frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 08 November 2019
doi: 10.3389/fnhum.2019.00386

The Dynamics of Attention Shifts Among Concurrent Speech in a Naturalistic Multi-speaker Virtual Environment

*Keren Shavit-Cohen and Elana Zion Golumbic**

The Gonda Multidisciplinary Brain Research Center, Bar Ilan University, Ramat Gan, Israel

Focusing attention on one speaker on the background of other irrelevant speech can be a challenging feat. A longstanding question in attention research is whether and how frequently individuals shift their attention towards task-irrelevant speech, arguably leading to occasional detection of words in a so-called unattended message. However, this has been difficult to gauge empirically, particularly when participants attend to continuous natural speech, due to the lack of appropriate metrics for detecting shifts in internal attention. Here we introduce a new experimental platform for studying the dynamic deployment of attention among concurrent speakers, utilizing a unique combination of Virtual Reality (VR) and Eye-Tracking technology. We created a Virtual Café in which participants sit across from and attend to the narrative of a target speaker. We manipulated the number and location of distractor speakers by placing additional characters throughout the Virtual Café. By monitoring participant's eye-gaze dynamics, we studied the patterns of overt attention-shifts among concurrent speakers as well as the consequences of these shifts on speech comprehension. Our results reveal important individual differences in the gaze-pattern displayed during selective attention to speech. While some participants stayed fixated on a target speaker throughout the entire experiment, approximately 30% of participants frequently shifted their gaze toward distractor speakers or other locations in the environment, regardless of the severity of audiovisual distraction. Critically, preforming frequent gaze-shifts negatively impacted the comprehension of target speech, and participants made more mistakes when looking away from the target speaker. We also found that gaze-shifts occurred primarily during gaps in the acoustic input, suggesting that momentary reductions in acoustic masking prompt attention-shifts between competing speakers, in line with "glimpsing" theories of processing speech in noise. These results open a new window into understanding the dynamics of attention as they wax and wane over time, and the different listening patterns employed for dealing with the influx of sensory input in multisensory environments. Moreover, the novel approach developed here for tracking the locus of momentary attention in a naturalistic virtual-reality environment holds high promise for extending the study of human behavior and cognition and bridging the gap between the laboratory and real-life.

Keywords: speech processing, auditory attention, eye-tracking, virtual reality, cocktail party effect, distractability

OPEN ACCESS

Edited by:
Pietro Cipresso,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Kristina C. Backer,
University of California, Merced,
United States
Carlo Sestieri,
Università degli Studi G. d'Annunzio
Chieti e Pescara, Italy

****Correspondence:***
Elana Zion Golumbic
elana.zion-golumbic@biu.ac.il

Specialty section:
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

Received: *02 May 2019*
Accepted: *16 October 2019*
Published: *08 November 2019*

Citation:
Shavit-Cohen K and Zion Golumbic E (2019) The Dynamics of Attention Shifts Among Concurrent Speech in a Naturalistic Multi-speaker Virtual Environment.
Front. Hum. Neurosci. 13:386.
doi: 10.3389/fnhum.2019.00386

INTRODUCTION

Focusing attention on one speaker in a noisy environment can be challenging, particularly in the background of other irrelevant speech (McDermott, 2009). Despite the difficulty of this task, comprehension of an attended speaker is generally good and the content of distractor speech is rarely recalled explicitly (Cherry, 1953; Lachter et al., 2004). Preferential encoding of attended speech in multi-speaker contexts is also mirrored by enhanced neural responses to attended vs. distractor speech (Ding and Simon, 2012b; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013b; O'Sullivan et al., 2015). However, there are also indications that distractor speech is processed, at least to some degree. Examples for this are the Irrelevant Stimulus Effect, where distractor words exert priming effect on an attended task (Treisman, 1964; Neely and LeCompte, 1999; Beaman et al., 2007), as well as occasional explicit detection of salient words in distractor streams (Cherry, 1953; Wood and Cowan, 1995; Röer et al., 2017; Parmentier et al., 2018). These effects highlight a key theoretical tension regarding how processing resources are allocated among competing speech inputs. Whereas Late-Selection models of attention posit that attended and distractor speech can be fully processed, allowing for explicit detection of words in so-called unattended speech (Deutsch and Deutsch, 1963; Duncan, 1980; Parmentier et al., 2018), Limited-Resources models hold that there are inherent bottlenecks for linguistic processing of concurrent speech due to limited resources (Broadbent, 1958; Lachter et al., 2004; Lavie et al., 2004; Raveh and Lavie, 2015). The latter perspective reconciles indications for occasional processing of distractor speech as stemming from rapid shifts of attention toward distractor speech (Conway et al., 2001; Escera et al., 2003; Lachter et al., 2004). Yet, despite the parsimonious appeal of this explanation, to date, there is little empirical evidence supporting and characterizing the psychological reality of attention switches among concurrent speakers.

