Journal of Speech, Language, and Hearing Research

Eye Gaze and Perceptual Adaptation to Audiovisual Degraded Speech

--Manuscript Draft--

Manuscript Number:	JSLHR-21-00106R1
Full Title:	Eye Gaze and Perceptual Adaptation to Audiovisual Degraded Speech
Article Type:	Research Article
Section/Category:	Speech
Corresponding Author:	Briony Banks Lancaster University Lancaster, Lancashire UNITED KINGDOM
Other Authors:	Emma Gowen
	Kevin J Munro
	Patti Adank
Funding Information:	Biotechnology and Biological Sciences Research Council — Dr Briony Banks
Keywords:	Speech Perception; audiovisual speech; eye tracking
Manuscript Classifications:	Cognition; Speech perception; Speech recognition
Abstract:	Purpose : Visual cues from a speaker's face may benefit perceptual adaptation to degraded speech, but current evidence is limited. We aimed to replicate results from previous studies to establish the extent to which visual speech cues can lead to greater adaptation over time, extending existing results to a real-time adaptation paradigm (i.e., without a separate training period). A second aim was to investigate whether eye gaze patterns towards the speaker's mouth were related to better perception, hypothesising that listeners who looked more at the speaker's mouth would show greater adaptation. Method: A group of listeners (N =30) were presented with 90 noise-vocoded sentences in audiovisual format while a control group (N =29) were presented with the audio signal only. Recognition accuracy was measured throughout and eye tracking was used to measure fixations towards the speaker's eyes and mouth in the audiovisual group. Results: Previous studies were partially replicated: the audiovisual group had better recognition throughout and adapted slightly more rapidly, but both groups showed an equal amount of improvement overall. Longer fixations on the speaker's mouth in the audiovisual group were related to better overall accuracy. An exploratory analysis further demonstrated that the duration of fixations to the speaker's mouth decreased over time. Conclusions: The results suggest that visual cues may not benefit adaptation to degraded speech as much as previously thought. Longer fixations on a speaker's mouth may play a role in successfully decoding visual speech cues, however this will need to be confirmed in future research to fully understand how patterns of eye gaze are related to audiovisual speech recognition. All materials, data, and code are available at https://osf.io/2wqkf/ .
Response to Reviewers:	Thank you for reviewing our manuscript "Eye Gaze and Perceptual Adaptation to Audiovisual Degraded Speech", manuscript no. JSLHR-21-00106. We appreciate the helpful and informative comments from the editor and the reviewers. We have addressed all of these in our revision of the manuscript; particularly, we have toned down some of our conclusions, provided extra details as requested, and revised the introduction and discussion accordingly, including additional literature where appropriate. We have also ensured that all data and materials are now publicly available on the OSF. We attach a response to reviewers addressing each point in detail. All changes are highlighted in yellow in the manuscript. We have also addressed the minor points highlighted by the editorial team, again highlighted in yellow. Thank you for considering the revised version of our paper for publication in the Journal of Speech, Language, and Hearing Research, and we look forward to hearing from you.

Eye Gaze and Perceptual Adaptation to Audiovisual Degraded Speech

Briony Banks [1], Emma Gowen [1], Kevin J Munro [2,3] and Patti Adank [4]

[1] Faculty of Biology, Medicine and Health, The University of Manchester

[2] Manchester Centre for Audiology and Deafness, Faculty of Biology, Medicine and Health, The University of Manchester

[3] Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre

[4] Speech, Hearing and Phonetic Sciences, University College London

Author Note

Briony Banks is now at Department of Psychology, Lancaster University.

We have no known conflict of interest to disclose.

BB was funded by a BBSRC research studentship and by The University of Manchester. KJM is supported by the NIHR Manchester Biomedical Research Centre.

Correspondence concerning this article should be addressed to Briony Banks, Department of Psychology, Fylde College, Lancaster University, Lancaster, LA1 4YF, UK, email: b.banks@lancaster.ac.uk

Abstract

Purpose: Visual cues from a speaker's face may benefit perceptual adaptation to degraded speech, but current evidence is limited. We aimed to replicate results from previous studies to establish the extent to which visual speech cues can lead to greater adaptation over time, extending existing results to a real-time adaptation paradigm (i.e., without a separate training period). A second aim was to investigate whether eye gaze patterns towards the speaker's mouth were related to better perception, hypothesising that listeners who looked more at the speaker's mouth would show greater adaptation.

Method: A group of listeners (*N*=30) were presented with 90 noise-vocoded sentences in audiovisual format while a control group (*N*=29) were presented with the audio signal only. Recognition accuracy was measured throughout and eye tracking was used to measure fixations towards the speaker's eyes and mouth in the audiovisual group.

Results: Previous studies were partially replicated: the audiovisual group had better recognition throughout and adapted slightly more rapidly, but both groups showed an equal amount of improvement overall. Longer fixations on the speaker's mouth in the audiovisual group were related to better overall accuracy. An exploratory analysis further demonstrated that the duration of fixations to the speaker's mouth decreased over time.

Conclusions: The results suggest that visual cues may not benefit adaptation to degraded speech as much as previously thought. Longer fixations on a speaker's mouth may play a role in successfully decoding visual speech cues, however this will need to be confirmed in future research to fully understand how patterns of eye gaze are related to audiovisual speech recognition. All materials, data, and code are available at https://osf.io/2wqkf/.

Key words: Speech perception, audiovisual speech, perceptual adaptation, eye tracking

Eye Gaze and Perceptual Adaptation to Audiovisual Degraded Speech

Human communication often takes place in suboptimal listening conditions such as in noisy environments, listening to a distorted phone or video signal, or encountering unfamiliar speech such as a foreign accent. Most listeners are adept at dealing with such difficult conditions by rapidly adapting to them – that is, undergoing a period where they learn and 'tune in' to the acoustic and perceptual differences in the particular listening condition. This perceptual adaptation to degraded or unfamiliar speech has been consistently and empirically demonstrated for a variety of adverse conditions, such as noise-vocoded (M. H. Davis et al., 2005; Hervais-Adelman et al., 2008), accented (Adank & Janse, 2010; Banks et al., 2015a, 2015b), and time-compressed speech (Peelle & Wingfield, 2005; Sebastian-Galles & Mehler, 2000). Artificially degrading the speech through noise-vocoding (Shannon et al., 1995) is particularly useful in such experiments due to the level of control that it offers the experimenter, particularly with regards to intelligibility (e.g., Dorman et al., 1997; Faulkner et al., 2000). Noise-vocoding distorts the spectral structure of speech while preserving the temporal structure, creating a speech signal that contains enough detail to be intelligible but with significantly less spectral, specifically harmonic, detail than the original (M. H. Davis et al., 2005). The relative intelligibility of the signal is associated with the number of channels initially used to divide the acoustic signal, with more channels resulting in higher levels of intelligibility (Loizou et al., 1999). Listeners can adapt to noise-vocoded sentences after relatively short exposure; for example, Davis et al. (2005) report a steady linear increase in recognition performance after listening to 30 sentences noise-vocoded into six channels, with participants improving from ~20% of words correctly reported to ~60%. Distortions such as noise-vocoding can reflect particularly challenging conditions that we might encounter in modern digital communication. However, the processes and individual strategies used during perceptual adaptation are still not fully understood, particularly the role of visual speech cues, as although we often communicate face-to-face with a speaker, the majority of research into perceptual adaptation of degraded speech has only examined auditory perception.

It is well established that access to visual cues from a speaker's face substantially improves speech recognition in difficult listening conditions; this *audiovisual benefit* has been demonstrated, for example, in the presence of background noise or with a distorted speech signal (Erber, 1975; MacLeod & Summerfield, 1987; Sommers et al., 2005; Sumby & Pollack, 1954). Listeners benefit from viewing articulatory cues, particularly from a speaker's mouth, integrating them with auditory cues and thus enhancing the overall speech signal and improving recognition (Summerfield, 1987). Attending to visual speech cues may thus improve or speed up the adaptation process required to adapt to unfamiliar or degraded speech, leading to greater improvements in speech recognition.

A handful of studies have investigated the benefits of visual speech cues in perceptual adaptation to degraded (noise-vocoded) speech, but with varying types of linguistic stimuli. At the syllable level, Bernstein, Auer, Eberhardt & Jiang (2013) found that the presence of visual speech cues leads to greater perceptual adaptation of noise-vocoded syllables. Kawase et al., (2009) extended this finding to individual noise-vocoded words, comparing perceptual adaptation with and without audiovisual speech cues (i.e., with and without the speaker's face visible), finding that listeners adapted a greater amount when visual speech cues were available to listeners compared to when they were not. However, listening to individual syllables or words, without any additional linguistic context, is not representative of everyday communication. Pilling and Thomas (2011) therefore tested auditory recognition of degraded sentences. Participants listened to 3 blocks of 76 noise-vocoded sentences, whereby the middle block was a training condition with either audiovisual, audio-only or non-degraded sentences. They observed a greater improvement in performance after training with visual cues compared to without (i.e., after exposure to audiovisual compared to audio-only sentences during training). Wayne & Johnsrude (2012) also assessed the contribution of training with visual speech information, comparing several training conditions during adaptation to noise-vocoded sentences. They found that training with audiovisual cues resulted in no more adaptation than training with non-degraded feedback – i.e., training where the listener heard the sentences both with and without noise-vocoding. However, the paradigm did not directly compare adaptation to noise-vocoded speech with and without visual speech cues as in Pilling & Thomas (2011), and it is therefore impossible to ascertain the amount of improvement that visual cues contributed to adaptation over and above the auditory signal alone. Moreover, if one is listening to speech in adverse conditions (e.g., a degraded phone or video signal) it is not always possible to obtain the type of clear (i.e. non-degraded) feedback as used in the training conditions by Wayne & Johnsrude, and visual cues may thus provide a more readily accessible source of perceptual information that can help listeners adapt to difficult listening conditions.

