Face viewing behavior predicts multisensory gain during speech perception

Johannes Rennig, Kira Wegner-Clemens & Michael S. Beauchamp

Psychonomic Bulletin & Review

ISSN 1069-9384
Volume 27
Number 1

Psychon Bull Rev (2020) 27:70-77
DOI 10.3758/s13423-019-01665-y

Springer

Psychonomic Bulletin & Review (2020) 27:70–77
https://doi.org/10.3758/s13423-019-01665-y

BRIEF REPORT

Face viewing behavior predicts multisensory gain during speech perception

Johannes Rennig[1] · Kira Wegner-Clemens[1] · Michael S. Beauchamp[1]

Published online: 16 December 2019

Abstract

Visual information from the face of an interlocutor complements auditory information from their voice, enhancing intelligibility. However, there are large individual differences in the ability to comprehend noisy audiovisual speech. Another axis of individual variability is the extent to which humans fixate the mouth or the eyes of a viewed face. We speculated that across a lifetime of face viewing, individuals who prefer to fixate the mouth of a viewed face might accumulate stronger associations between visual and auditory speech, resulting in improved comprehension of noisy audiovisual speech. To test this idea, we assessed interindividual variability in two tasks. Participants (n = 102) varied greatly in their ability to understand noisy audiovisual sentences (accuracy from 2–58%) and in the time they spent fixating the mouth of a talker enunciating clear audiovisual syllables (3–98% of total time). These two variables were positively correlated: a 10% increase in time spent fixating the mouth equated to a 5.6% increase in multisensory gain. This finding demonstrates an unexpected link, mediated by histories of visual exposure, between two fundamental human abilities: processing faces and understanding speech.

Keywords Audiovisual · Face · Multisensory · Speech perception · Eye tracking

Introduction

Two important human skills are the ability to extract visual information from a viewed face and the ability to extract auditory information from a heard voice. These abilities converge in audiovisual speech perception, when information from the face and the voice of a talker are integrated to improve perception. The independent source of information about speech content provided by the talker's mouth movements is especially useful under circumstances in which the auditory signal is degraded, as in a noisy environment (Bernstein & Liebenthal, 2014; Hickok & Poeppel, 2007). While the ability of visual speech to enhance the intelligibility of noisy auditory speech is well documented (Grant et al. 1998; Sumby & Pollack 1954; for a review, see Peelle & Sommers 2015), published studies report high interindividual variability across all tested stimulus types, including consonants, words, and meaningful or anomalous sentences (Grant et al., 1998; Sommers, Tye-Murray, & Spehar, 2005; Van Engen, Phelps, Smiljanic, & Chandrasekaran, 2014; Van Engen, Xie, & Chandrasekaran, 2017); across all types of auditory noise (Sommers et al., 2005; Tye-Murray, Spehar, Myerson, Hale, & Sommers, 2016; Van Engen et al., 2014, 2017); and across all populations, including young and old adults (Sommers et al., 2005; Tye-Murray et al., 2016). In every study, some participants show a small benefit for visual speech while others show a large benefit.

Another axis of individual variability is found in the eye movements made by humans viewing faces. As first described by Yarbus (1967), individuals viewing identical images make very different eye movements. Recent work has extended this finding to individual differences in face viewing. A preference to fixate the mouth or eye region of the face is found for both static and dynamic faces, is consistent across face exemplars (Gurler, Doyle, Walker, Magnotti, & Beauchamp, 2015; Mehoudar, Arizpe, Baker, & Yovel, 2014), and is stable across testing sessions as long as 18 months apart (Mehoudar et al., 2014). Yarbus (1967) was also the first to document the sensitivity of eye-movement behavior to task demands. During a task requiring recognition of noisy audiovisual speech, participants primarily fixate the mouth of the talker, reflecting the

✉ Michael S. Beauchamp
Michael.Beauchamp@bcm.edu

[1] Department of Neurosurgery and Core for Advanced MRI, Baylor College of Medicine, 1 Baylor Plaza Suite S104, Houston, TX 77030, USA

increased behavioral relevance of visual speech under these circumstances (Buchan, Paré, & Munhall, 2008; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). The contributions of interindividual and inter-task differences to eye-movement behavior have been integrated using Bayesian ideal observer models (Peterson & Eckstein, 2012, 2013).