Establishing whether and when rapid shifts of attention towards distractor stimuli occur is operationally challenging since it refers to individuals' internal state that researchers do not have direct access to. Existing metrics for detecting shifts of attention among concurrent speech primarily rely on indirect measures such as prolongation of reaction times on an attended task (Beaman et al., 2007) or subjective reports (Wood and Cowan, 1995). Given these limitations, the current understanding of the dynamics of attention over time, and the nature and consequences of rapid attention-shifts among concurrent speech is extremely poor. Nonetheless, gaining insight into the dynamics of internal attention-shifts is critical for understanding how attention operates in naturalistic multi-speaker settings.

Here, we introduce a new experimental platform for studying the dynamic deployment of attention among concurrent speakers. We utilize Virtual Reality (VR) technology to simulate a naturalistic audio-visual multi-speaker environment, and track participant's gaze-position within the Virtual Scene as a marker for the locus of overt attention and as a means for detecting attention-shifts among concurrent speakers. Participants experienced sitting in a "Virtual Café" across from a partner (avatar; animated target speaker) and were required to focus attention exclusively towards this speaker. Additional distracting speakers were placed at surrounding tables, with their number and location manipulated across conditions. Continuous tracking of gaze-location allowed us to characterize whether participants stayed focused on the target speaker as instructed or whether and how often they performed overt glimpses around the environment and toward distractor speakers. Critically, we tested whether shifting one's gaze around the environment and away from the target speaker impacted comprehension of target speech. We further tested whether gaze-shifts are associated with salient acoustic changes in the environment, such as onsets in distractor speech that can potentially grab attention exogenously (Wood and Cowan, 1995) or brief pauses that create momentary unmasking of competing sounds (Lavie et al., 2004; Cooke, 2006).

Gaze-shifts are often used as a proxy for attention shifts in natural vision (Anderson et al., 2015; Schomaker et al., 2017; Walker et al., 2017), however this measure has not been utilized extensively in dynamic contexts (Marius't Hart et al., 2009; Foulsham et al., 2011). This novel approach enabled us to characterize the nature of momentary attention-shifts in ecological multi-speaker listening conditions, as well as individual differences, gaining insight into the factors contributing to dynamic attention shifting and its consequences on speech comprehension.

MATERIALS AND METHODS

Participants

Twenty-six adults participated in this study (ages 18–32, median 24; 18 female, three left handed), all fluent in Hebrew, with self-reported normal hearing and no history of psychiatric or neurological disorders. Signed informed consent was obtained from each participant prior to the experiment, in accordance with the guidelines of the Institutional Ethics Committee at Bar-Ilan University. Participants were paid for participation or received class credit.

Apparatus

Participants were seated comfortably in an acoustic-shielded room and viewed a 3D VR scene of a café, through a head-mounted device (Oculus Rift Development Kit 2). The device was custom-fitted with an embedded eye-tracker (SMI, Teltow, Germany; 60 Hz monocular sampling rate) for continuous monitoring of participants' eye-gaze position. Audio was presented through high-quality headphone (Sennheiser HD 280 pro).

Stimuli

Avatar characters were selected from the Mixamo platform (Adobe Systems, San Jose, CA, USA). Soundtracks for the avatars' speech were 35–50 s long segments of natural Hebrew speech taken from podcasts and short stories[1]. Avatars' mouth and articulation movements were synced to the audio to

[1] www.icast.co.il

create a realistic audio-visual experience of speech (LipSync Pro, Rogo Digital, England). Scene animation and experiment programming was controlled using an open-source VR engine (Unity Software[2]). Speech loudness levels (RMS) were equated for all stimuli, in 10-s long bins (to avoid biases due to fluctuations in speech time-course). Audio was further manipulated within Unity using a 3D sound algorithm, so that it was perceived as originating from the spatial location of the speaking avatar, with overall loudness decreasing logarithmically with distance from the listener. Participant's head movements were not restricted, and both the graphic display and 3D sound were adapted on-line in accordance with head-position, maintaining a spatially-coherent audio-visual experience.