Both Pilling & Thomas (2011), and Wayne & Johnsrude (2012), used a training paradigm whereby adaptation was measured by testing participants *after* being exposed to audiovisual speech; however, adaptation to unfamiliar or degraded speech most likely occurs in real time – that is, we adapt to the listening conditions we are exposed to at the time, integrating useful visual cues as we adapt. Furthermore, the sentences used in both Pilling & Thomas (2011) and Wayne & Johnsrude (2012) were relatively simple in terms of vocabulary and structure. Such sentences may be relatively easy to perceive and adapt to compared to more challenging and less predictable sentences; for example, the more challenging IEEE sentences (e.g., 'Sickness kept him home the third week', 'The hog crawled under the high fence'; Rothauser et al., 1969) result in poorer recognition than the BKB sentences (e.g., 'A cat sits on the bed', 'The ice cream was pink'; Bench et al., 1979) used by Pilling & Thomas (2011), when presented in fluctuating masking (Schoof & Rosen, 2015). It is therefore possible that an equivalent audiovisual benefit to perceptual adaptation may not be present for different linguistic stimuli.

The benefit gained from visual speech cues has potential applications for listeners adapting to a variety of difficult listening conditions – whether these originate from the environment (for example background noise or a distorted phone line) or from listeners themselves in the form of a hearing impairment (Mattys et al., 2012). Nevertheless, current evidence of an audiovisual benefit to adaptation using naturalistic stimuli (i.e., sentences) comes essentially from a single study (Pilling & Thomas, 2011). The first aim of the present study was thus to replicate and extend the finding by

Pilling and Thomas (2011) that visual speech cues improve perceptual adaptation to degraded sentences, using a more naturalistic and real-time (i.e., continuous) adaptation paradigm whereby participants were continually exposed to noise-vocoded sentences with and without visual speech cues, and where recognition was measured throughout the task, rather than after a period of training. Additionally, we used the IEEE sentences (Rothauser et al., 1969), which are more complex than the BKB sentences, and thus potentially more challenging for listeners to integrate the auditory and visual signals, to more strongly test the effects of visual speech cues.

A second aim of the present study was to examine the role of eye gaze in comprehending and adapting to audiovisual degraded speech. Interest in listeners' eye gaze during speech perception has seen a recent increase (e.g., Barenholtz et al., 2016; Birulés et al., 2020; Lusk & Mitchel, 2016; Morin-Lessard et al., 2019; Wang Jianrong et al., 2020; Worster et al., 2018), with some studies suggesting a link between where and how listeners view a speaker's face and their resulting comprehension (Lusk & Mitchel, 2016; Worster et al., 2018). Adult listeners normally show a preference for looking at a speaker's eyes during communication (Morin-Lessard et al., 2019; Yarbus, 1967), which is likely for social reasons (Birmingham & Kingstone, 2009). Indeed, speech recognition studies employing eye-tracking have shown that in optimal listening conditions (i.e., in quiet and with a clear auditory signal), adults look more towards a speaker's eyes than the mouth (Buchan et al., 2007, 2008; Vatikiotis-Bateson et al., 1998). However, when listening conditions are challenging, e.g., when background noise is present, listeners look more often at a speaker's mouth (Buchan et al., 2007, 2008; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998). This pattern has also been found for artificial (Lusk & Mitchel, 2016) and non-native language (Barenholtz et al., 2016; Birulés et al., 2020). Indeed, the more challenging the condition (e.g., as background noise increases), the more frequently listeners look towards a speaker's mouth (Vatikiotis-Bateson et al., 1998) and the more attentional weighting is given to visual over auditory cues (Hazan et al., 2010). Although some useful speech cues can be gained from extra-oral areas such as the upper face and eye region (e.g., Preminger et al, 1998; Scheinberg, 1980), visible mouth movements are considerably more important for successful audiovisual speech comprehension in challenging listening conditions (Thomas & Jordan, 2004). Thus, in such conditions, listeners likely shift their attention (and thus their eye gaze) more frequently towards the speaker's mouth to benefit from the most useful visual cues (i.e., articulatory mouth movements), potentially to improve lexical segmentation (Lusk & Mitchel, 2016; Mitchel & Weiss, 2014). These observations fit well with the cognitive relevance framework of visual attention (Henderson et al., 2009), which stipulates that the weight allocated to a particular visual feature is dependent on the cognitive needs of the perceiver. Accordingly, gaze patterns towards facial features during audiovisual speech perception have been shown to vary depending on the task (Buchan et al., 2007; Malcolm et al., 2008) and the type of stimuli presented (Lansing & McConkie, 2003; Vo et al., 2012).

Observations that listeners look more towards the speaker's mouth in adverse listening conditions would suggest a direct relationship between listeners' patterns of eye gaze and successful recognition of audiovisual degraded speech – i.e., listeners' performance. Indeed, in both deaf and hearing children, the amount of time spent looking at a speaker's mouth has been related to better speech-reading (i.e., lip-reading) accuracy (Worster et al., 2018), although the same relationship was not observed in normal-hearing adults (Lansing & McConkie, 2003; Wilson, Alsius, Pare, & Munhall, 2016). Perception of the McGurk effect has also been related to listeners' patterns of eye gaze, whereby significantly more time is spent looking at a speaker's mouth in trials when it is perceived (Stacey et al., 2020), and stronger perceivers of the effect spend overall more time looking at the speaker's mouth than their eyes (Gurler et al., 2015). Nevertheless, the relevance of the McGurk illusion to audiovisual speech recognition is unclear (Alsius et al., 2018), and an equivalent relationship between patterns of eye gaze and audiovisual speech recognition has still not been found.

Two studies have reported correlational analyses between measurements of eye gaze and audiovisual speech recognition (Buchan et al., 2007; Everdell et al., 2007), but no significant correlations were observed. However, these analyses were not the main aim of the above studies, and certain aspects of their methodology may explain the lack of observed correlations, namely ceiling effects in recognition accuracy which likely reduced variability in the measure. Furthermore, different measures of eye gaze have been used between studies; while some have focused on the length of time spent fixating on the eyes and mouth (Worster et al., 2018), others have measured the number of fixations (Lansing & McConkie, 2003) or trials (Buchan et al., 2007) spent looking at the speaker's mouth, or even left-right asymmetry of eye gaze on the eyes and mouth (Everdell et al., 2007), so it is unclear if one particular pattern of eye movements is particularly important during speech perception.

More recently, Lusk & Mitchell (2016) demonstrated that, after a period of familiarisation, better speech segmentation of an artificial language (i.e., strings of non-words) was related to greater shifts in attention between the eyes and mouth during familiarisation – however, these shifts took place in either direction (i.e., participants looked more or less at the mouth over time), so it is unclear if a particular eye gaze strategy was directly related to learning the new language. Lewkowicz & Hansen-Tift (2012) demonstrated that infants shift their eye gaze more towards a speaker's mouth when learning to speak, but look more at the eyes at a later stage of development when they have become more proficient, indicating that looking at a speaker's mouth is important during language acquisition. Conversely, Birulés, Bosch, Pons & Lewkowicz (2020) demonstrated that non-native adult listeners look more at a speaker's mouth than native speakers regardless of their language proficiency, suggesting that eye gaze towards the mouth is not necessarily linked to learning or performance. In summary, evidence in support of a relationship between eye gaze patterns and language learning are mixed, and nevertheless, the mechanisms of learning a language (as investigated in the above studies), may differ from the mechanisms of adapting to unfamiliar speech in one's native language.

The following questions therefore remain unanswered with regards to eye gaze and perception of audiovisual degraded speech: first, are measures of eye gaze on a speaker's mouth related to i) listeners' speech recognition accuracy, and ii) amount of adaptation to the unfamiliar speech? Secondly, if such a relationship exists, is there a particular pattern of eye gaze on the speaker's mouth (for example, longer or more frequent fixations) that is related to better speech recognition and adaptation? Using eye tracking to investigate patterns of eye gaze towards a speaker's eyes and mouth during a relatively challenging speech recognition task, that avoids ceiling effects and where performance has room to improve over time, may reveal a direct relationship between eye gaze towards a speaker's mouth and audiovisual speech recognition.

The current study therefore had two aims: 1) To replicate and extend previous findings that the presence of visual speech cues improves perceptual adaptation to degraded speech, and 2) to examine the relationship between eye gaze on a speaker's mouth and speech recognition, as well as amount of adaptation (i.e., improvements in speech recognition over time). To address these aims, we measured recognition of degraded sentences in a real-time adaptation paradigm (i.e., where adaptation occurs during continuous exposure rather than after a training period), with and without visual speech cues. We recorded audiovisual sentences spoken from a single speaker and degraded these sentences using noise-vocoding; thus, we could create a relatively challenging speech recognition task that would avoid the ceiling and floor effects found in previous studies.

In a between-subjects design, we exposed a test group to audiovisual degraded speech stimuli, and a control group to audio-only degraded speech stimuli, using eye-tracking to measure participants' eye gaze. The control group was included to allow for direct comparison of speech recognition with and without visual speech cues. For consistency in our methods, we carried out eye tracking in both conditions, but presented the audio-only group with a static image of the speaker's face, therefore offering no dynamic visual cues that could be used to benefit speech recognition (see Methods for full details). To analyse eye gaze patterns during audiovisual speech recognition, we selected two commonly used eye-tracking variables in line with previous studies of audiovisual speech recognition: fixation duration and percentage fixations (Buchan et al., 2007; Everdell et al., 2007; Lansing & McConkie, 2003). Fixations (i.e., any period of time when eye gaze is relatively still; see Methods for full details) reflect the perceiver's foveal field of vision and thus the area of greatest visual acuity. The frequency and duration of fixations can indicate where and to what extent a perceiver's visual attention is primarily directed at any given time (Christianson et al., 1991), and so are a good indicator of when listeners are attending to visual speech cues.

We predicted that perceptual adaptation would be greater when visual speech cues were visible – that is, recognition of the noise-vocoded speech would improve more in the audiovisual group compared to the audio-only group. Secondly, we predicted that recognition accuracy and adaptation in the audiovisual group would be related to the percentage and duration of fixations to the speaker's mouth, with more and longer fixations on the mouth relating to better performance (i.e., higher accuracy and a greater amount of improvement over time).

