Psychobiological Responses Reveal Audiovisual Noise Differentially Challenges Speech Recognition

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

Objectives: 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 noises 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 toward 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 versus eyes of the face also depended on the gender of the talker.

Conclusions: Collectively, results suggest listeners (1) depend heavily on the auditory over visual channel when seeing and hearing speech and (2) alter their visual strategy from viewing the mouth to viewing the eyes of a talker with signal degradations, which negatively affects speech perception.

Key words: Audiovisual speech perception, Eyetracking, Listening strategy, Speech-in-noise (SIN).

(Ear & Hearing 2020;41;268–277)

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 (Sumby & Pollack 1954; Erber 1975; Lalonde & Holt 2016), demonstrating audiovisual (AV) cues enhance perceptual processing. AV enhancements are particularly salient for speech-in-noise (SIN) perception (Sumby & Pollack 1954; MacLeod & Summerfield 1987; Vatikiotis-Bateson et al. 1998; Xie et al. 2014). Indeed, previous behavioral studies show an average ~10 to 15 dB improvement in speech recognition threshold from the addition of visual cues to speech especially in challenging listening conditions (MacLeod & Summerfield 1987). Similarly, SIN perception is improved by lipreading (Bernstein et al. 2004; Navarra & Soto-Faraco 2007), and tracking visual

movements augments second language perception by way of multisensory integration (Navarra & 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 (Legault et al. 2010; Galatas et al. 2011; Atcherson et al. 2017). 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 with 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 Face-Time). Given that cross-modal influences between vision and audition are bidirectional (McGurk & MacDonald 1976; Lippert et al. 2007; Maddox et al. 2014; Bidelman 2016; Bidelman & Heath in press; Bidelman & Heath 2019), the first aim of this study was to investigate how noise in different modalities (i.e., the acoustic versus 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 AV processing (e.g., Tanenhaus et al. 1995; Ben-David et al. 2011). 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 & Mitchel 2016). Gaze velocity is also greater for eye movements to visual versus 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 2016; Bidelman & Yellamsetty 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

<zdoi; 10.1097/AUD.0000000000000755>

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 & Tharpe 2002).

To this end, we measured eye gaze fixations during speech recognition tasks to investigate 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 ratios (SNRs). This allowed us to compare different gaze strategies that listeners use when auditory versus 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 & Gagné 2011). Thus, we reasoned that AV 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 & Gillen 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 versus visual noise stressors impact behavioral performance and, more importantly, how listeners might adapt different covert perceptual strategies depending on 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 versus 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).

MATERIALS AND METHODS

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 to 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 postexperiment 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.

AV Speech Stimuli

Stimuli were AV sentences from the TCD-TIMIT database (https://sigmedia.tcd.ie/TCDTIMIT/) (Harte & 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 0º 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 & Gillen 2015).*

We selected 60 unique sentences from among two talkers of the TCD-TIMIT (30 from one male [Talker No.19] and 30 from one female [Talker No.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 (Fig. 2).

The acoustic noise was a multitalker 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/).† Sound and video SNRs were 6 dB and 6.8 to 7.8 dB SNR, respectively, depending on the exact calculation of image SNR.‡ This ensured that the physical SNR was equated between the auditory and

