1   **The relevance of the availability of visual speech cues during adaptation to noise-vocoded speech**

2

3   Anthony S Trotter [a], Briony Banks [b], and Patti Adank [a]

4

5   [a] Speech, Hearing and Phonetic Sciences, Chandler House, Wakefield Street, University College

6   London, London, UK

7   [b] Department of Psychology, Fylde College, Lancaster University, Lancaster, LA1 4YF, UK.

8 **Abstract**

9 **Purpose:** This study first aimed to establish whether viewing specific parts of the speaker's face (eyes

10 or mouth), compared to viewing the whole face, affected adaptation to distorted - noise-vocoded -

11 sentences. Second, this study also aimed to replicate results on processing of distorted speech from

12 lab-based experiments in an online setup.

13

14 **Method:** We monitored recognition accuracy online while participants were listening to noise-

15 vocoded sentences. We first established if participants were able to perceive and adapt to audiovisual

16 4-band noise-vocoded sentences when the entire moving face was visible (AV Full). Four further

17 groups were then tested: a group in which participants viewed the moving lower part of the speaker's

18 face (AV Mouth), only see the moving upper part of the face (AV Eyes), could not see the moving lower

19 or upper face (AV Blocked), and a group where participants saw an image of a still face (AV Still).

20

21 **Results:** Participants repeated around 40% of key words correctly and adapted during the experiment

22 but only when the moving mouth was visible. In contrast, performance was at floor level, and no

23 adaptation took place, in conditions when the moving mouth was occluded.

24

25 **Conclusions:** The results show the importance of being able to observe relevant visual speech

26 information from the speaker's mouth region, but not the eyes/upper face region, when listening and

27 adapting to distorted sentences online. Second, the results also demonstrated that it is feasible to run

28 speech perception and adaptation studies online, but that not all findings reported for lab studies

29 replicate.

30

31

32 **Key words:** Adaptation, audiovisual speech, noise-vocoded speech; speech perception.

33

Introduction

We often interact with others in suboptimal listening situations, e.g., in a crowded cafeteria, at a busy railway station, or when interacting online over a poor audio and/or video connection. Indeed, most of us can cope with these distortions, although speech recognition performance tends to be attenuated compared to clear listening conditions. For example, listeners can adapt to distortions of the speech signal. Such perceptual adaptation can occur in a relatively short time frame: listeners can improve their response speed and accuracy after exposure to fewer than 30 distorted sentences. Such rapid adaptation has, for example, been reported for noise-vocoded speech (Davis et al., 2005; Hervais-Adelman et al., 2008) accented (Adank et al., 2010; Banks et al., 2015b, 2015a; Brown et al., 2020), and time-compressed speech (Peelle & Wingfield, 2005; Sebastián-Gallés et al., 2000).

Much of the research on rapid adaptation used noise-vocoded speech, which is an artificial distortion of the speech signal in which harmonic components are replaced with bands of noise. The distorted signal has lost much of the final spectral and harmonic detail, but amplitude modulation information is largely preserved (Shannon et al., 1995). The speech signal is first divided into separate frequency bands (generally between 4 and 32). Next, the amplitude envelope is extracted, which is subsequently used to manipulate a broadband carrier signal. This type of distortion has been used as a simulation of how speech and other sounds are transmitted in people with a cochlear implant (Faulkner et al., 2000; Rosen et al., 1999), which is an implanted device that restores hearing in those who have severe or complete hearing loss. When normal-hearing listeners are exposed to noise-vocoded speech they generally show adaptation (i.e., improvement in speech perception performance over time). It is generally more difficult to understand noise-vocoded speech with a lower number of frequency bands (4 or 6) than a higher number of bands (Dorman et al., 1997; Faulkner et al., 2000; Sohoglu et al., 2014).

Most studies on adaptation to distorted speech published to date focused on adaptation to noise-vocoded speech used auditory-only stimuli. However, being able to see as well as hear the speaker can considerably improve perception of different types of distorted speech (e.g., speech in

60   background noise), a phenomenon referred to as the *audiovisual benefit* (Erber, 1975; MacLeod &

61   Summerfield, 1987; Sumby & Pollack, 1954). Listeners benefit from the availability of visual cues and

62   are thought to integrate them with auditory speech cues, which then in turn improves speech

63   perception performance. The audiovisual benefit has also been studied for perceptual adaptation to

64   noise-vocoded speech (Banks et al., 2020; Bernstein et al., 2013; Kawase et al., 2009; Pilling & Thomas,

65   2011; Wayne & Johnsrude, 2012). Pilling & Thomas (2011) and Banks et al. (2020) compared

66   adaptation to noise-vocoded sentences with and without audiovisual speech cues. When visual

67   speech cues were made available, listeners adapted more than for auditory-only conditions, although

68   the audiovisual benefit was smaller and earlier in Banks et al., peaking after exposure to 75 out of 90

69   sentences. Similar results were reported by Bernstein et al. (3013), who report that the presence of

70   visual speech cues leads to more adaptation to noise-vocoded syllables. Wayne & Johnsrude (2012)

71   also investigated adaptation to noise-vocoded sentences providing audiovisual cues as feedback

72   during a period of training and found that audiovisual feedback didn't benefit adaptation any more

73   than clear (i.e., not noise-vocoded) feedback; however, they did not directly compare degraded

74   audiovisual and audio-only conditions as in Pilling & Thomas (2011) and Banks et al (2020). Current

75   evidence thus indicates that concurrent visual speech cues can thus benefit listeners during rapid

76   adaptation to distorted speech. However, it remains unclear whether it is only visual cues from the

77   mouth that benefit listeners, or whether cues from other parts of the face (e.g., eyes), or the whole

78   face, are also useful in helping listeners adapt.

79   Several speech perception studies using eye-tracking demonstrated that listeners look more at

80   a speaker's mouth during perception of speech in noise (Buchan et al., 2007, 2008; Lansing &

81   McConkie, 2003) and noise-vocoded speech (Banks et al., 2020). Notably, fixations on the mouth

82   increase for poorer signal-to-noise ratios (Vatikiotis-Bateson et al., 1998). These findings suggest that

83   cues from a speaker's mouth are more important than other potential cues from a speaker's face –

84   for example, movements from the eyebrows or forehead. In addition, it may also be the case that

85   directing visual attention specifically to the speaker's mouth can benefit adaptation. Indeed, Banks et

86   al. (2020) observed a relationship between the duration of fixations on a speaker's mouth and speech

87   perception accuracy for noise-vocoded sentences, whereby longer fixations were related to more

88   accurate perception, but the evidence for this relationship was relatively weak. Furthermore, when a

89   listener directs their foveal vision towards (i.e., fixates or looks directly at) a speaker's mouth, other

90   cues from the speaker's face are still accessible in peripheral vision and may contribute to overall

91   improvements in speech perception. Although foveal vision provides the greatest visual acuity, (K. G.

92   Munhall et al., 2004) have shown that high spatial frequency is unnecessary for visual speech cues to

93   benefit perception of speech in noise. Similarly, Paré, Richler, ten Hove & Munhall (2003)

94   demonstrated that direct fixation of a speaker's mouth is neither required nor related to the presence

95   of a McGurk effect. Thus, the importance of specifically viewing a speaker's mouth in difficult listening

96   conditions is still not fully clear. Listeners may benefit from viewing a speaker's face as a whole, as

97   they can integrate multiple visual cues from a speaker's face with auditory cues. Conversely, it might

98   be more beneficial if observers can *only* look at the speaker's mouth during adaptation to noise-

99   vocoded speech, as their visual attention would be fully directed to the most salient visual speech

100 cues. That is, listeners might be able to benefit more from focusing solely on the mouth if the eyes

101 region is inaccessible to them.

102   A recent study tested to what extent listeners relied on information from the mouth region

103 while listening to noise-vocoded sentences (Drijvers & Özyürek, 2017). Drijvers and Özyürek's primary

104 aim was to establish how co-speech gestures contribute to information from visible speech to enhance

105 noise-vocoded speech perception, but their design also included conditions in which the speaker's

106 mouth region was obscured. They presented 20 normal hearing native speakers of Dutch with videos

107 of a female speaker producing an action verb in a free-recall task. Specifically, there were three audio-

108 only conditions created by blurring the speaker's mouth (clear (undegraded), 6-band noise-vocoded

109 speech, and 2-band noise-vocoded). The design also included three speech plus visual speech

110 conditions with clear, 6-band and 2-band degraded speech. Moreover, there were three conditions

111 pairing clear, 6-band and 2-band degraded speech with visual speech and an iconic gesture. Finally,

112   two visual-only control conditions were created by removing the audio in the visual and visual plus

113   iconic gesture conditions, see Figure 1 in Drijvers and Özyürek for a visual representation of all

114   conditions). Participants were tested in a within-group design and completed all conditions, however,

115   we will focus on the results relevant to the present study, omitting the effects of the presence of the

116   iconic gesture. Compared to the conditions in which the mouth region was blurred, participants

117   performed on average 10-20% better for the two vocoding conditions when full audiovisual

118   information was available. However, as Drijvers and Özyürek did not test whether and how availability

119   of visual speech information displayed by the mouth affected adaptation to noise-vocoded speech,

120   this question remains unaddressed.

121   The current study aimed to establish to what extent the audiovisual benefit during perception

122   of and adaptation to audiovisual noise-vocoded sentences relies on viewing visual cues from different

123   parts of the face. Although movements from the speaker's mouth provide the greatest and most

124   informative cues, movements in extra-oral areas (for example the upper and outer face and eye

125   region) may also contribute to speech perception, albeit to a lesser extent, especially as not all acoustic

126   elements of speech have equivalent mouth movements. For example, Scheinberg (1980) found that

127   cheek puffiness could help observers identify consonants that are not discriminable based on mouth

128   movements, while Preminger et al. (1998) found that certain consonants can be identified when

129   viewing the upper part of the face only (i.e., with the mouth region masked). Lansing & McConkie

130   (1999) also found that the upper face region can provide observers with information for sentence

131   intonation. Accordingly, facial and head movements have been found to be closely related to, and

132   predictive of, the acoustics features of speech (Munhall & Vatikiotis-Bateson, 1998; Yehia et al., 1998).

133   Thomas & Jordan (2004) tested perception of congruent and incongruent words in noise while

134   manipulating movements in different areas of the speaker's face (namely the mouth and outer face),

135   while also manipulating the visibility of the mouth and eye region. They found that mouth movements

136   were the most important for perception, but that information from extra-oral movements (from the

137   outer face and upper eye region) also contributed to observers' perception. The present study did not

138 aim to identify the exact extra-oral facial regions that may contribute to perception of noise-vocoded

139 speech; nevertheless, based on the above findings, we predicted that some information may be

140 gained by observers from our speaker's upper facial region when only this region was visible (i.e.,

141 when the mouth was obscured), compared to when the upper eye region was not visible (Hypothesis

142 3).

