64% over stimuli). We therefore used the item-specific data presented in their appendix for an item-based analysis, which showed that all effects are far from significant (all *ps* > 0.2).

<sup>2</sup> Note that such an exact match in duration between an auditory and visual word would not be necessary, as observers consistently fail to notice such small audiovisual asynchronies (Vatakis & Spence, 2007) and tend to integrate information from asynchronous audio and video within a time window of up to 200 ms (van Wassenhove, Grant, & Poeppel, 2007).

mostly used in compound words and rarely used in isolation in German. For the remaining ten pairs, these audio-files were then dubbed onto the videos by replacing the original audio (where the original audio was used to time-align the new audio). In this way we produced videos in which the auditory cues more or less matched the visual cues (e.g., a video with the utterance *Tanne* and an audio identified as *Tanne* at 80%) and videos in which the auditory cues mismatched the visual cues (i.e., a video with the utterance *Tanne* and an audio identified as *Tanne* at 20%).

For each of the twenty words from the ten minimal pairs, we performed a Google image search and selected an image to represent that word (see Figure 1 for some examples). Images were scaled to 200x200 pixels. Images and videos were combined to displays in two different conditions. The first condition was similar to the typical visual-world paradigm, in which, on each trial, four different pictures were presented on the screen (visual-world condition). The video of the speaker appeared centered on the screen and the pictures were placed to the right and left of the video. Pictures included a target (e.g. *Panne*), its competitor (*Tanne*) and a distractor pair that had the critical sound in different word position (here e.g., *gelb-Geld*; where the critical contrast is in word-final position). The distance between the center of the screen and the middle of the pictures was about 9° horizontally and about 7.5° vertically. The images on the screen changed for every trial (as is usual in the visual-world paradigm) and the randomization was done independently for each participant, with the constraint that the target and its competitor appeared equally often in all four possible positions. The pictures were presented 500 ms prior to the onset of the video, so that in total, the preview of the pictures was 1 s relative to the start of the audio. Participants were instructed to move a visible mouse cursor on the picture matching the word they heard and click on it.

The second condition was similar to the typical format of audiovisual speech perception studies with little trial-to-trial variation (minimal-variability condition). Each minimal pair was repeated for twelve trials, and only the two response options for this pair were displayed on the screen. The two pictures were presented in the upper right and upper left position (see Figure 1 for the exact positions) and did not switch sides for a set of twelve stimuli. The timing was the same as in the visual-world condition. Participants were asked to press the left mouse button if the utterance matched the picture on the left of the speaker and the right mouse button if the utterance matched the picture on the right of the speaker. No mouse cursor was visible and no mouse movement was hence required, following the standard procedure in audiovisual speech experiments. The experimental procedure was implemented using the SR Research Experiment Builder.



*Figure 1.* A still of an experimental trial in the visual-world condition of Experiment 1. The task of the participants was to click on the picture corresponding to the word uttered by the speaker in the video. In this display, the potential targets are *kalt*, Engl., *cold* (upper left), *Kalb*, Engl., *calf* (upper right), *Tanne*, Engl., *fir* (lower left), and *Panne*, Engl., *breakdown* (lower right).

Participants were first familiarized with the pictures and their names. Next, they were seated in front of a computer screen and an Eyelink SR 1000 eye-tracker in desktop set-up was calibrated. They were instructed that they would see a speaker in the center, two or four pictures scattered over the four quadrants of the computer screen and hear a word over headphones. They were asked to decide which word they thought the speaker in the video had uttered. How the response was given differed between the two conditions. In the visual-world condition, participants were asked to move a visible mouse cursor over the picture matching the utterance and click on it. In the minimal-variability condition, there was no mouse cursor visible, and participants simply clicked the left mouse button if they thought the utterance matched the picture on the left and the right mouse button if the utterance matched the picture of the right. Condition (visual-world vs. minimal variability) was manipulated within participants. It was blocked with the order counterbalanced across participants. Each condition contained 120 trials.

For each minimal pair, there were four possible audio-visual stimuli that arose by crossing the two auditory stimuli with the two visual stimuli. Since our main question about the timing of the use of auditory and visual information was addressed via the incongruent trials, these were presented twice as often as the congruent ones. Each participant hence saw—for each of the ten minimal pairs—the two



congruent stimuli (i.e., audio & video = labial and audio & video = alveolar) twice (= four trials) and the two incongruent trials four times (= eight trials).

In the visual-world condition, the stimuli were presented randomly, with the constraint that the same pair could not be used as a target on two consecutive trials. In the minimal-variability condition, the twelve stimuli for one minimal pair were presented consecutively, and the transition from one to the next minimal pair was indicated by a screen which showed the two pictures for the upcoming minimal pair. It informed participants that they would have to make a choice between these two for the next set of trials. In both conditions, there was a break halfway through the 120 trials; participants continued by pressing a mouse button.

### Data Processing

The output of the eye-tracker in terms of events (saccades, fixations, and blinks) was analyzed with a PERL script to generate a time line of looks for each trial. Saccades and Fixations were merged into looks to the position of the fixation (cf. McMurray et al., 2008). Blinks were replaced with the last preceding fixation position. If there was a fixation position outside the screen on a given trial, the data from this trial was discarded, as such fixation positions often indicate faulty eye-tracking. This led to the rejection of 181 trials (4.8%), of which 117 were from one participant, who was difficult to calibrate due to wearing glasses. The data from this participant was not used for the eye-tracking analysis. For the rest of the participants, the rejection rate was 2.1%.

For the remaining trials, the fixation positions were categorized as being on the face, on other parts of the video, or on one of the pictures. Fixations were only counted as on a picture, if the fixation position was on a pixel that was occupied by that picture. This was necessary, because the video and the picture were quite close to each other. Similarly, fixations were counted as being on the face if they were within a rectangular region between the eye-lashes and the chin in the vertical dimension and between the outer edges of the eye sockets in the horizontal dimensions. Previous research has indicated that the use of visual speech cues is invariably strong when this region is fixated (see, Paré et al., 2003, Experiments 1 and 2).

## Results

### Overall gaze patterns

First, we provide an overview where participants looked during the different tasks. This provides a frame of reference for interpreting the behavioral data regarding how often participants fixated on the face of the speaker versus on other parts of the visual display. Gaze patterns are shown separately for

minimal pairs differing in the *initial* phoneme and minimal pairs differing in their *final* phoneme, because the timing of visual versus auditory information is different between these two cases. With stop consonants, the visual cues precede the auditory cues in the word-initial case (since the closing gesture precedes the release with no acoustic trace), but the visual cues and auditory cues go hand in hand in the final condition, in which the visual closing gesture also leads to audible formant transitions. For all eye-tracking data, we used the time of the consonant release as the time anchor, indicated as zero milliseconds in all eye-tracking figures.

Figure 2 shows that participants fixated mostly on the face<sup>3</sup> up to the point of the release of the critical stop consonant. Only in the visual-world condition and there only for the stimuli with the critical consonant in word-final position, participants started moving their eye-gaze away from the video towards the pictures already around the release. That is, at the time of the release there were about 50% fixations on the video, and that trend was rapidly falling. However, 50% fixations on the video should be sufficient to expect an influence of the visual signal on the perceptual identifications.