---

*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/.
†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."
‡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 = $10 \log (\mu/\sigma)$, 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 & Woods 2008). The MATLAB function psnr gave a similar SNR estimate of 7.8 dB. By some estimates, the 6 to 7 dB visual SNR used here equates to 60% to 70% performance accuracy in facial recognition paradigms (Meytlis & Sirovich 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.
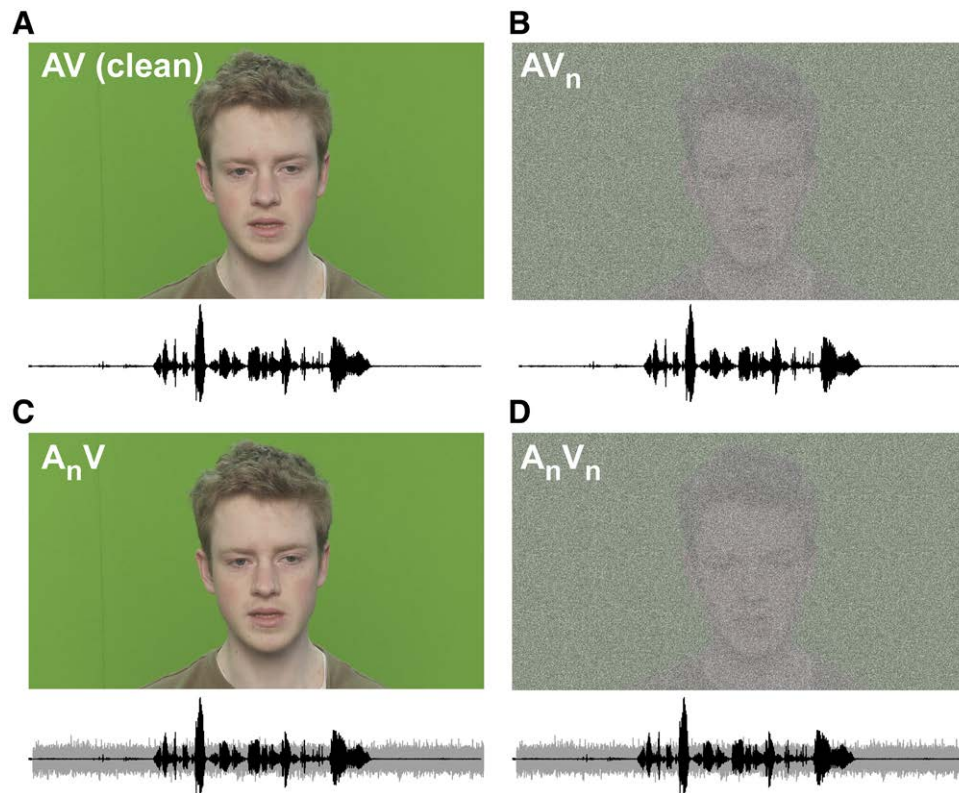
Fig. 1. Audiovisual speech stimuli with multimodal noise. A, Raw audiovisual speech containing clear audio and visual channels. B, AV speech with a degraded visual channel containing static visual noise overlaid on the talker's face (AV_n). C, AV speech with a degraded auditory channel containing acoustic speech plus noise babble (A_nV). D, AV speech with audiovisual noise containing both degraded sound and video channels (A_nV_n). AV, clear audiovisual speech; AV_n, audiovisual speech plus visual noise; A_nV, audiovisual speech plus acoustic noise; A_nV_n, audiovisual speech plus audiovisual noise.

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

Procedure

Listeners were seated in a double-walled sound attenuating chamber (Industrial Acoustics, Inc.) ~90 cm from a computer
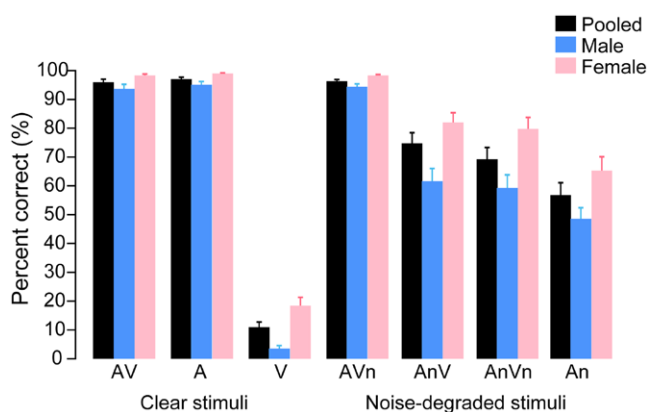

Fig. 2. Behavioral speech recognition for sentences presented with auditory, visual, or multimodal (audiovisual) noise. Responses are shown for stimuli separated and pooled (mean male and female talkers) across speaker gender. AV, clear audiovisual speech; A, auditory only speech; V, visual only speech; 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.

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 before 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)§. Breaks were provided between blocks to avoid fatigue. During the speech recognition task, continuous eye gaze locations were recorded using an eyetracker (detailed below).

Behavioral data were scored on a percent correct basis. Keywords (those carrying meaning) were preselected from each

§The effectiveness of the counterbalance in canceling out possible order/learning effects was confirmed by the (expected) nonsignificant effect of task order when this variable was included in our models [$F_{7,126} = 0.68$, $p > 0.05$].

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 (Pearsons $r = 0.99$, $p < 0.001$), 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 × 3 Talker Types; see Fig. 2). Percent correct scores were transformed via rationalized arcsine units (RAU) transform to account for possible ceiling/floor effects (Studebaker 1985).