143   We tested our three hypotheses using five conditions in an experiment in which we tested

144 perception of 4-band audiovisual noise-vocoded sentences also used in Banks et al. (2020) for five

145 groups of participants in a between-group design. In condition AV Full, participants were exposed to

146 audiovisual stimuli with the whole face of the speaker visible. The next three conditions were included

147 to establish the relative relevance of different parts of the face for adaptation to and perception of

148 audiovisual noise-vocoded sentences, so we tested a group of participants who could not see the eye

149 region, (AV Mouth) but who could see the mouth region, and a group had access to the eye region,

150 but not the mouth region (AV Eyes). Another group of participants was exposed to a video of the

151 speaker with the mouth and eyes obscured from view (AV Blocked), and a final group was shown a

152 still image of the speaker while being tested (AV Still), so it contained no useful visual cues at all, per

153 Banks et al. (2020).

154   We predicted a main effect of condition and a two-way interaction between condition and trial

155 which would indicate differences in perception and adaptation between conditions. Specifically,

156 hypothesis 1 is supported if conditions where the mouth is visible (AV Full and AV Mouth) show better

157 perception and greater adaptation than conditions where the mouth is not visible (AV Eyes, AV

158 Blocked and AV Still). Support for hypothesis 2 would require significant differences between the AV

159 Full and AV Mouth conditions, with better perception and greater adaptation in the AV Mouth

160 condition. Hypothesis 3 would be supported if we find significantly better perception and adaptation

161 in the AV Eyes condition (i.e., when only the eye region was visible), than in the AV Still, and AV Blocked

162 conditions (when the eye region was not visible).

163   In addition, we aimed to replicate the behavioural results reported in Banks et al. (2020) in an

164   online experimental paradigm to demonstrate that participants were able to adapt outside the lab.

165   Finally, we also asked participants to give us an indication of their perceived effort as different

166   circumstances in which people process distorted speech have been shown to affect performance in

167   similar ways yet be associated with different levels of perceived effort (McGarrigle et al., 2014, 2017;

168   Pichora-Fuller et al., 2016). Finally, to ensure participants attended to the speaker's face, we queried

169   them afterwards about how much attention they paid to the speaker's face and how much they

170   thought being able to see the speaker's face helped them during the task.

171

172   **Method**

173   *Participants*

174   We tested 150 participants (18-30 years of age (Y), 125F and 25M), who all declared to be native

175   monolingual speakers of British English and be resident in the UK at the time of the experiment. All

176   declared to have good hearing and vision, and to not have any neurological or psychiatric disorders

177   (including dyslexia). All participants were recruited through the online platform Prolific.co, and the

178   experiment was hosted on Gorilla.sc. We tested 30 participants per condition. Participants were

179   randomly allocated to each condition and were restricted from participating to more than one

180   condition/group in the experiment. Our minimal sample size, per group as well as the ratio of female

181   and male participants was based on Banks et al. (2020), and we tested 30 participants with a ratio of

182   25F:5M participants (see the Analysis section for further justification of the selected sample size). We

183   replaced one male participant in condition AV Blocked and one male participant in condition Eyes,

184   both for not engaging with the task (i.e., not giving a single response). The demographics were as

185   follows across the five conditions: AV Full 25F|5M, mean 24.6Y, standard deviation (SD) 3.8SY, AV

186   Mouth 25F|5M, mean 23.9Y, SD 3.6Y, AV Blocked 25F|5M, mean 22.8Y, SD 3.2Y, AV Still 25F|5M,

187   mean 24.5Y, SD 3.8Y, AV Eyes 25F|5M, mean 24.3Y, SD 4.0Y. Participants and the speaker all

188   consented to take part and were paid upon completion of the experiment at a rate corresponding to

189   £7.50 per hour (participants). The speaker consented to having her image published and was not paid.

190   The experiment was approved by UCL's Research Ethics Committee (UREC, #0599.001).

191   *Materials*

192   We used the same materials as in Banks et al. (2020) and adapted them to create the stimuli for the

193   specific conditions in the present study. Banks et al. originally used 91 randomly selected Institute of

194   Electrical and Electronics Engineers Harvard sentences (IEEE, 1969). Stimuli were recorded in a

195   soundproofed laboratory using a Shure SM58 microphone and a High-Definition Canon HV30 camera.

196   A 26-year-old female native British English speaker recited the sentences, and was asked to look

197   directly at the camera, to remain still and to maintain a neutral facial expression throughout the

198   recordings to minimise head movement (see Figure 1). Video recordings were subsequently imported

199   into iMovie 11, running on an Apple MacBook Pro, as large (960 x 540) high-definition digital video

200   (.dv) files. Video recordings were then edited to create a video clip per sentence. The audio tracks for

201   each clip were extracted as audio (.wav) files, then normalised by equating the root mean square

202   amplitude. Next, they were resampled at 22kHz in stereo, cropped at the nearest zero crossings at

203   voice onset and offset, and vocoded using Praat speech processing software (Boersma & Weenink,

204   2012) and custom scripts. Speech recordings were noise-vocoded (Shannon et al., 1995) using four

205   frequency bands (cut-offs: 50Hz → 369Hz → 1160Hz → 3124Hz → 8000Hz), selected to represent

206   equal spacing along the basilar membrane (Greenwood, 1990). Of the 91 sentences that were

207   originally recorded, we randomly selected a subset of 60 for inclusion in the online experiment. To

208   ensure that timing of the audio and video was synchronous, we attached the noise-vocoded audio

209   stimuli as an audio track to the video stimuli using Final Cut Pro as a mono track to be played over

210   both channels of a participant's headphones. We repeated the same procedure for an additional single

211   sentence in quiet, to be used in the practice trial presented prior to the main experiment. Audiovisual

212   stimuli were saved as MPeg-4 movie (MP4) files with a resolution of 1920x1080. We also created white

213   rectangular shapes that were used to cover (parts of) the speaker's face in four conditions. The

214   rectangle used to cover the eyes or the mouth in the conditions AV Mouth, AV Eyes, and AV Blocked

215   was a width of 1920 pixels and a height of 720 pixels and a resolution of 300 pixels per inch. For the

216   condition AV Still, we used a screenshot of the speaker's face in PNG format with a width of 1907

217   pixels and a height of 1074 pixels and a resolution of 300 pixels per inch.

218   *Procedure*

219   The experiment was conducted online, via the Gorilla Experiment Builder (Gorilla.sc) (Anwyl-Irvine et

220   al., 2020) and participants were recruited via Prolific (Prolific.co). Upon receiving an email invitation

221   via Prolific, participants entered the online study and were linked through to the experiment hosted

222   in Gorilla. They were then given information on the study, before providing consent. Participants who

223   did not provide consent were rejected from the study. Next, they were asked to enable auto play of

224   video and audio on their internet browser, maximise their screen, and plug in their headphones

225   (Bluetooth headphones were excluded per participant report). The mean display resolution across

226   participants was 1474 (SD = 234) * 856 (SD = 125), and the mean resolution of the experiment display

227   (viewport) was 1466 (SD = 226) * 770 (SD = 123). They were subsequently routed to a page where

228   they could check their sound levels where they were played a short sound consisting of one second of

229   white noise. They were asked to replay this sound over their headphones and adjust their volume to

230   a comfortable level before progressing to the headphone check.

231   The next check was previously developed to allow for more control over sound presentation in

232   online experiments by providing a test to establish whether participants are wearing headphones

233   (Woods et al., 2017). This test was designed to be difficult to complete if the participant is not wearing

234   headphones, through the manipulation of anti-phase attenuation rather than differences in intensity

235   between the tones. The headphone check is designed as a 3AFC task in which six sets of three sine

236   wave tone stimuli are played. After participants clicked at the start button, a new page appeared

237   where three 200Hz tones were played with a duration of 1000ms, with 100ms on- and off-ramps, two

238   at -14dB (in-phase) and one at -20dB (180° out of phase). The stimulus duration per triad of tones was

239   four seconds (tone duration: 900ms, interstimulus interval: 600ms, time before first stimulus onset:

240   100ms, time after the last stimulus offset: 100ms). Participants listened to six trials in total. The

241   participants were to decide which tone they perceived as having the lowest intensity by selecting one

242   of three buttons labelled "FIRST sound is SOFTEST", "SECOND sound is SOFTEST", and "THIRD sound

243   is SOFTEST". They had to select the correct stimulus for five of the six trials (accuracy level of 83.3%),

244   or they were rejected from the study.

245   Participants who successfully completed the headphone check were subsequently routed

246   through to the instructions and a single undistorted and visually unobstructed practice sentence. For

247   the conditions AV Full, AV Mouth, AV Eyes, and AV Blocked, all groups saw the same MP4 video and

248   heard the corresponding undistorted sentence. For the condition AV Still, participants were presented

249   with the same still PNG image as used in the main experiment. Participants were asked attend to the

250   video and spoken sentence and to type into a response text box any words they thought they had

251   heard. After the single practice trial, they were shown a screen explaining what they should have typed

252   in the response box. Subsequently, they were told that the main experiment would start next and that

253   all trials would progress to the next trial automatically, so they would not be able to take a break until

254   the main task finished.

255   In the main task, participants transcribed 60 noise-vocoded sentences. Participants triggered

256   the start of the experiment and each subsequent trial by pressing the "Next" Button at the bottom of

257   each screen. In each trial, the audiovisual noise-vocoded sentence and corresponding visual stimulus

258   was presented. The noise-vocoded sentence was played only once per trial. The visual part of this

259   stimulus was different for the five conditions (see Figure 1). Participants in the Full condition saw the

260   unobstructed video. Participants in the AV Mouth and AV Eyes conditions were shown the video with

261   a white rectangle covering the eyes or mouth of the speaker, respectively. As can be seen in Figure 1,

262   the block covered either the upper or lower part of the face. The tip of the nose and chin were mostly

263   visible in the AV Eyes and AV Mouth videos, but sometimes not visible due to the speaker moving

264   while speaking. Participants in the AV Blocked condition were shown the video with a white block

265   covering the mouth, nose, and eyes of the speaker Here, the chin and forehead of the speaker were

266   visible, but the space in between was covered, so that only small head movements were visible.

267   Finally, participants in the AV Still condition were shown a still image of the speaker, where the entire

268   face vas visible, but no movement.

269   After the main task, the programme moved to a final response screen, where participants typed

270   in their response. They were asked to type in "/" if they could not decipher any words in the sentence.

271   After finishing typing, they could move to the next stimulus by pressing the "Next" button. If they did

272   not press this button, the experiment moved to the next trial automatically after 23 seconds. After

273   the main task was completed, participants were shown a screen with three response sliders and a

274   response text box. Participants were asked to provide ratings of their perceived effort as follows:

275   "*Question 1: Please indicate using the slider below how effortful you found it to understand the*

276   *sentences (0 = Not effortful, 100 very effortful):*". They were also asked to rate what proportion of the

277   time the video was presented they looked at the speaker's face as follows: "*Question 2: Please indicate*

278   *using the slider below what proportion of the time you spent looking at the speaker's face when the*

279   *video was presented (0% of the time - 100% of the time):*". A final question queried whether being able

280   to see the speaker's face helped their speech performance: "*Question 3: How much do you think*

281   *looking at the speaker's face helped you understand the sentences? (0 Not at all - 100 Very much):*".