Method

Participants

Seventy young adults (10 male, *Mdn* = 23 years, age range 19-30 years) were initially recruited from the University of Manchester to participate in the study, which was approved by the university ethics committee. All participants were native British English speakers with no history of neurological, speech or language problems (self-declared), and gave their written informed consent. Participants were included if their corrected binocular vision was 6/6 or better using a reduced Snellen chart, and their stereoacuity was at least 60 seconds of arc using a TNO test. Participants' hearing was measured using pure-tone audiometry for the main audiometric frequencies of speech (0.5, 1, 2, and 4 kHz) in each ear separately. Any participant with a hearing threshold level greater than 20dB for more than one frequency in either ear was excluded from participation. Eleven participants in total (one male) were excluded; two based on the hearing criteria, two based on the visual criteria, five due to data loss during the eye tracking procedure (see Data Analysis for full details), one due to poor eye tracking calibration, and one due to technical failure. 59 participants (nine male, *Mdn* = 23 years, age range 19-30 years) were thus included in the final analyses reported here. Our sample size was based on the expected effect size for the audiovisual benefit to adaptation. Pilling & Thomas (2011) observed a 'benefit' of 12% accuracy for adaptation to audiovisual compared to audio-only degraded sentences using a similar measure of keywords to the present study, although insufficient statistics were reported to obtain an effect size. Bernstein et al. (2013) observed a large effect size of d = 1.21 for adaptation to degraded syllables; as our task was more challenging, we predicted a medium-sized effect. Brysbaert & Stevens (2018) recommend a minimum of 1600 observations per cell for linear mixed effect models detecting medium-sized effects, which we achieved with 60 keywords per testing block, and at least 29 participants per group (i.e., we had at least 1740 observations per cell).

Materials

Experimental materials are available at https://osf.io/2wqkf/. Our stimuli consisted of 91 randomly selected Institute of Electrical and Electronics Engineers Harvard sentences (IEEE; Rothauser et al., 1969). As we wanted to compare our adaptation results as far as possible to Pilling & Thomas (2011), we selected 4 keywords per sentence to score participant accuracy. These were content and function words, selected by the experimenters, that were considered important to the meaning of each sentence. A list of the sentences and keywords used is available as supplemental materials at the above link. Recordings were carried out in a soundproofed laboratory using a Shure SM58 microphone and a High Definition Canon HV30 camera. A 26-year-old female native British English speaker recited the sentences, and was asked to look directly at the camera, to remain still, and to maintain a neutral facial expression throughout the recordings to minimise head movement. Video recordings were imported into iMovie 11 running on an Apple MacBook Pro, as large (960 x 540) high-definition digital video (.dv) files. Recordings were edited to create individual video clips for each sentence. These were checked by the experimenter and any that were not deemed suitable (for example due to mispronunciation) were re-recorded. The audio tracks for each clip were extracted as audio (.wav) files, then normalised by equating the root mean square amplitude, resampled at 22 kHz in stereo, cropped at the nearest zero crossings at voice onset and offset, and vocoded using Praat speech processing software (Boersma & Weenink, 2018). Speech recordings were noise-vocoded (Shannon et al., 1995) using four frequency bands (cut-offs: 50 Hz → 369 Hz →

1160 Hz → 3124 Hz → 8000 Hz), selected to represent equal spacing along the basilar membrane (Greenwood, 1990). In the audio-only (control) condition, a static image of the speaker's face with the mouth in different "speaking" positions was displayed congruently with the audio files so that a visual component was also present in this condition, but with no useful linguistic information. Static faces have previously been used as a control condition for analysing speech perception in dynamic faces (e.g., Calvert & Campbell, 2003; C. Davis & Kim, 2004; Jerger Susan et al., 2018). Using a static face as a control allowed us to assess the contribution of visible articulatory cues to speech recognition, whilst controlling for visual attention towards any salient features of the speaker's face, and also allowing for eye tracking to be conducted in both groups for consistency. To create the still images (one image per trial), screen shots saved as TIFF files were taken from the videos of the speaker displaying a variety of mouth positions, to make the mouth visually salient and to make it evident that she was speaking. The still images, video files and the noise-vocoded audio files were imported into Experiment Builder software (SR Research, Ontario, Canada) to create the experimental stimuli. In the audio-only condition, the still images of the speaker were displayed for the exact length of each audio file, and for the audiovisual condition the audio and video files were played congruously.

Procedure and apparatus

Data were collected in a soundproofed booth in a single test lasting approximately 40 minutes. Participants were randomly allocated into either the audiovisual (*N*=30) or audio-only (*N*=29) control group. In both conditions, participants sat facing the screen approximately 50 cm from the monitor, with their chin on a chin-rest. They were asked not to move their head during the experiment and to look continuously at the screen. Before starting the experiment, the eye-tracker was calibrated for each participant (see 'Data analysis' for details). Participants first listened to one practice sentence (a clear version and a noise-vocoded version) that was not included in the experiment, to prepare them for hearing the unusual distortion. They then completed 90 trials with the remaining noise-vocoded sentences. Participants triggered the start of the experiment and each subsequent trial by pressing the space bar on the keyboard; there were no structured breaks and all 90 trials were presented in a single continuous session. All stimuli were presented through Sennheiser HD 25-SP II headphones. The experimenter set the volume for all stimuli at a comfortable level for the first participant, and kept it at the same level for all participants thereafter. A Panasonic lapel microphone attached to the chin-rest recorded their verbal responses.

To measure speech recognition, we asked participants to repeat out loud as much of each sentence as they could. The experimenter retrospectively scored participants' responses according to how many keywords they correctly repeated out of a maximum of four. Responses were scored as correct despite incorrect suffixes (such as -s, -ed, -ing) or verb endings; however if only part of a word (including compound words) was repeated this was scored as incorrect (Dupoux & Green, 1997; Golomb et al., 2007).

We used a desktop-mounted Eyelink 1000 eye-tracker with Experiment Builder software (SR Research, Ontario, Canada) to present all stimuli, and to record participants' eye movements. The pupil and corneal reflection of each participant's right eye were tracked at a sample rate of 1000 Hz, with a spatial resolution of 0.01° RMS and average accuracy of 0.25°–0.5°. Calibration was carried out for each participant before the experiment using a standard nine-point configuration, and again five minutes after the experiment began. Each calibration was validated for accuracy, and accepted if the average error was <1° and the maximum error was <1.5°. A drift check preceded each trial using a fixation point presented in the centre of the screen, and if the error between the computed fixation position and the on-screen target was >1.5°, calibration was repeated to correct this drift.

Data analysis

The dependent variables were recognition accuracy, fixation duration, and percentage fixations. Recognition accuracy was calculated as the percentage of keywords correctly repeated in each trial. To analyse recognition accuracy over time, we divided all consecutive trials into six blocks of 15 trials, and calculated mean percentage accuracy per testing block based on the number of correctly repeated keywords[1]. Fixations were defined as any period that was not a saccade (saccades were defined as eye movements with velocity >30°/sec, acceleration >8000°/sec^2, and motion >0.1°). Fixations were evaluated in relation to one of two regions of interest (ROIs). For each video clip, we created two elliptical ROIs (see Figure 1) based on the first video frame. These comprised the eye area (extending from just below the speaker's eyebrows to the tip of the nose) and the mouth area (from the septum to just below the bottom lip). Fixation duration and percentage fixations in these regions were then analysed to compare patterns of eye gaze between the two ROIs. We also created a third interest area that surrounded the speaker's face that was used to verify the proportion of eye gaze directed to the speaker's face rather than peripheral areas of the screen. *Fixation duration* was calculated as the mean duration of fixations in milliseconds. *Percentage fixations* was calculated as the percentage of all fixations in a trial falling in the current ROI. We selected these variables to indicate where listeners were allocating their attention at particular time points. Measurements of eye gaze were computed using Data Viewer (SR Research, Ontario, Canada), and we calculated the mean of each variable per testing block, and per interest area.

Data were analysed using linear mixed effects hierarchical regression models in the lmerTest package (Kuznetsova et al., 2017), which uses the lme4 package, running in R v3.4.1. All models included the random effect of participant to account for individual differences in baseline speech recognition. Fixed effects of group, ROI and testing block (i.e., time) were tested by comparing models pairwise using likelihood ratio tests and Bayes Factors calculated using the BIC (e.g., Wagenmakers, 2007). For effects of individual predictors within the model, beta (*B*) coefficients and estimated *p*-values are reported. The variable of fixation duration was rescaled (ms/1000) to make the coefficient more interpretable; estimates of this variable are therefore expressed in seconds.

--Include Figure 1 about here--

[1] Trials were only divided into testing blocks during data analysis – i.e., participants were not aware of the testing blocks during the procedure.

Results

Perceptual adaptation to noise-vocoded speech

Figure 2 shows mean recognition accuracy for the noise-vocoded speech across the six testing blocks for each group. We first tested for group effects against the baseline random effect of participant. Recognition was overall significantly better in the audiovisual group (M = 54%, SD = 2.0%) compared to the audio-only group (M = 35%, SD = 1.6%), B = 19.48, SE = 2.53, $p < 0.001$; χ^2 = 41.02, p <.001, BF_{10} = 42952865. We then added a group * testing block interaction to the model to test whether the audiovisual group improved more over the six testing blocks than the audio-only group. The comparison was significant, χ^2 = 145.45, p <.001, and the large Bayes Factor indicated strong evidence in favour of including the interaction in the model, BF_{10} = 6.911289e+18. However, across the whole experiment (i.e., between block 1 and block 6), recognition accuracy increased equally in both groups by approximately 19%, B = 18.68, SE = 1.56, $p < .001$.

Exploratory Analysis: Rate of Adaptation

Although we observed a group*testing block interaction, results of the mixed effects model described above indicated that the only significant difference in adaptation occurred between block 1 and block 5, where the audiovisual group adapted by 18.47% compared to 12.51% in the audio-only group, B = 6.69, SE = 3.08, p = 0.031. This suggested that listeners adapted more rapidly in the audiovisual group. To examine the rate of adaptation across the experiment in more detail, we conducted exploratory analyses of the amount of adaptation between groups for each consecutive pair of testing blocks. Figure 3 shows that the rate of adaptation was not consistent between blocks or groups. Most adaptation occurred during exposure to the first 30 sentences, when both groups showed ~9% improvement in recognition accuracy. Between blocks 2-5 adaptation slowed in both groups, but the audiovisual group consistently adapted slightly faster, improving by approximately 9% compared to only 2% in the audio-only group. However, between blocks 5 and 6 the audio-only group adapted more than the audiovisual group, improving by 6.4% compared to <1% in the audiovisual group.