Eyetracking

Listeners' gaze fixations on the talkers' face were acquired using a Gazepoint GP3 eyetracker.¶ This device provides precise measurement of the location of ocular gaze and 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 msec (i.e., 60 Hz sampling rate). Data from the GP3 were logged via a program interface with MATLAB. To ensure continued alignment with the screen, the eyetracker was recalibrated 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 before 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 (Lansing & McConkie 2003; Buchan et al. 2007; Van Belle et al. 2010; Russo et al. 2011). 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 the talkers (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

---

¶Although pupillometry (dilation) data were recorded, they are not analyzed in this report as the pupil dilates to a myriad of 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).

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.

RESULTS

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 analysis of variance (ANOVA) (7 Stimulus Conditions × 2 Talker Genders; subjects = random factor) conducted on RAU-transformed percent-correct scores revealed a main effect of gender [$F_{1, 273} = 114.54$, $p < 0.001$; $\eta_p^2 = 0.29$] 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 [$F_{6, 273}=329.11$, $p < 0.001$; $\eta_p^2 = 0.88$]. However, more critically, we found a Stimulus × Gender interaction, indicating that behavioral performance depended on both the specific noise condition and gender of the talker [$F_{6, 273}=2.55$, $p = 0.0204$; $\eta_p^2 = 0.05$]. Separate one-way ANOVAs conducted by talker gender confirmed a main effect of stimulus when viewing both the female [$F_{6, 126} = 128.07$, $p < 0.001$; $\eta_p^2 = 0.86$] and male talker [$F_{6, 126} = 197.04$, $p < 0.001$; $\eta_p^2 = 0.90$]. 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, $p < 0.001$) but still superior to the V-only condition ($A_n > V$; $p < 0.001$). 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 ($AV > A_nV$; $p < 0.001$). Auditory and visual noise also reduced intelligibility relative to clean AV speech ($AV > A_nV_n$; $p < 0.001$). 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 × 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 × Participant Gender × Talker Gender; subjects = random factor) showed no main effects of participants' gender
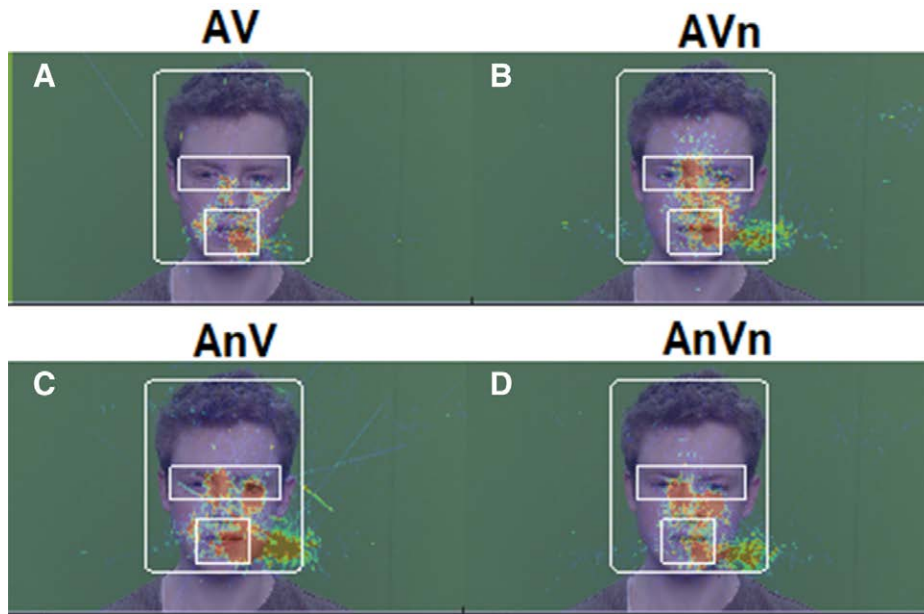
Fig. 3. Spatial distribution of eye gaze on the talker's face as a function of audiovisual noise. Hotter colors = more frequent fixations. Shown here are eye data from a representative subject overlaid onto a single frame of the male talker video. Fixations for (A) clean, (B) visual noise, (C) auditory noise, (D) auditory and visual noise conditions. Boxes demarcate analysis ROIs (eyes and 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). AV, clear audiovisual speech; AV$_n$, audiovisual speech plus visual noise; A$_n$V, audiovisual speech plus acoustic noise; A$_n$V$_n$, audiovisual speech plus audiovisual noise; ROI, regions of interest.