Since humans typically accumulate thousands of face-viewing hours during development, idiosyncratic preferences to fixate the mouth or eyes might lead to increased experience and expertise for the most-viewed face part. Individuals who prefer to fixate the mouth of the face could accumulate greater expertise in decoding the visual speech information present in talkers' mouth movements and realize a greater benefit of visual speech when attempting to understand noisy audiovisual speech.

To test this idea, participants performed two independent tasks within the same testing session. The first task was easy: discriminating audiovisual speech syllables without any added auditory noise. Since this task can be done using the auditory modality alone, it does not place any task constraints on face viewing behavior and has been previously demonstrated to elicit high levels of individual variability (Gurler et al., 2015). The second task was difficult: repeating sentences with high levels of auditory noise added, rendering visual speech information from the talker's mouth is very important for perception. Then, we compared the face-viewing behavior measured in the first experiment with the noisy speech perception measured in the second experiment to determine if the two axes of variability – individual differences in face looking and individual differences in noisy audiovisual speech perception – were linked.

Methods

Participants, stimuli and tasks

One hundred and two native English speakers (66 female, mean age 21 years, range 18–45) provided written informed consent under an experimental protocol approved by the Committee for the Protection of Human Participants of the Baylor College of Medicine, Houston, TX, USA. The sample size was selected because a previous study using similar methods found a correlation between eye movements and behavior of $r = 0.34$ (Gurler et al., 2015). To detect this effect size with 95% power requires a sample size of 102 according to G*Power (version 3.1, http://www.gpower.hhu.de; test family = *t*-tests; statistical test = point biserial correlation model; type of power analysis = *a priori* power analysis for sample size). The data in this study were compiled from two different groups of participants. Group 1 consisted of 33 participants and group 2 consisted of 69 participants, for a total *n* of 102; there were small experimental differences, described below, between the two groups.

Participants' eye movements were monitored using an infrared eye tracker (Eye Link 1000 Plus, SR Research Ltd., Ottawa, Ontario, Canada) as they viewed recordings of audiovisual speech presented using Matlab (Mathworks, Inc., Natick, MA, USA) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). Visual speech was presented on a high-resolution screen (Display++ LCD Monitor, 32-in., 1,920 × 1,080, 120 Hz, Cambridge Research Systems, Rochester, UK). Auditory speech was presented through speakers on either side of the screen at a constant sound pressure level of 60 dB, a value chosen to approximate the level of human speech. In order to ensure high-quality eye tracking, participants' heads were immobilized using a chin rest placed 90 cm from the display.

The task 1 stimuli consisted of 2-s audiovisual recordings of clear syllables (no added noise). Each trial began with a fixation crosshair presented outside of the location of where the face would appear in order to simulate natural viewing conditions in which faces rarely appear at the center of gaze (Gurler et al., 2015). As soon as the audiovisual speech video began playing, the fixation crosshairs disappeared, and participants were free to fixate anywhere on the screen. After the speech video ended, participants reported the identity of the syllable with a button press. Syllables were identified with 98% accuracy (SD: 5%). Group 1 participants viewed 190 randomly interleaved clear-syllable trials (divided into two runs): 20 repetitions × four talkers × two audiovisual syllables ("ba," "ga") and 10 repetitions × one talker × three audiovisual congruent syllables ("ba," "da," "ga"). Group 2 participants viewed 240 randomly interleaved clear-syllable trials (divided into four runs): 20 repetitions × four talkers × three audiovisual syllables ("ba," "da," "ga").