We conducted exploratory Bayesian hierarchical regression analyses of adaptation to quantify the evidence for group differences in adaptation rate between consecutive testing blocks. We used forward difference coding whereby a contrast variable was calculated for each pair of consecutive blocks (e.g., B1-B2, B2-B3 etc.), representing differences in recognition accuracy between each pair of blocks. The resulting five coded variables were added as fixed effects to a baseline model that also included group as a main fixed effect, and participant as a random effect. The interaction between each coded variable and group (e.g., B1-B2*group, which represents group differences in adaptation between blocks 1 and 2) was added individually and compared to the baseline model to test for group differences in adaptation at different time points. As these were exploratory analyses we report Bayes Factors and effect sizes only (see Table 1). The baseline model of adaptation between each consecutive pair of testing blocks, and a main effect of group, accounted for approximately 46% variance in recognition accuracy. Bayes factors indicated that there was either no evidence (BF < 0.3), or inconclusive evidence (BF > 0.3 <1), of a difference in adaptation between groups for each consecutive pair of testing blocks, and indeed, adding the interaction variables increased the explained variance by a maximum of just 0.3% (for the B2-B3*Group interaction).

--Include Figures 2 and 3, and Table 1 about here--

Patterns of Eye Gaze

We first examined overall patterns of eye gaze in both groups, to establish whether our eye tracking methods and stimuli had successfully replicated the patterns of eye gaze frequently seen in studies of audiovisual speech perception and when viewing static faces; particularly, to confirm that there were no unusually salient features in our stimuli that attracted viewer's visual attention. In the audiovisual group, 99% of all fixations fell on the speaker's face and 98% fell on the eyes and mouth. In line with previous studies of audiovisual speech recognition in difficult listening conditions (Buchan et al., 2007, 2008; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998), fixations on the speaker's mouth (*M* = 984.32ms, *SD* = 405ms) were significantly longer than fixations on the eyes (*M* = 363.37ms, *SD* = 164ms), χ^2 = 350.83, p < .001, BF_{10} = 8.024141e+74, *B* = 0.621, SE = 0.02, confirming that, as expected, listeners attended more to the speaker's mouth than the eyes. However, there was no difference in percentage fixations on the mouth (*M* = 49%, *SD* = 18%) and eyes (*M* = 49%, *SD* = 18%), χ^2 = 0, p = .988, BF_{10} = 0.05.

In the audio-only group, 83% of fixations were located on the speaker's face, with 74% on the eyes and mouth. The duration of fixations on the eyes (*M* = 443.46ms, *SD* = 179ms) and mouth (*M* = 443.30ms, *SD* = 189ms) did not differ, χ^2 = 0, p = .980, BF_{10} = 0.05. However, a higher percentage of fixations fell on the eyes (*M* = 65%, *SD* = 21%) than on the mouth (*M* = 18% *SD* = 17%), χ^2 = 315.59, p < .001, BF_{10} = 1.818774e+67, *B* = -0.47, SE = 0.02, p < 0.001, in line with previous results from viewing static faces (e.g., Birmingham & Kingstone, 2009). As there were no useful visual cues available in the audio-only group that could benefit speech recognition, and the stimuli was not dynamic, we did not analyse this data in relation to speech recognition; however all data is available as supplemental material here: https://osf.io/2wqkf/.

Are audiovisual speech recognition and perceptual adaptation related to patterns of eye gaze?

To test this hypothesis, we analysed speech recognition data from the audiovisual group, first establishing a baseline model of adaptation with testing block as a predictor; compared to a random effects model of participants' baseline accuracy, there was strong evidence for the baseline model of adaptation to the noise-vocoded speech: χ^2 = 84.53, p < .001, BF_{10} = 5.22229e+12. We then compared this baseline model to four experimental models, each of which included one of the following eye tracking measures as a predictor variable: 1) duration of fixations on the mouth; 2) duration of fixations on the eyes; 3) percentage fixations on the mouth, and 4) percentage fixations on the eyes (see Table 2 for models and corresponding R^2 values). Only the model including duration of fixations on the mouth was significantly different to the baseline model, χ^2 = 5.47, p = 0.019; longer fixations on the speaker's mouth were related to better recognition of the noise-vocoded sentences, *B* = 7.68, SE = 3.21, p = 0.018, however, evidence in support of this relationship was relatively weak (BF_{10} = 1.15). We then tested for an interaction between testing block and the duration of fixations on the mouth to ascertain whether the duration of fixations could predict adaptation. The results did not support the presence of an interaction, $\chi^2 = 9.17$, $p = 0.102$, $BF_{10} = 0.0002$, indicating that there was no overall relationship between eye gaze and adaptation over the course of the experiment.

--Include Table 2 about here--

Exploratory Analyses: Changes in Eye Gaze Over Time

As speech recognition and adaptation rate varied across the time course of the experiment, we conducted exploratory analyses to examine whether patterns of eye gaze in the audiovisual group, as well as their relationship with speech recognition, varied over time. As before, we used Bayesian hierarchical linear mixed effects models, comparing the inclusion of each experimental predictor to a baseline model with participant as a random effect. As these were exploratory analyses we report descriptive statistics, effect sizes and Bayes Factors only. Figure 4 shows the mean duration of fixations and percentage fixations over the time course of the experiment. There was strong evidence that the duration of fixations on the mouth decreased over time by an average of 268.77ms between block 1 and block 6 ($BF_{10} = 7522.16$, $B = -0.26877$, $SE = 0.04256$, marginal $R^2 = 0.05$). There was no evidence that the duration of fixations on the eyes changed over time ($BF_{10} = 0.0002$, marginal $R^2 = 0.01$), nor percentage fixations on the mouth ($BF_{10} = 0.0003$, marginal $R^2 = 0.01$) or the eyes ($BF_{10} = 0.0004$, marginal $R^2 = 0.01$).

Based on the variability in speech recognition, amount of adaptation and the duration of fixations on the speaker's mouth over time, it was possible that longer fixations on the speaker's mouth were more useful at particular time points of the experiment than others, for example during earlier testing blocks. We therefore explored whether the duration of fixations on the speaker's mouth were related to speech recognition in early (blocks 1-2), middle (blocks 3-4) or late (blocks 5-6) testing blocks. For each time period, we compared a model including the duration of fixations on the mouth to the baseline random effects model. We found evidence for a relationship between speech recognition and the duration of fixations on the mouth for middle testing blocks (blocks 3-4) only, BF_{10} = 19.90, B = 18.08, SE = 5.42, marginal R^2 = 0.21; conversely, we found evidence *against* a relationship between speech recognition and the duration of fixations on the mouth in early (blocks 1-2: BF_{10} = 0.13, marginal R^2 = 0.001), and late blocks (blocks 5-6: BF_{10} = 0.18, marginal R^2 = 0.02).

--Figure 4 about here--

Discussion

We investigated perceptual adaptation to noise-vocoded speech with and without visual speech cues, aiming to replicate and extend previous findings (Bernstein et al., 2013; Kawase et al., 2009; Pilling & Thomas, 2011) that being able to view a speaker's face can lead to greater improvement in recognition over time. We used a real-time (i.e., continuous) adaptation paradigm to better reflect real-life adaptation, and eye tracking to investigate eye gaze patterns during audiovisual speech recognition. We tested the relationship between performance and the duration and percentage of fixations on the speaker's eyes and mouth, predicting that looking more at the speaker's mouth would be related to better recognition accuracy and greater adaptation.

We partially replicated previous studies which found an audiovisual benefit to perceptual adaptation, but our observations are somewhat more complex. There was a clear overall benefit to speech recognition from the visual speech cues, with accuracy in the audiovisual group consistently ~20% better than in the audio-only group. However, we found no overall difference in the amount of adaptation between groups as expected – by the final testing block (i.e., after exposure to all 90 sentences), both groups had improved by ~19% accuracy overall. Instead, we only observed a difference between blocks 1 and 5 (after exposure to 75 sentences). Exploratory analyses suggested that the rate of adaptation between blocks varied across the experiment, with the greatest amount of adaptation within the first 30 trials in both groups, who initially adapted at an equal rate despite different baseline levels of accuracy. After this point, the audiovisual group adapted slightly faster until testing blocks 5 and 6, when the audio-only group improved more quickly. However, in Bayesian terms, there was no evidence for group differences in adaptation rate between most blocks, although evidence was inconclusive between testing blocks 2 and 3. Overall, the benefit from visual speech cues to adaptation to degraded speech in our data is smaller and less clear than expected; particularly, we expected the audiovisual group to adapt more overall than the audio-only group.

Our findings are in contrast to studies which found a clear audiovisual benefit to adaptation for noise-vocoded syllables (Bernstein et al., 2013), words (Kawase et al., 2009), and sentences (Pilling & Thomas, 2011); these studies all found greater overall adaptation when the speaker's face was visible compared to when it was not. However, there is some similarity between our findings and those of Pilling & Thomas (2011); we found that adaptation was greater in our audiovisual group following exposure to 75 sentences (testing block 5), while Pilling & Thomas observed the same effect after a similar amount of exposure (76 sentences) during an audiovisual training period. Nevertheless, we did not predict that the audiovisual benefit to adaptation would only be limited to the fifth testing block, and this finding could therefore be due to chance.