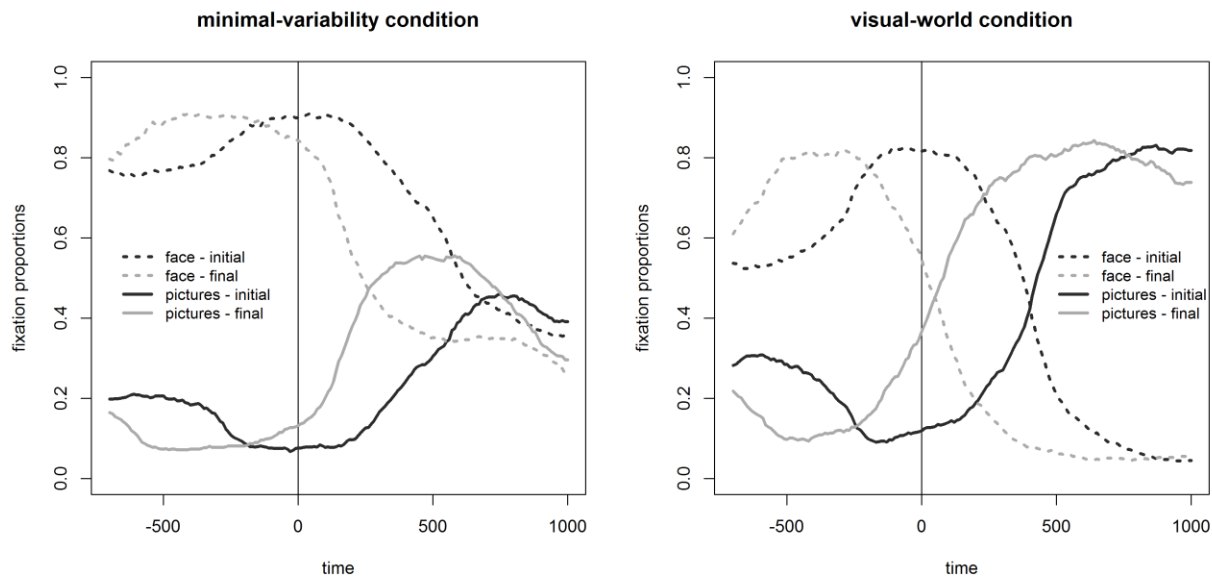
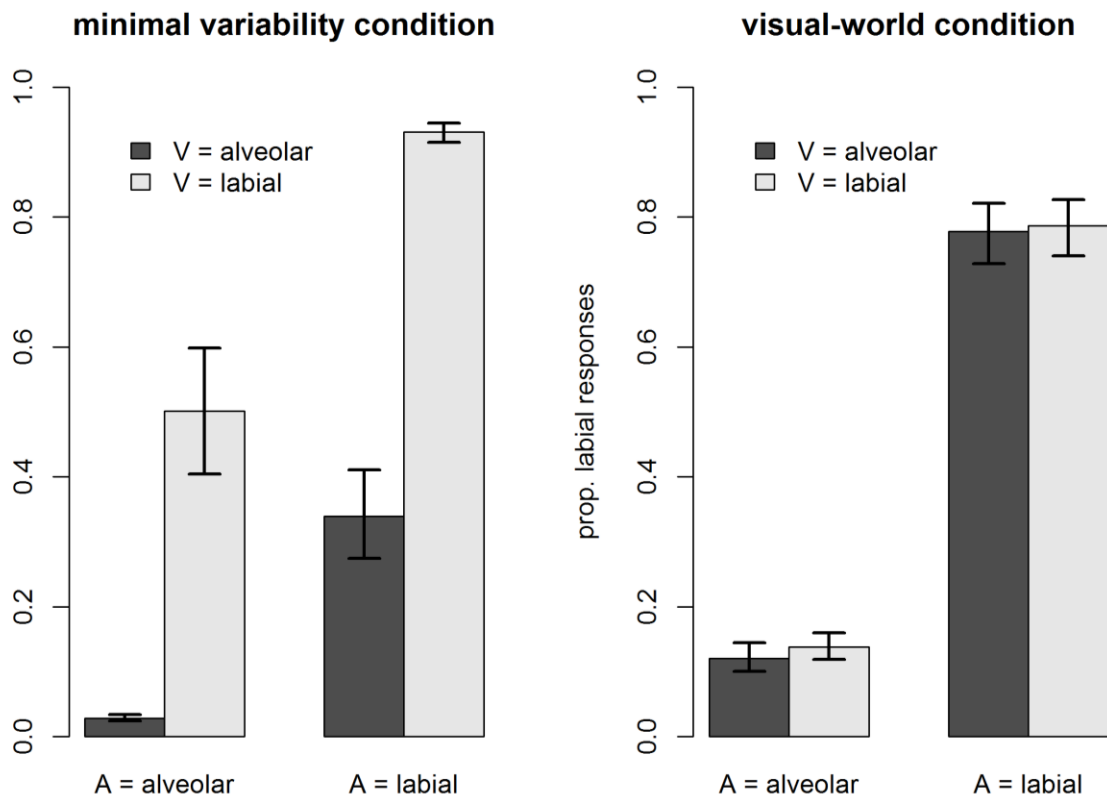


Figure 2. Eye-tracking data showing the proportion of fixations on the face of the speaker versus the pictures in both conditions. In both conditions, the video was fixated on more than the pictures up to the release of the critical stop consonant.

<sup>3</sup> During the presentation of the speech stimuli, there were hardly any fixations that were on the video but not on the face area (< 1% overall).

### Perceptual identifications

Figure 2 shows the proportion of cases in which the stimuli were identified as *labial*, that is, responses in which the labial member of the minimal pair was chosen for all combinations of auditory and visual cues. The left panel shows that in the minimal-variability condition, both auditory and visual cues influence the likelihood of labial responses. The data for the visual-world condition (right panel), in contrast, indicate that only the auditory cues mattered. This difference was confirmed by a generalized linear mixed-effects model using the package lme4 (v.1.1.10) in R (v.3.2.5). In this analysis, response (0 = alveolar, 1 = labial) was the dependent variable and the fixed effect predictors were visual and auditory cue (contrast coded as -0.5 = alveolar, 0.5 = labial), condition (contrast coded as -0.5 = minimal-variability and +0.5 = visual-world) and the two-way interaction of cues and condition.



*Figure 3.* Perceptual decisions—coded as the proportion of trials in which participants chose the labial member of the pair—in the two identification tasks in Experiment 1, depending on the auditory (x-axis) and visual cues (light bars: visual cues for labial, dark bars: visual cue for alveolar). The data are shown in proportion but are based on averaging individual proportions in logOdds. The error bars represent standard errors of the mean in logOdds, following the method of Morey (2008).

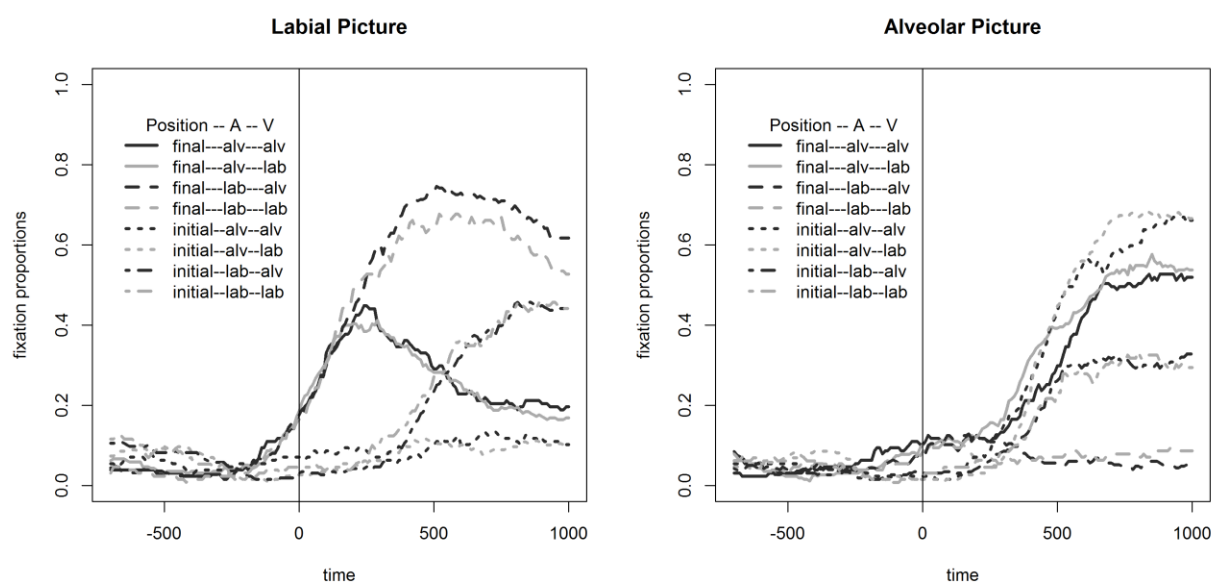
We did not specify a full-factorial model, because smaller models require fewer random effects and as a consequence are less likely to lead to convergence problems. Participant and item (i.e., video file) were entered as random effects, with a maximal random effect structure (Barr, Levy, Scheepers, & Tily, 2013). The analysis gave rise to a significant effect of auditory cue ( $b = 4.352$ ,  $SE = 0.636$ ,  $z = 6.842$ ,  $p < .001$ ) that was marginally moderated by condition ( $b = 1.435$ ,  $SE = 0.805$ ,  $z = 1.783$ ,  $p = .074$ ) and an effect of visual cue ( $b = 2.392$ ,  $SE = 0.572$ ,  $z = 4.184$ ,  $p < .001$ ) that was strongly affected by condition ( $b = -4.667$ ,  $SE = 0.7458$ ,  $z = -6.261$ ,  $p < .001$ ). To further investigate this interaction, we ran separate models for both conditions, always using a maximal-random effect structure. In the minimal-variability condition, there was an effect of auditory cue ( $b = 3.443$ ,  $SE = 0.641$ ,  $z = 5.373$ ,  $p < .001$ ) and visual cue ( $b = 4.451$ ,  $SE = 0.641$ ,  $z = 6.944$ ,  $p < .001$ ). In contrast, in the visual-world condition, there only was a significant effect of auditory cue ( $b = 5.381$ ,  $SE = 0.822$ ,  $z = 6.545$ ,  $p < .001$ ) but not of visual cue ( $b = 0.067$ ,  $SE = 0.748$ ,  $z = 0.090$ ,  $p = .928$ ).

Given that the null effect of the visual speech cue was unexpected, we further investigated this in two ways. First of all, we calculated a visual effect measure for each participant by subtracting the logOdds of labial responses given an alveolar visual speech cue from the logOdds of labial responses given a labial visual speech cue, and ran a Bayesian one-sample t-test on these data. For this analysis we used the function *ttestBF* from the R package *BayesFactor* (version 0.9.12) with its default priors. This produces as test statistic a Bayes Factor (Rouder, Speckman, Sun, Morey, & Iverson, 2009), which provides evidence for the null hypothesis if below one third. The Bayes Factor for the visual effect in the visual-world condition was 0.211 and hence provides evidence for the hypothesis that the visual cue is not used in the visual-world condition.