Experiment Design

In the Virtual Café setting, participants experienced sitting at a café table facing a partner (animated speaking avatar) telling a personal narrative. They were told to focus attention exclusively on the speech of their partner (target speaker) and to subsequently answer four multiple-choice comprehension questions about the narrative (e.g., "What computer operating system was mentioned?"). Answers to the comprehension questions were evenly distributed throughout the narrative, and were pre-screened in a pilot study to ensure accuracy rates between 80% and 95% in a single-speaker condition. The time-period containing the answer to each question was recorded and used in subsequent analysis of performance as a function of gaze-shift behaviors (see below). Additional pairs of distracting speakers (avatars) were placed at surrounding tables, and we systematically manipulated the number and location of distractors in four conditions: No Distraction (NoD), Left Distractors (LD), Right Distractors (RD), Right and Left Distractors (RLD; **Figure 1**). Each condition consisted of five trials (∼4 min per condition) and was presented in random order, which was different for each participant. The identity and voice of the main speaker were kept constant throughout the experiment, with different narratives in each trial, while the avatars and narratives serving as distractors varied from trial to trial. The allocation of each narrative to the condition was counter-balanced across participants, to avoid material-specific biases. Before starting the experiment itself, participants were given time to look around and familiarize themselves with the Café environment and the characters in it. During this familiarization stage, no audio was presented and participants terminated it when they were ready. They also completed two training-trials, in the NoD and RLD conditions, to familiarize them with the stimuli and task as well as the type of comprehension questions asked. This familiarization and training period lasted approximately 3-min.

Analysis of Eye-Gaze Dynamics

Analysis of eye-gaze data was performed in Matlab (Mathworks, Natick, MA, USA) using functions from the fieldtrip toolbox[3] as well as custom-written scripts. The position of eye-gaze position in virtual space coordinates (x, y, z) was monitored continuously throughout the experiment. Periods surrounding eye-blinks were removed from the data (250 ms around each blink). Clean data from each trial were analyzed as follows.

First, we mapped gaze-positions onto specific avatars/locations in the 3D virtual scene. For data reduction, we used a spatial clustering algorithm (k-means) to combine gaze data-points associated with similar locations in space. Next, each spatial cluster was associated with the closest avatar, by calculating the Euclidean distance between the center of the cluster and the center of each avatar presented in that condition. If two or more clusters were associated with looking at the same avatar, they were combined. Similarly, clusters associated with the members of the distractor avatar-pairs (left or right distractors) were combined. If a cluster did not fall within a particular distance-threshold from any of the avatars, it was associated with looking at "The Environment." This resulted in a maximum of four clusters capturing the different possible gaze locations in each trial: (1) Target Speaker; (2) Left Distractors (when relevant); (3) Right Distractors (when relevant); and (4) Rest of the Environment. The appropriateness of cluster-to-avatar association and distance-threshold selection was verified through visual inspection.

Based on the clustered data, we quantified the percent of time that participants spent focusing at each location (*Percent Gaze Time*) in each trial, and detected the times of *Gaze-Shifts* from one cluster to another. Gaze-shifts lasting less than 250 ms were considered artifacts and removed from the analysis, as they are physiologically implausible (Bompas and Sumner, 2009; Gilchrist, 2011). The number of Gaze-shifts as well as the Percent Gaze Time spent at each of the four locations—Target Speaker, Left Distractors, Right Distractors and Environment—were averaged across trials, within condition. Since conditions differed in the type and number of distractors, comparison across conditions focused mainly on metrics pertaining to gazing at/away-from the target speaker.

Mixed linear regression models were used in all analyses to fit the data and test for effects of Condition on gaze patterns (both Percent Gaze-Time Away and Gaze-Shifts), as well as possible correlations with speech comprehension accuracy measures. These analyses were conducted in R (R Development Core Team, 2012) and we report statistical results derived using both regular linear (lme4 package for R; Bates et al., 2015) and robust estimation approaches (robustlmm package for R; Koller, 2016), to control for possible contamination by outliers. The advantage of mixed-effects models is that they account for variability between subjects and correlations within the data, as well as possible differences in trial numbers across conditions (Baayen et al., 2008), which makes them particularly suitable for the type of data collected here.

Analysis of Speech Acoustics Relative to Gaze-Shifts

A key question is what prompts overt gaze-shifts away from the target speakers, and specifically whether they are driven by changes in the acoustic input or if they should be considered more internally-driven. Two acoustic factors that have been suggested as inviting attention-shifts among concurrent speech