282   At the bottom of the page was a response box where they were invited to type in any comments. After

283   they clicked next, they were returned to Prolific for payment.

284   All data were collected in a single session lasting approximately 20 minutes. However, as the

285   experiment was in part self-paced, durations differed across participants, although the main

286   transcription part of the study lasted maximally 30min if participants did not manually progress each

287   trial. A single participant (in the condition AV Blocked) took 30min for the main task, but as their

288   responses were within the ranges specified for accuracy in their group, we included their data in the

289   final analysis. Average durations for the entire session was 21min and 4s across all 150 participants

290   (SD 7min and 14s). The session timed out automatically after 90 minutes. The average duration for

291   the main transcription part of the session task was 12min and 30s (SD 3min and 18s) and as follows

292   for individual conditions: AV Full 13min and 18s (SD 4min and 19s), AV Mouth 13mins and 27s (SD

293   2min and 18s), AV Blocked 13min and 31s (SD 2min and 18s), AV Still 10min and 28s (SD 3min and

294   19s), and AV Eyes took 13 minutes and 31 seconds (SD 2 minutes and 18 seconds). Data from all

295   participants in the Condition AV full was collected first, followed by the AV Mouth condition, the AV

296   Blocked condition, the AV Still condition, and the AV Eyes condition. Online testing took place in May-

297   June 2020.

298   ---Include Figure 1 about here ---

299   *Design and Analysis*

300   The experiment measured speech perception performance as the by-trial percentage of words

301   accurately entered as the dependent variable. The independent variables were Trial and Condition.

302   Trial was the stimulus number ranging from 1-60. The use of trial as an index of exposure contrasts

303   with the proposed analysis in the pre-registration. Upon reflection, we opted to use trial as it would

304   give a more fine-grained and accurate analysis of adaptation patterns. To support this choice we

305   calculated the BF_{10} for models utilising blocks and trials for both the AV Full (Blocks BF_{10} =8.659 x 10^{+16},

306   Trials BF_{10} = 4.256 x 10^{+20}) and all conditions (Blocks BF_{10} = 2.515 x 10^{+30}, Trials BF_{10} = 1.112 x 10^{+36})

307   analysis, both of which supported the by-trials analysis. As such, we will hitherto present the by-trials

308   analyses. The pre-registered by-blocks analysis (https://osf.io/2w6j4/) is presented in the

309   supplementary materials (see Supplementary Analysis 1, Table S2, and figure S1 for the AV Full analysis

310   and Supplementary Analysis 2, Table S3, and figure S2 for the by-blocks and conditions analysis). The

311   factor Condition had five levels: AV Full, AV Mouth, AV Eyes, AV Blocked, AV Still to test our three

312   hypotheses outlined in the introduction. Hypothesis 1 predicted that being able to see the moving

313   mouth improves adaptation and perception compared to when it is not visible. Hypothesis 2 predicted

314   that having to focus on the mouth region (i.e., when only the mouth region is visible) improves

315   adaptation and perception compared to when the full face is visible. Hypothesis 3 predicted that being

316   able to see only the eye region improves perception and adaptation compared to when it is not visible.

317   The AV Mouth condition was included to establish to what extent forcing participants to focus on the

318   speaker's mouth affects perception of and adaptation to noise-vocoded sentences. The AV Eyes

319   condition was included to test if and how being able to see only the eye region supports perception

320   of and adaptation to noise-vocoded speech. The AV Blocked condition was included to determine if

321   and how removing information conveyed by the speaker's mouth and eyes affected

322   perception/adaptation. The AV Still condition was included to test if the presence of moving visual

323   information affected perception/adaptation and to test if results for this condition show the same

324   effects as reported in Banks et al. (2020), who included this condition as a control. If Hypothesis 1 is

325   correct, then participants in condition AV mouth and AV Full should show better speech perception

326   performance and greater adaptation, than participants in conditions AV Blocked, AV Eyes, and AV Still.

327   If Hypothesis 2 is correct, then participants in condition AV Mouth should show better

328   perception/adaptation, than participants in condition AV Full. Finally, if Hypothesis 3 is correct, then

329   participants in condition AV Eyes should show better perception/adaptation than participants in

330   conditions AV Blocked and AV Still.

331   We retrospectively scored participants' responses according to how many key words (content

332   or function words) they correctly repeated out of a maximum of four following Banks et al. (2020).

333   Banks et al. chose four keywords as the sentences they were all of varying duration, and therefore

334   using four keywords made perception accuracy comparable across all sentences. We

335   included/excluded (typed) responses as follows. Responses were scored as correct despite incorrect

336   suffixes (such as -s, -ed, -ing) or verb endings; however, if only part of a word (including compound

337   words) was repeated, this response was scored as incorrect following Banks et al. (2015, 2020). It

338   should be noted that Banks et al. audio-recorded participants' verbal responses, and these responses

339   were subsequently judged by an experimenter. In contrast, as we asked participants to type in their

340   responses, we also included homonyms (e.g., "weak" instead of "week"), compound words separated

341   by a space (e.g., "door knob" instead of "doorknob", as well obvious typos (e.g., "whire" instead of

342   "wire"). Moreover, we excluded participants as follows: participants who had an average % error rate

343   greater than three standard deviations (3SD) away from the group mean were excluded from further

344   analysis and replaced. Participants were excluded if they failed to provide responses to a number of

345   trials >2SD from the group mean.

346   As condition AV Full was intended to closely replicate the design of the audiovisual condition in

347   Banks et al. (2020), we initially decided to collect 30 participants as a minimum sample and then used

348   sequential hypothesis testing with Bayes Factors to determine our final sample size (Schönbrodt et

349   al., 2017). After collecting the initial 30 participants, we calculated BF_{10} to assess whether we reached

350   a pre-defined level of evidence ($BF_{10} > 3$ in favour of the alternative hypothesis, and $BF_{10} < 0.2$ in favour

351   of the null hypotheses). BF_{10} indicates how likely the data are to occur under the alternative

352   hypothesis. If $BF_{10} > 0.2$ and < 3.0, we aimed to collect additional participants. After collecting an

353   additional participant for each group, we would calculate BF_{10} until we met the conditions noted

354   above. In the case more participants were required, we planned to minimise the risk of type 1 and

355   type 2 errors by graphing BF_{10} after running each additional participant to assess whether any changes

356   in the BF were stable. When the BF was stable for four consecutive participants, we planned to cease

357   data collection. However, the BF_{10} exceeded the criterion value of 3.0 after collecting 30 participants

358   for each condition. As such, additional data collection was not necessary.

359   To calculate BF_{10}, we utilised Bayes Information Criterion (BIC) values obtained during model

360   comparison of linear mixed effects (LME). The *step* function of *lmerTest* utilises a backward model-

361   selection strategy to find the best fitting model. *Step* takes as input an *lmer* model. First, the random

362   effects structure is subjected to backwards elimination, where random effects are either reduced or

363   removed utilising log-likelihood tests. Random effects are removed from the model where it

364   significantly improves model fit ($p < .05$). Next, this procedure is repeated for main effects, however,

365   in this stage χ^2 tests of model fit are used after the removal of each model term, starting with the most

366   complex interactions. Next, we performed a hierarchical comparison of the best fitting model (H1, e.g.

367   accuracy ~ (1|participant) + trial) with a model excluding the effect of interest (H0, e.g. accuracy ~

368   (1|participant)) using the *anova* function to obtain BIC values for each. We used the difference in BIC

369   to compute the Bayes Factor (BF_{10}) using the following equation (Jarosz & Wiley, 2014):

370
$$BF_{01} = e^{\Delta BIC/2}$$

371
$$BF_{01} = e^{(BICH1-BICH0)/2}$$

372
$$BF_{10} = 1/BF_{01}$$

373 We initially collected and analysed the data from condition AV Full only, as all hypotheses relied on

374 whether it is possible to measure perceptual adaptation to distorted speech in an online paradigm. In

375 this first stage, we tested whether accuracy increases over the course of the experiment, as measured

376 over the course of the 60 trials. In this case, the H1 BIC value corresponded to a model predicting

377 accuracy including main effect of trial, whilst H0 BIC was for a model only including random by-

378 participants and by-items slopes. In the second stage, we analysed data for all five conditions and

379 tested main effects of trial (as a linear and polynomial) and condition, and their interaction using LMEs

380 as described above. The H1 BIC therefore modelled the critical two-way interaction between trial (see

381 section 3.1.2) and condition, while the H0 BIC included the main effects only.

382 We also analysed the data collected in the questionnaire presented to participants in the online

383 study after the main task. However, as this dataset was comprised of only one observation per

384 question per participant, we utilised simple linear models to analyse the effort questionnaire data.

385 The design of conditions AV Full, AV Mouth, and AV Blocked was preregistered on

386 www.AsPredicted.org under number #41527 *"Transcribing distorted audiovisual speech."* The

387 inclusion and design of conditions AV Still and AV Eyes was preregistered on www.AsPredicted.org

388 under number #42910 *"Transcribing distorted audiovisual speech, a follow-up study."* In all analyses

389 we discuss results for the five conditions in the order they were collected. All raw data plus analysis

390 scripts can be found on the Open Science Framework: https://osf.io/2w6j4/.

391

392 **Results**

393 *Accuracy*

394 For 12 trials (0.13%), stimulus materials could not be loaded by Gorilla across all 150 participants. In

395 seven cases for the Full condition, two for the Mouth condition, two for the Eyes condition, the video

396   mp4 file could not be loaded (which occurred to a different sentence every time and seemed to be

397   due to a random occurrence or glitch in Gorilla). In a single case for the Still condition, the audio file

398   could not be loaded. These 12 cases were therefore removed from the data set.

399

400   *Accuracy: AV Full*

401   Participants in the AV Full condition reported a mean of 1.7 (SD = 1.4) key words correct across the 60

402   sentences. We first examined the effect Trial to test our hypothesis that participants could adapt to

403   noise-vocoded sentences. In this analysis, inclusion of the by-participants (p = .296) and by-items (p =

404   .767) slopes did not significantly improve model fit. The best fitting model therefore included only by-

405   participants and by-items random intercepts, and the main effect of trial (see Table S1 in the

406   supplementary materials for the full model summary). In this case, the alternative hypothesis states

407   that participant performance would increase over trials. Therefore, we compared the best fitting

408   model (BIC = 17197) against a model including only the random effects (BIC = 17275). BF_{10} was > 150,

409   indicating that the evidence in favour of the alterative hypothesis – that adaptation will occur across

410   trials – was very strong (Raftery, 1995). The model outcomes for the linear effect of trial was significant

411   (t = 10.387, p < .0001), indicating that participants in the AV Full condition adapted to the masked

412   speech over trials. Whilst the quadratic effect of trial also reached significance (t = -2.329, p = .02), the

413   smaller t- and p-values indicate that the effect of trial was better modelled as a linear effect. This

414   effect is illustrated in Figure 2 below - generated using the *effects* package (Fox et al., 2019) to extract

415   model estimates at five moments of the distribution for the linear function - which displays the model

416   estimates of performance by Trial. To conclude, participants showed an increase in performance

417   across the trials for the AV Full condition. Following these results, we decided to collect and analyse

418   data for the four follow-up conditions.