$[F_{1, 284} = 0.20, p = 0.66; \eta_p^2 = 0.0007]$ or 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.

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‖. 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 versus D). Quantitative analyses of eye fixations at each ROI are shown in Figure 4. An omnibus three-way ANOVA (Stimulus × ROI × Gender) revealed interactions between ROI × Stimulus $[F_{3, 311} = 8.05, p < 0.001; \eta_p^2 = 0.072]$ and Gender × ROI $[F_{1, 311} = 13.62, p = 0.0003; \eta_p^2 = 0.042]$. To parse these effects, we conducted separate two-way ANOVAs (ROI × Stimulus) by gender. For the male talker, the ROI × Stimulus Interaction was significant $[F_{3,145} = 3.99, p = 0.0091; \eta_p^2 = 0.076]$ (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

‖Figure 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 across 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). The nearly identical pattern of results for the male and female talker (Fig. 4) also confirm a consistent calibration across talker conditions.

auditory channel remained intact (i.e., AV $[p < 0.001]$ and AV$_n$ $[p < 0.001]$). Fixations at the eyes increased in conditions containing auditory noise (i.e., A$_n$V, $p = 0.027$).

Similarly, we found a ROI × Stimulus Interaction for the female talker $[F_{3, 145} = 3.90, p = 0.0102; \eta_p^2 = 0.075]$ (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 × 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
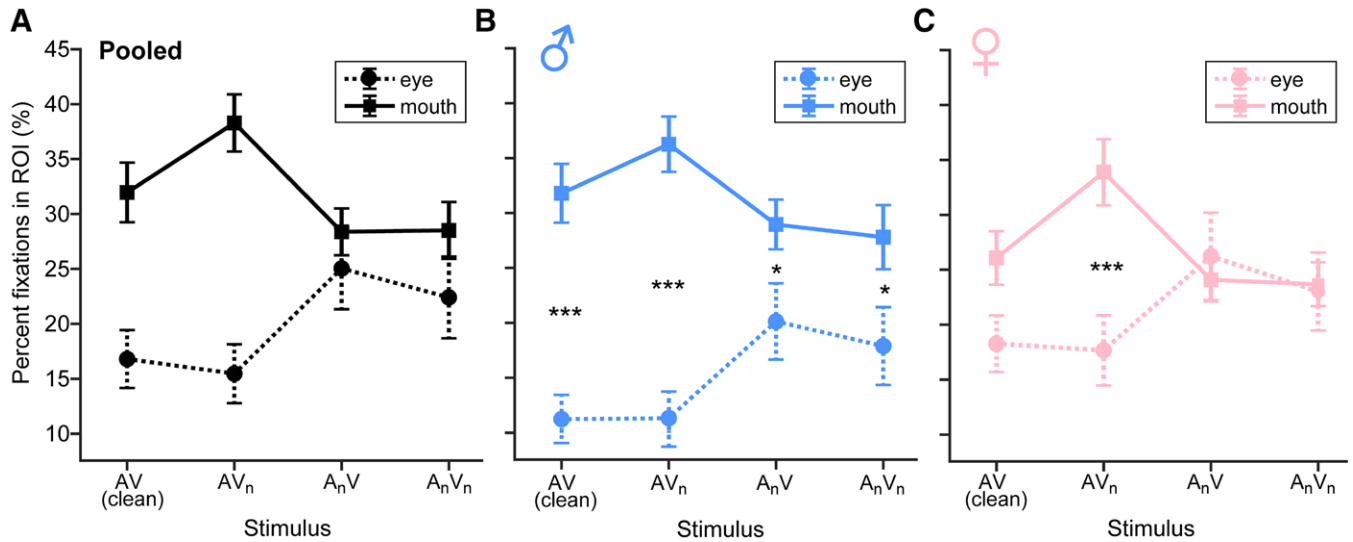
Fig. 4. Gaze fixations on the eyes and mouth are modulated by AV noise and gender of the talker. A, Fixations pooled over male and female speakers. B, Fixations for male speaker sentences. C, Fixations for female speaker sentences. Gaze is fixated more on the mouth during clean and visual noise conditions 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. AV, clear audiovisual speech; AV_n, audiovisual speech plus visual noise; A_nV, audiovisual speech plus acoustic noise; A_nV_n, audiovisual speech plus audiovisual noise.

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 strategies that draw attention toward the eyes negatively affects behavioral performance.

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.