[2] unity3d.com

[3] fieldtriptoolbox.org

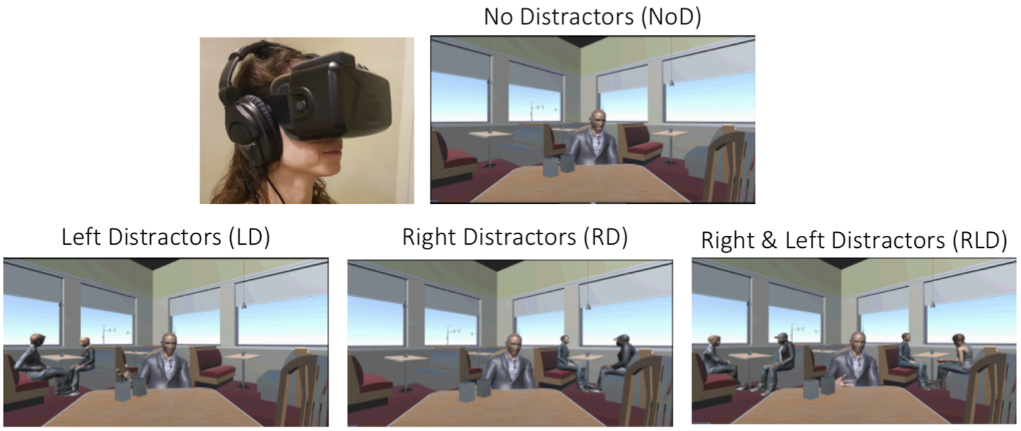

FIGURE 1 | Manipulation of distraction in the Virtual Café. Participants are instructed to attend to the narrative of the target speaker facing them. The number and location of distractor speakers was manipulated in four conditions: only the target-speaker presented and No Distractors (NoD), a single distractor-pair sitting to the left (LD) or right (RD) of the target speaker, and two distractor-pairs sitting to the right and the left of the target speaker (RLD). Top-Left: demonstration of a participant experiencing the Virtual Café (written informed consent was obtained from the participant for publication of this photograph).

are: (a) onsets/loudness increases in distractor speech that can potentially grab attention exogenously (Wood and Cowan, 1995); and (b) brief pauses that create momentary unmasking of competing sounds (Lavie et al., 2004; Cooke, 2006). To test whether one or both of these factors account for the occurrence of gaze-shifts away from the target speaker in the current data, we performed a gaze-shift time-locked analysis of the speech-acoustics of target speech (in all conditions) and distractor speech (in the LD, RD and RLD conditions).

To this end, we first calculated the temporal envelope of the speech presented in each trial using a windowed RMS (30 ms smoothing). The envelopes were segmented relative to the times where gaze-shifts *away from the target speaker* occurred in that particular trial (−400 to +200 ms around each shift). Given that the initiation-time for executing saccades is ∼200 ms (Gilchrist, 2011), the time-window of interest for looking at possible influences of the acoustics on gaze-shifts is prior to that, i.e., 400–200 ms prior to the gaze-shift itself.

Since the number of gaze-shifts varied substantially across participants, we averaged the gaze-shift-locked envelope-segments across all trials and participants, within condition. The resulting average acoustic-loudness waveform in each condition was compared to a distribution of non-gaze-locked loudness levels, generated through a permutation procedure as follows: the same acoustic envelopes were segmented randomly into an equal number of segments as the number of gaze-shifts in each condition (sampled across participants with the same proportion as the real data). These were averaged, producing a non-gaze-locked average waveform. This procedure was repeated 1,000 times and the real gaze-shift locked waveform was compared to the distribution of non-gaze-locked waveforms. We identified time-points where the loudness level fell above or below the top/bottom 5% tile of the non-gaze-locked distribution, signifying that the speech acoustics were particularly quiet or loud relative (relative to the rest of the presented speech stimuli). We also quantified the signal-to-noise ratio (SNR) between the time-resolved spectrograms of target and distractor speech surrounding gaze-shifts, according to: $SNR(f,t) = \log\left(\frac{P_{\text{target}}(f,t)}{P_{\text{distractor}}(f,t)}\right)$, with *P(f,t)* depicting the power at frequency *f* at time *t*. This was calculated for target-distractor combinations surrounding each gaze-shift, and averaged across shifts and trials.

RESULTS

Gaze-Patterns and Speech Comprehension

On an average, participants spent -7.6% of each trial (-3 s in a 40-s-long trial) looking at locations other than the target speaker and they performed an average of 2.5 gaze-shifts per trial. **Figure 2A** shows the distribution of eye-gaze location in two example trials taken from different participants, demonstrating that sometimes gaze was fixated on the target-speaker throughout the entire trial, and sometimes shifted occasionally towards the distractors. The distribution of Gaze-shifts was relatively uniform over the course of the entire experiment (**Figure 2B**, left). Twenty-three percentage of gaze-shifts were performed near the onset of the trial, however, the majority of gaze-shifts occurred uniformly throughout the entire trial (**Figure 2B**, right).