There are several possible conclusions from our data. First, that providing a specific audiovisual training period (as in Pilling & Thomas, 2011) is more effective than real-time adaptation; this may, for example, be due to participants attending more to audiovisual speech cues during a separate period of training, in comparison to continuous exposure which may result in lessened attention or fatigue; indeed, the rate of adaptation slowed considerably for our audiovisual group between the final two testing blocks. Second, the amount of benefit to adaptation gained from visual speech cues may depend on the type of stimuli, whereby a greater benefit is possible with simpler and more predictable linguistic items, or from particular speakers (Blackburn et al., 2019). Indeed, using the linguistically more complex IEEE sentences, we observed less improvement in our audiovisual condition (19%) than with the BKB sentences used by Pilling & Thomas (26%) even after greater exposure, although this difference could also be explained by the different speakers used in each study. Lastly, visual speech cues may in fact lead to *faster* adaptation rather than greater overall improvement; that is, without visual cues listeners can still adapt equally well but require more exposure to do so, as was the case for our audio-only group. Our exploratory analyses of adaptation rate seem to support this, as speech recognition rapidly improved in both groups initially, but then slowed in the audio-only condition; however, this group difference was small, and the Bayesian evidence from our data didn't support a clear difference in adaptation rate. The amount of adaptation observed may thus depend on exactly *when* it is measured, and how much exposure participants have had to the degraded speech.

Overall, our results indicate that the benefits of visual speech cues to adaptation are not as great or clearcut as results from previous studies suggest. Instead, the benefits potentially depend on factors such as the linguistic items used (i.e., the specific linguistic characteristics of the stimuli such as length, syntactic complexity or semantic predictability), speaker, and amount of exposure, and the contribution of these factors will need to be confirmed in future studies. The small advantage to adaptation in the audiovisual group during middle testing blocks suggests that benefits from visual cues could further be related to participants' attention or energy levels, whereby visual cues are particularly beneficial to learning at points where attention and motivation are low – such as in the middle of a challenging laboratory experiment. The benefits of visual cues in real-life contexts may thus depend on the type of communication taking place; while these cues do not necessarily lead to greater adaptation early on, they may be particularly useful in contexts where longer periods of sustained adaptation are required, for instance, listening to a lecture or when participating in a longer conversation. The interaction between use of visual speech cues and attention or fatigue may thus be an interesting line for future research into speech recognition in adverse listening conditions. Nevertheless, the small audiovisual benefit that we observed during middle testing blocks could just have been an anomaly – i.e., it could have occurred by chance.

It should be noted that recognition of noise-vocoded sentences (with or without visual cues) varies considerably between studies. We observed mean performance of 35% accuracy in our audio-only condition, but similar studies have found differing levels of performance. For example, using 4-band noise-vocoding and the IEEE sentences (as in the present study), McGettigan et al., (2014) observed approximately 40% mean accuracy for recognition of only 10 sentences; however, this was following exposure to 70 noise-vocoded BKB sentences, perhaps accounting for the higher level of accuracy than in the present study. In comparison, using 6-band noise-vocoding, Paulus et al. (2020) observed approximately 60% accuracy after exposure to 48 IEEE sentences. Using the simpler BKB sentences, Scott et al., (2006) observed approximately 40% accuracy using 4-band noise-vocoding, but after exposure to only 16 sentences, while Rosen et al., (1999) observed 64% mean accuracy after exposure to 112 sentences also vocoded with 4 channels. Thus, recognition of noise-vocoded speech can vary greatly depending on the amount of exposure, the type of linguistic stimuli, and the exact vocoding transformation. In the present study, we specifically chose to use the IEEE sentences and 4-band noise-vocoding to create a more challenging task (and particularly to prevent ceiling effects in the audiovisual condition). Nevertheless, the intelligibility of our stimuli may also have been affected by the speaker we used (e.g., Bradlow & Bent, 2008). Indeed, specific acoustic-phonetic features (namely vowel space dispersion and mean energy in mid-range frequencies) can account for differing levels of intelligibility between speakers for noise-vocoded speech, although these features do not necessarily impact listeners' amount of adaptation (Paulus et al., 2020). Furthermore, the amount of benefit that visual cues can provide also varies between speakers (Blackburn et al., 2019). As changing speakers can interfere with adaptation (e.g., Dupoux & Green, 1997), we used the same speaker throughout our study. However, we note that a limitation of the current findings is that we cannot confirm whether mean levels of performance in either condition, or indeed the benefit that listeners obtained from the speaker's visual cues, would be the same for other speakers.

The second aim of our study was to examine patterns of eye gaze during adaptation to audiovisual degraded speech, and specifically to test whether there is a direct relationship between eye gaze towards a speaker's mouth movements and speech recognition. We found that longer fixations on the speaker's mouth were related to better recognition, but not to the amount of adaptation. This supports findings from speechreading (Worster et al., 2018) which found that longer time spent fixating the speaker's mouth was related to better speechreading in both deaf and normal-hearing children. Two previous studies have also directly tested the relationship between eye gaze patterns and speech recognition (Buchan et al., 2007; Everdell et al., 2007), but found no significant relationship. However, methodological differences can potentially account for the different results reported here. First, audiovisual speech recognition was at ceiling in both studies, i.e., 86% (Buchan et al., 2007) and 90% (Everdell et al., 2007), compared to 41-61% in the present study. Second, neither study analysed the duration of fixations (as in the present study), or time spent fixating the speaker's mouth (as in Worster et al., 2018). Everdell et al. (2007) analysed an index of left-right asymmetry of eye gaze on the eyes and mouth, while Buchan et al. (2007) analysed percentage trials spent looking at the speaker's mouth, but neither observed correlations between these measures and speech recognition. Current evidence thus suggests that measurements of the *time* spent fixating a speaker's mouth is indicative of effective use of visual speech cues, rather than the frequency or proportion of fixations; indeed, we found no correlation between percentage fixations on the speaker's mouth and speech recognition, similar to Lansing & McConkie (2003) who found no relationship between the number of fixations on the mouth and speechreading. More recently, Lusk & Mitchell (2016) observed a positive relationship between changes in the amount of eye gaze on a speaker's mouth during passive listening to an artificial language, and subsequent segmentation of non-words from this language. However, note that Lusk & Mitchell's finding only partially supports the current findings, as the relationship was irrespective of direction – i.e., the shift could involve looking more or less at the mouth. Thus, to our knowledge, ours is the first study to observe a direct relationship between looking more at a speaker's mouth and audiovisual speech recognition.

The results add to a growing body of literature indicating that patterns of eye gaze – that is, where and how listeners look at a speaker's face – are important for successfully understanding unfamiliar or degraded audiovisual speech. Thus, it is not merely the presence of visual speech cues, but also the particular visual strategies employed by listeners, that relate to successful speech recognition. As we compared two measures of eye gaze commonly used in eye tracking studies, we can further conclude that the *duration* of fixations on a speaker's mouth are likely more important than the proportion of fixations. Longer fixations on the mouth likely reflected a greater focus of attention on this region, particularly as visual perception is reduced during eye movements (Matin & Ethel, 1974). Thus, with longer fixations and less eye movement, listeners could better or more efficiently decode articulatory cues from a speaker's mouth, improving recognition. The duration of fixations on a speaker's mouth is thus potentially a useful measure when assessing the use or relevance of visual speech cues. Indeed, longer fixations on a speaker's mouth have indicated increased use of visual cues in other studies of adverse listening conditions (Buchan et al., 2007, 2008), although the measure has not previously been related to performance. The importance of this measure was indirectly supported by our exploratory observation that the duration of fixations on the speaker's mouth decreased over time, as performance improved (while no such change was observed for percentage fixations). This decrease would suggest that participants' use of visual cues from the speaker's mouth decreased as they adapted to the degraded speech. A similar observation was made by Lusk & Mitchel (2016) who noted a decrease in overall gaze time on a speaker's mouth, but not on the eyes or nose, during a period of familiarisation to an artificial language (i.e., passive listening/viewing), prior to listeners being tested on non-word recognition. The duration of fixations on a speaker's mouth may thus be an important indicator of effective use of visual speech cues when learning or adapting to unfamiliar speech – for example helping word segmentation (Mitchel & Weiss, 2014); however, we did not observe a correlation between the duration of fixations and amount of adaptation.

Another interpretation of our finding is that the decrease in fixation durations indicates changes in attention or effort. After the period of rapid adaptation between testing blocks 1 and 2, decoding the noise-vocoded speech perhaps no longer required as much cognitive effort, or attention, from participants. Listening effort (as measured by relative pupil size) is greater during perception of noise-vocoded speech compared to undegraded speech in quiet (Paulus et al., 2020); furthermore, it has been shown to decrease during a period of adaptation to unfamiliar accented speech (Brown et al., 2020), just as the duration of fixations decreased in our study. An interpretation of our results related to cognitive effort is compatible with those of Birulés et al. (2020), who found that listeners looked more towards a speaker's mouth (measured as proportion of total gaze time) during recognition of non-native speech than native, regardless of linguistic ability; that is, the cognitive demands required to understand non-native speech were consistently greater – indicated by more time spent looking at the speaker's mouth (and potentially greater reliance on visual speech cues). Outside of the speech perception literature, changes in eye gaze patterns have also been associated with cognitive load; for example, fewer and longer fixations during scene viewing are observed with greater memory loads (Cronin et al., 2020), again suggesting that greater cognitive demands can influence patterns of eye gaze. Although the present results cannot confirm this interpretation of our data, they nonetheless offer an interesting avenue for future research.

Some limitations to the current findings should be noted. First, the evidence for a relationship between eye gaze and speech recognition was relatively weak in Bayesian terms. Exploratory analyses suggested that the relationship was in fact only present in middle testing blocks, but why this would be the case is unclear; the pattern somewhat matches our observation that audiovisual cues were most beneficial to adaptation during middle testing blocks, rather than in early or later blocks, and so could indicate a particular reliance on visual cues during this time. Visual cues from the speaker's mouth could potentially serve to compensate for decreasing attention or motivation, resulting in a stronger relationship between longer fixations and performance during this period. Nevertheless, the results require further testing. A second limitation is that the result was correlational, and we therefore cannot ascertain whether longer fixations on the speaker's mouth resulted in better recognition, or whether participants who performed better looked more steadily at the speaker's mouth. Again, this correlational result would benefit from further testing whereby particular eye gaze strategies are manipulated to observe the effects on performance. Finally, we note that using a static face as a control condition for the audio-only condition is less naturalistic than, for example, providing no visual information at all, and thus does not have an exact 'real-world' equivalent (except, perhaps, a frozen screen during a video call). Our motivation in including this condition was to equate the procedure for both groups as far as possible, including visual information and eye tracking in both. However, we are confident that performance in this condition was not significantly worse than would be expected without a visible static face (for an online replication see Trotter et al., 2020), and thus that it was a valid comparison for speech adaptation.