The task 2 stimuli consisted of sentences recorded from a single male talker combined with auditory pink noise at a signal-to-noise ratio (SNR) of -16 dB, as used in a previous study (Van Engen et al., 2017). Pink noise, defined as noise with decreasing energy at increasing frequency, was added as it is commonly used in studies of auditory function as it is less aversive than white noise (Mesgarani, Cheung, Johnson, & Chang, 2014; Van Engen et al., 2017). The sentences were presented either alone (noisy auditory-only, A) or paired with a video recording (noisy auditory + visual, AV). Each trial began with a fixation crosshair presented outside of the location of where the face would appear in audiovisual trials. During auditory-only trials, the fixation crosshair shifted to the center of the screen when auditory playback began. In audiovisual trials, the fixation crosshair disappeared when the video began. After the sentence ended (3 s duration) participants repeated the sentence. Responses were scored for number of correct keywords (e.g., "The **hot sun warmed** the **ground**," keywords in bold). For each participant, the total number of keywords recognized was divided by the total number of keywords to generate a "percentage words recognized" score for the auditory-only and audiovisual conditions.

Each participant was presented with 80 sentence trials, consisting of randomly interleaved presentations of 40 auditory-only and 40 audiovisual sentences. For the group 1 participants, the sentences were not counterbalanced, so that one group of 40 sentences was heard only in auditory-only form and one group of 40 sentences was heard only in audiovisual form. For the group 2 participants, stimulus counterbalancing was used. To determine if the stimulus counterbalancing affected the main result of a relationship between mouth fixation time and multisensory gain, we constructed a regression model with multisensory gain as the dependent variable, mouth fixation time as the independent variable, and a categorical variable for subject group. The model showed a significant main effect for percent mouth fixation time ($F = 9.10$, $p = 0.003$) but not for subject group ($F = 1.03$, $p = 0.312$) or the interaction ($F = 1.24$, $p = 0.269$; $r^2 = 0.01$). Therefore, both groups were combined for the analyses reported in this paper.

Data and analysis code are available for download at https://osf.io/qm7tj. Two kinds of audiovisual speech were presented: syllables presented *without* any added auditory noise (in task 1) and sentences presented *with* added auditory noise (in task 2).

Eye tracking and analysis

Eye tracking was performed with a sampling rate of 500 Hz. Before each run of each task, a 9-target array was presented for eye-tracker calibration and validation. Three times within each run, participants fixated a centrally presented crosshair. The difference between the measured eye position during these epochs and the screen center was applied to correct the eye-tracking data in the preceding stimulus epoch. Two regions of interest (ROIs) were defined for each video, consisting of an eye ROI and a mouth ROI (Fig. 1A). Blinks and saccades were excluded from the analysis and the percentage of fixation time spent within each ROI was calculated. Our primary eye-tracking measure was the percentage of fixation time spent on the mouth (percent mouth looking) that was collected in the first task (presentation of clear audiovisual syllables).

Bayesian model comparison

To understand the relationship between the experimental variables, models were constructed that consisted of different combinations of the measured variables and their interactions. Bayesian model comparison was used to determine the model that best predicted performance of individual participants. Overfitting with excess parameters was penalized using the Bayesian information criterion (BIC) using the *BIC* function as implemented in R. This function calculated a log-likelihood for every model using the number of predictors and observations in each model. Then, the log-likelihoods for all models were ranked relative to the most likely model in order to evaluate their relative performance.

Results

In the first task, participants identified the syllable spoken by an audiovisual talker. Performance was at ceiling (mean of 98% accuracy, SD 1%), similar to a previous study in which auditory-only versions of the same stimuli were presented (97% accuracy in Mallick et al., 2015). The high accuracy for clear audiovisual and auditory-only syllables demonstrates that distinguishing a restricted set of syllables is an easy task that does not require visual speech information. In contrast to the uniformly high perceptual accuracy for the speech identification task, there was substantial variability in eye-fixation behavior. Some participants spent as little as 3% of total fixation time fixating the mouth of the talker, while others spent as much as 98%, with a mean of 49% (SD: 24%; Fig. 1A).

In the second task, participants listened to noisy auditory sentences presented with a fixation crosshair (auditory-only) or paired with a video of the talker's face (audiovisual) (Fig. 1B). A high level of auditory noise was used (SNR: -16 dB), with the result that in the auditory-only condition participants recognized only a few words (mean 9%, SD: 4%, range: 1–20%). In the audiovisual condition, participants recognized significantly more words (mean 38%, SD: 14%, range: 4–74%; paired t-test $t_{(101)} = 23.58$, $p = < 2.2 \times 10^{-16}$, $d = 1.63$), demonstrating the benefit of visual speech in enhancing the intelligibility of noisy auditory speech (Fig. 1B).