Figures 3A,B show how the average *Gaze Time Away* from the target speaker (i.e., time spent looking at distractor avatars or other locations in the Environment) and the number of *Gaze-Shifts* away from the target speaker, varied across the four conditions. To test whether gaze patterns (number of *Gaze-Shifts* and/or proportion *Gaze-Time Away*) differed across conditions, we estimated each of them separately using linear mixed effect model with the factor Condition as a fixed effect (Gaze-Shifts' Condition and Gaze-Time–Condition), where each of the three distraction conditions (RD, LD and RLD) was compared to the NoD condition. By-subject intercepts were

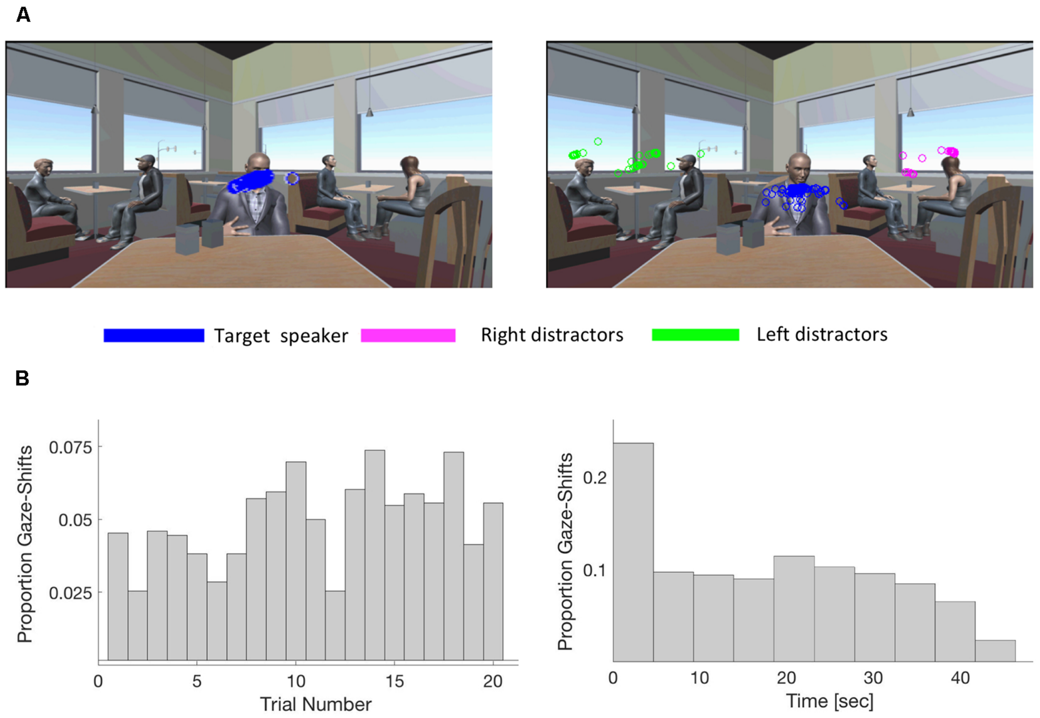

FIGURE 2 | Characterization of gaze-shift patterns. **(A)** Illustration of the variability in gaze-patterns across individuals. The figure depicts all gaze data points in a specific trial in the RLD condition for two example participants. While the participant shown in the left panel remained focused exclusively on the target speaker throughout the trial (blue dots), the participant in the right panel spent a substantial portion of the trial looking at the distractor speakers on both the left (green) and the right (magenta). **(B)** Left: distribution of all gaze-shifts across the duration of the experiment, collapsed across participants. Gaze-shifts occurred throughout the experiment and were not more prevalent the beginning/end. Right: distribution of gaze-shifts over the course of a trial, collapsed across all trials and participants. Twenty-three percentage of gaze-shifts occurred during the first 5 s of each trial, and the remainder could occur with similar probability throughout the entire trial.

included as random effects. No significant effects of Condition were found on *Gaze-Time*, however, participants performed significantly more *Gaze-Shifts* in the RLD condition relative to the NoD condition (lmer: $\beta = 0.8$, $t = 2.5$, $p = 0.01$; robustlmm: $\beta = 0.54$, $t = 2.5$).