419   *--- Include Figure 2 about here ---*

420   *Accuracy for all five conditions*

421 Mean accuracy was 43.069 (SD = 36.03) in the AV Full condition, 43.806 (SD = 36.13) in the AV Mouth

422 condition, 5.486 (SD =13.625) in the AV Eyes condition, 5.333 (SD = 14.122) in the AV Blocked

423 condition, and 3.861 (SD = 11.323) in the AV still condition. Figure 3 displays a locally estimated

424 smoothed scatterplot (LOESS) of accuracy over the 60 trials. The LOESS function from *ggplot2*

425 (Wickham, 2016) fits simple linear models to local subsets of the data to describe its variance, point

426 by point. Taken together, the descriptive statistics suggest that performance was almost identical

427 when participants were able to see the speaker's mouth movements (i.e., in the AV Full and AV Mouth

428 conditions).

429   *--- Include Figure 3 about here ---*

430 To analyse all five conditions, we followed the same procedure as the analysis for the AV Full

431 condition, while including testing condition (factor-coded) as an additional main effect, and the two-

432 way interaction between condition and trial. The maximal model upon which we conducted the

433 backwards stepwise model comparison therefore included by-item and by-participant random

434 intercepts, a random intercept for participant nested within condition, the main effects of trial and

435 condition, and the two-way interactions between trial and condition. Random slopes were excluded

436 from the analysis, as their inclusion resulted in issues of singular model fit. The backwards stepwise

437 model selection indicated that the inclusion of the main effect of block (p = .183), the interaction

438 between block and condition (p = .766) and the simple by-participants random effect did not improve

439 model fit (p = 1). As a result, the final model included a by-items random intercept, a random-intercept

440 for participant nested within condition, the main effects of trial and condition, and the two-way

441 interaction between condition and trial (see Table S4 in the supplementary materials for full model

442 syntax and model summary). The analysis including the effect of block can also be found in the

443 supplementary materials (Supplementary Analysis 2, Table S3, and figure S2) on the Open Science

444 Framework: https://osf.io/2w6j4/.

445   To assess the likelihood of the alternative hypothesis – that different conditions would elicit

446 different levels of adaptation – we compared a model including the interaction (H_1, BIC = 80876)

447   against a null model only including the main effects (H$_0$, BIC = 80976). BF10 was therefore 5.185 x

448   10^{+21}, vastly exceeding Raftery's (1995) threshold for strong evidence (> 150) in favour of the

449   alternative hypothesis. This reflects the floor performance seen in the AV Blocked, AV Eyes, and AV

450   Still conditions relative to the AV Full, and AV Mouth conditions.

451   The outcomes of the linear model indicated that perception differed between conditions. Here,

452   perception is reflected by the main effect of condition; the model tests whether mean performance

453   differed significantly from zero. The results indicated that accuracy in AV Full (t = 18.739, p < .0001),

454   AV Mouth (t = 20.077, p < .0001), AV Blocked (t = 2.444, p = .015), and the AV Eyes (t = 2.514, p = .012)

455   conditions differed from zero. In contrast, accuracy did not differ from zero in the AV Still condition (t

456   = 1.77, p = .078). Critically, performance did not significantly differ between the AV Full and AV Mouth

457   conditions (t = 0.289, p = .77), indicating similar levels of accuracy. Both the AV Full (t = - 14.737, p <

458   .0001) and AV Mouth (t = -15.026, p < .0001) conditions differed significantly from the AV Eyes

459   condition, and both differed significantly from the AV Blocked and AV Still conditions (all t-values > 2,

460   all p-values < .05). The AV Blocked, AV Still and AV Eyes conditions failed to differ from one another

461   (all t-values < 2, all p-values > .05), indicating similar performance at floor in these conditions.

462   Adaptation is measured by the two-way interaction between trial and condition. The interaction

463   term was significant for the AV Full (t = 12.659, p < .0001), AV Mouth (t = 12.657, p < .0001), and AV

464   Eyes conditions (t = 2.092, p = .037). However, participants in the AV Blocked (t = 0.876, p = .381), and

465   AV Still (t = 0.664, p = .507) conditions did not show adaptation with increased exposure. The two-way

466   interaction between trial and condition did not differ significantly between the AV Full and AV Mouth

467   conditions (t = -0.008, p = .994), indicating similar adaptation in these conditions. Both AV Full and AV

468   Mouth differed significantly from the AV Blocked and AV Still conditions (all t-values > 2, all p-values

469   < .05). Adaptation in the AV Block, AV Still and AV Eyes conditions did not significantly differ (all t-

470   values < 2, all p-values > .05). This suggests that while a small amount of adaptation did occur in the

471   AV Eyes condition, it remained indistinguishable from AV Blocked and AV Still conditions, suggesting

472   the adaptation was minimal. For each condition, the data was better described by a linear function of

473   trial (see figure 4 below). Two of the quadratic estimates (between trials 0 to 20, and 50 to 60) for the

474   AV Mouth condition differ from the linear estimates, suggesting that the largest increase in

475   performance occurred in the first twenty trials, and tailed off slightly in the last ten trials. However,

476   the model estimates demonstrate the fit was better for the linear (t = 12.657, p < .0001) relative to

477   the quadratic (t = -5.104, p < .0001) term. This demonstrates that adaptation largely proceeded

478   linearly across trials, with only minor deviations from this trend over training.

479   *--- Include Figure 4 about here ---*

480   In summary, participants in the AV Full and AV Mouth condition showed increased speech

481   perception performance and demonstrated adaptation (i.e., better accuracy for later trials). In

482   contrast, when participants were unable to see the speaker's mouth (AV Block, AV Still, AV Eyes)

483   speech perception was impaired, and participants were unable to adapt to the vocoded speech. Whilst

484   participants in the AV Eyes condition showed adaptation, it was significantly smaller than that seen in

485   the AV Full and AV Mouth conditions, and failed to significantly differ from AV Blocked and AV Still

486   conditions, indicating that the effect was minimal. In comparison to the AV Mouth condition, in the

487   AV Full condition, participants were able to see the speaker's eyes and upper face/head. As perception

488   and adaptation did not differ statistically between these conditions, the results suggest that focusing

489   specifically on the speaker's mouth does not benefit perception or adaptation any more than being

490   able to see the speaker's full face. Taken together, the results support Hypothesis 1 - being able to see

491   the moving mouth improves adaptation and perception - as adaptation and perception did not differ

492   in conditions where participants were able to see the speaker's moving mouth. Hypothesis 2 - that

493   having to focus on the mouth region improves adaptation and perception - did not receive support,

494   however, as participant performance in the AV Mouth and AV Full conditions did not differ, despite

495   being able to see the eyes in the latter. Hypothesis 3 - being able to see the speaker's eyes while the

496   moving mouth is not visible improves adaptation and perception - was also not supported; there were

497   no statistical differences in adaptation or perception between the AV Eyes, AV Still, or AV Blocked

498   conditions.

499

500   *Effort questionnaire*

501   Ratings from two participants in the AV Mouth condition, and from one in the AV Blocked condition

502   were removed as they were >3SD separated from the average for that respective condition.

503   Participants in the condition AV Full provided on average an effort score of 91.2% (SD = 9.7%) and

504   estimated that they had looked at the speaker's face 91.7% (SD = 14%) of the time, and 65% (SD =

505   28.7%) stated that being able to see the speaker's face helped speech perception. Participants in the

506   condition AV Mouth provided on average an effort score of 85.6% (SD = 10.3%), rated they looked at

507   the face 94.6% (SD = 8.8%) of the time, and 61.3% (SD = 31.9%) stated that seeing the speaker's face

508   helped speech perception. Participants in the condition AV Blocked gave an average effort score of

509   97.2% (SD = 11.2%), rated they looked at the face 66.2% (SD = 28.8%) of the time, and 17.2% (SD =

510   22.2%) stated that seeing the speaker's face helped speech perception. Participants in the condition

511   AV Still gave an average effort score of 98.8% (SD = 4.2%), rated they looked at the face 41.9% (SD =

512   30.6%) of the time and 2.1% (SD = 4.7%) stated that seeing the speaker's face helped speech

513   perception. Participants in the condition AV Eyes gave an average effort score of 97.9% (SD = 5.1%),

514   rated they looked at the face 69.4% (SD = 26.3%) of the time and 26.1% (SD = 22.7%) stated that seeing

515   the speaker's face helped speech perception.

516   Three separate models were conducted of the dependent variables Effort (perceived effort

517   score), Face (estimated proportion of time spent looking at the face), and Face (estimation of how

518   much being able to see the face was helpful) with condition (AV Full, AV Mouth, AV Blocked, AV Still,

519   AV Eyes) as a factor. In each case, AV Full was taken as the reference level for the condition factor.

520   The Effort model revealed that participants reported significantly lower effort in the AV Mouth

521   compared to the AV Full condition (t = -2.490, p = .014), whilst the AV Blocked (t = 2.716, p = .007), AV

522   Eyes (t = 3.481, p = .003) and AV Still (t = 3.481, p = .0007) conditions reported significantly higher

523   effort, suggesting that participants found the AV Mouth condition the least effortful (see figure 4). The

524   Face model indicated that participants spent a similar amount of time looking at the speaker's face in

525   the AV Full and AV Mouth conditions ($t = 0.471$, $p = .638$), however participants in the AV Blocked ($t =$

526   -4.159, $p < .0001$), AV Eyes ($t = -3.679$, $p = .0003$), and AV Still ($t = -7.941$, $p < .0001$) conditions spent

527   significantly less time looking at the face. The Face Helped model suggested that participants in the

528   AV Full and AV Mouth condition found a similar benefit from seeing the face ($t = -0.561$, $p = .575$),

529   while participants found the face helped significantly less in the AV Blocked ($t = -7.678$, $p < .0001$), AV

530   Eyes ($t = -6.292$, $p < .0001$) and AV Still conditions ($t = -10.211$, $p < .0001$).

531   Overall, the participant ratings on these three factors align well with the experimental results;

532   participants who could see the speaker's mouth reported lower required effort. Participants reported

533   lower effort for the AV Mouth relative to the AV Full condition. This pattern offers some degree of

534   support for Hypothesis 2 (being forced to focus on the speaker's mouth should improve perception

535   and adaptation); removing the information provided by the eyes in the AV Mouth condition was

536   associated with reduced perceived effort. This effect was not reflected in the accuracy data. When

537   participants could not see the speaker's mouth, effort was increased. Participants who were able to

538   see the speaker's mouth (AV Full, AV Mouth) reported the highest benefit from being able to see the

539   face, and that seeing the face assisted, in contrast to participants who could not (AV Block, AV Eyes,

540   AV Still). As a result, it appeared that in this online testing environment, being able to see the speaker's

541   moving mouth improved both objective (adaptation and perception) and subjective measures of

542   performance (perceived effort).

