HHS Public Access

Author manuscript

Ear Hear. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

Ear Hear. 2020 ; 41(2): 268–277. doi:10.1097/AUD.0000000000000755.

Psychobiological responses reveal audiovisual noise differentially challenges speech recognition

Gavin M. Bidelman[1,2,3,†], **Bonnie Brown**[1], **Kelsey Mankel**[1,2], **Caitlin Nelms Price**[1,2]

[1]School of Communication Sciences & Disorders, University of Memphis, Memphis, TN, USA

[2]Institute for Intelligent Systems, University of Memphis, Memphis, TN, USA

[3]University of Tennessee Health Sciences Center, Department of Anatomy and Neurobiology, Memphis, TN, USA

Abstract

Objective—In noisy environments, listeners benefit from both hearing and seeing a talker, demonstrating audiovisual (AV) cues enhance speech-in-noise (SIN) recognition. Here, we examined the relative contribution of auditory and visual cues to SIN perception and the strategies used by listeners to decipher speech in noise interference(s).

Design—Normal-hearing listeners (n=22) performed an open-set speech recognition task while viewing audiovisual TIMIT sentences presented under different combinations of signal degradation including visual (AV_n), audio (A_nV), or multimodal (A_nV_n) noise. Acoustic and visual noise were matched in physical signal-to-noise ratio. Eyetracking monitored participants' gaze to different parts of a talker's face during SIN perception.

Results—As expected, behavioral performance for clean sentence recognition was better for A-only and AV compared to V-only speech. Similarly, with noise in the auditory channel (A_nV and A_nV_n speech), performance was aided by the addition of visual cues of the talker regardless of whether the visual channel contained noise, confirming a multimodal benefit to SIN recognition. The addition of visual noise (AV_n) obscuring the talker's face had little effect on speech recognition by itself. Listeners' eye gaze fixations were biased towards the eyes (decreased at the mouth) whenever the auditory channel was compromised. Fixating on the eyes was negatively associated with SIN recognition performance. Eye gazes on the mouth vs. eyes of the face also depended on the gender of the talker.

Conclusions—Collectively, results suggest listeners (i) depend heavily on the auditory over visual channel when seeing and hearing speech and (ii) alter their visual strategy from viewing the mouth to viewing the eyes of a talker with signal degradations which negatively affects speech perception.

Keywords

audiovisual speech perception; eyetracking; speech-in-noise (SIN); listening strategy

[†]**Address for editorial correspondence:** Gavin M. Bidelman, PhD, School of Communication Sciences & Disorders, University of Memphis, 4055 North Park Loop, Memphis, TN, 38152, TEL: (901) 678-5826, FAX: (901) 525-1282, gmbdlman@memphis.edu.

1. INTRODUCTION

Successful communication requires more than favorable audibility. Speech perception is a multisensory experience. In naturalistic conversation, listeners benefit from both hearing and seeing a talker (Erber 1975; Lalonde et al. 2016; Sumby et al. 1954), demonstrating audiovisual (AV) cues enhance perceptual processing. AV enhancements are particularly salient for speech-in-noise (SIN) perception (MacLeod et al. 1987; Sumby and Pollack 1954; Vatikiotis-Bateson et al. 1998; Xie et al. 2014). Indeed, previous behavioral studies show an average ~10–15 dB improvement in speech recognition threshold from the addition of visual cues to speech especially in challenging listening conditions (MacLeod and Summerfield 1987). Similarly, SIN perception is improved by lip-reading (Bernstein et al. 2004; Navarra et al. 2007), and tracking visual movements augments second language perception by way of multisensory integration (Navarra and Soto-Faraco 2007). It is clear from behavioral studies that combining multisensory (AV) cues represents an important way the brain overcomes noise and facilitates the perception of degraded speech.

While the effects of acoustic noise on speech intelligibility are well established, how impoverished *visual* information influences SIN processing is not well documented (cf. Atcherson et al. 2017; Galatas et al. 2011; Legault et al. 2010). This is important given that sensory declines are often comorbid across the lifespan, as is the case with concomitant hearing and visual impairments (e.g., wearing hearing aids and corrective lenses) (Brennan et al. 2005). Indeed, reduced visual acuity (e.g., blurred vision, 6/60 acuity) tends to exacerbate SIN perception in older adults (Legault et al. 2010). However, even within reduced visual information, speech recognition is still better with visual cues than performance with the auditory modality alone (Legault et al. 2010). This suggests that sight helps to enhance speech understanding even when visual acuity is suboptimal as might be the case with visual impairments or a poor connection during videotelephony (e.g., Skype, Apple FaceTime). Given that cross-modal influences between vision and audition are bidirectional, (Bidelman 2016; Bidelman et al. in press; Bidelman et al. 2019; Lippert et al. 2007; Maddox et al. 2014; McGurk et al. 1976), the first aim of this study was to investigate how noise in different modalities (i.e., the acoustic vs. visual channel) impact spoken word recognition.

There is growing evidence to suggest that eyetracking measures might offer an important objective proxy of clear and degraded spoken word recognition, dynamic lexical analysis, and audiovisual processing (e.g., Ben-David et al. 2011; Tanenhaus et al. 1995). For example, studies demonstrate that while listeners' gaze concentration is directed to the eyes and mouth of a talker they alter their fixations as they learn word boundaries and become more familiar with a speaker (Lusk et al. 2016). Gaze velocity is also greater for eye movements to visual vs. auditory targets, revealing modality-specific influences in perceptual efficiency (Goldring et al. 1996). Recent translational studies further show that in hearing aid patients, eye gaze "steering" toward relevant cues of a target talker enhances speech intelligibility (Favre-Félix et al. 2018). Testing visual components of degraded speech recognition might also offer ways to assess modality-aspecific cognitive skills that are difficult to disentangle in conventional (auditory) assessments of SIN perception (e.g., Zekveld et al. 2007). Relevant to our questions on "decoding" the physiological mechanisms

of cocktail party listening (Bidelman et al. 2016; Bidelman et al. 2017), we used eye tracking techniques here to investigate speech processing online and reveal covert (perhaps even unconscious) listening strategies (e.g., Ben-David et al. 2011) that are not captured by self-reports or behavioral measures alone (cf. Wendt et al. 2016). This is important since listeners presumably compensate in certain (degraded) listening conditions to achieve comparable levels of behavioral performance while employing very different task strategies (Broadbent 1958; Goldring et al. 1996; Hick et al. 2002).

To this end, we measured eye gaze fixations during speech recognition tasks to investigate different online perceptual and gaze strategies listeners' use to cope with different forms of noise interference to the speech signal (cf. Ben-David et al. 2011). We compared the relative impact of *auditory* and *visual* noise presented at comparable signal-to-noise-ratio (SNRs). This allowed us to compare different gaze strategies that listeners use when auditory vs. visual interference are matched in overall *physical clarity*. While visual cues improve recognition, they also place higher demands on the resources needed for speech processing (Gosselin et al. 2011). Thus, we reasoned that audiovisual noise would compound such effects and potentially manifest in different patterns of gaze fixation on the talker depending on the noise characteristics.

Listeners performed an open-set sentence recognition task while viewing AV TIMIT sentences (Harte et al. 2015) presented in different combinations of *multimodal* speech degradation including auditory and/or visual noise. We also compared these AV conditions to unimodal conditions where sentences contained only sound or visual cues along with clear (no noise) conditions. Eyetracking monitored participants' gaze to different parts of the talker's face during SIN perception and noise-related changes with degradations to the visual and/or auditory sensory modality. We included both male and female talkers in light of evidence that females are more intelligible than males (Bradlow et al. 1996) and studies showing that the gender of a talker (actor) being observed can influence gaze patterns on the face (Coutrot et al. 2016). This experimental design allowed us to assess how acoustic vs. visual noise stressors impact behavioral performance, and more importantly, how listeners might adapt different covert perceptual strategies depending which modality contained degraded speech cues and the gender of the talker. We predicted that if listeners differentially weight the auditory and visual modality during SIN perception (e.g., Hirst et al. 2018), behavioral performance would differ for visual vs. auditory noise, even when matched in physical SNR. We also hypothesized that changes in perceptual performance would be accompanied by different looking patterns on a talker's face and that gaze distributions might depend on which modality contained noise and/or gender of the talker (Coutrot et al. 2016).