Of critical interest is whether the presence of distractors and gaze-shifts towards them impacted behavioral outcomes of speech comprehension. Accuracy on the multiple-choice comprehension questions of the target speaker was relatively good in all conditions (mean accuracy 82% ± 3; **Figure 3C**). A mixed linear model estimating Accuracy $\sim$ Condition did not reveal any significant differences in Accuracy between conditions (lmer: all t's < 0.199, $p > 0.6$; robustlmm: all t's < 0.05). However, adding Percent Gaze-Time as a second fixed effect to the Accuracy $\sim$ Condition model, improved the model significantly ($\chi^2 = 9.14$, $p < 10^3$), with Percent Gaze-Time showing a significant correlation with Accuracy (lmer: $\beta = -0.19$, $t = -3.13$, $p = 0.001$; robustlmm: $\beta = -0.23$, $t = -3.77$; **Figure 3D**). Adding Number of Shifts to the Accuracy $\sim$ Condition model, however, did not yield any additional significant advantage (likelihood ratio test $\chi^2 = 2.4$, $p > 0.1$; **Figure 3E**), suggesting that the number of gaze-shifts performed *per se* did not affect speech comprehension.

To further assess the link between performance on the comprehension questions and gaze-shifts, we tested whether participants were more likely to make mistakes on specific questions if they happened to be looking away from the target-speaker when the critical information for answering that question was delivered. Mistake rates were slightly lower when participants fixated on the target speaker when the critical information was delivered (16% miss-rate) vs. when they looked away (18% miss-rate). To evaluate this effect statistically, we fit a linear mixed model to the accuracy results on individual questions testing whether they were mediated by the presence of a gaze-shift when the answer was given, as well as the condition [Accuracy $\sim$ Shift (yes/no) + Condition as fixed effects], with by-subject intercepts included as random effects. This analysis demonstrated a small yet significant effect of the presence of a gaze-shift during the period when the answer was given (lmer $\beta = -0.05$, $t = -2.16$, $p < 0.04$; robustlmm $t = -3$; **Figure 3F**), however there was no significant effect of Condition (all t's < 0.5).

Individual Differences in Gaze Patterns

When looking at gaze-patterns across participants, we noted substantial variability in the number of gaze-shift performed and percent time spent gazing away from the target speaker. As illustrated in **Figures 2A**, **4**, some participants stayed completely focused on the main speaker throughout the entire experiment, whereas others spent a substantial portion of each trial gazing