Every single participant showed improved performance between the auditory-only and audiovisual speech conditions. The amount of improvement, referred to as multisensory gain (calculated as % words recognized during audiovisual noisy speech – % words recognized during auditory-only noisy speech) quantifies the benefit provided by viewing the talker's face. The mean multisensory gain was 29% but there was a high degree of variability. Some participants improved as little as 2% while others improved as much as 58% (SD: 13%; Fig. 1B).

We observed high interparticipant variability in two measures: eye movements during viewing of clear syllables in task 1 (% mouth fixation time, Fig. 1A) and multisensory gain for noisy audiovisual speech in task 2 (% multisensory gain, Fig. 1B). To determine if there was a relationship between these measures, we plotted them against each other (Fig. 1C). Participants who spent more time fixating the mouth during viewing of clear syllables had higher multisensory gain (Fig. 1C; r = 0.288, p =

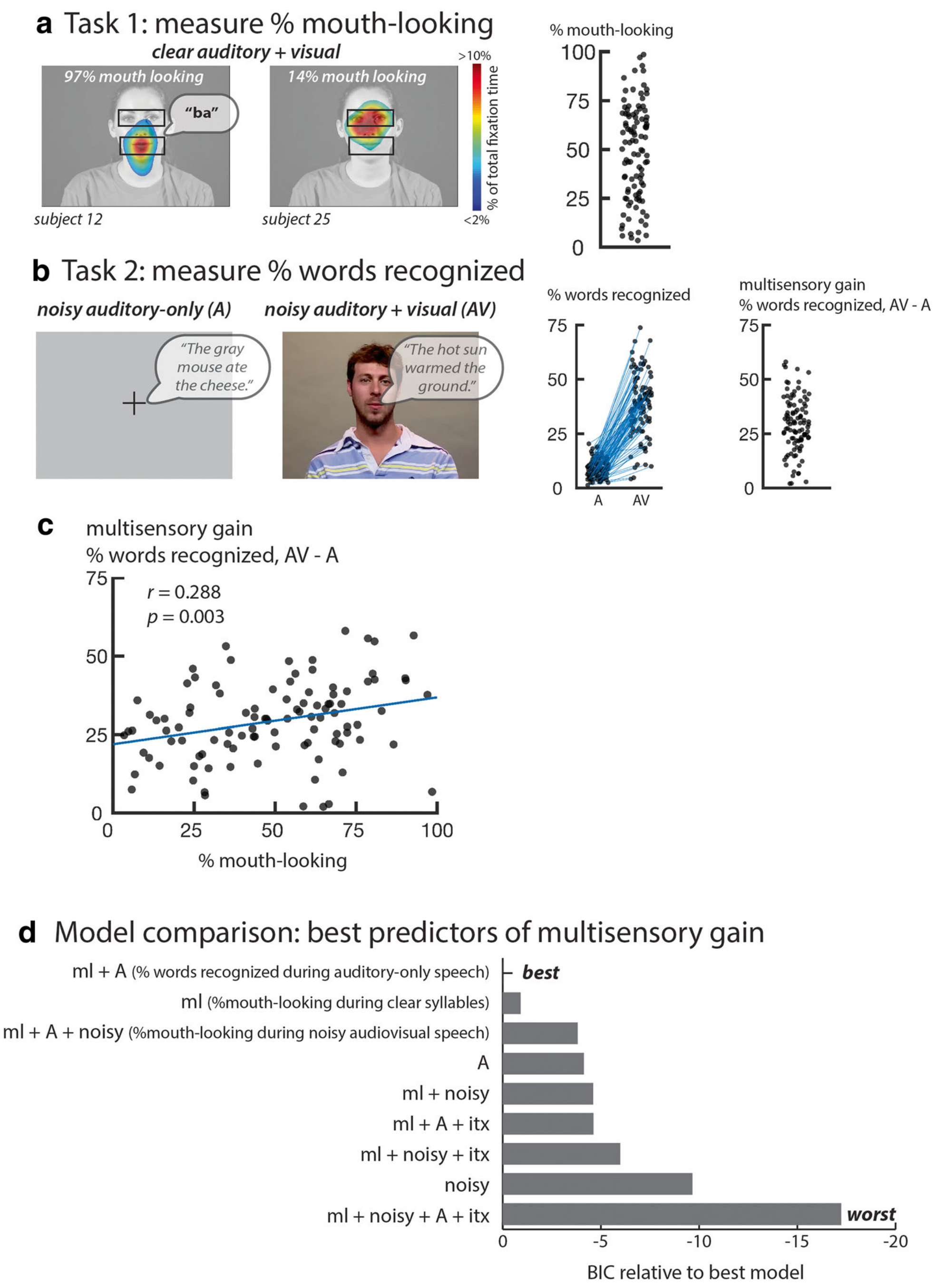

Fig. 1 (**A**) In the first task, eye movements were measured during face viewing. Colors (overlaid on a still frame from the stimulus video) show the time spent fixating each location in the display as a percentage of total fixation time for two sample participants. For each participant, the fixation maps were converted into a single value, the percentage of time spent fixating the mouth region of interest (ROI; lower black box, not present in actual display, the upper box illustrates the complementary eye ROI). The speech bubble shows the auditory component of the stimulus. The plot at right shows this value for all participants, one symbol per participant. (**B**) In the second task, participants reported the words in noisy sentences presented without a video of the talker's face (auditory-only, A) or with the video (audiovisual, AV). The speech bubble shows the auditory component of the stimulus. The left scatter plot shows the percentage of key words recognized as a fraction of total words in each condition (two symbols for each participant connected by a blue line). The right scatter plot shows multisensory gain, calculated as the difference between the two conditions (one symbol per participant). (**C**) Correlation across participants between mouth-looking time (plot in **A**) and multisensory gain (right plot in **B**), with one