2. MATERIALS & METHODS

2.1 Participants

Twenty-two young adults [age (*mean±SD*): 24.9±2.6 years; 15 female; 7 male] participated in the experiment. All were native speakers of English, had normal hearing based on audiometric testing (i.e., thresholds ≤ 25 dB HL; 250–8000 Hz), a similar level of education (18.7±2.2 years), and reported no previous history of neuropsychiatric illnesses. All but one

participant was right handed (73.6±41.5% laterality) based on the Edinburgh Handedness Survey (Oldfield 1971). Vision was not formally screened, which is a limitation of this study. However, all participants self-reported normal or corrected-to-normal vision and, where necessary, were allowed to wear corrective lenses in the form of contacts. All confirmed the screen and visual stimuli were clearly visible during a post-experiment debrief. On average, participants had 4.2±4.5 years of formal musical training. Each gave written informed consent in compliance with a protocol approved by the IRB of the University of Memphis.

2.2 Audiovisual speech stimuli

Stimuli were AV sentences from the TCD-TIMIT database (https://sigmedia.tcd.ie/TCDTIMIT/) (Harte and Gillen 2015). The TCD-TIMIT consists of high-quality audio and video footage of 62 speakers reading a total of 6913 phonetically rich sentences based on the original TIMIT audio corpus (Garofolo et al. 1993). Video footage was originally captured using a Sony PMW-EX3s camera (MPEG-2 format) with an angle of 0° azimuth to the talker's face, zoomed such that the frame contained only the actor's head, shoulders and a green screen in the shot (for details, see Harte and Gillen 2015)[1].

We selected 60 unique sentences from among two talkers of the TCD-TIMIT [30 from one male (Talker #19) and 30 from one female (Talker #11); random selection]. Average clip duration was 4.7 ± 0.87 sec. This same sentence list was presented in seven blocks (counterbalanced), each containing different combinations of acoustic or visual noise degradation (Fig. 1). In addition to a clean AV condition, multimodal sentences were presented in degraded conditions containing visual (AV_n), audio (A_nV), or multimodal (A_nV_n) noise interference. Throughout, we use the notation A_nV_n where capital letters denote which modality is present and the subscript (*n*) denotes the inclusion of noise to that modality. Additionally, auditory only (A), visual only (V), and unimodal auditory plus noise (A_n) conditions were included to assess how each modality by itself impacts speech recognition. A unimodal visual noise condition was not included since clean V-only stimuli elicited floor performance even before the addition of noise (see Fig. 2).

The acoustic noise was a multi-talker noise babble adopted from the QuickSIN (cf. Killion et al. 2004), which contains one male and three female talkers. Babble onset/offset was gated with the sentence audio. The visual interference was an image of white noise (static commonly used in image and video processing, e.g., TV salt and pepper) that was overlaid onto the video track using FFmpeg (http://ffmpeg.org/)[2]. Sound and video signal-to-noise ratios (SNRs) were 6 dB and 6.8–7.8 dB SNR, respectively depending on the exact calculation of image SNR[3]. This ensured that the physical SNR was equated between the

[1]Actors in the TCD-TIMIT database were consented on the reuse of their video images for academic purposes. The consent form can be found in the original thesis (Gillen 2014), available at https://sigmedia.tcd.ie/TCDTIMIT/.

[2]The choice of "TV static" was due principally to practical constraints on what types of noise FFmpeg could render onto our videos. Blurring was another option but we did not explore this mode of degradation since visual blurs are typically defined based on percentages (Zekveld et al. 2007) and there would be no explicit way to control "SNR."

[3]Calculating SNR of a video is nontrivial given the time-varying nature of images on the screen and two-dimensional nature of the signal image (i.e., x-y pixel values). We adopted the general definition of SNR for images, computed as SNR = 10log(μ/σ), where μ is the signal mean (e.g., the male talker's image) and σ is the standard deviation of the noise (i.e., the static) (González et al, 2008). The

auditory and visual noise. Pilot testing confirmed a nominal 6 dB sound and image SNR avoided ceiling performance yet partially masked the sound and visual channel.

2.3 Procedure

Listeners were seated in a double-walled sound attenuating chamber (Industrial Acoustics, Inc.) ~90 cm from a computer monitor. Stimulus delivery and response data collection were achieved using VLC media player (www.videolan.org) controlled by MATLAB 2013b (The MathWorks, Inc.), respectively. AV stimuli were displayed at the center of the screen on a black background, subtending a 6.3° visual angle (Samsung SyncMaster S24B350HL; nominal 75 Hz refresh rate). The auditory channel was presented binaurally using high-fidelity circumaural headphones (Sennheiser HD 280 Pro) at a comfortable level (75 dB SPL). On each trial, participants watched and/or heard a single sentence produced by the male or female talker. Between trials, a fixation cross-hair (+) centered on the screen was presented to center participants' gaze prior to viewing each video. Following each AV sentence presentation, a black screen appeared in which participants provided a typed, open-set response via computer keyboard. They were encouraged to respond as accurately as possible, recalling as many keywords as they could remember. Presentation order of the different conditions was counterbalanced across participants according to a Latin square sequence to control possible learning/repetition effects (Bradley 1958)[4]. Breaks were provided between blocks to avoid fatigue. During the speech recognition task, continuous eye gaze locations were recorded using an eyetracker (detailed in Section 2.4).

Behavioral data were scored on a percent correct basis. Keywords (those carrying meaning) were preselected from each of the 60 sentences (average= 5.1±1.6 keywords/sentence). Two raters scored the percentage of keywords participants correctly recalled in their typed response. Common typographical errors and misspellings were generally accepted as deemed appropriate by the raters assuming the intended target word was apparent. Inter-rater reliability confirmed highly consistent ratings (Pearson's-r = 0.99, p<0.0001), Hence, the two scores were averaged across raters. Scores were computed separately for the male and female talkers, as well as the pooled list. This resulted in a total of 21 speech perception scores per participant (7 noise conditions x 3 talker types; see Fig. 2). Percent correct scores were transformed via a rationalized arcsine-transform (RAU) to account for possible ceiling/floor effects (Studebaker 1985).

2.4 Eyetracking

Listeners' gaze fixations on the talkers' face were acquired using a Gazepoint GP3 eyetracker[5]. This device provides precise measurement of the location of ocular gaze and

MATLAB function psnr gave a similar SNR estimate of 7.8 dB. By some estimates, the 6–7 dB visual SNR used here equates to 60–70% performance accuracy in facial recognition paradigms (Meytlis et al. 2007). SNR was computed for a screen frame at the midpoint of a representative video. While video SNR is actually a time-varying (frame-by-frame) quantity, the talker's head remained fixed in position within the camera view so that the mean signal pixel values (and hence SNR) were on average, constant throughout the clip. SNR estimates for the male and female videos were within 1 dB.

[4]The effectiveness of the counterbalance in canceling out possible order/learning effects was confirmed by the (expected) non-significant effect of task order when this variable was included in our models [$F_{7,126}$=0.68, p=0.68].

[5]Though pupillometry (dilation) data were recorded, they are not analyzed in this report as the pupil dilates to a myriad f stimulus attributes unrelated to speech perception (e.g., subjective salience, novelty, and task uncertainty; Liao et al. 2016; Preuschoff et al. 2011; Wang et al. 2014).

pupil diameter with an accuracy of ~1° visual angle via an infrared, desktop mounted camera. Consequently, the IAC booths' lights remained off during the task. Continuous eye data were collected from the left and right eyes every 16.6 ms (i.e., 60 Hz sampling rate). Data from the GP3 were logged via an API interface with MATLAB. To ensure continued alignment with the screen, the eyetracker was re-calibrated before each stimulus block using the GP3's internal routine where the eyes were calibrated at 9-points across the horizontal/vertical dimensions of the screen.