We report several exploratory analyses in the current paper to support interpretation of the findings, and these are intended as hypothesis-generating observations rather than hypothesis-testing, whereby our aim is to open up further lines of enquiry regarding adaptation to unfamiliar speech and related patterns of eye gaze. For example, the decrease in the duration of fixations during adaptation may be further investigated by comparing eye gaze during audiovisual speech recognition to a control condition with non-informative mouth movements, or compared to measures of listening effort. Furthermore, differences in the rate of adaptation to unfamiliar speech with and without visual cues should be investigated in more detail to establish the exact parameters that determine when visual cues offer a clear benefit to listeners. The analyses and observations presented here will thus be beneficial to the research fields of audiovisual speech perception and, more broadly, communication in difficult listening conditions.

Conclusion

We have demonstrated that the benefit of visual speech cues to adaptation to degraded (noise-vocoded) speech is more limited than previously thought – potentially resulting in slightly faster adaptation only after a period of initial exposure and rapid adaptation, but not resulting in an overall greater amount of improvement after a longer period of exposure. Longer fixations on the speaker's mouth were related to better overall recognition accuracy of the audiovisual speech, adding to a growing body of evidence that patterns of eye gaze are related to effective use of visual speech cues. Nevertheless, evidence for this relationship was relatively weak and will need further testing to be fully confirmed and understood. We further observed that the duration of fixations on the speaker's mouth decreased over time; future research will need to determine the relevance of this finding, as well as whether particular patterns of eye gaze can intentionally bring benefits to listeners in adverse listening conditions.

Bibliography

Adank, P., & Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychology and Aging*, *25*(3), 736–740. https://doi.org/10.1037/a0020054

Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty Years After Hearing Lips and Seeing Voices: The McGurk Effect Revisited. *Multisensory Research*, *31*(1–2), 111–144. https://doi.org/10.1163/22134808-00002565

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015a). Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. *Frontiers in Human Neuroscience*, *9*(AUGUST). https://doi.org/10.3389/fnhum.2015.00422

Banks, B., Gowen, E., Munro, K. J., & Adank, P. (2015b). Cognitive predictors of perceptual adaptation to accented speech. *Journal of the Acoustical Society of America*, *137*(4). https://doi.org/10.1121/1.4916265

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*. https://doi.org/10.1016/j.cognition.2015.11.013

Bench, J., Kowal, Å., & Bamford, J. (1979). The Bkb (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. *British Journal of Audiology*, *13*(3), 108–112. https://doi.org/10.3109/03005367909078884

Bernstein, L. E., Auer, E. T., Eberhardt, S. P., & Jiang, J. (2013). Auditory Perceptual Learning for Speech Perception Can be Enhanced by Audiovisual Training. *Frontiers in Neuroscience*, *7*, 34. https://doi.org/10.3389/fnins.2013.00034

Birmingham, E., & Kingstone, A. (2009). Human Social Attention. *Annals of the New York Academy of Sciences*, *1156*(1), 118–140. https://doi.org/10.1111/j.1749-6632.2009.04468.x

Birulés, J., Bosch, L., Pons, F., & Lewkowicz, D. J. (2020). Highly proficient L2 speakers still need to attend to a talker's mouth when processing L2 speech. *Language, Cognition and Neuroscience*, *0*(0), 1–12. https://doi.org/10.1080/23273798.2020.1762905

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual Speech Benefit in Clear and Degraded Speech Depends on the Auditory Intelligibility of the Talker and the Number of Background Talkers. *Trends in Hearing*, *23*, 2331216519837866. https://doi.org/10.1177/2331216519837866

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer. In *Retrieved October 2017 from http://www.praat.org/* (6.0.31).

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology*, 1747021820916726. https://doi.org/10.1177/1747021820916726

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.10

Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, *2*(1), 1–13. https://doi.org/10.1080/17470910601043644

Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research*, *1242*, 162–171. https://doi.org/10.1016/j.brainres.2008.06.083

Calvert, G. A., & Campbell, R. (2003). Reading Speech from Still and Moving Faces: The Neural Substrates of Visible Speech. *Journal of Cognitive Neuroscience*, *15*(1), 57–70. https://doi.org/10.1162/089892903321107828

Christianson, S. a, Loftus, E. F., Hoffman, H., & Loftus, G. R. (1991). Eye fixations and memory for emotional events. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *17*(4), 693–701. https://doi.org/10.1037/0278-7393.17.4.693

Cronin, D. A., Peacock, C. E., & Henderson, J. M. (2020). Visual and verbal working memory loads interfere with scene-viewing. *Attention, Perception, & Psychophysics*, *82*(6), 2814–2820. https://doi.org/10.3758/s13414-020-02076-1

Davis, C., & Kim, J. (2004). Audio–Visual Interactions with Intact Clearly Audible Speech. *The Quarterly Journal of Experimental Psychology Section A*, *57*(6), 1103–1121. https://doi.org/10.1080/02724980343000701

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. https://doi.org/10.1037/0096-3445.134.2.222

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, *102*(4), 2403–2411. https://doi.org/10.1121/1.419603

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology. Human Perception and Performance*, *23*(3), 914–927. https://doi.org/10.1037/0096-1523.23.3.914

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, *40*(4), 481–492. https://doi.org/10.1044/jshd.4004.481

Everdell, I. T., Marsh, H., Yurick, M. D., Munhall, K. G., & Paré, M. (2007). Gaze behaviour in audiovisual speech perception: Asymmetrical distribution of face-directed fixations. *Perception*, *36*(10), 1535–1545. https://doi.org/10.1068/p5852

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, *108*(4), 1877–1887. https://doi.org/10.1121/1.1310667

Golomb, J. D., Peelle, J. E., & Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *The Journal of the Acoustical Society of America*, *121*(3), 1701–1708. https://doi.org/10.1121/1.2436635

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, *87*(6), 2592–2605. https://doi.org/10.1121/1.399052

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, & Psychophysics*, *77*(4), 1333–1341. https://doi.org/10.3758/s13414-014-0821-1

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication*, *52*(11–12), 996–1009. https://doi.org/10.1016/j.specom.2010.05.003

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856. https://doi.org/10.3758/PBR.16.5.850

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460–474. https://doi.org/10.1037/0096-1523.34.2.460

Jerger Susan, Damian Markus F., Karl Cassandra, & Abdi Hervé. (2018). Developmental Shifts in Detection and Attention for Auditory, Visual, and Audiovisual Speech. *Journal of Speech, Language, and Hearing Research*, *61*(12), 3095–3112. https://doi.org/10.1044/2018_JSLHR-H-17-0343

Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., & Kobayashi, T. (2009). Bimodal audio-visual training enhances auditory adaptation process. *Neuroreport*, *20*(14), 1231–1234. https://doi.org/10.1097/WNR.0b013e32832fbef8

Kuznetsova, A., Brockhoff, B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, *65*(4), 536–552. https://doi.org/10.3758/BF03194581

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(5), 1431–1436. https://doi.org/10.1073/pnas.1114783109

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *The Journal of the Acoustical Society of America*, *106*(4), 2097–2103. https://doi.org/10.1121/1.427954

Lusk, L. G., & Mitchel, A. D. (2016). Differential Gaze Patterns on Eyes and Mouth During Audiovisual Speech Segmentation. *Frontiers in Psychology*, *7*, 52. https://doi.org/10.3389/fpsyg.2016.00052

MacLeod, a, & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(October), 131–141. https://doi.org/10.3109/03005368709077786

Malcolm, G. L., Lanyon, L. J., Fugard, A. J. B., & Barton, J. J. S. (2008). Scan patterns during the processing of facial expression versus identity: An exploration of task-driven and stimulus-driven effects. *Journal of Vision*, *8*(8), 2–2. https://doi.org/10.1167/8.8.2

Matin, E., & Ethel. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, *81*(12), 899–917. https://doi.org/10.1037/h0037368

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

McGettigan, C., Rosen, S., & Scott, S. K. (2014). Lexico-semantic and acoustic-phonetic processes in the perception of noise-vocoded speech: Implications for cochlear implantation. *Frontiers in Systems Neuroscience*, *8*. https://doi.org/10.3389/fnsys.2014.00018

Mitchel, A. D., & Weiss, D. J. (2014). Visual speech segmentation: Using facial cues to locate word boundaries in continuous speech. *Language, Cognition and Neuroscience*, *29*(7), 771–780. https://doi.org/10.1080/01690965.2013.791703

Morin-Lessard, E., Poulin-Dubois, D., Segalowitz, N., & Byers-Heinlein, K. (2019). Selective attention to the mouth of talking faces in monolinguals and bilinguals aged 5 months to 5 years. *Developmental Psychology*, *55*(8), 1640–1655. https://doi.org/10.1037/dev0000750

Paulus, M., Hazan, V., & Adank, P. (2020). The relationship between talker acoustics, intelligibility, and effort in degraded listening conditions. *The Journal of the Acoustical Society of America*, *147*(5), 3348–3359. https://doi.org/10.1121/10.0001212

Peelle, J. E., & Wingfield, A. (2005). *Dissociations in Perceptual Learning Revealed by Adult Age Differences in Adaptation to Time-Compressed Speech*. https://doi.org/10.1037/0096-1523.31.6.1315

Pilling, M., & Thomas, S. (2011). Audiovisual Cues and Perceptual Learning of Spectrally Distorted Speech. *Language and Speech*, *54*(4), 487–497. https://doi.org/10.1177/0023830911404958

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, *106*(6), 3629–3636. https://doi.org/10.1121/1.428215

Rothauser, E. H., Chapman, N. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058

Schoof, T., & Rosen, S. (2015). High sentence predictability increases the fluctuating masker benefit. *The Journal of the Acoustical Society of America*, *138*(3), EL181–EL186. https://doi.org/10.1121/1.4929627

Scott, S. K., Rosen, S., Lang, H., & Wise, R. J. S. (2006). Neural correlates of intelligibility in speech investigated with noise vocoded speech—A positron emission tomography study. *The Journal of the Acoustical Society of America*, *120*(2), 1075–1083. https://doi.org/10.1121/1.2216725

Sebastian-Galles, N., & Mehler, J. (2000). Adaptation to time-compressed speech: Phonological determinants. In *Perception & Psychophysics* (Vol. 62, Issue 4).