Additionally, we used the eye-tracking record to focus on trials in which the face of the speaker was fixated for more than 90% of the 200ms interval around the closure release. This was the case for about 59% of the trials (894 out of 1521 trials with valid responses and good eye-tracking). An additional eleven trials from one participant were rejected, because this participant looked away from the face on the majority of trials and had no data for one cell of the design. For this subset, the effect of visual speech cue was also small (1%) and not significant ( $b = 0.111$ ,  $SE = 1.071$ ,  $z = 0.104$ ,  $p > .2$ ). A Bayesian t-test on the individual measures of the visual speech cue (defined as above) provided a Bayes Factor of 0.327, which is again evidence for the assumption that there is no effect of the visual speech cue, even when the participants focused on the face during the release of the critical consonant.

### Time course of auditory- and visual-cues effects

The main rationale of this experiment was to track the time course of the effect of auditory and visual speech cues on fixations on referent pictures in the visual-world condition. The effectiveness of a cue is reflected in more looks to labial vs. alveolar referent pictures when audio or video indicate a labial rather than an alveolar consonant. With eye-tracking, we can see when these effects start emerging. The relevant data are displayed in Figure 4 and show a clear difference between lines depending on the auditory cues but no clear difference between lines only differing in visual cues. The effects of the auditory cues arise between 200 and 300ms after the stop release. This can be seen by comparing the solid versus dashed lines for the word-final condition and the dotted versus dashed-dotted lines for the word-initial condition. Additionally, the figure shows an overall preference for the labial interpretation at the onset of the stop release for the stimuli with word-final stops. For these conditions there are very few looks to the alveolar pictures before about 300ms after the stop release.



*Figure 4.* Fixation proportions to the labial target pictures (left panel) and alveolar target pictures (right panel) depending on position of the critical sounds in the words, auditory cue (line type) and visual cue (line color). The comparison of dark (visual cue = alveolar) and light (visual cue = labial) lines shows the (absence of the) effect of the visual cue.

To statistically test the time course of effects in such data, two methods have been previously used. First, following methods used in electrophysiology (see, e.g., Turenout, Hagoort, & Brown, 1998), moving time windows are used to establish in which time window the effect first reaches significance<sup>4</sup> (Altmann, 2011; Mitterer & Reinisch, 2013; Salverda, Kleinschmidt, & Tanenhaus, 2014). Secondly, a jackknife method has been used to estimate when an effect reaches a certain percentage of its maximum (see, e.g., McMurray et al., 2008). The latter method has the advantage to be insensitive to effect size, and is hence preferable when effects differ in size. However, this method cannot be applied to the current data, because it requires an effect to be present with a clearly defined maximum. This is not the case for the visual cues.

It is conceivable, however, that there might be at least a transitory effect of the visual cues. To test whether this is the case, we used the moving window method and ran linear mixed-effects models on sequences of a moving 200 ms time window starting 100 ms before and leading up to 900 ms after the stop release with the center of the window being shifted in steps of 100 ms. For each of these time windows, we calculated the preference for fixating on the picture of the referent with the labial by subtracting the logOdds of the fixation proportion for the alveolar picture from the logOdds of the fixation proportion for the labial picture (proportion of zero and one were replaced by  $1/n+2$  and  $(2*n-1)/n$ , as recommended by Macmillan and Creelman (2004)). This measure was used as the dependent variable and the predictors were auditory and visual cue (again contrast coded so that a labial cue is coded as positive).

Figure 5 shows the outcome of the analysis. Fixation proportions for the labial picture were significantly influenced by the auditory cue from about 300ms after the stop release. However, there was no discernable effect of visual cue. Above we noted that there was an overall preference for a labial interpretation for the word-final minimal pairs. In this time-course analysis, the dependent variable was the fixation proportion to labial minus the fixation proportion to alveolar pictures. In the statistical analysis, the preference for a labial interpretation was hence reflected in a significantly positive intercept. This intercept, indicating an overall preference for labial referents was significantly larger than zero in the time windows from 0 to 500ms after stimulus onset (not displayed in the figure to prevent clutter).

---

<sup>4</sup> In such time-window methods, usually there is no correction for multiple comparisons, as this would be overly conservative. The time-window analysis is not meant to show that an effect is substantial; this is asserted using the behavioural data or an overall measure of the eye-tracking record. Rather, a time-window analysis is used to track the onset of an effect.

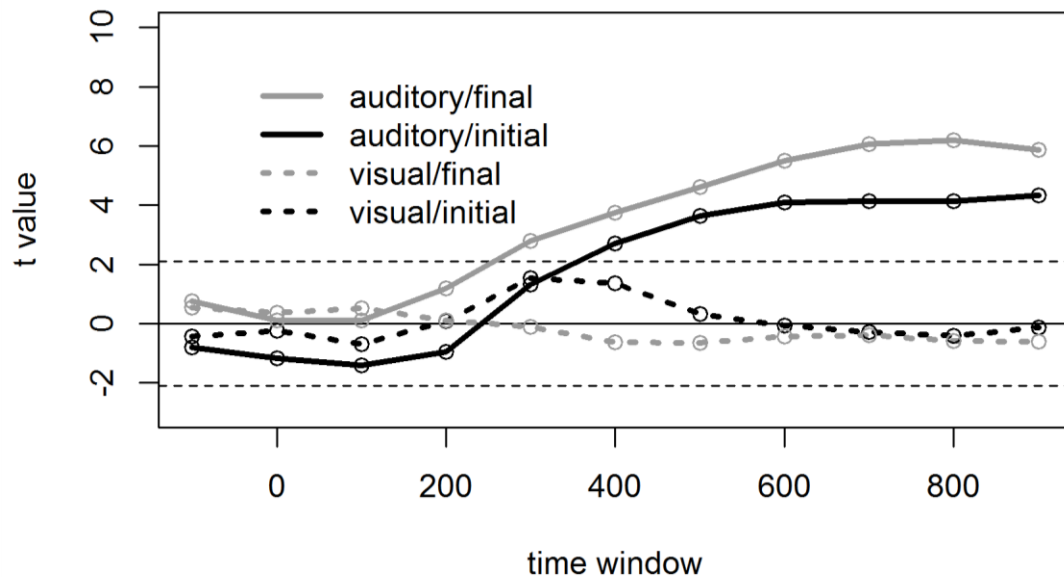


Figure 5. Results of the time-window analysis in Experiment 1. The two dotted lines indicate an approximate threshold of significance, based on the median number of degrees of freedom ( $df = 18$ ) estimated by lmerTest. There is an effect of the auditory cues on the eye-movements but not of the visual cues, just as observed for the click responses.

## Discussion

The aim of this experiment was to test the time course of the use of visual and auditory cues in a visual-world paradigm and to compare these effects to a setting that closely mimics typical experiments on the use of audiovisual information in phonetic categorization (i.e., minimal-variability condition). The minimal-variability condition served as a control that the videos contain visual cues to the place of articulation of the critical consonants and showed that this was the case. When participants saw a labial closure on the video, they were more likely to perceive the corresponding word as containing a labial. Or stated simpler, our stimuli gave rise to a McGurk effect.

However, this McGurk effect disappeared in the visual-world condition in which participants had to click on one of four target pictures, which appeared one second before the onset of the speech stimulus but differed from trial-to-trial in their position. The eye-tracking data was used to show that this was also the case when participants focused on the speaker during the critical consonant release. It is also important to note that the visual angle between fixation and center of the screen with our display rarely went beyond  $10^\circ$ , a distance which hardly affects the McGurk effect (Paré et al., 2003).

It hence seems that the processing of the visual display—largely independent of eye-gaze—interferes with the processing of the visual speech. This extends the literature which shows that the

McGurk effect—though resilient to variation in visual acuity—is rather vulnerable when there is an additional visual load (Alsius et al., 2005; Navarra et al., 2010). Our results showed that this is even the case when the stimuli are neither concurrently appearing with the visual speech (note that there was a 1 s preview of the pictures before the onset of speech) nor overlapping spatially with the visual speech.

It is possible that the use of visual speech cues would have been stronger if the auditory stimuli had been more ambiguous. Note, however, that the McGurk effect and indeed the majority of studies on audiovisual speech processing has relied on unambiguous auditory stimuli. As such, strong ambiguity of the auditory information does not seem to be necessary for effects of visual speech to occur. By using slightly ambiguous auditory stimuli, we already provided a better situation for the visual signal to have some leverage over the final percept than most studies on audio-visual integration in speech perception.