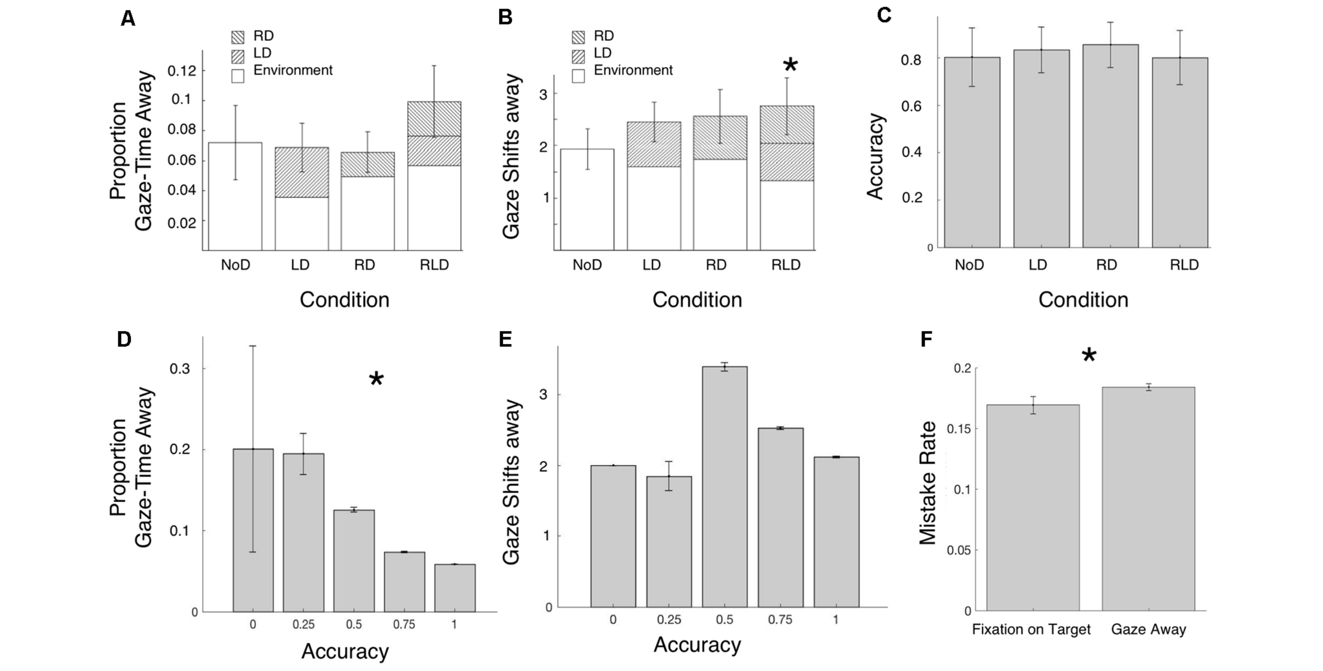

FIGURE 3 | Summary of gaze-shift patterns and behavioral outcomes across conditions. **(A,B)** Proportion of Gaze-Time and Number of Gaze-Shifts Away from target speaker, per trial and across conditions. Results within each condition are broken down by gaze-location (Right Distractors, Left Distractors or Environment in blank, left and right diagonals, respectively). There was no significant difference between conditions in the total Gaze-time away from the target speaker or number of gaze-shifts. Significantly more Gaze-Shifts were performed in the RLD condition relative to the NoD condition. No other contrasts were significant. **(C)** Mean accuracy on comprehension questions, across condition. Difference between conditions was not significant. **(D,E)** Analysis of Accuracy as a function of Gaze-Shift Patterns, at the whole trial level. Trials where participants spent a larger proportion of the time looking away from the target-speaker were associated with lower accuracy rates. No significant correlation was found between accuracy rates and the number of Gaze-Shifts performed. **(F)** Analysis of Accuracy on single question as a function of Gaze-Shift Patterns. Mistake rates were significantly higher if participants were looking away from the target speaker vs. fixating on the target speaker during the time-window when the information critical for answering the question was delivered. Error bars indicate Standard Error of the Mean (SEM). *$p < 0.05$.

around the environment (*range across participants*: 0–18 average number gaze-shifts per trial; 0–34.52% average percent of trial spent looking away from the target speaker). This motivated further inspection of gaze-shift behavior at the individual level. Specifically, we tested whether individual behavior of performing many or few gaze-shifts away from the target was stable across conditions. We calculated Cronbach's α between conditions and found high internal consistency across conditions in the number of gaze-shifts performed as well as in the percent of gaze-time away from the target speaker ($\alpha = 0.889$ and $\alpha = 0.832$, respectively). This was further demonstrated by strong positive correlations between the percent time spent gazing away from the target speaker in No Distraction condition vs. each of the Distraction conditions (lmer: all r's > 0.5; robustlmm all r's > 0.6) as well as the number of gaze-shifts (lmer and robustlmm: all r's > 0.5; **Figures 4C,D**). This pattern suggests that individuals have characteristic tendencies to either stay focused or gaze-around the scene, above and beyond the specific sensory attributes or degree of distraction in a particular scenario.

Gaze-Locked Analysis of Speech Acoustics

Last, we tested whether there was any relationship between the timing of gaze-shifts and the local speech-acoustics. To this end, we performed a gaze-shift-locked analysis of the envelope of the target or distractor speech (when present). Analysis of distractor speech envelope consisted only of eye-gaze shifts *toward that distractor* (i.e., excluding shifts to other places in the environment). **Figure 5** shows the average time-course of the target and distractor speech envelopes relative to the onset of a gaze-shift. For both target speech (top row) as well as for distractor speech (bottom row), gaze-shifts seem to have been preceded by a brief period of silence (within the lower 5% tile; red shading) between 200 and 300 ms prior to the shift.

Frequency-resolved analysis of the SNR between target and distractor speech similarly indicates low SNR in the period preceding gaze-shifts. A reduction in SNR prior to gaze-shifts was primarily evident in the 3–8 kHz range (sometimes considered the "unvoiced" part of the speech spectrum; Atal and Hanauer, 1971), whereas SNR in the lower part of the spectrum (0–2 kHz) was near 1 dB both before and after gaze-shifts. Although SNR does not take into account the overall loudness-level of each speaker but only the ratio between the speakers, the observed SNR modulation is consistent with momentary periods of silence/drops in the volume of both concurrent speakers.

This pattern is in line with an acoustic release-from-masking account, suggesting that gaze-shifts are prompted by momentary gaps in the speech, and particularly when gaps in concurrent