543   *--- Include Figure 5 about here ---*

544   **Discussion**

545   This study aimed to establish if viewing the mouth or eyes (i.e. the upper or lower part) of the

546   speaker's face affected perception of and adaptation to noise-vocoded sentences when compared to

547   viewing their whole face. We ran an online experiment with five listener groups, who could either see

548   the full moving face of the speaker (AV Full), see the moving face with the eyes blocked (AV Mouth),

549   see the moving face with the mouth blocked (AV Eyes), see the face with the eyes and mouth blocked

550   (AV Blocked), or were presented with a still image of the speaker's face (AV Still). We tested three

551   hypotheses: Hypothesis 1 predicted that being able to see the moving mouth improves adaptation

552   and perception, Hypothesis 2 predicted that having to focus on the mouth region improves adaptation

553   and perception, and Hypothesis 3 predicted that being able to see only the eye region would improve

554   perception and adaptation compared to when the eye region was not visible. All groups transcribed

555   60 4-band noise-vocoded sentences. The results showed clear differences between the five

556   conditions, with participants in the conditions AV Full and AV Mouth showing considerably better

557   overall accuracy scores than the participants in the other three groups, where performance was

558   effectively at floor level. There was no difference in overall accuracy between conditions AV Full and

559   AV Mouth, and no differences were found between AV Eyes, AV Block, and AV Still. Second, the results

560   showed an interaction between condition and trial, indicating that participants in the conditions AV

561   Full and AV mouth improved their accuracy scores over the course of the experiment, while no such

562   pattern was found for the other three conditions. Therefore, perceptual adaptation to noise-vocoded

563   speech was only found when the moving mouth area of the speaker's face was visible.

564   *AV Full and AV Still conditions*

565   The results for the AV Full condition in part replicate the results from the audiovisual condition in

566   Banks et al. (2020) as participants adapted to the noise-vocoded sentences over the four blocks.

567   However, participants performed overall worse than in Banks et al.'s audiovisual condition, as our

568   participants showed an average overall accuracy of 43% correct, and accuracy improved from 33.7%

569   to 50% when comparing how performance improved over the 60 sentences when split into four blocks

570   of 15 sentences, in analogy with Banks et al. Participants in Banks et al.'s audiovisual condition

571   repeated an average of 54% of key words correctly, and this accuracy percentage improved from 42%

572   to 61% over their six testing blocks (participants improved between 42% to 59% over the first four

573   blocks, i.e., over the first 60 sentences). In contrast, the results for the condition AV Still, which

574   replicates the audio-only condition in Banks et al., show a very different picture. Banks et al. report an

575   average performance of 35% of key words correctly repeated, with performance increasing from 24%

576   to 43% over their six blocks (participants improved to 37% over the first four blocks, i.e., over the first

577   60 sentences). We found that average performance was 4% for our AV Still condition on average, with

578   performance remaining largely stable. Therefore, while we mostly replicated the (patterns in) the

579   results for the AV Full condition, such replication was clearly not found for the AV Still condition.

580   Furthermore, baseline and overall accuracy was lower in the AV Full condition in the present study

581   compared to the audiovisual condition in Banks et al.

582   It is not clear what factors can account for the differences in results between our and Banks et

583   al.'s results for the AV Full and AV Still conditions. It seems plausible that this difference might be

584   accounted for by differences across both studies, the most prominent of which is the difference in

585   testing platform. Banks et al. tested their participants in a sound-proofed, light-controlled lab, and

586   participants were tested using the same stimulus delivery parameters (e.g., headphones, intensity and

587   sound card, screen size and resolution) and in the absence of any other distractions. In contrast,

588   participants in the current study were tested online. They all wore headphones, but these headphones

589   varied in quality, and by our estimation only a very small number (two out of 150 participants) of

590   headphones listed by our participants could match the audio quality delivered by the headphones

591   used in Banks et al. (Sennheiser HD 25-SP II). In Banks et al. participants were tested in a more

592   controlled environment in terms of focusing their attention on the task, while in our experiment, we

593   could not control their testing environment and whether they were refraining from engaging in other

594   distractions (e.g., looking at their phone). Also, Banks et al. recorded participants' eye gaze using eye-

595   tracking while participants adapted and could therefore closely monitor whether and where

596   participants looked at the speaker while listening to the audiovisual sentences.

597   As participants were tested in their own environment and eye gaze was not monitored, we

598   cannot be certain that participants attention was focused on the task alone or that they always looked

599   at the video in the audiovisual conditions. However, all participants were asked in a final questionnaire

600   whether and how much they looked at the speaker's face after the main task ended. On average,

601   participants in the condition AV Full estimated that they had looked at the speaker's face 91.7% of the

602   time and 65% stated that being able to see the speaker's face helped speech perception. For the

603   condition AV Still, participants on average looked at the face 98.8% (even though it displayed no

604   movement) of the time even though only 2% stated that being able to see the speaker's face helped

605   speech perception. Second, we presented participants with 30 fewer stimuli than Banks et al., who

606   exposed them to 90 sentences in total. However, it does to seem likely that this issue can explain the

607   observed difference in the results for the AV Full and AV Still conditions, as baseline accuracy between

608   the two studies was vastly different. A final reason might be due to differences in participant sample,

609   particularly given that participants in Banks et al were recruited from a University (and were therefore

610   mostly undergraduate students), whereas the sample in the present study was drawn from the

611   general population, or as far as participants on Prolific represent this population. However, as we

612   included two conditions where both the mouth and eyes were blocked (and participants could only

613   take part in one condition/group), using slightly different visual stimuli, i.e., a still image compared to

614   the video of the speaker with eyes and mouth obscured, both of which had similarly poor overall

615   accuracy, this explanation also seems unlikely. It is thus plausible that differences between our results

616   for condition AV Full and AV Still and Banks et al.'s were mostly related to the differences in testing

617   conditions: online versus lab-based.

618   *AV Mouth, Eyes, and Block conditions*

619   We included the AV Mouth, Eyes, and Blocked conditions to test the three hypotheses of this study.

620   Hypothesis 1 stated that being able to see the moving mouth region (AV Full and AV Mouth) will show

621   better speech perception performance, and greater adaptation, than when the mouth region is not

622   visible (AV Eyes, AV Still and AV Blocked). AV Mouth was also included to test whether participants

623   would perform better and adapt more if their attention was focused on the mouth region per

624   Hypothesis 2. The condition AV Eyes and AV Blocked were included to establish whether information

625   from the eyes was useful per hypothesis 3. The results from AV Mouth, in which participants could

626   see the mouth moving but the eyes were blocked, were nearly identical to those reported for the AV

627   Full condition, and therefore also replicate in part the results for the audiovisual condition in Banks et

628   al. Better speech perception performance in AV Full and AV Mouth compared to the other three

629   conditions (AV Eyes, AV Still and AV Blocked), and an interaction between condition and trial, confirm

630   Hypothesis 1. Next, we predicted that participants in the AV Mouth condition might show better

631   overall speech perception performance, and greater adaptation than the AV Full condition, as their

632   attention would be focused specifically on the speaker's articulatory mouth movements. This

633   prediction was not supported by speech accuracy results, as performance did not differ between the

634   AV Mouth and the AV Full group, and there was also no difference in the rate and amount of

635   perceptual adaptation. Thus, the results refute Hypothesis 2 with respect to objective task

636   performance. However, we found that being able to see the speaker's moving mouth without the eyes

637   improved a subjective measure of performance (perceived effort) compared to the AV Full condition.

638   Banks et al (2020) found a relationship between longer fixations on a speaker's mouth and better

639   perception of noise-vocoded speech, although evidence for this was weak. However, as we did not

640   specifically account for subjective performance when we designed the experiment (and it was not

641   included in the preregistration), the present results do not confirm Hypothesis 2. Yet, we are planning

642   to explore the subjective performance differences between distortion conditions further in future

643   online studies. Nevertheless, we cannot exclude the possibility that in the AV full and the AV mouth

644   conditions participants focused only on the mouth region, and that this is the reason for the similar

645   performance in the two conditions. Future studies could elucidate this issue by combining online

646   perceptual adaptation to noise-vocoded speech with eye-tracking using the participant's webcam. A

647   recent study has shown that it is feasible to collect eye gaze data online and establish whether

648   participants look more at the mouth or the eyes of a moving face (Semmelmann & Weigelt, 2018).

649   Using a setup similar to the one used in Semmelman & Weigelt could further clarify whether

650   participants looked at the mouth equally in the conditions tested in the present study.

651   Finally, it can be concluded that not being able to see the speaker's eyes and the upper part

652   of the speaker's face, did not benefit speech perception or adaptation as we predicted for Hypothesis

653   3. It was found to be unhelpful for participants to be only able to see the speaker's eyes and upper

654   face during perception of noise-vocoded audiovisual speech, i.e., any movements from the speaker's

655   eye region offered no benefit to perception of the noise-vocoded speech. Furthermore, overall

656   accuracy and adaptation were almost identical in the AV Full and AV mouth conditions, indicating that

657   viewing the speaker's entire face offered no additional benefits over and above viewing only their

658   mouth. Thus, our results do not support Hypothesis 3 that information from the speaker's eye region

659   may contribute to perception of degraded speech.

660   The results for the conditions AV Eyes and AV Blocked mirrored those for AV Still. For all three

661   conditions a floor effect was found, with participants on average reporting 4.9% of key words

662   correctly. In addition, participants did not improve their performance over the course of the

663   experiment, although there was a small improvement over trials in the AV Eyes condition. These

664   results were unexpected as they do not replicate findings reported by other studies using noise-

665   vocoded sentences. The majority of studies examining adaptation to noise-vocoded speech were

666   conducted using audio-only stimuli, yet still manage to find evidence that participants adapt after

667   short-term exposure to a low number of sentences or words (Davis et al., 2005; Huyck & Johnsrude,

668   2012; Kennedy-Higgins et al., 2020; Paulus et al., 2020). Studies using audiovisual stimuli report

669   adaptation for their audio-only conditions. For instance, Pilling & Thomas (2011) presented two

670   groups of participants with noise-vocoded sentences in audiovisual and audio-only training conditions.

671   Participants in both groups adapted readily to the noise-vocoded sentences, although participants in

672   the audiovisual group adapted more than those in the auditory-only condition (participants were

673   exposed to three blocks of 76 sentences (pre-training, training, post-training) and reported key

674   words). Nevertheless, Pilling & Thomas used an 8-band noise-vocoder with a pitch-shift that aimed to

675   approximate a cochlear implant with a 6 mm place mismatch, while we used 4-band noise-vocoded

676   speech without a pitch shift. The degradations are therefore different, and it is likely that the

677   degradation used in Pilling & Thomas resulted in overall more intelligible speech.

678   Our results for the AV Blocked, AV Still, and AV Eyes conditions are instead similar to those

679   reported in (Drijvers & Özyürek, 2017), as they report close to floor-level performance for their

680   conditions with 2- and 6-band noise-vocoded speech where the mouth area was not clearly visible.