There is one oddity to discuss in our results. As pointed out in the results section, participants had an overall preference for a labial interpretation for the word-final minimal pairs independent of the experimental condition. It is important to note that for these stimuli, there were two temporally separated cues that listeners could use in pairs such as *Kalb – kalt* (Engl., *calf – cold*). First, there is the formant transition into the stop closure, which is then followed by the release burst (note that stops in German are canonically released). The auditory stimuli were selected based on a pre-test, and were identified as labial in 20 or 80% of the cases. In doing so, we apparently selected items in which the formant transition was biased towards a labial interpretation, because, for released stops, the final percept is mostly determined by the burst (see, e.g., Dahan & Tanenhaus, 2004). Therefore, the stimulus with an 80% alveolar interpretation based on the combination of transition and release still had a formant transition that was biased towards a labial interpretation. This may be due to the fact that the alveolar release burst is typically louder than a labial release burst, so the mix requires a good deal of the labial release burst to be perceptually ambiguous. This result shows that our eye-tracking data are sensitive enough to reflect the online processing of such fine phonetic detail in a highly time-locked fashion. The fact that even with this measure, no effect of visual speech was observed strengthens the argument that the visual speech information is not used under a visual load.

While the data clearly show that visual cues are used in the minimal-variability but not in the visual-world condition, it is difficult to say what caused this difference. The two conditions differ in many respects such as the variability of the visual environment, response format, etc. We will return to this issue in the General Discussion, because the data of Experiment 2 provide further constraints on when visual speech cues are utilized.



## Experiment 2

In this experiment, we set out to measure audio-visual integration in a visual-world paradigm, in which participants move the mouse on a target picture in a different set-up than in Experiment 1. In order to lessen the perceptual load, we used a design in which trials were blocked by minimal pair, as in the *minimal-variability condition* of Experiment 1. Also, the visual display (i.e., the position of the pictures on the screen) stayed constant for a series of twelve trials. What differed, however, was that participants saw four pictures per trial, two pictures for each word, one to the left and one to the right of the speaker (see Figure 7). To focus attention on the speaker, the speaker produced sentences such as *Click on this comb* and pointed towards the right or left, indicating which of the pictures should be clicked on.

The reason to use a video with a gesture next to visual speech was twofold. First of all, Experiment 1 showed that a variable visual display creates an attentional load that impairs and even prevents audio-visual integration. This gives rise to the question whether *any* additional visual stimulus is sufficient to impair audio-visual integration. If using a gesture already impaired audio-visual integration, this would indicate that audio-visual speech perception might in fact not be as important in everyday interaction—peppered with gestures—than previously thought on the basis of laboratory experiments (for a similar argument for orthography, see Mitterer & Reinisch, 2015). Therefore, knowing whether the processing of visual speech information is impaired by a concurrent gesture is interesting in itself. The second rationale for using a hand gesture was our initial motivation to test the time course of the use of visual speech from lipreading. We hypothesized that using a gesture might in fact enhance the use of visual speech, as it focuses attention on the speaker. If this is the case, we are in a good position to compare the relative timing of utilization of visual and auditory speech.

This experiment was performed in English (with Maltese English speakers as participants), because in German very few minimal pairs could be used with a sentence such as *Click on this .....*. This is because, first of all, in this paradigm, target words must be nouns and, secondly, many German minimal pairs in which both members are nouns (such as *Bach – Dach*, Engl., *creek – roof*), do not give rise to minimal sentences like *Click on this ...* due to case marking on the determiner (i.e., *Klick auf dieses Dach* vs. *Klick auf diesen Bach*).

## Method

### Participants

Twenty-five native speakers of Maltese English participated in the Experiment. They were all students at the University of Malta, and used English on a daily basis. They all had normal hearing and normal or corrected-to-normal vision.

### Stimuli and Procedure

We audio-visually recorded a native speaker of Maltese English producing ten English minimal pairs in the sentence frame *Click on this...* (720x576 pixels, 25fps). The minimal pairs were differentiated by the place of articulation of stop or nasal consonants, which was either labial or alveolar (e.g., *beam* – *bean*, Table A2 provides the complete list of minimal pairs used in this experiment). Half of the words differed word-initially, the other half word-finally.

The speaker was seated on a chair in front of a white wall. During the production of the sentence she produced a pointing gesture to the left or right. For each minimal pair, four recordings were selected—one for each combination of target word and left vs. rightward pointing gesture. To facilitate further editing, care was taken that the timing of the speech stream of the selected videos was similar across the four recordings. These short clips were cut to a duration of 1.2s with the audio starting between the seventh and eighth frame. A 200 ms fade-in and fade-out was added with a 5-frame transition between a black frame and the first and last video frame respectively. The video was then cropped to a size of 430 x 576 pixels, which was the minimal size so that the head and the pointing gesture were fully visible in all videos.

The audio was extracted from these video clips. Based on these recordings, the MBORLA diphone speech synthesizer (voice us1) was used to generate a version of the sentence based on the mean duration of each segment and following the intonation contour of one of the items. We chose the use of a speech synthesizer, because this made it easier to generate multiple versions of the preceding sentences that did not convey coarticulatory information from the word-final /s/ in *this* regarding the upcoming target. This was achieved by introducing a 10 ms silence between the /s/ and the following segment, leading to the selection of an /s/-silence diphone for all utterances.

From these resynthesized sentences, the target words were spliced out and audio-only continua were generated using the Straight audio morphing algorithm (Kawahara et al., 1999). Analogous to Experiment 1, six speakers then categorized these stimuli as either the labial or alveolar member of the

minimal pairs. Audio files receiving 20% and 80% of labial responses were selected, spliced back into the sentences, and then combined with the video files. In these recombined videos, the gesture started to be visible in the videos on average after 510 ms (range 360 ms – 640 ms), which is after the onset of the sentence (442 ms, range 418 ms-461 ms) and reached its apex on average at 752 ms (range 560 ms – 920 ms), which is before the onset of the target word (range 962 ms – 975 ms, mean = 917 ms).

For each of the twenty words, we performed a Google image search and selected two images that represented each word (see Figure 6 for examples). Images were scaled to 200x250 pixels. The video of the speaker appeared on the center of the screen. There were two different conditions in which the video was combined with the pictures.

In the first condition, four different pictures appeared on the screen. The video of the speaker appeared centered on the screen and the pictures were placed to the right and left of the video (see Figure 6). The distance between the middle of the pictures and the center of the screen was horizontally about 9.3° and vertically about 4.9°. The arrangement of the images on the screen was constant for twelve consecutive trials on which participants were presented the same minimal pair. Within these blocks, the order of audio-visual stimuli was randomized independently for each participant. Images were presented 0.5 s before the onset of the video, so that in total, there was a preview of 1 s before the audio started. There were two pictures for each word of a minimal pair, one on each side, with the “labial” picture always on top. The hand gesture of the speaker then determined the side of the target picture, so that it remained clear for the participant which picture was the target. They were asked to use the mouse cursor and click on the target. We will call this the face&gesture condition.

The second condition was similar to the minimal-variability condition in Experiment 1. Only two response options for a pair were displayed on the screen. The two pictures were presented on the upper right and upper left position (see Figure 6), and participants were asked to press the left or right mouse button to indicate whether the utterance better matched the left or right picture. As in the minimal-variability condition of Experiment 1, no mouse cursor was visible. Importantly, the lower part of the video was occluded by a square in the grey background color. Participants could hence not see the gesture but only the face. This condition will be referred to as the face-only condition. The timing was the same as in the face&gesture condition. The experimental procedure was implemented using the SR Research Experiment Builder.



Figure 6. An example of the experimental display for the minimal pair *comb* – *cone*. The hand gesture indicates whether participants have to click on the target of the right or left set of pictures.

Participants were first familiarized with the pictures and their names. Next, participants were seated in front of a computer screen and an Eyelink SR 1000 eyetracker in desktop set-up was calibrated. For the face&gesture condition, participants were informed that they would see a speaker and four pictures, one in each of the four quadrants of the computer screen, and hear a sentence over headphones. They were asked to click with the computer mouse on the picture corresponding to the word they heard on the side indicated by the hand gesture of the speaker. They were explicitly instructed to ignore the vertical position of the gesture. That is, the hand gesture determined only whether the pictures on the right side or left side of the speaker should be clicked on, but was not diagnostic with regard to whether the upper or lower picture on that side should be clicked on. For the face-only condition, participants were instructed to click on the left mouse button if the last word in the sentence matched the left picture on the screen, and on the right mouse button if it matched the right picture on the screen.

All participants performed both conditions with 120 trials each. The order of conditions was counterbalanced across participants. For each minimal pair, there were eight possible audio-visual stimuli that arose by crossing the two auditory stimuli with the two visual stimuli and two directions of the pointing gesture. Since our main question was answered using incongruent trials, the incongruent trials

were presented twice as often as the congruent ones. Each participant hence saw—for each of the ten minimal pairs—the four congruent stimuli once (i.e., audio & video = labial and audio & video = alveolar; 4 trials) and the four incongruent stimuli twice (8 trials). In both conditions, the twelve stimuli for one minimal pair were presented consecutively, and the transition from one to the next minimal pair was indicated by a screen, which showed one of the pictures for each member of this minimal pair. Participants were informed that they would have to make a choice between these pictures for the next set of trials. For each condition, there was a break halfway through the 120 trials; participants continued by pressing a mouse button.