symbol per participant. (**D**) In order to determine the factors influencing multisensory gain, models were constructed that included all measured experimental variables. The models were compared using the Bayesian Information Criterion (BIC), which adjusts the total variance explained by the number of predictors, penalizing overfitting. Models are ordered from best to worst

0.003). The regression slope was $m = 0.56$: for each 10% extra time a participant spent observing the mouth during clear syllables, their multisensory gain increased by 5.6%.

Mouth-looking behavior during presentation of clear syllables might simply reflect mouth looking during noisy speech perception. This explanation was not supported by the data: during viewing of noisy sentences, nearly all participants primarily fixated the mouth (Fig. 2A; mean of 92% of total fixation time, SD: 14%). Reflecting this lack of variability, there was no significant correlation between the percent total time spent fixating the mouth during noisy sentences and multisensory gain (Fig. 2B; $r = 0.032$, $p = 0.747$).

To test different combinations of the explanatory factors, we constructed nine different models that used all of our measured variables and their interactions to predict multisensory gain. As shown in Fig. 1D, the best model predicted multisensory gain using two factors: mouth-looking during clear syllables and words recognized during auditory-only speech. To compare the importance of these two factors, we examined the performance of models that used each factor in isolation. The model that used only the mouth-looking variable was the second-best model overall and was four times more likely to explain the observed data than the model using only the auditory speech variable.

In a control analysis, we measured the percent of total fixation time that each participant spent fixating the eyes of the talker during presentation of clear syllables. Since the mouth and eyes are the primary fixation targets in viewed faces, there was a strong negative correlation between the time spent fixating the eyes and the time spent fixating the mouth ($r = -0.78$, $p = < 2.6 \times 10^{-16}$). This resulted in a significant negative correlation between time spent fixating the eyes and multisensory gain during noisy speech perception ($r = -0.22$, $p = 0.028$), the opposite of the positive correlation between mouth fixations and multisensory gain.