Shannon, R. V, Zeng, F.-G. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science (New York, N.Y.)*, *270*(5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing*, *26*(3), 263–275. https://doi.org/10.1097/00003446-200506000-00003

Stacey, J. E., Howard, C. J., Mitra, S., & Stacey, P. C. (2020). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. *Attention, Perception, & Psychophysics*, *82*(7), 3544–3557. https://doi.org/10.3758/s13414-020-02042-x

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Summerfield, Q. A. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In *Hearing by Eye: The Psychology of Lip-reading* (pp. 3–51). https://doi.org/citeulike-article-id:795404

Thomas, S. M., & Jordan, T. R. (2004). Contributions of Oral and Extraoral Facial Movement to Visual and Audiovisual Speech Perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(5), 873–888. https://doi.org/10.1037/0096-1523.30.5.873

Trotter, A., Banks, B., & Adank, P. (2020). *Effects of the availability of visual cues during adaptation to noise—Vocoded speech*. PsyArXiv. https://doi.org/10.31234/osf.io/jxaeb

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, *60*(6), 926–940. https://doi.org/10.3758/BF03211929

Vo, M. L.-H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision*, *12*(13), 3–3. https://doi.org/10.1167/12.13.3

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wang Jianrong, Zhu Yumeng, Chen Yu, Mamat Abdilbar, Yu Mei, Zhang Ju, & Dang Jianwu. (2020). An Eye-Tracking Study on Audiovisual Speech Perception Strategies Adopted by Normal-Hearing and Deaf Adults Under Different Language Familiarities. *Journal of Speech, Language, and Hearing Research*, *63*(7), 2245–2254. https://doi.org/10.1044/2020_JSLHR-19-00223

Wayne, R. V., & Johnsrude, I. S. (2012). The role of visual speech information in supporting perceptual learning of degraded speech. *Journal of Experimental Psychology: Applied*, *18*(4), 419–435. https://doi.org/10.1037/a0031042

Worster, E., Pimperton, H., Ralph-Lewis, A., Monroy, L., Hulme, C., & MacSweeney, M. (2018). Eye movements during visual speech perception in deaf and hearing children. *Language Learning*, *68*(Suppl 1), 159–179. https://doi.org/10.1111/lang.12264

Yarbus, A. L. (1967). *Eye Movements and Vision*. Springer US. https://doi.org/10.1007/978-1-4899-5379-7

874

Figure Captions

Figure 1

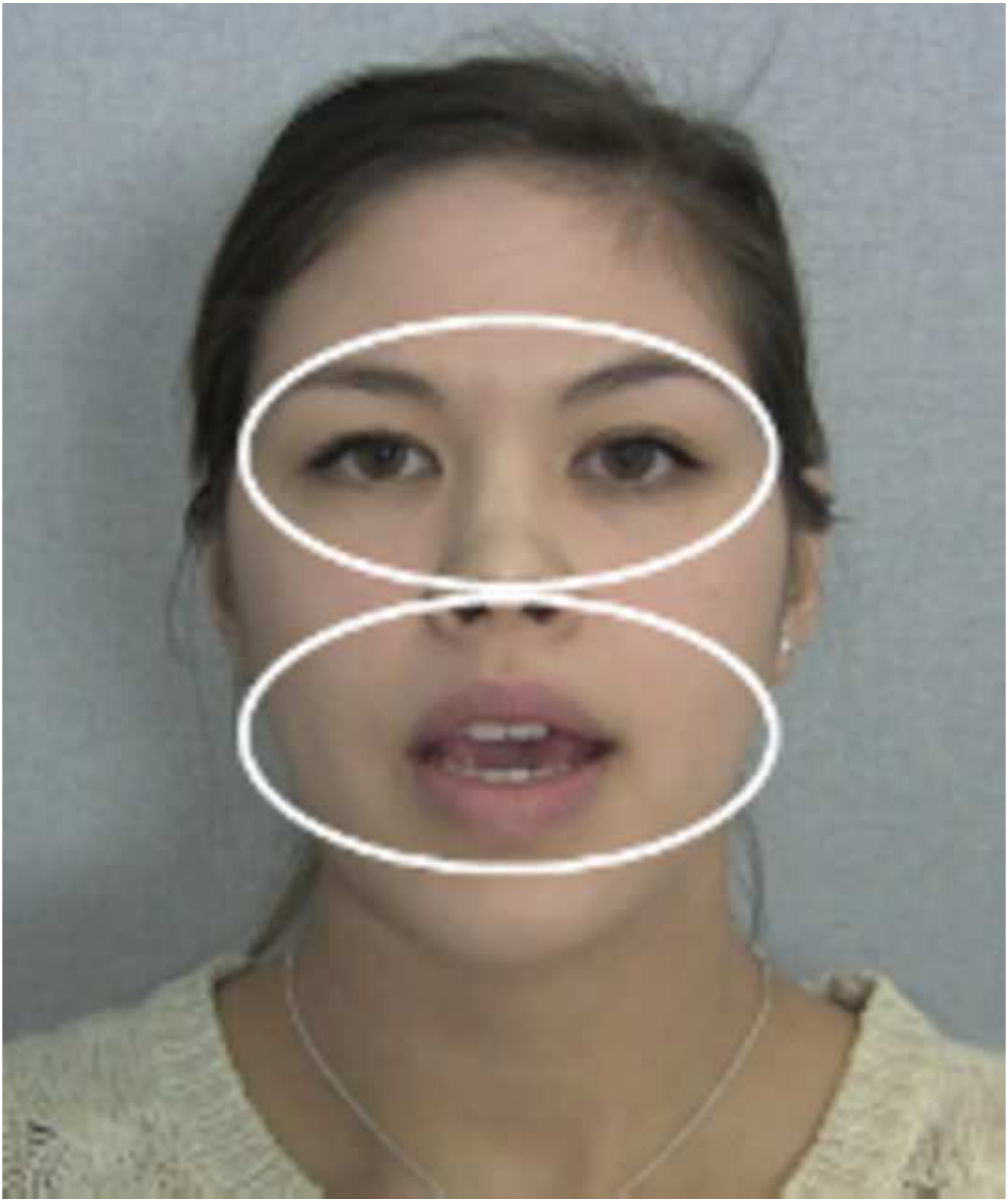
Image of the speaker with regions of interest ('mouth' and 'eyes').

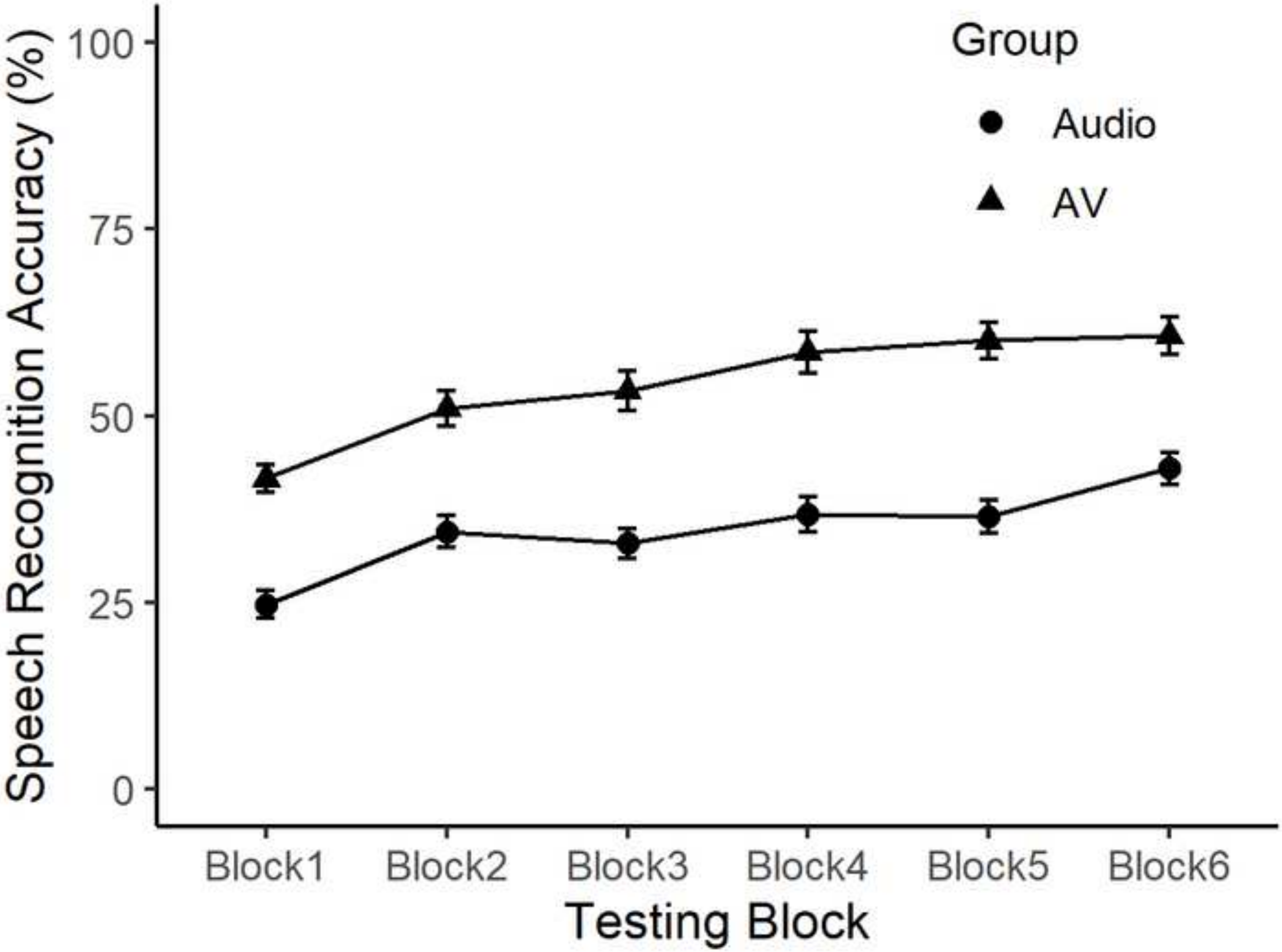
Figure 2.