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
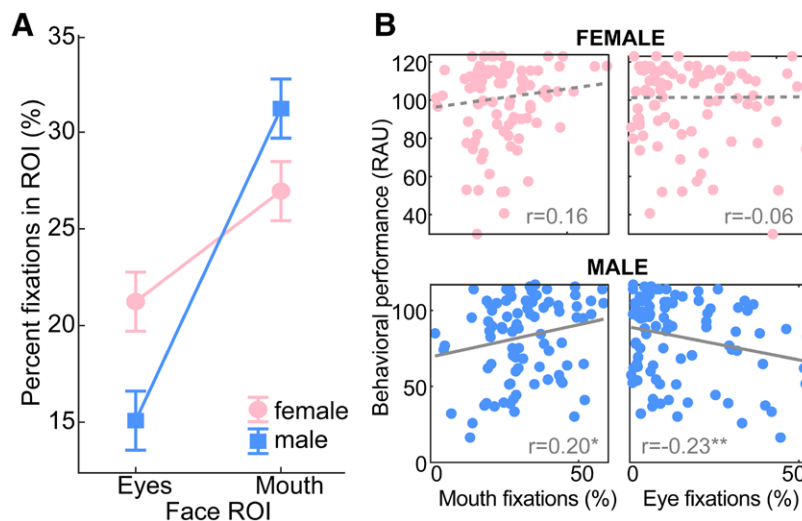

Fig. 5. ROI × Talker Gender Interaction in gaze fixations and correlations with behavior. A, Pooled across stimulus conditions, participants gaze more at the eyes of the female talker and mouth of the male talker. B, Correlations between percentage of gaze fixations in each ROI and behavioral performance for each talker gender. Errorbars = ±1 S.E.M., *$p < 0.05$, **$p \leq 0.01$. RAU, rationalized arcsine units; ROI, regions of interest.

auditory sensation in AV processing (especially for spatial tasks; McGurk & MacDonald 1976; Maddox et al. 2014), 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 and sensory impairments), sound can trump vision to maintain robust perception (Alais & Burr 2004; Narinesingh et al. 2015; Myers et al. 2017). Reaction times are also faster in response to auditory versus visual stimuli (Shelton & Kumar 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 AV interaction in the time course of brain activity (van Wassenhove et al. 2005; Alsius et al. 2014). 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 (Murphy et al. 2011; Zekveld & Kramer, 2014; Eckstein et al. 2017). 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.

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. It is important to note that gaze patterns differed despite relatively similar behavioral performance across AV conditions (cf. Fig. 2 versus Fig. 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 & 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 & 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 & 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 Vatikiotis-Bateson et al. 1998;

Lansing & McConkie 2003; Lusk & Mitchel 2016). 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 toward 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 versus 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 & Heath in press; Bidelman & Heath 2019) but poorer SIN perception in their second (less familiar) language (Rogers et al. 2006; Xie et al. 2014; Bidelman & Dexter 2015; Reetzke et al. 2016). 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.

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, for example, viewing faces of the opposite gender in terms of sexual and social selection (Little et al. 2011; Scott et al. 2014). Adults do tend to focus on the eyes to glean social cues (Lewkowicz & Hansen-Tift 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 versus 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. 1996; Bradlow & Bent 2002). 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 & McConkie 2003).

In addition, while we attempted to match the SNR of visual and auditory noise (~6 to 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 nontrivial 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 versus 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 (contrast sensitivity, motion perception, etc.) are more problematic for older adults (Legault et al. 2010; Daffner et al. 2013), 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.

Future Directions

Interestingly, certain clinical disorders (e.g., autism) are associated with the opposite pattern of gaze fixations observed here, that is, increased fixations on the mouth versus 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 AV 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 (Narinesingh et al. 2015; Myers et al. 2017) 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 & Bernstein 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 suggests eye gaze patterns might be important to incorporate into auditory rehabilitation and assistive hearing technologies.

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 AV speech.

REFERENCES

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol, 14*, 257–262.

Alsius, A., Möttönen, R., Sams, M. E., et al. (2014). Effect of attentional load on audiovisual speech perception: Evidence from ERPs. *Front Psychol, 5*, 727.

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, S. R., Mendel, L. L., Baltimore, W. J., 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.

Banks, B., Gowen, E., Munro, K. J., et al. (2015). Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. *Front Hum Neurosci, 9*, 422.