Discussion

We hypothesized that individual differences in face viewing behavior would influence a person's ability to make use of visual speech information. Individuals with a preference for fixating the mouth of a talking face might accumulate more expertise in the association between specific mouth movements and their associated vocalizations, resulting in improved comprehension of noisy audiovisual speech. We found this to be true: participants who spent more time fixating the mouth when it was not necessary (during presentation of clear audiovisual speech) made better use of the visual speech information when it was necessary (during presentation of noisy audiovisual speech).

Yarbus (1967) first demonstrated that presentation of the same visual scene can evoke very different eye movements. This variability has two components. The task component results from humans modifying their eye-movement behavior based on task demands. For instance, when asked to identify joy, an emotion that is primarily represented in the mouth region of the face, humans are more likely to fixate the mouth (Schurgin et al., 2014). Similarly, when perceiving noisy auditory speech, humans are more likely to fixate the mouth than when perceiving clear auditory speech, because the noisy speech task benefits from visual speech information (Buchan et al., 2008; Vatikiotis-Bateson et al., 1998). We replicated this finding in our study, with the average time spent fixating the mouth increasing from 49% during the clear speech task to 92% in the noisy speech task.

The inter-individual component of eye-movement variability is less well understood. Recent studies have shown that different individuals have idiosyncratic preferences in how they view faces (Gurler et al., 2015; Kanan, Bseiso, Ray, Hsiao, & Cottrell, 2015; Mehoudar et al., 2014; Peterson & Eckstein, 2012, 2013). Even when performing the identical task, some individuals prefer to fixate the mouth of the talker while others fixate the eyes, a preference that is unchanged

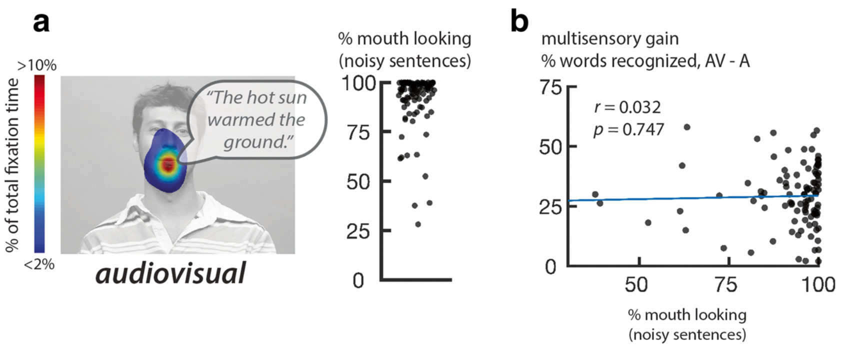

Fig. 2 (**A**) Eye movement measurements from the second task, perception of noisy audiovisual sentences. Colors (overlaid on a still frame from the stimulus video) show the time spent fixating each location in the display as a percentage of total fixation time, averaged across participants (speech bubble shows auditory stimulus). Plot shows the percentage of time fixating the mouth region of the face, one symbol per participant. (**B**) Correlation between eye-movement behavior in task 2 (from Fig. 2A) and multisensory gain (from Fig. 1B)

when tested up to 18 months apart (Mehoudar et al., 2014; Peterson & Eckstein, 2012, 2013). Interindividual differences in face preference have been shown for both static faces (Mehoudar et al., 2014; Peterson & Eckstein, 2012, 2013) and dynamic talking faces (Gurler et al., 2015), and the contributions of interindividual and inter-task differences to eye-movement behavior have been integrated using Bayesian ideal-observer models (Peterson & Eckstein, 2012, 2013). In our first task, we measured face-viewing behavior using an easy task: recognizing clear speech in a quiet room. Since this task can be performed with near perfect accuracy even without visual speech (Mallick et al., 2015), face-viewing behavior was driven by participants' internal preferences rather than by task demands. Consistent with previous studies, we observed substantial interindividual variability, with mouth fixation time ranging from 3% to 99% of total fixation time.