Continuous eye data were recorded online while participants performed the behavioral speech recognition task. Time stamps were triggered in the data file demarcating the onset of each stimulus presentation. This allowed us to analyze time-locked changes in eye data for each stimulus (Beatty 1982; Eckstein et al. 2017). Blinks were automatically logged by the eye tracking system and epochs contaminated with these artifacts were discarded prior to analysis. X-Y coordinate positions of the eyes on the screen were recorded during the AV sentence presentations to track the dynamics of participants' gaze on the face. Fixations were first converted to a 2D heat map, representing a histogram of gaze concentration on different parts of the screen throughout the entirety of each video clip. Using the hist3 function in MATLAB, this quantified the number of times a listener gazed to each pixel with the dimensions of the screen [size= 1080 ×1920]. Ellipses drawn on the face demarcated several regions of interest (ROI) including the entire head, mouth, and eyes (Buchan et al. 2007; Lansing et al. 2003; Russo et al. 2011; Van Belle et al. 2010). The size of the ROIs were identical between the male and female talker. ROIs covered the head (spanning ~6.3° visual angle) as well as the eye (4.6° visual angle) and mouth (2.21° visual angle) regions of both talkers (see Fig. 3). We then computed the percentage of fixation counts within the eyes and mouth ROIs among the total count of gaze fixations falling within the head region. This allowed us to investigate the concentration of gaze on talkers' faces and how listeners might selectively monitor different facial features with auditory and/or visual degradation to AV speech. Gaze distributions were pooled across video tokens separately for the male and female talker, allowing us to investigate differences in participants' visual search strategy as a function of talker gender.

3. RESULTS

3.1 Behavioral data

Behavioral speech recognition performance (raw percent correct scores) is shown for the various AV noise conditions in Figure 2. A two-way, mixed-model ANOVA (7 stimulus conditions x 2 talker genders; subjects=random factor) conducted on RAU-transformed percent-correct scores revealed a main effect of gender [$F_{1, 273}$=114.54, p<0.0001; η_p^2=0.29] such that the female talker was, on average, more intelligible than the male talker. There was also a main effect of stimulus on behavioral scores [$F_{6, 273}$=329.11, p<0.0001; η_p^2=0.88]. However, more critically, we found a stimulus x gender interaction, indicating that behavioral performance depended on both the specific noise condition and gender of the talker [$F_{6, 273}$=2.55, p=0.0204; η_p^2=0.05]. Separate one-way ANOVAs conducted by talker gender confirmed a main effect of stimulus when viewing both the female [$F_{6, 126}$=128.07,

$p<0.0001$; $\eta_p^2=0.86$] and male talker [$F_{6, 126}=197.04$, $p<0.0001$; $\eta_p^2=0.90$]. Tukey-Kramer adjusted multiple comparisons revealed a nearly identical pattern of responses (in terms of which contrasts were significant) for both the male and female stimuli. As expected, speech recognition was better for AV and A-only speech compared to V-only speech (*ps*<0.0001). When listening to speech (audio channel only), performance was worse with acoustic noise ($A > A_n$, $p<0.0001$) but still superior to the V-only condition ($A_n>V$; $p<0.0001$). Additionally, auditory performance was aided by the addition of visual cues of the talker regardless of whether the visual channel contained noise (i.e., $AV>A_n$, $A_nV >A_n$, and $AV_n >A_n$; *ps*<0.01). These results confirm visual cues benefit SIN recognition, even if they are degraded by noise.

Considering next the effects of noise on AV speech perception, we found that degradation in the visual channel had no effect on recognition performance ($AV = AV_n$; $p=1.0$). However, this might be expected given the spared (clean) auditory channel in both of these conditions. Contrasting the visual modality, the addition of acoustic noise to the auditory channel hindered speech intelligibility ($AV > A_nV$; $p<0.0001$). Auditory and visual noise also reduced intelligibility relative to clean AV speech ($AV > A_nV_n$; $p<0.0001$). Lastly, A_n speech was less intelligible than A_nV_n for the female ($p=0.0034$) but not male talker ($p=0.112$), which accounts for the stimulus x gender interaction. Collectively, these results suggest that regardless of talker gender, noise in the acoustic channel had a stronger effect on speech recognition than SNR-matched visual noise masking the talker's face. That is, listeners relied more heavily on the auditory over visual input when seeing and hearing degraded speech. Additionally, the visual degradation was more deleterious for the male talker (or stated conversely, V cues were more helpful when viewing the female).

Lastly, we tested the possibility that the gender of the talker (stimulus) might interact with the gender of the *participants* to affect behavioral performance. A mixed-model ANOVA (noise stimulus x participant gender x talker gender; subjects = random factor) showed no main effects of participants' gender [$F_{1, 284}=0.20$, $p=0.66$; $\eta_p^2=0.0007$] nor interaction with talker gender [$F_{1, 284}=0.09$, $p=0.76$; $\eta_p^2=0.0003$]. This suggests that differences in behavioral speech recognition between the male and female stimuli may have been talker specific (i.e., related to the stimuli) and not due to the gender of our cohort, *per se.*

3.2 Eye tracking (gaze fixation) data

Fixations, reflecting the spatial distribution of eye gaze on the talker's face, are shown as heat maps for each AV noise condition in Figure 3[6]. Hotter colors represent more frequent fixations at a particular location on the screen. Listeners tended to shift their gaze away from the talker's mouth to their eyes in more challenging stimulus conditions (cf. Fig. 3 A vs. D). Quantitative analysis of eye fixations at each ROI are shown in Figure 4. An omnibus three-way ANOVA (stimulus x ROI x gender) revealed interactions between ROI x stimulus

[6]Fig. 3 data are from a representative subject overlaid onto a *single frame* of the male talker video. Any apparent misalignment is due to the fact that the heatmap reflects the aggregate distribution cross all trials of a given condition. While the talker's heads were fixed within the viewing frame, they were not immobile nor did the ROIs move with respect to the head. Importantly, we calibrated the eyetracker after each block to ensure continued calibration with the screen (see Methods, Sect. 2.4). The nearly identical pattern of results for the male and female talker (Fig. 4) also confirm a consistent calibration across talker conditions.

[$F_{3, 311}$=8.05, p<0.0001; η_p^2=0.072] and gender x ROI [$F_{1, 311}$=13.62, p=0.0003; η_p^2=0.042]. To parse these effects, we conducted separate two-way ANOVAs (ROI x stimulus) by gender. For the *male* talker, the ROI x stimulus interaction was significant [$F_{3,145}$=3.99, p=0.0091; η_p^2=0.076] (Fig. 4B), meaning that the location of eye gaze on the male's face depended on the specific type of AV noise. Multiple comparisons revealed fixations were more frequent at the male talker's mouth than eyes for conditions where the auditory channel remained intact [i.e., AV (p<0.0001) and AV_n (p<0.0001)]. Fixations at the eyes increased in conditions containing auditory noise (i.e., A_nV, p=0.027).

Similarly, we found a ROI x stimulus interaction for the *female* talker [$F_{3, 145}$=3.90, p=0.0102; η_p^2=0.075] (Fig. 4C). Participants made fewer fixations at the eyes than mouth when speech contained visual noise (AV_n; p=0.0001). All other stimuli produced a similar distribution of gaze fixations on the female face. Collapsing across stimuli, we also found that eye fixations were more frequent when viewing the female talker (p=0.0028) but more frequent at the mouth when viewing the male (p=0.0377). This ROI x gender interaction is shown in Figure 5.