### Data Processing

The data were processed in the same way as in Experiment 1. Rejection of trials with a fixation position outside the screen area lead to the rejection of 69 trials (1.2 % of the data). The maximal rejection rate for a given participant was 7.5%, and the data from all participants were retained. For the remaining trials, the fixations positions were categorized as being on the face, on the arms/hands, on other parts of the video, or on one of the pictures. Fixations were only counted as on a picture, if the fixation position was on a pixel that was occupied by that picture. This was necessary, because the video and the picture were close to each other (though not overlapping).

## Results

### Overall gaze patterns

First of all, we provide an overview on which parts of the display participants looked in the different conditions. For the face&gesture condition, we distinguished looks to the face, the hands, and the pictures. Since the hands were not visible during the face-only condition, only looks to the face and the pictures were considered for this condition. Figure 7 provides the relevant results. As in Experiment 1, the task which required participants to move the mouse to a referent picture to click on it led to more looks away from the face. Similar as in Experiment 1, participants looked away from the face earlier when the critical phoneme was word final. They also looked less at the pictures in the condition in which they only had to use the left versus right mouse button (face-only condition) than when they had to move a mouse cursor to a picture to click on it (face&gesture condition).

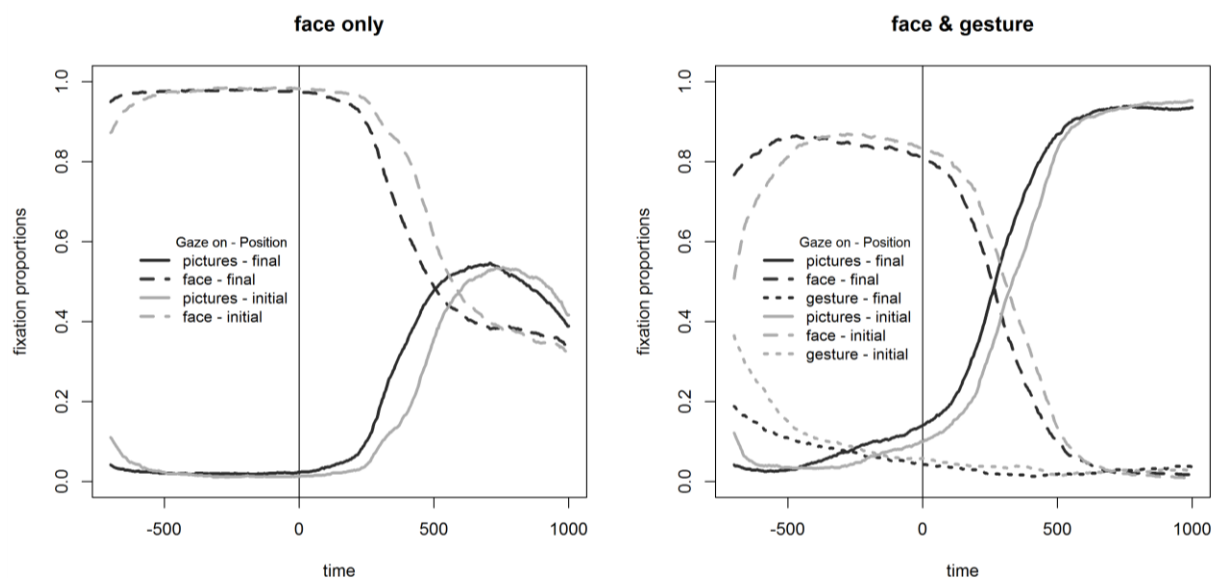
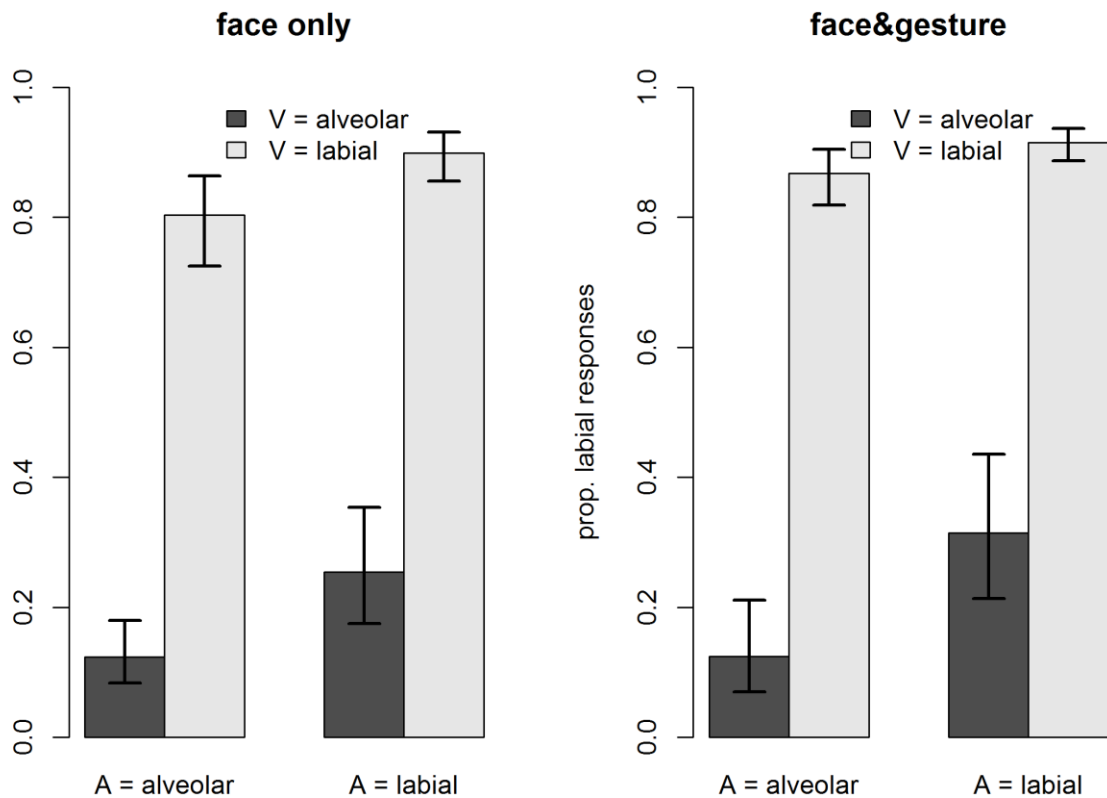


Figure 7. Eye-tracking data showing the amount of fixation on the speaker compared with fixations on the pictures in Experiment 2.

### Perceptual identifications

Figure 8 shows the proportion of cases in which the stimuli were identified as *labial*, that is, responses in which the labial member of the minimal pair was chosen, for each combination of auditory and visual cues. There was a strong influence of the visual cue in both conditions, and a smaller but consistent influence of the auditory cue as well.

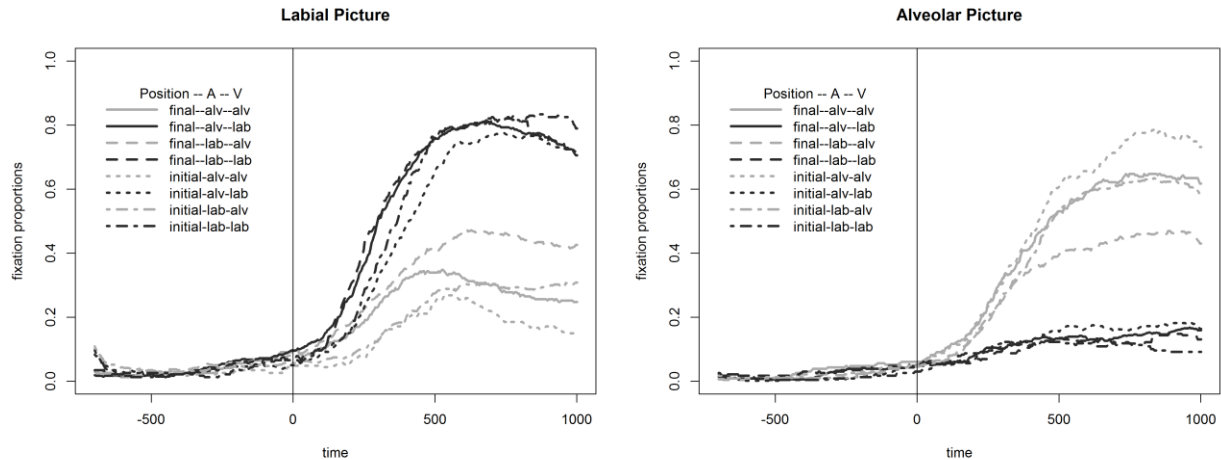
This was confirmed by a linear mixed effects model using participant and wav-file as random effects and visual and auditory cues and their interaction with condition as fixed effects. All possible random slopes were included. No other higher-order interactions were specified. All possible random slopes were specified. Data were coded as in Experiment 1 (response: 0 = alveolar, 1 = labial; visual and auditory cues: -0.5 = alveolar, 0.5 = labial). The analysis revealed an effect of visual cue ( $b = 3.811$ ,  $SE = 0.377$ ,  $z = 10.108$ ,  $p < 0.001$ ) that was not qualified by an interaction with condition ( $b = 0.133$ ,  $SE = 0.300$ ,  $z = 0.443$ ,  $p = 0.657$ ) and an effect of auditory cue ( $b = 0.987$ ,  $SE = 0.387$ ,  $z = 2.54$ ,  $p = 0.011$ ) that was also not qualified by an interaction with condition ( $b = 0.011$ ,  $SE = 0.196$ ,  $z = 0.056$ ,  $p = 0.956$ ). That is, in this experiment the visual and auditory cues were effective in both conditions.