Another poorly understood axis of individual variability is the perceptual benefit provided by viewing a talker's face. While the large benefit of seeing the face in understanding noisy speech is incontrovertible (Grant et al. 1998; Sumby & Pollack 1954; for a review, see Peelle & Sommers 2015), there is no explanation in the literature for the high interindividual variability in this benefit that is observed across all published experiments (Grant et al., 1998; Sommers et al., 2005; Tye-Murray et al., 2016; Van Engen et al., 2014, 2017). Consistent with these reports, we observed large individual variability, with audiovisual gain ranging from 2% to 58% across our 102 participants.

Possible explanations for the observed correlation

We observed a significant correlation between individual differences in face viewing and noisy audiovisual speech perception. This finding could not be explained by eye movements made during noisy audiovisual speech perception, as noisy speech induces all participants to fixate the talker's mouth. Instead, participants who fixated the mouth when it was not important (during the first task with clear audiovisual syllables) received more benefit from fixating the mouth when it was important (during the second task with noisy speech). Conversely, even fixating the mouth when it was important (during the noisy speech task) provided less benefit to participants who did not fixate the mouth when it was not important.

Our working explanation for this correlation is that a lifetime of viewing faces provides a large training set on which to learn the correspondence between visual and auditory speech features. Visual attention is thought to serve as a gating mechanism for perceptual learning (Ahissar & Hochstein, 2004; Roelfsema, van Ooyen, & Watanabe, 2010). Attention and fixation location are tightly linked. Since the retina has a region of very high acuity in the fovea, fixating a specific region of the face provides higher quality information about that feature. Fixating (and attending to) the talker's mouth may greatly enhance the learned association between visual and auditory speech features, allowing for enhanced perception of noisy audiovisual speech (greater audiovisual gain).

The discovery of a link between eye movements and speech perception could be useful for treating patients with an impaired ability to understand noisy speech. While it may not be possible to improve the auditory sensory signal from the cochlea, eye movements are under cognitive control and hence are presumably amenable to training programs.

Other ramifications

We found that a predilection for fixating the mouth correlated with better noisy audiovisual speech perception. It is likely that this relationship extends, in both positive and negative directions, to other tasks besides speech perception. For instance, viewing some emotional facial expressions, such as joy, drive fixations to the mouth, while others, such as anger, drive fixations to the eyes (Schurgin et al., 2014). We might expect that participants who preferentially fixate the mouth might be better at detecting joy, while eye-lookers might be better at detecting anger. In a different domain, information about facial identity is most available in the eye region of the face; mouth-lookers might be better at noisy speech perception at the cost of performing worse at facial recognition than eye-lookers.

It is important to point out that individual differences in face-viewing preferences explains only a small amount of the variance in noisy speech perception (10%). Both face-viewing and speech perception are complex cognitive processes and it would be surprising indeed if there was a perfect mapping between them. Understanding noisy audiovisual speech involves many computations, including extracting visual speech information from the mouth region of the talker's face (lipreading or speechreading); extracting auditory speech information from a noisy background; combining the two sources of speech information using multisensory integration; and comparing the results of integration with prior knowledge about the likelihood of particular speech content. Our study does not allow us to determine which of these computations is most influenced by face-viewing preferences.

Neural substrates

An intriguing question for future research concerns the neural substrates of the observed correlation (Bernstein & Liebenthal, 2014; Hickok & Poeppel, 2007). The posterior superior temporal cortex contains anatomically distinct regions that are more responsive to viewing eye movements or viewing mouth movements (Pelphrey, Morris, Michelich, Allison, & McCarthy, 2005; Zhu &

Beauchamp, 2017), and during free viewing of a talking face, these regions are differentially active, depending on whether the observer fixates the eye or mouth region of the talking face (Rennig & Beauchamp, 2018). The mouth-preferring regions of the posterior temporal cortex responds strongly to auditory voices, while the eye-preferring regions does not (Rennig & Beauchamp, 2018; Zhu & Beauchamp, 2017). Mouth-looking experience might strengthen synaptic connections between neurons in the mouth-preferring regions of posterior temporal cortex representing seen mouth movements and neurons representing heard speech sounds. Supporting this idea, Hickok et al. (2018) examined stroke patients and found a strong correlation between damage to the superior temporal gyrus and decreased susceptibility to the McGurk effect, an index of audiovisual speech integration.