Because mouth gazes were more frequent when participants showed better behavioral speech recognition (Fig. 2), a natural question that arises is whether this type of mouth-centric visual strategy is advantageous for perception. To address this question, we conducted correlations between mouth and eye fixation counts and listeners' (RAU-transformed) behavioral speech perception scores. Pooled across stimuli, we found that higher percentages of gaze fixations to the *mouth* were positively associated with increased behavioral recognition [r=0.20, p=0.032]. In contrast, fixations at the *eyes* were not correlated with behavioral performance [r=−0.14, p=0.09]. Significant correlations were also observed when considering each talker separately. For the male, increased fixations at the mouth were associated with better behavioral recognition [r=0.20, p=0.03] and conversely, fixations toward the eyes predicted poorer performance [r=−0.23, p=0.015] (Fig. 5B). These correlations were marginal for female talker [mouth: r=0.16, p=0.07; eye: r=−0.06, p=0.30]. These results suggest that listeners alter their visual gaze strategy from the mouth to the eyes when monitoring a talker's face, especially in noisy listening scenarios. However, this change in visual search also seems to depend on the gender of the talker (*eye fixations*: female > male; *mouth fixations*: male > female; Fig. 5A). Collectively, these findings suggest visual cues from a speaker's mouth drive successful AV speech recognition and that strategies that draw attention toward the eyes negatively affects behavioral performance.

4. DISCUSSION

By measuring behavioral recognition and eyetracking responses to acoustically and visually degraded speech, results of this study relate to two main observations: (1) listeners depend heavily on the auditory over visual channel when seeing and hearing clear and degraded speech; (2) listeners alter their visual gaze strategy from monitoring a talker's mouth to fixating on their eyes as the availability of speech cues diminish, which results in poorer speech recognition.

4.1 Noise across sensory modalities differentially challenges speech perception

Behaviorally, we found that acoustic noise corrupting the sound channel severely limited speech intelligibility. In contrast, visual noise masking a talker's face had a negligible impact on perception. These findings suggest that individuals rely more heavily on the auditory over visual input when seeing and hearing speech. Interestingly, this dominance of sound information occurred despite visual cues being generally more reliable than the corresponding acoustic information (i.e., the visual channel had a higher SNR). While vision is often assumed to dominate auditory sensation in AV processing (especially for spatial tasks; Maddox et al. 2014; McGurk and MacDonald 1976), recent studies demonstrate that sound can dominate bimodal perception even when the auditory signal is weak (e.g., contains noise), is entirely ignored, or is matched in discriminability to the visual portion of the signal (Burr et al. 2009; Ortega et al. 2014). Similarly, under conditions where visual cues are deemed unreliable (e.g., noise, sensory impairments), sound can trump vision to maintain robust perception (Alais et al. 2004; Myers et al. 2017; Narinesingh et al. 2015). Reaction times are also faster in response to auditory vs. visual stimuli (Shelton et al. 2010). Our data are consistent with these latter studies and suggest that audition dominates AV speech recognition under noise conditions affecting sight and sound.

We found that the combined effect of visual and auditory noise had a deleterious effect on speech processing compared to clean AV speech. EEG studies have demonstrated that earlier sensory components of the auditory evoked potentials (N1-P2) peak earlier to AV compared to A-only speech, indicating early audiovisual interaction in the time course of brain activity (Alsius et al. 2014; van Wassenhove et al. 2005). However, this temporal facilitation is reduced when attention is loaded, suggesting interactions between audition and vision depend on proper deployment of attentional resources (Alsius et al. 2014). It is possible that the poorer behavioral performance (Fig. 2) we observed in speech especially with a degraded auditory channel ($A_nV \approx A_nV_n > A_n$) reflects a similar form of over arousal and/or attentional disengagement, either of which would lower one's intensity of cognitive processing in the task (Eckstein et al. 2017; Murphy et al. 2011; Zekveld et al. 2014). This notion is supported by the findings of Zekveld and Kramer (2014) where participants reported that they often gave up listening at low intelligibility levels (i.e., poor SNRs) and also had smaller pupil responses (a physiological marker of attentional engagement) in these conditions.

4.2 Eye gaze location on the face is differentially modulated by noise

Eye fixation data revealed listeners altered their visual gaze strategy by changing how they selectively monitored facial features during SIN perception. Importantly, gaze patterns differed despite relatively similar behavioral performance across AV conditions (cf. Figs. 2 vs. 3). For clean speech, it is thought that listeners spend more time monitoring a speaker's mouth than eyes (as seen here), likely to better segment the incoming speech signal (Lusk and Mitchel 2016). With degraded AV cues, we found that gaze shifted from the mouth to the eyes when monitoring a talker's face. Moreover, this move to the eyes was negatively associated with declines in behavioral performance; increased looking at the eyes in more challenging conditions was paralleled by poorer behavioral SIN recognition (Fig. 5B). This indicates that while listeners track landmark features of the face during perception, even a

modest increase in task difficulty affects the spatial distribution of gaze on the face (Lansing and McConkie 2003). Perceivers' focus of attention (gaze concentration) is drawn away from the mouth to the eyes under AV degradations (probably inadvertently) and this negatively affects speech recognition (cf. Lansing and McConkie 2003). Stated differently, difficulty in SIN perception is directly associated with the amount of time a perceiver's gaze is directed toward a talker's mouth (see also Lansing and McConkie 2003; Lusk and Mitchel 2016; Vatikiotis-Bateson et al. 1998). This effect was also evident in our correlational analyses, which showed that increased fixations at the mouth (but not eyes) were associated with improved speech perception. Interestingly, previous studies have shown that gaze moves from the mouth towards the eyes with increasing audibility when viewing a singer (Russo et al. 2011)—opposite the pattern observed for here for SIN. This suggests that where visual gaze is drawn on the face depends on the specific stimulus context and domain of information conveyed by the face (e.g., music vs. speech).

Previous studies have noted that the potential improvement in speech comprehension from integrating a speaker's visual cues with their sound utterance tends to be larger when information from the auditory modality is unreliable as might be the case for unfamiliar (e.g., nonnative or accented speech; Banks et al. 2015) or unpredictable speech (Maguinness et al. 2011). Evidence for this proposition also stems from bilinguals, who tend to show stronger perceptual binding of AV cues (Bidelman and Heath in press; Bidelman and Heath 2019) but poorer SIN perception in their second (less familiar) language (Bidelman et al. 2015; Reetzke et al. 2016; Rogers et al. 2006; Xie et al. 2014). Under this hypothesis, when speech is masked and becomes unreliable and/or unpredictable, individuals may disregard signals at the talker's mouth in favor of a broader visual pursuit of other relevant facial cues. Indeed, eye-gaze patterns on the face have been shown to change with stimulus uncertainty (Van Belle et al. 2010), and similar noise-induced gaze shifts to those observed here have been observed in previous eyetracking studies (Buchan et al. 2007; Van Belle et al. 2010). For instance, Buchan et al. (2007) showed that the inclusion of acoustic noise to AV speech caused listeners to focus their gaze more centrally on the face, perhaps to maximize the amount of visual information from a talker.

4.3 Study limitations

Although gender of our participants did not affect behavioral performance, we did find that the *talker's* gender modulated recognition. On average, the female was more intelligible than the male speaker (Fig. 2), consistent with previous reports (Bradlow et al. 1996). The female talker also elicited a different pattern of gaze fixations on the face (Fig. 5). Previous studies have shown that the gender of the participant (gazer) and the person being observed (actor) influence gaze patterns and face exploration (Coutrot et al. 2016). Female gazers tend to spend more time looking at the eyes of female talkers (Coutrot et al. 2016). Our results parallel these findings. Participants in our sample (2:1 females to males) fixated more at the eyes when viewing the female talker but spent more time looking at the mouth when viewing the males. It is tempting to suggest that these effects reflect some aspect of social psychology, e.g., viewing faces of the opposite gender in terms of sexual and social selection (Little et al. 2011; Scott et al. 2014). Adults do to tend to focus on the eyes to glean social cues (Lewkowicz et al. 2012). There is also some indication that females are better at

utilizing visual cues than males (Watson et al. 1996). However, given that male and female participants of our sample did not differ in behavioral performance, our data likely reflect a talker (stimulus) effect, rather than interactions between the gender of the gazer and actor, *per se* (cf. Coutrot et al. 2016). An imbalance in sample gender may have also contributed to this null effect.