*Figure 8.* Perceptual decisions—coded as the proportion of trials in which participants chose the labial member of the pair—in the two identification tasks in Experiment 2, depending on the auditory (x-axis) and visual cue (light bars: visual cues for labial, dark bars: visual cue for alveolar) in the stimulus. Mean and error bars (representing standard error of the mean following Morey’s, 2008, method for within-participants design) are calculated in logOdds but transformed back to proportions. The results indicate that both cues were effective in both conditions.

#### Time course of auditory- and visual-cue effects

Given that we observed an effect of the visual cue in the face&gesture condition, we are in a position to compare the time course of the effects that the auditory and visual cues have on the looks to the pictures. Figure 9 shows the fixation patterns on the pictures in the face&gesture condition. As in Experiment 1, line color is used to represent the visual cue, and it immediately becomes apparent that the visual cue influenced gaze positions. If the visual cue was labial, participants fixated mostly on the labial picture (black lines in the left panel of Figure 9), while they mostly fixated on the alveolar pictures when the visual cue was alveolar (grey lines in the right panel of Figure 9). There was also an effect of auditory cue, with more looks to the labial pictures when the auditory cue was labial (solid and dashed black line in the right panel of Figure 9).



*Figure 9.* Fixation proportions in Experiment 2 to the labial target pictures (left panel) and alveolar target pictures (right panel) depending on position and auditory cue (line type) and visual cue (line color). The comparison of dark (visual cue = labial) and light (visual cue = alveolar) lines shows the effect of the visual cue.

Since there was both an effect of visual and auditory speech cues, the time course of both effects can be compared. Since the effect of the visual speech cues is larger, it will naturally be significant earlier in a moving time window analysis; therefore we focus on the jackknife method, in which the effects are normalized (McMurray et al., 2008; Mitterer & Reinisch, 2013). For this analysis, we focused on the time window of -200-800 ms and calculated the auditory effect by adding up the preference for the labial picture over the alveolar picture when the auditory cue was labial and subtracted from that the preference for the labial over the alveolar picture when the auditory cue was alveolar. In the same fashion, we calculated an overall visual effect. These effects were baselined for the first 100ms and normalized so that the maximum effect was 1.0. A 50 ms sawtooth filter was used to low-pass filter the effect curves (cf. McMurray et al., 2008). We then calculated at what point in time the visual and auditory effects reached 10, 20, 30, and 40% of their maxima for subsamples of the data with each subsample having one participant removed (i.e., a jack-knife procedure, see McMurray et al., 2008; Mitterer & Reinisch, 2013) and compared these values for the visual and auditory effect.



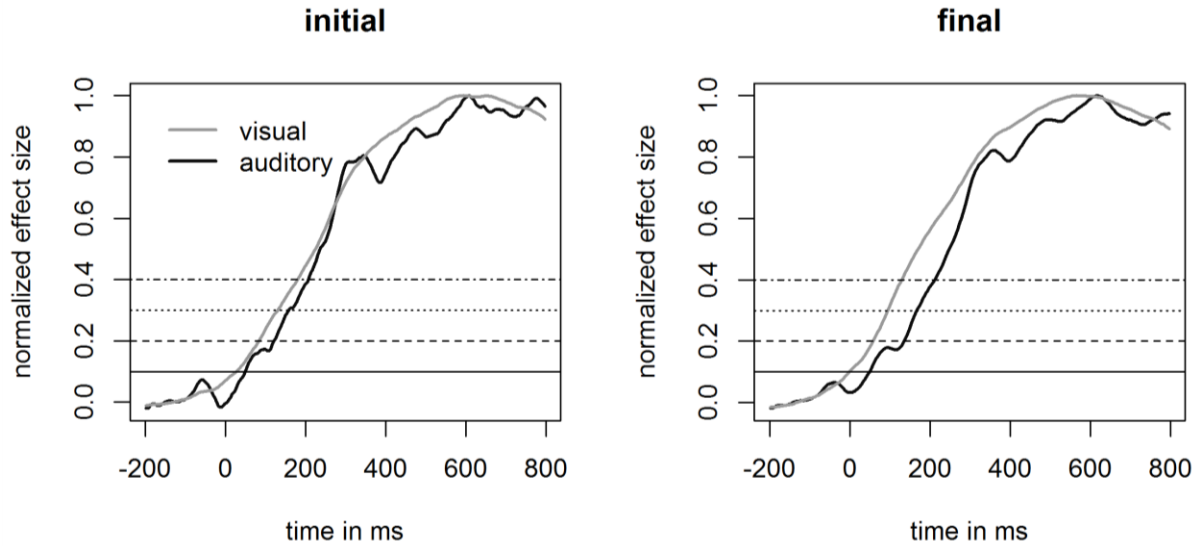


Figure 10. Time course of the normalized auditory and visual influences on the fixation proportions. The data show the visual (grey lines) and auditory effect (black lines) having a similar time course for the word-initial minimal pairs, but the visual leads the auditory effect for the word-final minimal pairs.

Figure 10 shows the time course for both effects separately for the word-initial and word-final minimal pairs. The onset latency for the auditory and visual effects were compared by means of a t-test using the data from the twelve subsamples with one participant removed. This t-value was corrected by dividing by twenty-four, correcting for the fact that the data from each participant was used twenty-four times ( $=n-1$ ). Table 1 shows the results of these analyses: in initial position, both cues were used at around the same time, while, in final position, the effect of the visual cues seemed to arise earlier than the effect of auditory cues. This result may be surprising, as a visual inspection of Figure 9 suggests that the visual cues are always earlier. However, this impression is influenced by the larger size of the visual effects, which makes them easier to pick out by visual inspection of Figure 9.

Table 1: *Jackknife estimates of onset latencies of the auditory and visual effects in Experiment 2*

Position	% max	Auditory	Visual	$t_{(\text{adjusted})}(24)$	$p_{(\text{adjusted})}$
Initial	10	149 ms	128 ms	-0.696	.247
	20	221 ms	184 ms	-0.628	.268
	30	262 ms	231 ms	-0.507	.308
	40	307 ms	280 ms	-0.218	.272
Final	10	89ms	82 ms	-0.042	.483
	20	244 ms	140 ms	-2.766	.005*
	30	264 ms	173 ms	-2.751	.006*
	40	291 ms	201 ms	-1.088	.143

Note: \*  $p < 0.05$ , the time points indicate when an effect reached a certain percentage (i.e., 10, 20, 30, or 40% of its maximum).

## Discussion

Experiment 1 showed that high visual load prevented the use of visual cues in speech perception. This second experiment tested the scope of this finding by asking whether even a simple gesture would be sufficient to hinder the use of visual speech cues. Part of the motivation to use a gesture was that this would constitute the minimal distraction that occurs in every face-to-face conversation. It could serve as a yardstick whether visual speech would matter in spontaneous conversation at all. Our results show that visual speech may well be used in face-to-face conversations, because a simple gesture did not suffice to block the use of visual speech cues. Notably in the face&gesture condition, the gesture was even task relevant. The results thus showed that participants made use of the visual cues even if they also had to attend to a deictic gesture.

Because there was a clear effect of the visual cues in the face&gesture condition, we were in a position to examine our main question about the time course of the use of visual and auditory cues using a jackknife analysis to determine the onset of the two effects (McMurray et al., 2008; Mitterer & Reinisch, 2013). This analysis showed that the onset of the visual effect only led the onset of the auditory effect for word-final pairs. This is surprising, because, if anything, we had expected the opposite. For word-initial minimal pairs, the visual cues clearly precede the auditory cues, because the place of articulation for the stops or nasals is clearly visible even before the onset of the sound. In contrast, for the word-final minimal pairs, the visual and auditory cues are simultaneous. That is, there is phonation concurrent to the closing gesture, which then gives rise to audible formant transitions. This seems to indicate that visual speech by itself is not used to predict speech sounds. If so, we should see that the visual cues lead the auditory cues more for the word-initial minimal pairs, which is contrary to fact. This finding hence suggests that the full use of visual speech requires a concurrent auditory signal.

It may seem surprising that the visual speech cues even precede the use of the auditory speech cues to labial vs. alveolar place of articulation for word-final pairs. Note, however, that the speech cues in question are formant transitions out of the vowel into the consonant. Such formant transitions from a vowel into a nasal or stop have been shown to be difficult to perceive (Repp & Svastikula, 1988; Steriade, 2001), so that the visual lip-closing gesture may be a more reliable cue than the concurrent auditory formant transition. Stronger auditory cues to place of articulation are to be found in the nasal murmur and especially in the stop release (see Experiment 1), but those become available later and are hence likely to exert an influence later than the visual cues.