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

Ben-David, B. M., Chambers, C. G., Daneman, M., et al. (2011). Effects of aging and noise on real-time spoken word recognition: Evidence from eye movements. *J Speech Lang Hear Res, 54*, 243–262.

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

Bidelman, G. M. (2016). Musicians have enhanced audiovisual multisensory binding: Experience-dependent effects in the double-flash illusion. *Exp Brain Res, 234*, 3037–3047.

Bidelman, G. M., & 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 Lang, 143*, 32–41.

Bidelman, G. M., & Heath, S. (in press). *Enhanced temporal binding of audiovisual information in the bilingual brain. Biling-Lang Cogn*, doi: 10.1017/S1366728918000408, 1–11.

Bidelman, G. M., & Heath, S. T. (2019). Neural correlates of enhanced audiovisual processing in the bilingual brain. *Neuroscience, 401*, 11–20.

Bidelman, G. M., & Howell, M. (2016). Functional changes in inter- and intra-hemispheric cortical processing underlying degraded speech perception. *Neuroimage, 124*(Pt A), 581–590.

Bidelman, G. M., & Yellamsetty, A. (2017). Noise and pitch interact during the cortical segregation of concurrent speech. *Hear Res, 351*, 34–44.

Bradley, J. V. (1958). Complete counterbalancing of immediate sequential effects in a Latin square design. *J Am Stat Assoc, 53*, 525–528.

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *J Acoust Soc Am, 112*, 272–284.

Bradlow, A. R., Torretta, G. M., Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Commun, 20*, 255–272.

Brennan, M., Horowitz, A., Su, Y. P. (2005). Dual sensory loss and its impact on everyday competence. *Gerontologist, 45*, 337–346.

Broadbent, D. E. (1958). *Perception and Communication*. London: Pergamon.

Buchan, J. N., Paré, M., Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Soc Neurosci, 2*, 1–13.

Burr, D., Banks, M. S., Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Exp Brain Res, 198*, 49–57.

Coutrot, A., Binetti, N., Harrison, C., et al. (2016). Face exploration dynamics differentiate men and women. *J Vis, 16*, 16.

Daffner, K. R., Haring, A. E., Alperin, B. R., et al. (2013). The impact of visual acuity on age-related differences in neural markers of early visual processing. *Neuroimage, 67*, 127–136.

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., et al. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Dev Cogn Neurosci, 25*, 69–91.

Erber, N. P. (1975). Auditory-visual perception of speech. *J Speech Hear Disord, 40*, 481–492.

Favre-Félix, A., Graversen, C., Hietkamp, R. K., 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, J. S., Lamel, L. F., Fisher, W. M., et al. (1993). *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*. Retrieved 2015 from https://catalog.ldc.upenn.edu/LDC93S1.

Gillen, E. (2014). *TCD-TIMIT: A New Database for Audio-Visual Speech Recognition*. Department of Electronic and Electrical Engineering: Trinity College Dublin.

Goldring, J. E., Dorris, M. C., Corneil, B. D., et al. (1996). Combined eye-head gaze shifts to visual and auditory targets in humans. *Exp Brain Res, 111*, 68–78.

González, R. C., & Woods, R. E. (2008). *Digital Image Processing*. Upper Saddle River, New Jersey: Prentice Hall.

Gosselin, P. A., & Gagné, J. P. (2011). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *Int J Audiol, 50*, 786–792.

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

Hick, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *J Speech Lang Hear Res, 45*, 573–584.

Hirst, R. J., Stacey, J. E., Cragg, L., et al. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Sci Rep, 8*, 12372.

Irwin, J., Brancazio, L., Volpe, N. (2017). The development of gaze to a speaking face. *J Acoust Soc Am, 141*, 3145.

Jordan, T. R., McGowan, V. A., Paterson, K. B. (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.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., 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. *J Acoust Soc Am, 116*(4 Pt 1), 2395–2405.

Klatte, M., Bergström, K., Lachmann, T. (2013). Does noise affect learning? A short review on noise effects on cognitive performance in children. *Front Psychol, 4*, 578.

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.

Knecht, H. A., Nelson, P. B., Whitelaw, G. M., et al. (2002). Background noise levels and reverberation times in unoccupied classrooms: Predictions and measurements. *Am J Audiol, 11*, 65–71.

Lalonde, K., & Holt, R. F. (2016). Audiovisual speech perception development at varying levels of perceptual processing. *J Acoust Soc Am, 139*, 1713.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Percept Psychophys, 65*, 536–552.