We did not attempt to control for differences in intrinsic linguistic/emotional variability when selecting the male vs. female sentences. Thus, one explanation of the observed talker gender effect is due to paralinguistic cues (Pisoni 1993). For example, the female' productions may have been perceived as more "clear speech", less variable, or more expressive than the male talker which could increase intelligibility (Bradlow et al. 2002; Bradlow et al. 1996). However, we find this explanation unlikely given that the TCD-TIMIT videos were recorded with neutral affect with very little variation in extrinsic emotional/ expressive content. Still, it remains to be seen if the aforementioned eyetracking effects generalize to a larger, more diverse set of male and female stimuli beyond the pair used here. Presumably, speech intelligibility and/or eye gaze could interact in a gender-specific manner, depending on the relative sex of two interlocutors (e.g., Lansing and McConkie 2003).

Additionally, while we attempted to match the SNR of visual and auditory noise (~6–7 dB), similar auditory and visual SNRs does not imply that stimuli were equated in their *perceptual* severity. A full titration of task difficulty to balance A and V noise levels for perceptual equivalency is non-trivial and would require an extensive psychophysical mapping study beyond the scope of this report. Still, performance matching would be critical in studies attempting to equate listening effort across modalities. This type of manipulation would be of interest for future research as it could reveal whether perceptually-matched noise in the auditory vs. visual modality evokes different degrees of perceptual effort and/or visual search strategies. Additionally, even with corrected vision, subtle variations in visual acuity could affect perception (Jordan et al. 2011), particularly under the greater demands of noise. Visual acuity problems (and contrast sensitivity, motion perception, etc.) are more problematic for older adults (Daffner et al. 2013; Legault et al. 2010) so limiting our sample to younger participants helped control this potential variability. Still, we argue that our effects likely represent an underestimate of actual AV benefits since this additional noise would tend to weaken (rather than drive) observed effects. At the very least, our behavioral data suggest that auditory and visual interferences are not strictly additive in terms of their effects of speech perception but instead show a complex interaction.

4.4 Future directions

Interestingly, certain clinical disorders (e.g., autism) are associated with the opposite pattern of gaze fixations observed here, i.e., increased fixations on the mouth vs. eyes (Klin et al. 2002). Modern classrooms are also inherently noisy environments (Knecht et al. 2002). Thus, in addition to understanding speech and figure-ground perception in clinical populations, it would be interesting to investigate the influence of AV noise in educational settings and how visual gaze might promote or deny learning in the classroom. While the negative impact of acoustic noise on classroom learning is perhaps self-evident—and important enough to require architectural standards (ANSI/ASA 2010)—task-irrelevant

sounds can also interfere with reading and verbal recall (reviewed by Klatte et al. 2013). Consequently, understanding how different forms of noise interference affect the use of multisensory cues might be important to optimize learning the bustling classroom. Similarly, our correlational findings suggest that a decrement in performance in noise might be related to listeners inherently shifting their gaze from the mouth to the eyes (Fig. 4). Thus, another logical question stemming from these data is whether training listeners to override these tendencies and gaze at the mouth even in noise might help increase SIN recognition.

Other studies have suggested children undergo a shift in sensory dominance in early childhood (Hirst et al. 2018). For example, children show reduced susceptibility to the audiovisual McGurk illusion before age 10 (Hirst et al. 2018), suggesting a dominance of the auditory compared to visual sense. Visual influences on speech perception, as indexed by the McGurk effect, assume adult-like levels by 10 years of age (Hirst et al. 2018). A similar time course from auditory to visual dominance was reported by Tremblay et al. (2007). Interestingly, it has been suggested that this developmental increase in visual cue influence on heard speech is due to an increase in gazes to the mouth of a speaker that occur between ages 5 and 10 (Irwin et al. 2017). In adults, we similarly find a heavier reliance on cues from the mouth which critically, change as a function of noise. Thus, an interesting avenue for future research would be to extend the present study and investigate differences in visual search strategies and sensory dominance for degraded AV speech from a developmental standpoint.

Sensory dominance is also modulated by the reliability of the visual and auditory input (present study; Hirst et al. 2018) and can be reweighted (to the unimpaired modality) in individuals with visual (Myers et al. 2017; Narinesingh et al. 2015) or hearing-based (Schorr et al. 2005) deficits. Thus, natural gaze patterns like the ones used here might provide an objective assay to monitor rehabilitative interventions for auditory or visual impairments in cases where behavioral benefits do not reveal improved speech-understanding scores (e.g., Sheffield et al. 2018). Additionally, in hearing aid patients, eye gaze “steering” toward the relevant cues of a target talker can enhance speech intelligibly (Favre-Félix et al. 2018), This suggest eye gaze patterns might be important to incorporate into auditory rehabilitation and assistive hearing technologies.

5. CONCLUSIONS

Our data suggest human listeners depend more heavily on the auditory over visual channel when seeing and hearing speech. The fact that degradations to sound have a more egregious effect on speech recognition than visual interferences suggests that auditory information dominates noise-degraded speech perception. To cope with AV noise, listeners alter their visual strategy when monitoring a talker’s face, shifting their gaze patterns from the mouth to the eyes as the signal becomes progressively more challenging at the expense of behavioral recognition. Collectively, our findings suggest that listeners produce a differential pattern of behavioral performance and task strategies when deciphering audiovisual speech.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number NIH/NIDCD R01DC016267 (G.M.B.).

References

Alais D, Burr D (2004). The ventriloquist effect results from near-optimal bimodal integration. Current Biology, 14, 257–262. [PubMed: 14761661]

Alsius A, Möttönen R, Sams ME, et al. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. Frontiers in Psychology, 5.

ANSI/ASA. (2010). ANSI/ASA S12.60–2010/Part 1. American National Standard Acoustical Performance Criteria, Design Requirements, and Guidelines for Schools, Part 1: Permanent Schools. In.

Atcherson SR, Mendel LL, Baltimore WJ, et al. (2017). The Effect of Conventional and Transparent Surgical Masks on Speech Understanding in Individuals with and without Hearing Loss. J Am Acad Audiol, 28, 58–67. [PubMed: 28054912]

Banks B, Gowen E, Munro KJ, et al. (2015). Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. Frontiers in Human Neuroscience, 9, 422. [PubMed: 26283946]

Beatty J (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol Bull, 91, 276–292. [PubMed: 7071262]

Ben-David BM, Chambers CG, Daneman M, et al. (2011). Effects of aging and noise on real-time spoken word recognition: evidence from eye movements. Journal of Speech, Language, and Hearing Research, 54, 243–262.

Bernstein LE, Auer ET Jr, Takayanagi S (2004). Auditory speech detection in noise enhanced by lipreading. Speech Communication, 44, 5–18.

Bidelman GM (2016). Musicians have enhanced audiovisual multisensory binding: Experience-dependent effects in the double-flash illusion. Experimental Brain Research, 234, 3037–3047. [PubMed: 27334887]

Bidelman GM, Dexter L (2015). Bilinguals at the “cocktail party”: Dissociable neural activity in auditory-linguistic brain regions reveals neurobiological basis for nonnative listeners’ speech-in-noise recognition deficits. Brain and Language, 143, 32–41. [PubMed: 25747886]

Bidelman GM, Heath S (in press). Enhanced temporal binding of audiovisual information in the bilingual brain. Bilingualism: Language and Cognition, doi:10.1017/S1366728918000408, 1–11.