681 Our results are also in line with those reported by Rosen et al. (1999), who examined perception of

682 and adaptation to 4-band spectrally shifted noise-vocoded speech in a live setup. Participants were

683 trained to the distorted speech with a connected discourse tracking task. In this task they were to

684 repeat words when communicating with a live speaker who could be seen though a glass partition and

685 whose speech was vocoded and pitch-shifted in real-time. Participants were exposed to eight blocks

686 of five minutes of speech and reported back what they could understand. After the first block, they

687 could only understand around 1% of the key words, but after the training had finished, they could

688 understand over 40%, showing a clear adaptation effect. It therefore appears that our participants'

689 performance was comparable to that of the participants in Rosen et al. after the first training block.

690 However, we do not know if our participants would subsequently also improve, as our experiment

691 ended after the presentation of 60 sentences. It would be interesting to repeat our study with a more

692 longitudinal design, e.g., like Rosen et al.'s study, to enable establishing the extent to which learning

693 continues and to also gain insights into individual patterns of learning.

694 *Limitations*

695 Despite similarities with previous studies of noise-vocoded speech perception, it is unclear why

696 participants were unable to understand or adapt to the noise-vocoded sentences in the AV Blocked,

697 AV Still, and AV Eyes conditions. It seems likely that participants simply 'gave up' as they were not able

698 to understand most of the sentence when they first heard them. It could be that participants in fact

699 need to be able to understand at least part of the sentence upon the first encounter for perception to

700 improve over time. However, this explanation cannot fully explain our results, as participants in Rosen

701 et al. and Pilling & Thomas also started out at a similarly low performance level of <5% correct, and

702 participants all improved over the course of both studies. It should be noted that their participants

703 were provided with a significantly larger number of sentences/utterances than was the case in our

704 study: Pilling & Thomas exposed participants to 228 noise-vocoded sentences, and participants in

705 Rosen et al. listened to a speaker whose speech was noise-vocoded and spectrally shifted for a total

706 of 40 minutes in eight five-minute blocks. In follow-up studies, it could be considered to provide

707 participants with substantially more training sentences and to present these to them in separate

708 sessions to further facilitate learning and avoid potential fatigue effects. In addition, it might be useful

709 to examine if participants would show better performance and adaptation in similar online conditions

710 to AV Block, AV Still, and AV Eyes for 6, 9, or 10-channel noise-vocoded sentences; the online (and

711 anonymous) setting may have particularly affected participants' motivation to understand the speech

712 compared to a laboratory setting where an experimenter is present. Moreover, a final possibility could

713 be that participants in the online conditions simply did not adapt because they were not made aware

714 that they actually *could* adapt to this type of distortion. We did not mention in the instructions that

715 we expected them to adapt and the title of the study on Prolific was "Transcribing distorted

716 audiovisual speech". Perhaps participants would adapt more if they were 'primed' to learn in the

717 instructions or if learning or adaptation was mentioned in the name of the experiment. We asked

718 participants to provide comments after the main task ended, and most comments could be

719 summarised as they all found the task very difficult and the speech near-impossible to understand.

720 Second, it seems possible that participants may not have attended the stimulus materials sufficiently

721 to correctly perceive them. Huyck & Johnsrude (2012) showed that perceptual adaption to noise-

722 vocoded speech only occurred when attention was selectively directed to the speech task, rather than

723 concurring auditory and visual distractors in their task. Therefore, it seems likely that the current result

724 may in part be explained by participants not paying their full attention to the stimuli. It is not

725 straightforward to control for this issue in an online design. However, or suggestion to combine our

726 online design with eye-tracking to monitor the extent to which participants fixated on the face would

727 likely provide more insights into this issue.

728 Third, we stress that our results may be modest in scope due to the specific manipulation used,

729 which involved occluding of parts of the face using superimposed white blocks. While this

730 manipulation was very effective in establishing the intended aim of preventing the participants from

731 viewing specific parts of the face, it was somewhat lacking in ecological validity. A more ecologically

732 valid manipulation would be to record stimuli while the speaker was actually wearing a face mask to

733   cover the mouth and/or eyes. Yet, when using a face mask to obscure parts of the speaker's face, care

734   should be taken to avoid a potential confound between speech production and occlusion site (mouth

735   or eyes). Wearing a face mask over the mouth might alter speech production. For instance, the face

736   mask could impede speech articulation, making speech intrinsically less understandable, e.g., due to

737   the speaker articulating less clearly (although there is some evidence that the effect of wearing a face

738   covering is relatively minor (Llamas et al., 2009). Alternatively, the speaker could aim to compensate

739   for the face mask's presence by articulating more clearly (Smiljanić & Bradlow, 2009). Thus, any future

740   study aiming to use a more ecologically valid approach than the current study should therefore ensure

741   to control for possible confounds.

742   Fourth, even though noise vocoded-speech is a useful model to study adaptation to

743   degradations of the speech signal in normal-hearing listeners, noise-vocoded speech is not a perfect

744   approximation of the type of speech signal experienced by someone with a cochlear implant. For

745   instance, due to the way the electrodes are placed on the auditory nerve, the transformed speech

746   signal will also be pitch-shifted (Rosen et al., 1999). In addition, as the number of frequency bands

747   decrease, the amount of fine-grained spectral information decreases accordingly (Shannon et al.,

748   1995). In addition, noise-vocoding the speech signal does not adequately simulate the representation

749   of phonetic-acoustic cues in a real cochlear implant. For example, depending on the specific

750   configuration of the vocoder (e.g., carrier filter widths), normal hearing people listening to vocoded

751   speech may rely more on formant transitions for differentiating pairs of syllables, whereas cochlear

752   implant users are more inclined to benefit from spectral tilt when performing the same task (Winn &

753   Litovsky, 2015). Moreover, speech perception in normal hearing (vocoded) and cochlear implant

754   listeners differs when the spectral degradation is convolved with additional degradation in the input

755   signal, e.g., when the speech is accented. While the speech recognition performance is better in CI

756   over NH (vocoded) listeners when listening to unaccented speech, this pattern of performance

757   reverses when speech is accented (Tinnemore et al., 2020). Future studies could therefore aim to

758   address some of these issues by combining noise-vocoding with pitch shifting, to establish how

759   listeners perceive and adapt to a more direct representation of the percept likely experienced by

760   people with a cochlear implant.

761   *Conclusion*

762   The results from our study demonstrate that it is essential to be able to see a speaker's moving mouth

763   while trying to understand noise-vocoded sentences in an online setup. When the mouth was not

764   visible, participants could not understand the noise-vocoded sentences at all. Our prediction that

765   being able to see the speaker's mouth movements would benefit observers when listening to noise-

766   vocoded speech, compared to situations when these movements were not visible was confirmed.

767   However, our prediction that participants who were forced to focus more on the mouth rather than

768   the whole face would perform better, was not confirmed as participants in both AV Full and AV Mouth

769   conditions performed equally well. In addition, it also appears that being able to see the eyes region,

770   but not the mouth region, does not support speech perception or adaptation. Moreover, from our

771   results it can also be concluded that, while we partially replicate the lab-based results from Banks et

772   al., it should not be assumed that lab-based tasks will necessarily replicate in online designs, especially

773   when these tasks are particularly difficult. Finally, even though this study was conducted with normal-

774   hearing listeners, we expect that our results may have implications for those with hearing loss,

775   especially when communicating in adverse listening conditions. Our work has demonstrated the key

776   role of the availability of the mouth region when background noise is present or the speech signal is

777   degraded. We recommend to always ensure that listeners are able to observe the speaker's moving

778   mouth to optimise intelligibility and reduce perceived listening effort.

779

783

784   **References**

31

785 Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension.

786   *Psychological Science*, *21*(12), 1903–1909. https://doi.org/10.1177/0956797610389192 [doi]

787 Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst:

788   An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407.

789   https://doi.org/10.3758/s13428-019-01237-x

790 Banks, B., Gowen, E., Munro, K., & Adank, P. (2015a). Audiovisual cues benefit recognition of accented

791   speech in noise but not perceptual adaptation. *Frontiers in Human Neuroscience*, *9*(422), 1–

792   13. https://doi.org/10.3389/fnhum.2015.00422

793 Banks, B., Gowen, E., Munro, K., & Adank, P. (2015b). Cognitive predictors of perceptual adaptation

794   to accented speech. *Journal of the Acoustical Society of America*, *137*(4), 2015–2024.

795   https://doi.org/10.1121/1.4916265

796 Banks, B., Gowen, E., Munro, K., & Adank, P. (2020). Individual differences in eye gaze during

797   perceptual   adaptation   to   audiovisual   noise-vocoded   speech.   *PsyArXiv*.

798   https://doi.org/10.31234/osf.io/65849

799 Bernstein, L. E., Auer, E. T., Eberhardt, S. P., & Jiang, J. (2013). Auditory perceptual learning for speech

800   perception can be enhanced by audiovisual training. *Frontiers in Neuroscience*, *7*(34).

801   https://doi.org/10.3389/fnins.2013.00034

802 Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer (Version 5.3.05)*.

803 Brown, V. A., Mclaughlin, D. J., Strand, J. F., & Engen, K. J. V. (2020). Rapid adaptation to fully intelligible

804   nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental*

805   *Psychology*, *73*(9), 1431–1443. https://doi.org/10.1177/1747021820916726

806 Buchan, J. N., Paré, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face

807   processing. *Social Neuroscience*, *2*(1), 1–13. https://doi.org/10.1080/17470910601043644

808 Buchan, J. N., Paré, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening

809   conditions on gaze behavior during audiovisual speech perception. *Brain Research*, *25*, 162–

810   171. https://doi.org/10.1016/j.brainres.2008.06.083

811 Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A. G., Taylor, K., & McGettigan, C. (2005). Lexical

812   information drives perceptual learning of distorted speech: Evidence from the comprehension

813   of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241.

814   https://doi.org/10.1037/0096-3445.134.2.222

815 Dorman, M. F., Loizou, P. L., & Rainey, D. (1997). Speech intelligibility as a function of the number of

816   channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal*

817   *of the Acoustical Society of America*, *102*(4), 2403–2411. https://doi.org/10.1121/1.419603

818 Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures

819   and visible speech to degraded speech comprehension. *Journal of Speech, Language, and*

820   *Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101

821 Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*,

822   *40*(4), 481–492. https://doi.org/10.1044/jshd.4004.481

823 Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information

824   on the intelligibility of four-channel vo- coded speech: Implications for cochlear implants.

825   *Journal   of   the   Acoustical   Society   of   America*,   *108*,   1877–1887.

826   https://doi.org/10.1121/1.1310667

827 Fox, J., Weisberg, S., Friendly, M., Hong, J., Anderson, R., Firth, D., & Taylor, S. (2019). *Effects Package:*

828   *Effect Displays for Linear, Generalized Linear, and Other Models:. R package version 4.1-4.*

829 Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later.