Legault, I., Gagné, J. P., Rhoualem, W., et al. (2010). The effects of blurred vision on auditory-visual speech perception in younger and older adults. *Int J Audiol, 49*, 904–911.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proc Natl Acad Sci USA*, *109*, 1431–1436.

Liao, H. I., Kidani, S., Yoneya, M., et al. (2016). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychon Bull Rev, 23*, 412–425.

Lippert, M., Logothetis, N. K., Kayser, C. (2007). Improvement of visual contrast detection by a simultaneous sound. *Brain Res, 1173*, 102–109.

Little, A. C., Jones, B. C., DeBruine, L. M. (2011). Facial attractiveness: Evolutionary based research. *Philos Trans R Soc Lond B Biol Sci, 366*, 1638–1659.

Lusk, L. G., & Mitchel, A. D. (2016). Differential gaze patterns on eyes and mouth during audiovisual speech segmentation. *Front Psychol, 7*, 52.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *Br J Audiol, 21*, 131–141.

Maddox, R. K., Pospisil, D. A., Stecker, G. C., et al. (2014). Directing eye gaze enhances auditory spatial cue discrimination. *Curr Biol, 24*, 748–752.

Maguinness, C., Setti, A., Burke, K. E., et al. (2011). The effect of combined sensory and semantic components on audio-visual speech perception in older adults. *Front Aging Neurosci, 3*, 19.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Meytlis, M., & Sirovich, L. (2007). On the dimensionality of face space. *IEEE Trans Pattern Anal Mach Intell, 29*, 1262–1267.

Murphy, P. R., Robertson, I. H., Balsters, J. H., et al. (2011). Pupillometry and P3 index the locus coeruleus-noradrenergic arousal function in humans. *Psychophysiology, 48*, 1532–1543.

Myers, M. H., Iannaccone, A., Bidelman, G. M. (2017). A pilot investigation of audiovisual processing and multisensory integration in patients with inherited retinal dystrophies. *BMC Ophthalmology*, *17*, 1–13.

Narinesingh, C., Goltz, H. C., Raashid, R. A., et al. (2015). Developmental trajectory of McGurk effect susceptibility in children and adults with amblyopia. *Invest Ophthalmol Vis Sci, 56*, 2107–2113.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychol Res, 71*, 4–12.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113.

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

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Commun, 13*, 109–125.

Preuschoff, K., 't Hart, B. M., Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Front Neurosci, 5*, 115.

Reetzke, R., Lam, B. P., 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, C. L., Lister, J. J., Febo, D. M., et al. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, *27*, 465–485.

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

Schorr, E. A., Fox, N. A., van Wassenhove, V., et al. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proc Natl Acad Sci USA*, *102*, 18748–18750.

Scott, I. M., al., e. (2014). Human preferences for sexually dimorphic faces may be evolutionarily novel. *Proc Natl Acad Sci USA, 111*, 14388–14393.

Sheffield, S. W., & Bernstein, J. G. (2018). Assessing perceived listening difficulty using behavioral gaze patterns for audiovisual speech. *J Acoust Soc Am*, *143*, 1940–1940.

Shelton, J., & Kumar, G. P. (2010). Comparison between auditory and visual simple reaction times. *Neurosci Med, 1*, 30–32.

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *J Speech Hear Res, 28*, 455–462.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J Acoust Soc Am, 26*, 212–215.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., et al. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.

Tremblay, C., 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. *Front Psychol, 1*, 20.

van Wassenhove, V., Grant, K. W., Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc Natl Acad Sci USA*, *102*, 1181–1186.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., et al. (1998). Eye movement of perceivers during audiovisual speech perception. *Percept Psychophys, 60*, 926–940.

Wang, C. A., Boehnke, S. E., Itti, L., et al. (2014). Transient pupil response is modulated by contrast-based saliency. *J Neurosci, 34*, 408–417.

Watson, C. S., Qiu, W. W., Chamberlain, M. M., et al. (1996). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences in speech recognition. *J Acoust Soc Am, 100*(2 Pt 1), 1153–1162.

Wendt, D., Dau, T., Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Front Psychol, 7*, 345.

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, A. A., George, E. L., Kramer, S. E., et al. (2007). The development of the text reception threshold test: A visual analogue of the speech reception threshold test. *J Speech Lang Hear Res, 50*, 576–584.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology, 51*, 277–284.