Bidelman GM, Heath ST (2019). Neural correlates of enhanced audiovisual processing in the bilingual brain. Neuroscience, 401, 11–20. [PubMed: 30639306]

Bidelman GM, Howell M (2016). Functional changes in inter- and intra-hemispheric auditory cortical processing underlying degraded speech perception. Neuroimage, 124, 581–590. [PubMed: 26386346]

Bidelman GM, Yellamsetty A (2017). Noise and pitch interact during the cortical segregation of concurrent speech. Hearing Research, 351, 34–44. [PubMed: 28578876]

Bradley JV (1958). Complete counterbalancing of immediate sequential effects in a Latin square design. Journal of the American Statisical Association, 53, 525–528.

Bradlow AR, Bent T (2002). The clear speech effect for non-native listeners. Journal of the Acoustical Society of America, 112, 272–284. [PubMed: 12141353]

Bradlow AR, Torretta GM, Pisoni DB (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Communication, 20, 255–272. [PubMed: 21461127]

Brennan M, Horowitz A, Su YP (2005). Dual sensory loss and its impact on everyday competence. Gerontologist, 45, 337–346. [PubMed: 15933274]

Broadbent DE (1958). Perception and communication. London: Pergamon.

Buchan JN, Pare M, Munhall KG (2007). Spatial statistics of gaze fixations during dynamic face processing. Social Neuroscience, 2, 1–13. [PubMed: 18633803]

Burr D, Banks MS, Morrone MC (2009). Auditory dominance over vision in the perception of interval duration. Experimental Brain Research, 198, 49–57. [PubMed: 19597804]

Coutrot A, Binetti N, Harrison C, et al. (2016). Face exploration dynamics differentiate men and women. Journal of Vision, 16, 16.

Daffner KR, Haring AE, Alperin BR, et al. (2013). The impact of visual acuity on age-related differences in neural markers of early visual processing. Neuroimage, 67, 127–136. [PubMed: 23153966]

Eckstein MK, Guerra-Carrillo B, Miller Singley AT, et al. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? Developmental Cognitive Neuroscience, 25, 69–91. [PubMed: 27908561]

Erber NP (1975). Auditory-visual perception of speech. Journal of Speech and Hearing Disorders, 40, 481–492. [PubMed: 1234963]

Favre-Félix A, Graversen C, Hietkamp RK, et al. (2018). Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment. Trends in Hearing, 22, 2331216518814388.

Galatas G, Potamianos P, Papangelis A, et al. (2011). Audio visual speech recognition in noisy visual environments. In PETRA 2011, Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments (pp. 1–4). Crete, Greece.

Garofolo JS, Lamel LF, Fisher WM, et al. (1993). DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM. 2015 from https://catalog.ldc.upenn.edu/LDC93S1.

Gillen E (2014). TCD-TIMIT: A new database for audio-visual speech recognition. In Department of Electronic and Electrical Engineering: Trinity College Dublin.

Goldring JE, Dorris MC, Corneil BD, et al. (1996). Combined eye-head gaze shifts to visual and auditory targets in humans. Experimental Brain Research, 111, 68–78. [PubMed: 8891638]

González RC, Woods RE (2008). Digital image processing. Prentice Hall.

Gosselin PA, Gagne JP (2011). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. International Journal of Audiology, 50, 786–792. [PubMed: 21916790]

Harte N, Gillen E (2015). TCD-TIMIT: An audio-visual corpus of continuous speech. IEEE Transactions on Multimedia, 17, 603–615.

Hick CB, Tharpe AM (2002). Listening effort and fatigue in school-age children with and without hearing loss. Journal of Speech, Language, and Hearing Research, 45, 573–584.

Hirst RJ, Stacey JE, Cragg L, et al. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. Scientific Reports, 8, 12372. [PubMed: 30120399]

Irwin JR, Brancazio L, Volpe N (2017). The development of gaze to a speaking face. Journal of the Acoustical Society of America, 141, 3145–3145. [PubMed: 28599552]

Jordan TR, McGowan VA, Paterson KB (2011). Out of sight, out of mind: the rarity of assessing and reporting participants' visual abilities when studying perception of linguistic stimuli. Perception, 40, 873–876. [PubMed: 22128559]

Killion MC, Niquette PA, Gudmundsen GI, et al. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. Journal of the Acoustical Society of America, 116, 2395–2405. [PubMed: 15532670]

Klatte M, Bergström K, Lachmann T (2013). Does noise affect learning? A short review on noise effects on cognitive performance in children. Frontiers in Psychology, 4, 578–578. [PubMed: 24009598]

Klin A, Jones W, Schultz R, et al. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. Arch Gen Psychiatry, 59, 809–816. [PubMed: 12215080]

Knecht HA, Nelson PB, Whitelaw GM, et al. (2002). Background noise levels and reverberation times in unoccupie classrooms: Predictions and measurements. American Journal of Audiology, 11, 65–71. [PubMed: 12691216]

Lalonde K, Holt RF (2016). Audiovisual speech perception development at varying levels of perceptual processing. Journal of the Acoustical Society of America, 139, 1713. [PubMed: 27106318]

Lansing CR, McConkie GW (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. Perception & Psychophysics, 65, 536–552. [PubMed: 12812277]

Legault I, Gagne JP, Rhoualem W, et al. (2010). The effects of blurred vision on auditory-visual speech perception in younger and older adults. International Journal of Audiology, 49, 904–911. [PubMed: 20874052]

Lewkowicz DJ, Hansen-Tift AM (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. Proceedings of the National Academy of Sciences of the United States of America, 109, 1431–1436. [PubMed: 22307596]

Liao H-I, Kidani S, Yoneya M, et al. (2016). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. Psychonomic Bulletin & Review, 23, 412–425. [PubMed: 26163191]

Lippert M, Logothetis NK, Kayser C (2007). Improvement of visual contrast detection by a simultaneous sound. Brain Research, 1173, 102–109. [PubMed: 17765208]

Little AC, Jones BC, DeBruine LM (2011). Facial attractiveness: evolutionary based research. Philosophical Transactions of the Royal Society B: Biological Sciences, 366, 1638–1659.

Lusk LG, Mitchel AD (2016). Differential gaze patterns on eyes and mouth during audiovisual speech segmentation. Frontiers in Psychology, 7.

MacLeod A, Summerfield Q (1987). Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology, 21, 131–141. [PubMed: 3594015]

Maddox RK, Pospisil DA, Stecker GC, et al. (2014). Directing eye gaze enhances auditory spatial cue discrimination. Current Biology, 24, 748–752. [PubMed: 24631242]

Maguinness C, Setti A, Burke K, et al. (2011). The effect of combined sensory and semantic components on audio-visual speech perception in older adults. Frontiers in Aging Neuroscience, 3.

McGurk H, MacDonald J (1976). Hearing lips and seeing voices. Nature, 264, 746–748. [PubMed: 1012311]

Meytlis M, Sirovich L (2007). On the dimensionality of face space. IEEE Trans Pattern Anal Mach Intell, 29, 1262–1267. [PubMed: 17496382]

Murphy PR, Robertson IH, Balsters JH, et al. (2011). Pupillometry and P3 index the locus coeruleus-noradrenergic arousal function in humans. Psychophysiology, 48, 1532–1543. [PubMed: 21762458]

Myers MH, Iannaccone A, Bidelman GM (2017). A pilot investigation of audiovisual processing and multisensory integration in patients with inherited retinal dystrophies. BMC Ophthalmology, 17, 1–13. [PubMed: 28068950]

Narinesingh C, Goltz HC, Raashid RA, et al. (2015). Developmental trajectory of mcgurk effect susceptibility in children and adults with amblyopia. Invest Ophthalmol Vis Sci, 56, 2107–2113. [PubMed: 25744982]

Navarra J, Soto-Faraco S (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychological Research, 71, 4–12. [PubMed: 16362332]

Oldfield RC (1971). The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia, 9, 97–113. [PubMed: 5146491]

Ortega L, Guzman-Martinez E, Grabowecky M, et al. (2014). Audition dominates vision in duration perception irrespective of salience, attention, and temporal discriminability. Attention, Perception & Psychophysics, 76, 1485–1502.