830   *The   Journal   of   the   Acoustical   Society   of   America*,   *87*(6),   2592–2605.

831   https://doi.org/10.1121/1.399052

832 Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of

833   noise vocoded words: Effects of feedback and lexicality. *J Exp Psychol Hum Percept Perform*,

834   *34*(2), 460–474. https://doi.org/10.1037/0096-1523.34.2.460

835 Huyck, J. J., & Johnsrude, I. S. (2012). Rapid perceptual learning of noise-vocoded speech requires

836   attention. *The Journal of the Acoustical Society of America*, *131*(3), EL236-42.

837   https://doi.org/10.1121/1.3685511

838 IEEE. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on*

839   *Audio and Electroacoustics*, *AU-17*, 225–246.

840 Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes

841   factors. *The Journal of Problem Solving*, *7*(7), 2. https://doi.org/10.7771/1932-6246.1167

842 Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., & Kobayashi, T. (2009). Bimodal Audio-Visual

843   Training Enhances Auditory Adaptation Process. *NeuroReport*, *20*(14), 1231–1234.

844   https://doi.org/10.1097/WNR.0b013e32832fbef8

845 Kennedy-Higgins, D., Devlin, J. T., & Adank, P. (2020). Investigating the cognitive mechanisms

846   underpinning successful perception of different speech distortions. *The Journal of the*

847   *Acoustical Society of America*, *147*(4), 2728–2740. https://doi.org/10.1121/10.0001160

848 Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and

849   visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, *65*(4),

850   536–552. https://doi.org/10.3758/BF03194581

851 Lansing, Charissa R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic

852   visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, *42*(3),

853   526–539. https://doi.org/10.1044/jslhr.4203.526

854 Llamas, C., Harrison, P., Donnelly, D., & Watt, D. (2009). Effects of different types of face coverings on

855   speech acoustics and intelligibility. *York Papers in Linguistics*, *Series 2*.

856 MacLeod, A., & Summerfield, A. Q. (1987). Quantifying the contribution of vision to speech perception

857   in noise. *British Journal of Audiology*, *21*(October), 131–141.

858   https://doi.org/10.3109/03005368709077786

859   McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). *Pupillometry reveals*

860   *changes in physiological arousal during a sustained listening task*. *54*(2), 193–203.

861   https://doi.org/doi.org/10.1111/psyp.12772

862   McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014).

863   Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology

864   Cognition in Hearing Special Interest Group 'white paper. *International Journal of Audiology*,

865   *53*(7), 433–440. https://doi.org/10.3109/14992027.2014.890296

866   Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for

867   audiovisual speech perception. *Perception & Psychophysics*, *66*(4), 574–583.

868   https://doi.org/10.3758/BF03194902

869   Munhall, Kevin G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication.

870   *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*,

871   123–139.

872   Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech

873   perception: The influence of ocular fixations on the McGurk effect. *Perception &*

874   *Psychophysics*, *65*(4), 553–567. https://doi.org/10.3758/BF03194582

875   Paulus, M., Hazan, V., & Adank, P. (2020). The relationship between talker acoustics, intelligibility and

876   effort in degraded listening conditions. *Journal of the Acoustical Society of America*, *147*.

877   https://doi.org/10.1121/10.0001212

878   Peelle, J. E., & Wingfield, A. (2005). Dissociations in perceptual learning revealed by adult age

879   differences in adaptation to time-compressed speech. *Journal of Experimental Psychology:*

880   *Human Perception and Performance*, *31*(6), 1315–1330. https://doi.org/10.1037/0096-

881   1523.31.6.1315

882   Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U.,

883   Lunner, T., Matthen, M., Mack- ersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M.,

884   Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing Impairment and Cognitive

885   Energy: The Framework for Under- standing Effortful Listening (FUEL). *Ear and Hearing*, *37*,

886   5S-27S. https://doi.org/10.1097/AUD.0000000000000312

887   Pilling, M., & Thomas, S. (2011). Audiovisual Cues and Perceptual Learning of Spectrally Distorted

888   Speech. *Language and Speech*, *54*(4), 487–497. https://doi.org/10.1177/0023830911404958

889   Preminger, J. E., Lin, H.-B., Payen, M., & Levitt, H. (1998). Selective visual masking in speechreading.

890   *Journal of Speech, Language, and Hearing Research*, *41*(3), 564–575.

891   https://doi.org/10.1044/jslhr.4103.564

892   Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–

893   163. https://doi.org/10.2307/271063

894   Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts

895   of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America*,

896   *106*(6), 3629–3636. https://doi.org/10.1121/1.428215

897   Scheinberg, J. S. (1980). Analysis of speechreading cues using an interleaved technique. *Journal of

898   Communication Disorders*, *13*(6), 489–492. https://doi.org/10.1016/0021-9924(80)90048-9

899   Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis

900   testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2),

901   322–339. https://doi.org/10.1037/met0000061

902   Sebastián-Gallés, N., Dupoux, E., Costa, A., & Mehler, J. (2000). Adaptation to time-compressed

903   speech: Phonological determinants. *Perception & Psychophysics*, *62*, 834–842.

904   https://doi.org/10.3758/BF03206926

905   Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first

906   look. *Behavioral Research*, *50*, 451–465. https://doi.org/10.3758/s13428-017-0913-7

907   Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with

908   primarily temporal cues. *Science*, *270*, 303–304. https://doi.org/

909   10.1126/science.270.5234.303

910 Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in
911 speaking style changes. *Linguistics and Language Compass*, *3*, 236–264.
912 https://doi.org/10.1111/j.1749-818X.2008.00112.x

913 Sohoglu, E., Pelle, J. E., Carlyon, R. P., & Davis, M. H. (2014). Top-Down Influences of Written Text on
914 Perceived Clarity of Degraded Speech. *Journal of Experimental Psychology: Human Perception*
915 *and Performance*, *40*(1), 186–199. https://doi.org/10.1037/a0033206

916 Sumby, W. H., & Pollack, I. (1954). Visual contribution of speech intelligibility in noise. *Journal of the*
917 *Acoustical Society of America*, *26*, 212–215. https://doi.org/10.1121/1.1907309

918 Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual
919 and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception*
920 *and Performance*, *30*(5), 873. https://doi.org/10.1037/0096-1523.30.5.873

921 Tinnemore, A. R., Gordon-Salant, S., & Goupell, M. J. (2020). Audiovisual Speech Recognition With a
922 Cochlear Implant and Increased Perceptual and Cognitive Demands. *Trends in Hearing*, *24*,
923 2331216520960601. https://doi.org/10.1177/2331216520960601

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

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

930 Winn, M. B., & Litovsky, R. Y. (2015). Using speech sounds to test functional spectral resolution in
931 listeners with cochlear implants. *The Journal of the Acoustical Society of America*, *137*(3),
932 1430–1442. https://doi.org/10.1121/1.4908308

933 Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-
934 based auditory experiments. *Attention, Perception & Psychophysics*, *79*(7), 2064–2072.
935 https://doi.org/10.3758/s13414-017-1361-2

936   Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial

937   behavior. *Speech Communication*, *26*(1–2), 23–43. https://doi.org/10.1016/S0167-

938   6393(98)00048-X

939

940   **Figures**

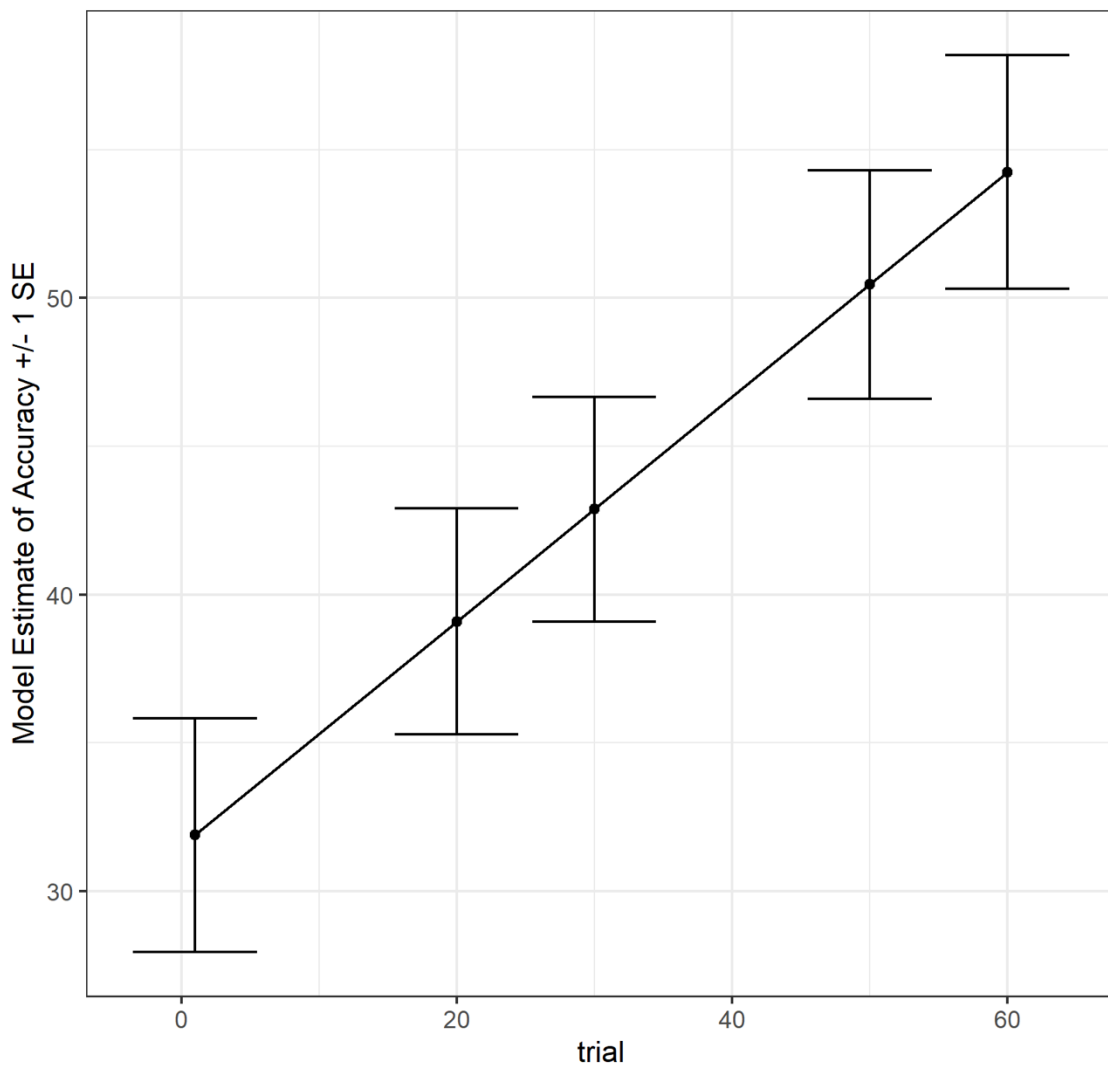AV Full   AV Mouth   AV Blocked   AV Still   AV Eyes