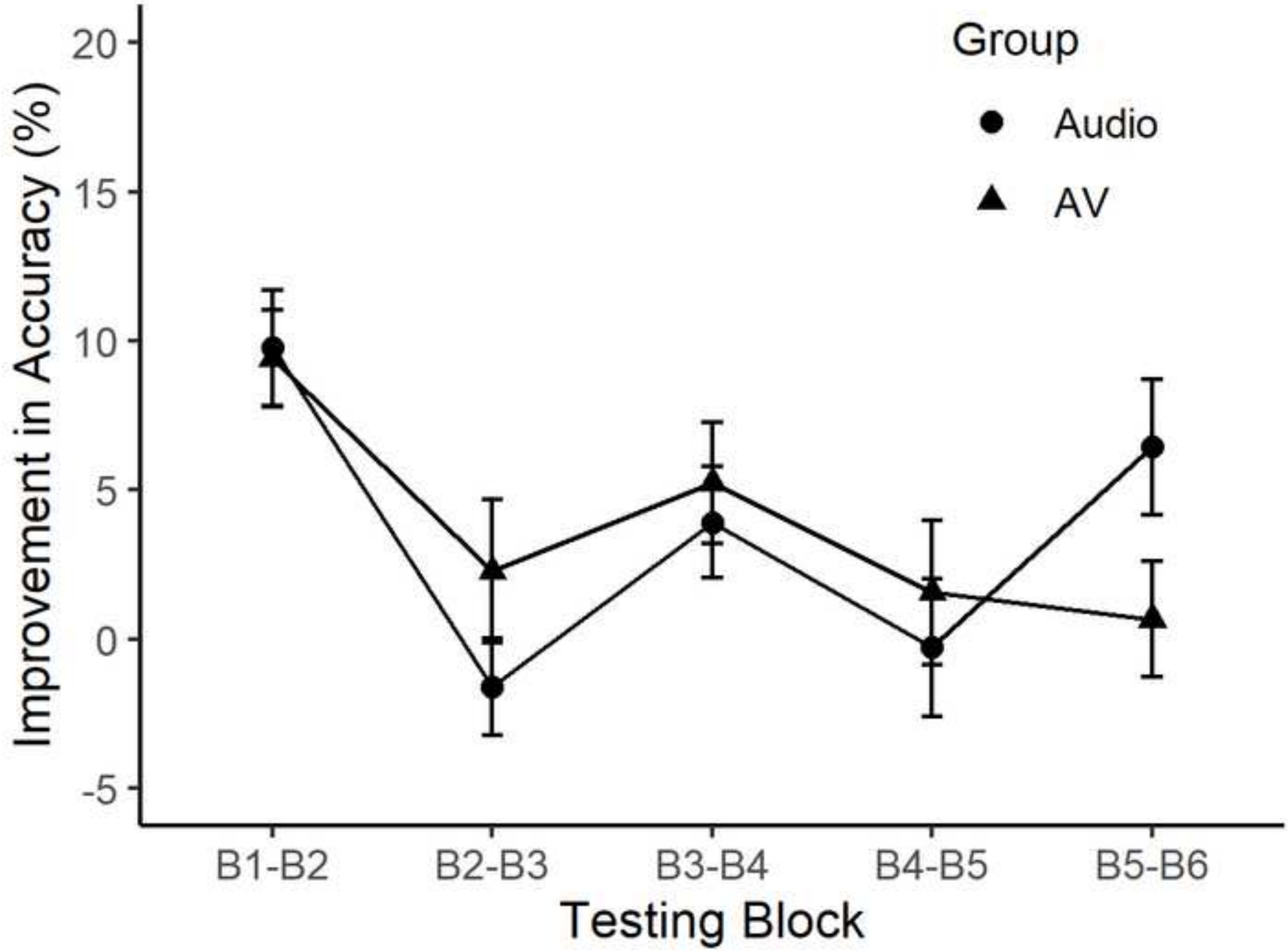
Mean recognition accuracy per testing block, per group. Error bars show ±1SE.

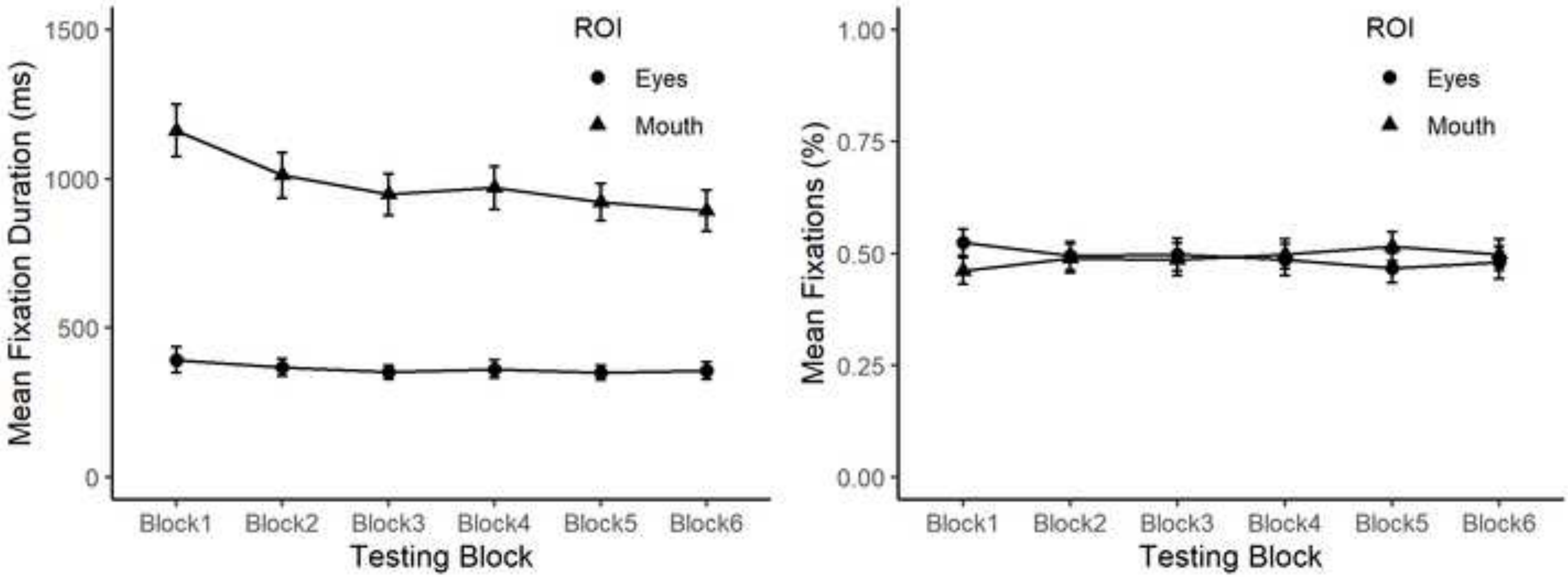
Figure 3. Adaptation (amount of improvement) between consecutive testing blocks per group. Error bars show ±1SE.

Figure 4. Duration of fixations (left panel) and percentage fixations (right panel) on the mouth and eyes, per testing block in the audiovisual group. Error bars ±1SE.




Speech Recognition Accuracy (%)
100
75
50
25
0
Group
Audio
AV
Block1
Block2
Block3
Block4
Block5
Block6
Testing Block


Group
Audio
AV
Improvement in Accuracy (%)
20
15
10
5
0
-5
B1-B2
B2-B3
B3-B4
B4-B5
B5-B6
Testing Block



ROI
Eyes
Mouth
Mean Fixation Duration (ms)
1500
1000
500
0
Block1
Block2
Block3
Block4
Block5
Block6
Testing Block
ROI
Eyes
Mouth
Mean Fixations (%)
1.00
0.75
0.50
0.25
0.00
Block1
Block2
Block3
Block4
Block5
Block6
Testing Block

Table 1. Exploratory Bayesian hierarchical regression analyses of group differences in adaptation rate.

Model	BF_{10}	R^2m	ΔR^2	Interpretation
Baseline model	-	.458	-	-
B1-B2*Group	0.12	.459	.001	No group difference
B2-B3*Group	0.59	.462	.003	Inconclusive
B3-B4*Group	0.23	.460	.001	No group difference
B4-B5*Group	0.08	.459	.000	No group difference
B5-B6*Group	0.08	.459	.000	No group difference

Note. R^2m = marginal R^2 (fixed effects only); ΔR^2 indicates change in marginal R^2 based on difference between baseline model and the addition of the model interaction. BF_{10} = Bayes Factor indicating evidence of a difference between groups in the amount of adaptation between each consecutive pair of testing blocks.

Table 2. Hierarchical mixed model comparisons for the audiovisual group predicting overall speech recognition by each measure of eye gaze.

Model	R^2	*p*-value	BF_{10}
Testing Block (baseline model of adaptation)	0.20	<.001**	5.22229e+12
Testing Block + Duration of Fixations on Mouth	0.25	.019*	1.15
Testing Block + Duration of Fixations on Eyes	0.20	.805	0.08
Testing Block + Percentage Fixations on Mouth	0.20	.496	0.09
Testing Block + Percentage Fixations on Eyes	0.20	.613	0.08
Testing Block * Duration of Fixations on Mouth (interaction)	0.20	1.00	0.07

Note: All models contain the random effect of participant. We report marginal R^2 representing the variance explained by fixed effects only.

* $p < .05$; ** $p < .001$