Pisoni DB (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. Speech Communication, 13, 109–125. [PubMed: 21461185]

Preuschoff K, ‘t Hart B, Einhauser W (2011). Pupil dilation signals surprise: evidence for noradrenaline’s role in decision making. Frontiers in Neuroscience, 5.

Reetzke R, Lam BPW, Xie Z, et al. (2016). Effect of simultaneous bilingualism on speech intelligibility across different masker types, modalities, and signal-to-noise ratios in school-age children. PloS ONE, 11, e0168048.

Rogers CL, Lister JJ, Febo DM, et al. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. Applied Psycholinguistics, 27, 465–485.

Russo FA, Sandstrom GM, Maksimowski M (2011). Mouth versus eyes: Gaze fixation during perception of sung interval size. Psychomusicology: Music, Mind and Brain, 21, 98.

Schorr EA, Fox NA, van Wassenhove V, et al. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. Proceedings of the National Academy of Sciences of the United States of America, 102, 18748–18750. [PubMed: 16339316]

Scott IM, al., e. (2014). Human preferences for sexually dimorphic faces may be evolutionarily novel. Proceedings of the National Academy of Sciences of the United States of America, 111, 14388–14393. [PubMed: 25246593]

Sheffield SW, Bernstein JG (2018). Assessing perceived listening difficulty using behavioral gaze patterns for audiovisual speech. Journal of the Acoustical Society of America, 143, 1940–1940.

Shelton J, Kumar GP (2010). Comparison between auditory and visual simple reaction times. Neuroscience & Medicine, 1, 30–32.

Studebaker GA (1985). A “rationalized” arcsine transform. Journal of Speech, Language, and Hearing Research, 28, 455–462.

Sumby WH, Pollack I (1954). Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, 26, 212–215.

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, et al. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632–1634. [PubMed: 7777863]

Tremblay CD, Champoux F, Voss P, et al. (2007). Speech and non-speech audio-visual illusions: A developmental study. PLoS One, 2, e742.

Van Belle G, Ramon M, Lefèvre P, et al. (2010). Fixation patterns during recognition of personally familiar and unfamiliar faces. Frontiers in Psychology, 1.

van Wassenhove V, Grant KW, Poeppel D (2005). Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences of the United States of America, 102, 1181–1186. [PubMed: 15647358]

Vatikiotis-Bateson E, Eigsti I-M, Yano S, Munhall KG (1998). Eye movement of perceivers during audiovisual speech perception. Perception & Psychophysics, 60, 926–940. [PubMed: 9718953]

Wang C-A, Boehnke SE, Itti L, et al. (2014). Transient pupil response is modulated by contrast-based saliency. Journal of Neuroscience, 34, 408–417. [PubMed: 24403141]

Watson CS, Qiu WW, Chamberlain MM, et al. (1996). Auditory and visual speech perception: confirmation of a modality-independent source of individual differences in speech recognition. Journal of the Acoustical Society of America, 100, 1153–1162. [PubMed: 8759968]

Wendt D, Dau T, Hjortkjær J (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. Frontiers in Psychology, 7.

Xie Z, Yi H-G, Chandrasekaran B (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. PloS ONE, 9, e114439.

Zekveld AA, George EL, Kramer SE, et al. (2007). The development of the text reception threshold test: a visual analogue of the speech reception threshold test. Journal of Speech, Language, and Hearing Research, 50, 576–584.

Zekveld AA, Kramer SE (2014). Cognitive processing load across a wide range of listening conditions: insights from pupillometry. Psychophysiology, 51, 277–284. [PubMed: 24506437]

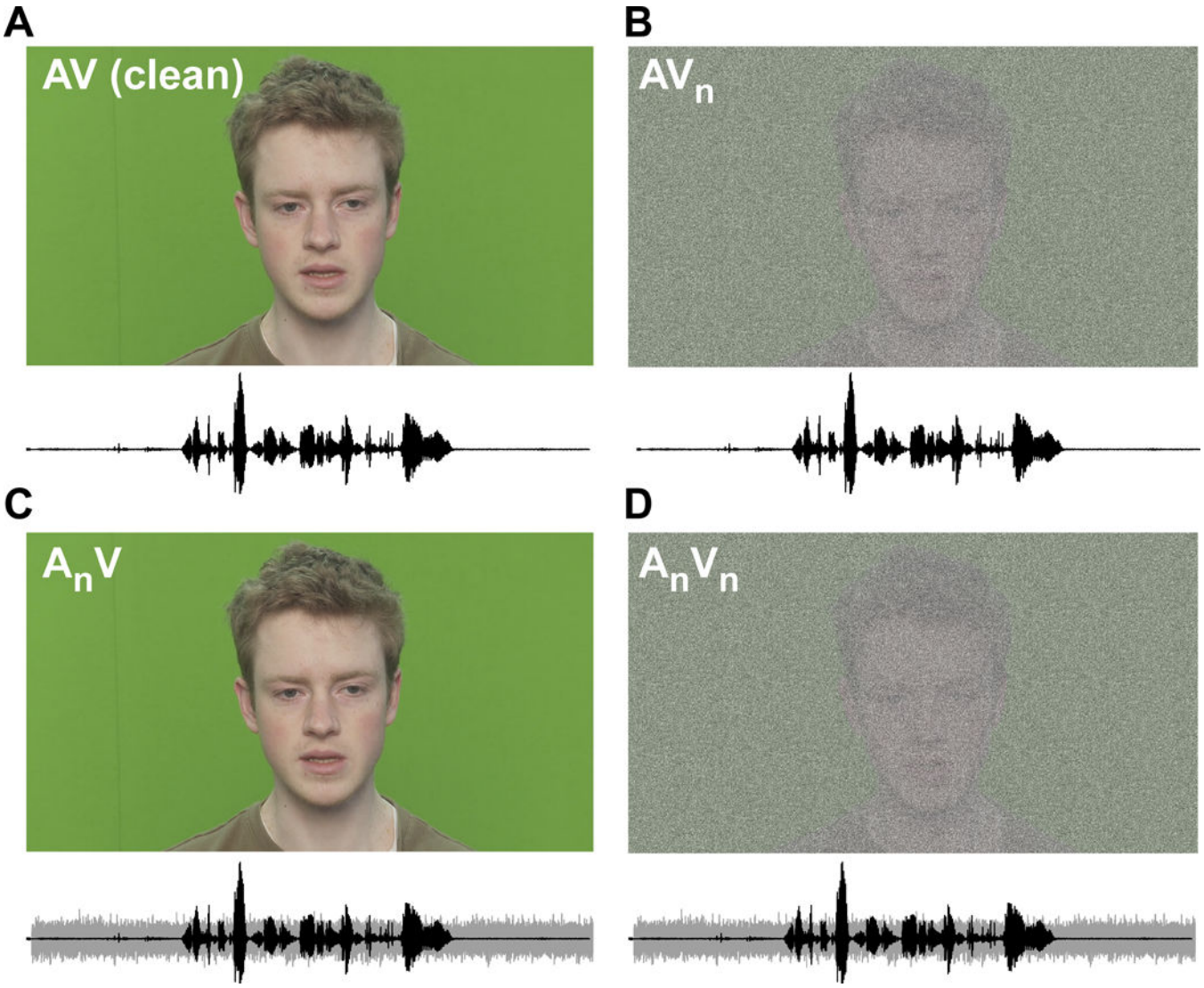

Figure 1: Audiovisual speech stimuli with multimodal noise.

(**A**) Raw audiovisual speech containing clear audio and visual channels. (**B**) AV speech with a degraded *visual* channel containing static visual noise overlaid on the talker's face (AV_n). (**C**) AV speech with a degraded *auditory* channel containing acoustic speech plus noise babble (A_nV). (**D**) AV speech with *audiovisual* noise containing both degraded sound and video channels (A_nV_n).