The results from Experiment 2 also further narrow down the possible explanations for why, in Experiment 1, an effect was found only in the minimal-variability condition but not in the visual-world

condition. The conditions differed not only in visual load but also in response complexity, because participants had to move the mouse only in the visual-world condition. The fact that an effect of visual speech cues was found in the face&gesture condition of the current experiment rules out that the response format is crucial. In the face&gesture condition, participants also had to move the mouse, showing that this does not interfere with the use of visual speech cues.

## General Discussion

Two experiments used eye tracking to evaluate when visual cues for speech perception are utilized in comparison to auditory cues. Experiment 1 showed that observers do not use visual speech when interacting with unpredictable visual displays containing four different pictures that change on every trial. A minimal-variability condition that acted as a control showed that the audio-visual speech stimuli were nevertheless capable of inducing a McGurk effect. These first results hence suggested that the visual environment determines the use of visual speech. Experiment 2 then reduced the visual load in the display by blocking critical words and pictures while asking participants to focus on the speaker by making a pointing gesture task relevant. With this set-up, we found an effect of visual speech. This indicates that a pointing gesture as a concurrent visual load is not sufficient to block the use of visual speech cues. This set-up then allowed us to investigate our main question about the time course of the use of visual and auditory information in speech perception. An analysis of the time-course of auditory and visual effects revealed that the visual signal is used early and, for word-final consonants, even before the auditory signal.

The two main results from these experiments are that visual speech cues are not always used, but if they are used, they are used immediately. The first of these results adds to the literature that the use of visual speech cues can falter under a processing load (Alsius et al., 2005; Navarra et al., 2010). Experiment 2 ruled out that the absence of the effect in the visual world condition was due to a response load (i.e., by asking participants to move the mouse onto one of the pictures and click on it). It is also unlikely that the larger size of the video in Experiment 2 can explain the difference in results, since the area covered by the face was larger in Experiment 1 (240x 410 pixels) than in Experiment 2 (110 x 200 pixels). Additional differences include that participants had to consider four lexical items in Experiment 1 but only two in Experiment 2, and that attention was taken away from the speaker in Experiment 1 through higher visual load but drawn to the speaker in Experiment 2 by means of the gesture. While we cannot rule out that the crucial difference is that the number of lexical candidates to be considered is crucial, these last two differences suggest that the attentional processing of the visual display—largely

independent of eye-gaze—interferes with the processing of the visual speech, even though we cannot completely rule out that considering four different lexical items is the culprit. Apparently, attention might be withdrawn from the speaker in Experiment 1 even when the picture stimuli are neither concurrently appearing with the visual speech (note that there was a 1 s preview of the pictures before the onset of speech) nor overlapping spatially with the visual speech. This suggests that the use of visual speech cues requires attention, rather than fixations onto the speaker's face, that is, the visual speech. Paré et al. (2003) tested whether participants would still use visual speech cues if they were forced to fixate on a spot that was 20° away from the speaker's mouth. Under these conditions, there still was a strong influence of visual speech cues that was only mildly reduced compared to a situation in which participants were focusing on the speaker. Note, however, that no other visual stimuli had to be processed in that experiment, so that participants were able to direct covert attention (note that attention is not necessarily at the point of fixation, see, e.g., van der Heijden, 1992) to the speech gestures independent of gaze position. Our experiments present the opposite situation: participants gazed at the speaker's mouth but had to process other visual information, leading to the allocation of attention away from the speaker's mouth. This suggests that it is not so much gaze position but rather allocation of visual attention which influences the utilization of visual speech cues. Such an explanation would also be in line with the findings of Baart and Vroomen (2010), who found no decline in the use of visual speech cues in a perceptual-recalibration paradigm (see, e.g., Bertelson, Vroomen, & de Gelder, 2003; Reinisch, Wozny, Mitterer, & Holt, 2014) when participants only had a memory load but did not have to process other visual stimuli during the presentation of audiovisual speech. Hence, participants could attend to the visual speech gestures, which, in turn, lead to their utilization.

In their review, Navarra et al. (2010) discussed two possibilities for how a processing load may affect the use of visual speech gestures. Either the processing of the visual speech gestures is compromised or the process of multisensory integration is compromised, even though visual processing is extracting the relevant features from the input. Our results would support the latter possibility, since the eye-tracking record shows that the participants were mostly focusing on the speaker during the critical visual release of the stop consonant. It is unlikely that the visual speech could not be processed when participants' gaze is focused on it. Even though participants were under concurrent visual load as they had to recognize the objects, this is unlikely to interfere with the perception of biological motion. Object recognition and the perception of biological motion seem to engage different neural substrates (Giese & Poggio, 2003). Both, pointing gestures and the speech gestures are examples of biological motion. Considering this, if the problem was a processing bottleneck, the pointing gestures should have been more

disruptive than a static array of pictures, since in the former but not the latter case two types of biological motion had to be processed. Moreover, Thornthorn and Vuong (2004) showed that biological motion can be processed outside the focus of attention. It hence seems that it is the *use* of visual speech cues that is suspended under attentional load rather than the processing of visual speech signal itself.

This brings us to the second main result, that visual speech cues influence processing early. At first glance, the eye-tracking data on the early use of visual speech lead to a different conclusion than the ERP data discussed in the introduction. Evidence from ERPs indicated that the visual speech signal—or more precisely, the phonetic context of the visual speech signal—is only integrated late with auditory speech information (Colin et al., 2002; Meyer et al., 2013; Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010). However, there is a way to reconcile these two sets of findings. The fact that visual speech is *used* early and immediately does not necessarily require that it is *integrated* early with the auditory information. It is possible that both auditory and visual streams are evaluated early but independently and only integrated late, as assumed by the FLMP (Massaro, 1998). The visual information could then still be used immediately but without influencing the processing of the auditory stream. This account fits well with a recent report by Baart and Samuel (2015), who investigated lexical and visual context effects in speech perception. They find relatively early visual effects and early lexical effects (<400ms), but the visual effects do not moderate the lexical effects. If the lexical effects arise in an auditory processing stream, these results would align with ours that the visual speech signal is used early but does not influence auditory processing.

It is in fact difficult to see how an influence of visual information on auditory processing itself would be beneficial for spoken-word recognition. To account for audio-visual benefits for speech-in-noise, it is sufficient to assume that the auditory and the visual signal are both used to decide the most likely input; such findings do not necessarily show that auditory processing itself is altered by the visual input. Note that similar arguments have been made regarding lexical contributions to spoken-word recognition (Frauenfelder & Peters, 1998; Norris, McQueen, & Cutler, 2000). As indicated by computational modeling simulations in Frauenfelder and Peters (1998), spoken-word recognition is not facilitated if lexical information influenced lower-level auditory perception. Similarly, it may be the case that optimal audiovisual speech perception is better achieved by an independent evaluation of the auditory and visual signal.

Interestingly, the effect of the visual cues only preceded the effect of the auditory cues in word-final position. As argued above, this is counterintuitive, because the lead of the visual signal in terms of when cues are available in the signal is larger in word-initial position. One way to explain this paradox is

that visual speech needs an auditory signal to be fully utilized. This clashes with claims that speech is fundamentally multimodal (or amodal, see Rosenblum, 2008) but would support the idea that speech perception is mainly auditory with contributions from visual speech if sufficient resources are available.

There are three lines of evidence that would support such a claim. First of all, Experiment 1 showed that audiovisual integration in speech perception falters in highly variable visual environments even when during the critical words this environment was static. This would suggest that in all interactions in which interlocutors are talking about the visual environment, speech perception proceeds without much influence from visual speech. Our data from Experiment 1 show that, despite looking at the speaker, covert attention to other parts of the display seems to be sufficient to render the visual signal inconsequential; hence, when the joint attention in the dialogue is on a visual referent, visual speech cues are unlikely to contribute much to speech communication. Indeed, the contrast between the use of visual speech in Experiment 1 and Experiment 2 indicates that there is still research to be done with regard to the question under what circumstances visual cues are utilized. Secondly, Moos and Trouvain (2007) documented that blind listeners achieve speech-perception capabilities well above those of sighted, normal-hearing listeners. This seems to indicate that the visual signal is not necessary for perceivers to become expert speech recognizers. However, even a relatively mild hearing impairment can cause problems in oral language acquisition, and profoundly deaf children acquire very little spoken language at all (Mogford, 1993). Thirdly, there is now strong empirical (rather than merely anecdotal) evidence that gaze patterns in conversation differ strongly between different cultures (Rossano, Brown, & Levinson, 2009). In contrast to this strong variation in gaze patterns, there is very little variation across languages in how conversations progress auditorily. Across different languages conversations proceed with invariably short silences between the start of the speakers' turns (Stivers et al., 2009). If speech was inherently multimodal, there would be little room for cross-cultural variation in gaze patterns.