Acknowledgements This work was supported by the National Institutes of Health (R01NS065395 to M.S.B) and the Deutsche Forschungsgemeinschaft (RE 3693/1-1 to J.R.). We acknowledge the Core for Advanced MRI at Baylor College of Medicine. We thank Bharath Chandrasekaran for providing the stimuli for the noisy sentence task.

Author contributions MSB and JR developed the study concept and study design. Data collection was performed by KWC. JR and KWC conducted data analysis. MSB and JR wrote the paper.

Compliance with ethical standards

Conflict of interest The authors declare no competing financial interests or other conflicts of interest.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464. https://doi.org/10.1016/j.tics.2004.08.011

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, *8*(DEC), 386. https://doi.org/10.3389/fnins.2014.00386

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436. https://doi.org/10.1163/156856897X00357

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

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*(5 Pt 1), 2677–2690. https://doi.org/10.1121/1.422788

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

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hickok, G., Rogalsky, C., Matchin, W., Basilakos, A., Cai, J., Pillay, S., … Fridriksson, J. (2018). Neural networks supporting audiovisual integration for speech: A large-scale lesion study. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *103*, 360–371. https://doi.org/10.1016/j.cortex.2018.03.030

Kanan, C., Bseiso, D. N. F., Ray, N. A., Hsiao, J. H., & Cottrell, G. W. (2015). Humans have idiosyncratic and task-specific scanpaths for judging faces. *Vision Research*, *108*, 67–76. https://doi.org/10.1016/j.visres.2015.01.013

Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, *22*(5), 1299–1307. https://doi.org/10.3758/s13423-015-0817-4

Mehoudar, E., Arizpe, J., Baker, C. I., & Yovel, G. (2014). Faces in the eye of the beholder: unique and stable eye scanning patterns of individual observers. *Journal of Vision*, *14*(7), 6. https://doi.org/10.1167/14.7.6

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science (New York, N.Y.)*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision, 10*, 437–442. https://doi.org/10.1163/156856897X00366

Pelphrey, K. A., Morris, J. P., Michelich, C. R., Allison, T., & McCarthy, G. (2005). Functional anatomy of biological motion perception in posterior temporal cortex: An fMRI study of eye, mouth and hand movements. *Cerebral Cortex*, *15*(12), 1866–1876. https://doi.org/10.1093/cercor/bhi064

Peterson, M. F., & Eckstein, M. P. (2012). Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(48), E3314-23. https://doi.org/10.1073/pnas.1214269109

Peterson, M. F., & Eckstein, M. P. (2013). Individual differences in eye movements during face identification reflect observer-specific optimal points of fixation. *Psychological Science*, *24*(7), 1216–1225. https://doi.org/10.1177/0956797612471684

Rennig, J., & Beauchamp, M. S. (2018). Free viewing of talking faces reveals mouth and eye preferring regions of the human superior temporal sulcus. *NeuroImage*, *183*(January), 25–36. https://doi.org/10.1016/j.neuroimage.2018.08.008

Roelfsema, P. R., van Ooyen, A., & Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attention. *Trends in Cognitive Sciences*, *14*(2), 64–71. https://doi.org/10.1016/j.tics.2009.11.005

Schurgin, M. W., Nelson, J., Iida, S., Ohira, H., Chiao, J. Y., & Franconeri, S. L. (2014). Eye movements during emotion recognition in faces. *Journal of Vision*, *14*(13), 14. https://doi.org/10.1167/14.13.14

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing*, *26*(3), 263–275. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15937408

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

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, *31*(4), 380–389. https://doi.org/10.1037/pag0000094

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: interactions among context, modality, speech style, and masker. *Journal of Speech, Language, and Hearing Research: JSLHR*, *57*(5), 1908–1918. https://doi.org/10.1044/JSLHR-H-13-0076

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception & Psychophysics*, *79*(2), 396–403. https://doi.org/10.3758/s13414-016-1238-9

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

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York: Plenum Press.

Zhu, L. L., & Beauchamp, M. S. (2017). Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(10), 2697–2708. https://doi.org/10.1523/JNEUROSCI.2914-16.2017

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.