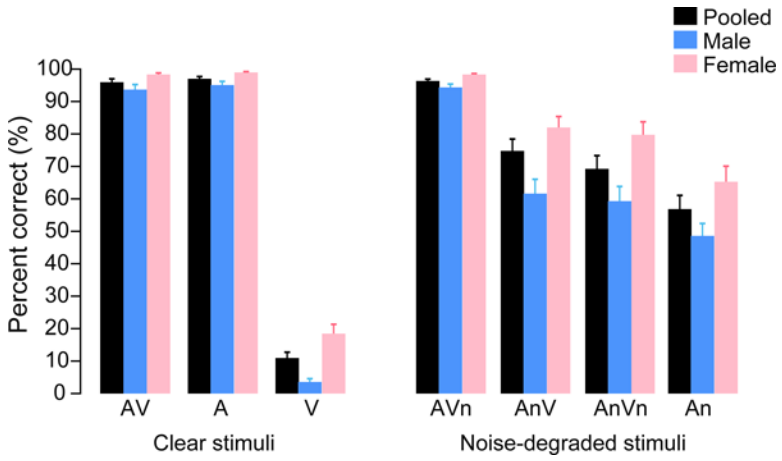

Figure 2: Behavioral speech recognition for sentences presented with auditory, visual, or multimodal (audiovisual) noise.

Responses are shown for stimuli separated and pooled (mean male and female talkers) across speaker gender. AV, clear audiovisual speech; A, auditory only speech; V, visual only speech; A_n, unimodal auditory speech plus acoustic noise; AV_n, audiovisual speech plus visual noise; A_nV, audiovisual speech plus acoustic noise; A_nV_n, audiovisual speech plus audiovisual noise. error bars = ±1 s.e.m.

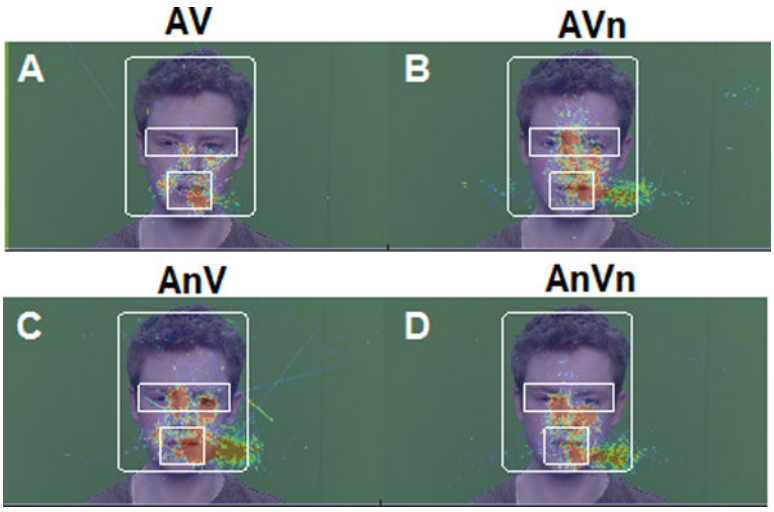

Figure 3: Spatial distribution of eye gaze on the talker's face as a function of audiovisual noise. Hotter colors = more frequent fixations. Shown here are eye data from a representative subject overlaid onto a single frame of the male talker video. Fixations for (**A**) clean (**B**) visual noise, (**C**) auditory noise, (**D**) auditory and visual noise conditions. Boxes demarcate analysis ROIs (eyes, mouth) within the space of the talker's head. In more difficult conditions, listeners shift their gaze way from the mouth to the eyes of the speaker (cf. A vs. D).

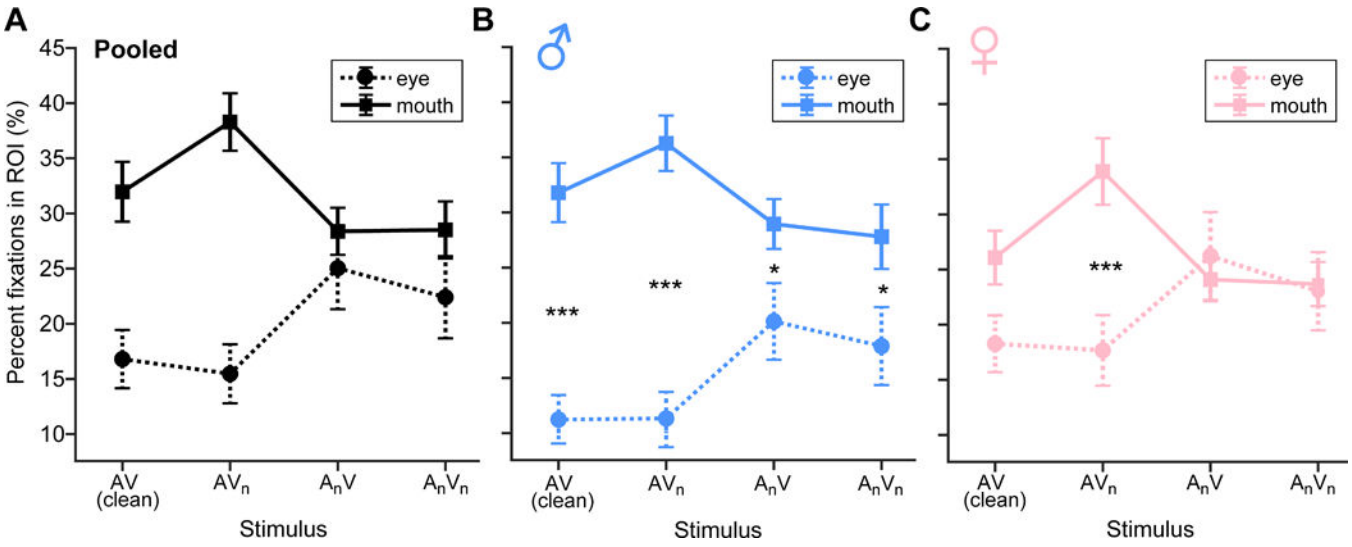

Figure 4: Gaze fixations on the eyes and mouth are modulated by AV noise and gender of the talker.

(**A**) Fixations pooled over male and female speakers. (**B**) Fixations for male speaker sentences. (**C**) Fixations for female speaker sentences. Gaze is fixated more on the mouth during clean and visual noise conditions and shifts toward the eyes whenever the *auditory* channel is degraded. This effect is more prominent when viewing the male talker. *p<0.05, ***p<0.001, errorbars = ±1 s.e.m.

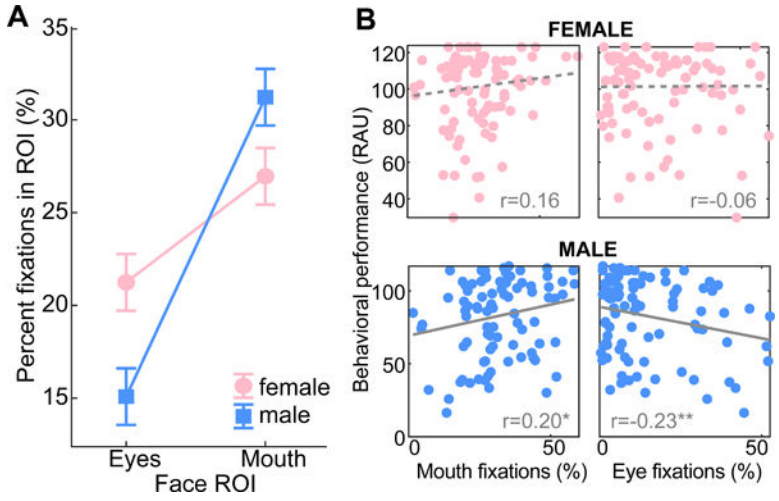

Figure 5: ROI x talker gender interaction in gaze fixations and correlations with behavior.

(A) Pooled across stimulus conditions, participants gaze more at the eyes of the female talker and mouth of the male talker. (**B**) Correlations between percentage of gaze fixations in each ROI and behavioral performance for each talker gender. errorbars = ±1 s.e.m., *p<0.05, **p≤0.01