To summarize, our data show that visual speech cues are not used automatically. If there is a concurrent processing load, visual speech cues, even though probably processed in parallel, are not used for speech perception. Nevertheless, when the visual speech cues are used, they are used immediately. This suggests that visual and auditory streams may be processed independently and only integrated at a later stage.

## References

- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419–439. <https://doi.org/10.1006/jmla.1997.2558>
- Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, 15(9), 839–843. <https://doi.org/10.1016/j.cub.2005.03.046>
- Altmann, G. T. M. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica*, 137(2), 190–200. <https://doi.org/10.1016/j.actpsy.2010.09.009>
- Baart, M., & Samuel, A. G. (2015). Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. *Journal of Memory and Language*, 85, 42–59. <https://doi.org/10.1016/j.jml.2015.06.008>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Barutchu, A., Crewther, S. G., Kiely, P., Murphy, M. J., & Crewther, D. P. (2008). When /b/ill with /g/ill becomes /d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology*, 20(1), 1–11. <https://doi.org/10.1080/09541440601125623>
- Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, 14, 592–597.
- Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 30(3), 445–463. <https://doi.org/10.1037/0096-1523.30.3.445>
- Brunellière, A., Sánchez-García, C., Ikumi, N., & Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *International Journal of Psychophysiology*, 89(1), 136–147. <https://doi.org/10.1016/j.ijpsycho.2013.06.016>
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58(5), 412–424. <https://doi.org/10.1027/1618-3169/a000123>
- Colin, C., Radeau, M., A., S., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology*, 113(495–506).
- Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology*, 6(1), 84–107. [https://doi.org/10.1016/0010-0285\(74\)90005-X](https://doi.org/10.1016/0010-0285(74)90005-X)
- Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 498–513.
- Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17, 816–828.
- Frauenfelder, U. H., & Peters, G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in TRACE. In J. Grainger & A. M. Jacobs (Eds.), *Localist*

- connectionist approaches to human cognition* (pp. 101–146). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Heuven, W. J. B. van, Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6), 1176–1190. <https://doi.org/10.1080/17470218.2013.850521>
- Huetting, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, 137(2), 151–171. <https://doi.org/10.1016/j.actpsy.2010.11.003>
- Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction. *Speech Communication*, 27, 187–207. [https://doi.org/10.1016/S0167-6393\(98\)00085-5](https://doi.org/10.1016/S0167-6393(98)00085-5)
- Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique*. Boston: MIT Press.
- Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21(2), 131–141. <https://doi.org/10.3109/03005368709077786>
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User's Guide* (2 Rev ed.). Mahwah, N.J: Lawrence Erlbaum.
- Massaro, D. W. (1998). *Perceiving talking faces: from speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.
- McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, 15(6), 1064–1071. <https://doi.org/10.3758/pbr.15.6.1064>
- Meyer, G. F., Harrison, N. R., & Wuerger, S. M. (2013). The time course of auditory–visual processing of speech and body actions: Evidence for the simultaneous activation of an extended neural network for semantic processing. *Neuropsychologia*, 51(9), 1716–1725. <https://doi.org/10.1016/j.neuropsychologia.2013.05.014>
- Mitterer, H. (2011). The mental lexicon is fully specified: Evidence from eye-tracking. *Journal of Experimental Psychology: Human Perception and Performance*, 37, 496–513. <https://doi.org/10.1037/a0020989>
- Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language*, 69(4), 527–545. <https://doi.org/10.1016/j.jml.2013.07.002>
- Mitterer, H., & Reinisch, E. (2015). Letters don't matter: No effect of orthography on the perception of conversational speech. *Journal of Memory and Language*, 85, 116–134. <https://doi.org/10.1016/j.jml.2015.08.005>
- Mogford, K. (1993). Oral language acquisition in the prelinguistically deaf. *Language Development in Exceptional Circumstances*, 110–131.
- Moos, A., & Trouvain, J. (2007). Comprehension of Ultra-Fast Speech–Blind vs. 'Normally Hearing' Persons. In *Proceedings of the 16th International Congress of Phonetic Sciences* (Vol. 1, pp. 677–680). Retrieved from <http://www.icphs2007.de/conference/Papers/1186/1186.pdf>
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau. In *Tutorials in Quantitative Methods for Psychology*, 4 (pp. 61–64).



- Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion*, 11(1), 4–11. <https://doi.org/10.1016/j.inffus.2009.04.001>
- Norris, D., McQueen, J. M., & Cutler. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(3), 299–325. <https://doi.org/10.1017/S0140525X00003241>
- Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What you see isn't always what you get: Auditory word signals trump consciously perceived words in lexical access. *Cognition*, 151, 96–107. <https://doi.org/10.1016/j.cognition.2016.02.019>
- Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, 65(4), 553–567.
- Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, 41(2), 101–116. <https://doi.org/10.1016/j.wocn.2013.01.002>
- Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics*, 45, 91–105. <https://doi.org/10.1016/j.wocn.2014.04.002>
- Repp, B. H., & Svastikula, K. (1988). The perception of the [m]–[n] distinction in VC syllables. *Journal of the Acoustical Society of America*, 83, 237–247.
- Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, 17(6), 405–409. <https://doi.org/10.1111/j.1467-8721.2008.00615.x>
- Rossano, F., Brown, P., & Levinson, S. C. (2009). Gaze, questioning and culture. In J. Sidnell (Ed.), *Conversation analysis: Comparative perspectives* (pp. 187–249). Cambridge: Cambridge University Press. Retrieved from <https://books.google.de/books?hl=de&lr=&id=c0tTAAFLBGoC&oi=fnd&pg=PA187&dq=Rossano+Levinson+gaze&ots=FHikyG2S7s&sig=AYsstebe3ar9PVLw9NnX4rdDO-A>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, 71(1), 145–163. <https://doi.org/10.1016/j.jml.2013.11.002>
- Sams, M., Manninen, P., Surakka, V., Helin, P., & Kättö, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, 26(1–2), 75–87. [https://doi.org/10.1016/S0167-6393\(98\)00051-X](https://doi.org/10.1016/S0167-6393(98)00051-X)
- Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19(12), 1964–1973.
- Steriade, D. (2001). Directional asymmetries in place assimilation: a perceptual account. In E. Hume & K. Johnson (Eds.), *The role of speech perception in phonology* (pp. 219–250). New York, NJ: Academic Press.
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26), 10587–10592. <https://doi.org/10.1073/pnas.0903616106>
- Thornton, I. M., & Vuong, Q. C. (2004). Incidental processing of biological motion. *Current Biology*, 14(12), 1084–1089. <https://doi.org/10.1016/j.cub.2004.06.025>

- Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*(3), 457–472. <https://doi.org/10.1080/09541440340000268>
- Turenout, M. van, Hagoort, P., & Brown, C. M. (1998). Brain activity during speaking: from syntax to phonology in 40 milliseconds. *Science, 280*(5363), 572–574. <https://doi.org/10.1126/science.280.5363.572>
- van der Heijden, A. H. C. (1992). *Selective attention in vision*. New York: Routledge.
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181–1186. <https://doi.org/10.1073/pnas.0408949102>
- Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience, 22*(7), 1583–1596.
- Windmann, S. (2004). Effects of sentence context and expectation on the McGurk illusion. *Journal of Memory and Language, 50*(2), 212–230. <https://doi.org/10.1016/j.jml.2003.10.001>

## Appendix

Table A1:

*List of German minimal pairs used in Experiment 1*

Place of Articulation					
Labial			Alveolar		
German	Translation	Frequency	German	Translation	Frequency
Alp	Alpine meadow	0.33	alt	old	160.12
Kalb	Calf	3.54	kalt	cold	80.51
halb	Half	47.68	Halt	stop	359.42
Laub	leaves	1.02	laut	loud	84.65
Gelb	yellow	8.35	Geld	money	642.7
Dock	dock	3.11	Bock	ram	7.21
Bühne	stage	28.98	Düne	dune	0.63
Bach	creek	9.76	Dach	roof	48.78
Panne	breakdown	2.68	Tanne	fir	0.43
Pier	pier	5.04	Tier	animal	34.45

Note: Frequencies are given as frequency per million based on the SUBTLEX-DE corpus (Brysbaert et al., 2011). The concept of *Alpine Meadow*, is a frequent one for inhabitants of Munich. Moreover, German uses final devoicing in stops, so that *Kalb* [kalp] and *kalt* [kalt]) are truly a minimal pair.

Table A2

*List of English minimal pairs used in Experiment 2*

Place of Articulation	
Labial	Alveolar
beer (29.82)	deer (18.17)
boar (3.72)	door (182.04)
bow (16.57)	dough (11.1)
mail (42.47)	nail (15.09)
mite (1.21)	night (447.97)
beam (9.24)	bean (11.44)
comb (3.89)	cone (4.65)
gum (5.19)	gun (39.75)
lab (19.07)	lad (27.08)
robe (1.29)	road (190.92)

Note: Numbers in brackets are frequencies per million based on the SUBTLEX-UK corpus (Heuven, Mandera, Keuleers, & Brysbaert, 2014).