941

942   *Figure 1. Still images from conditions AV Full, AV Mouth, AV Eyes, AV Blocked, and AV Still.*
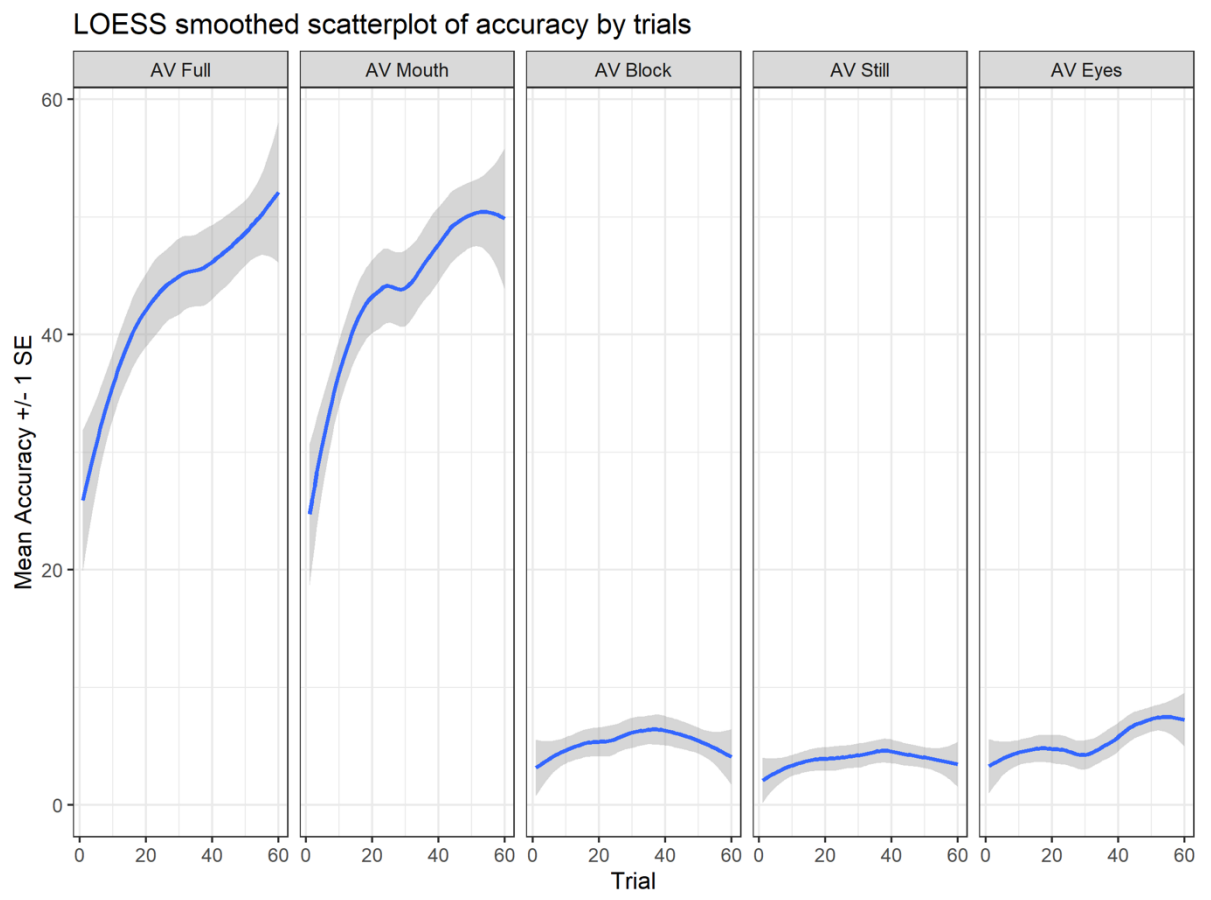
AV Full effect plot

943

Figure 2. Model estimates of percentage correctly reported key words across trials in condition AV Full,
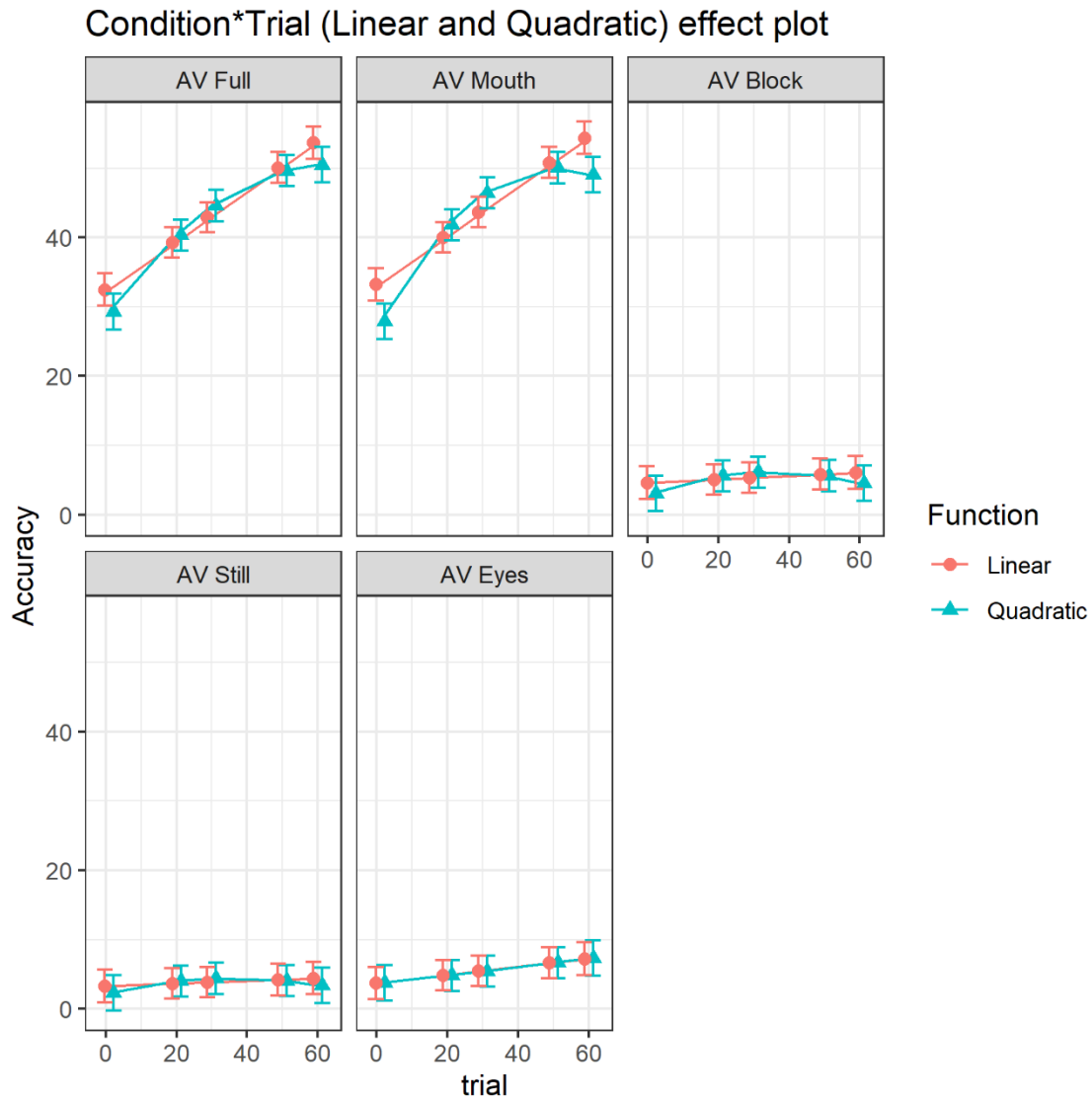
error bars represent one standard error.

946

LOESS smoothed scatterplot of accuracy by trials

947

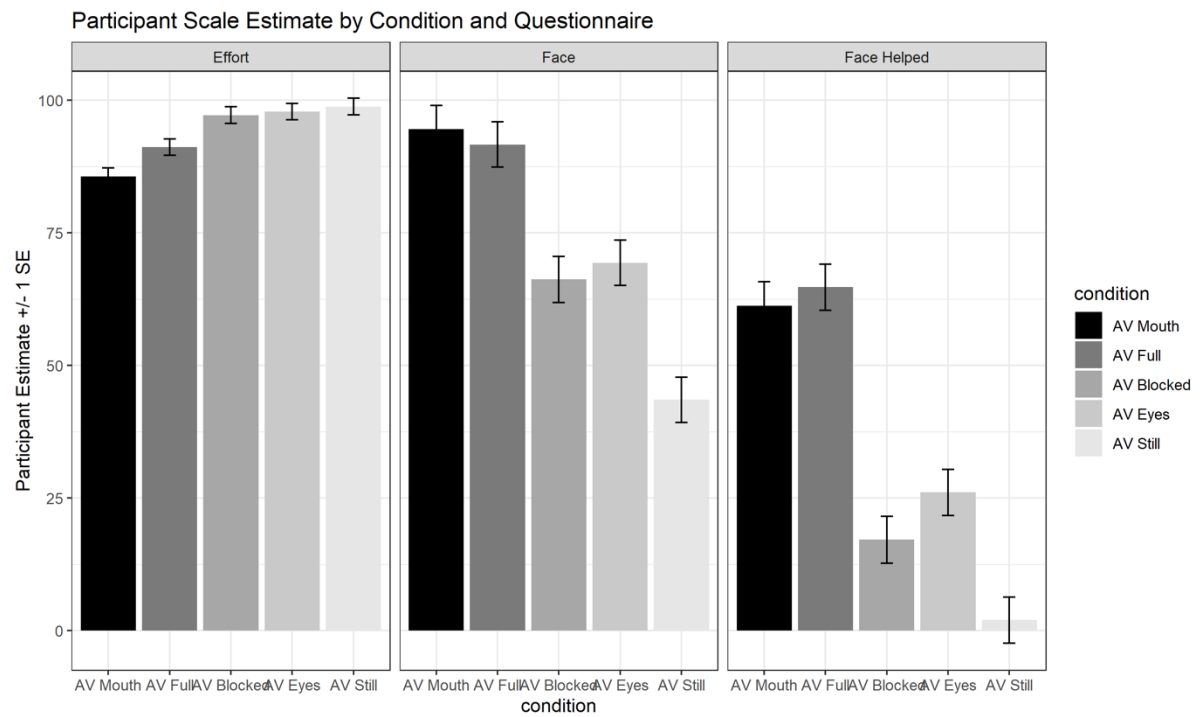*Figure 3. Locally Estimated Scatterplot Smoothing (LOESS) plot showing mean accuracy across individual trials all five conditions, borders represent one standard error.*

950

Figure 4. Model estimates of percentage correctly reported key words across trials – modelled as a

linear and quadratic relationship - in all conditions, error bars represent one standard error.

Figure 5. Participant estimates of perceived effort, time spent looking at the speaker's face (Face), and whether the face being visible helped participants during the task (Face Helped). Error bars indicate one standard error.