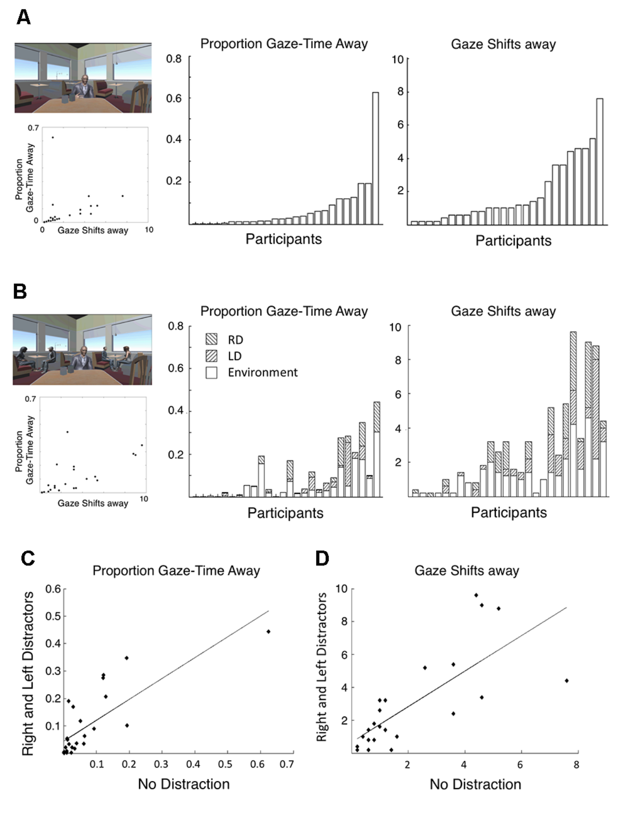

FIGURE 4 | Individual gaze-shift patterns. **(A,B)** Proportion of time spent gazing away from the target speaker (left) and average number of gaze-shifts per trial (right) in the NoD condition **(A)** and the RLD conditions **(B)**, across individual participants. In both cases, participant order is sorted by the NoD condition (top panels). Scatter plots on the left indicate the relationship between the number of gaze-shifts and the proportion gaze-time away, across all participants in each condition. **(C,D)** Scatter plots depicting the relationship between the proportion of time spent gazing away from the target speaker **(C)** and average number of gaze-shifts per trial **(D)**, in the two extreme conditions: NoD vs. RLD. Correlations were significant in both cases ($r > 0.5$).

speech coincide-temporally (as seen here in the Single and Two Distractor conditions). Conversely, the suggestion that attention-shifts are a product of exogenous capture by salient events in distracting speech does not seem to be supported by the current data, since the acoustics of the distractor speech that participants shifted their gaze towards did not seem to contain periods with consistently loud acoustics. We did, however, find increases in loudness of the target speech acoustics near gaze-shift onset (within the top 5% tile; red shading between −100 and +50 ms).

DISCUSSION

The current study is a first and novel attempt to characterize how individuals deploy overt attention in naturalistic audiovisual settings, laden with rich and competing stimuli. By monitoring eye-gaze dynamics in our Virtual Café, we studied the patterns of gaze-shifts and its consequences for speech comprehension. Interestingly, we found that the presence and number of competing speakers in the environment did not, on average, affect the amount of time spent looking at the target speaker, nor did it impair comprehension of the target speaker, although participants did perform slightly more gaze-shifts away in the two-distractor RLD condition. This demonstrates an overall resilience of the attention and speech-processing systems for overcoming the acoustic-load posed by distractors in naturalistic audio-visual conditions. This ability is of utmost ecological value, and likely benefits both from the availability of visual and spatial cues (Freyman et al., 2004; Zion Golumbic et al., 2013a) as well as the use of semantic context to maintain comprehension despite possible reductions in speech intelligibility (Simpson and Cooke, 2005; Vergauwe et al., 2010; Ding and Simon, 2012a; Calandruccio et al., 2018). At the same time, our results also suggest that the ability to maintain attention on the designated speaker under these conditions is highly individualized. Participants displayed characteristic patterns of either staying focused on a target speaker or sampling other locations in the environment overtly, regardless of the severity of the so-called sensory distraction. Critically, the amount of time that individuals spent looking around the environment and away from the target speaker was negatively correlated with speech comprehension, directly linking overt attention to speech comprehension. We also found that gaze-shifts away from the target speaker occurred primarily following gaps in the acoustic input, suggesting that momentary reductions in acoustic masking can prompt attention-shifts between competing speakers, in line with "glimpsing" theories of processing speech in noise. These results open a new window into understanding the dynamics of attention as they wax and wane over time, and the listening patterns exhibited by individuals for dealing with the influx of sensory input in complex naturalistic environments.

Is Attention Stationary?

An underlying assumption of many experimental studies is that participants allocate attention solely to task-relevant stimuli, and that attention remains stationary over time. However, this assumption is probably unwarranted (Weissman et al., 2006; Esterman et al., 2013) since sustaining attention over long periods of time is extremely taxing (Schweizer and Moosbrugger, 2004; Warm et al., 2008; Avisar and Shalev, 2011), and individuals spend a large proportion of the time mind-wandering or "off-task" (Killingsworth and Gilbert, 2010; Boudewyn and Carter, 2018; but see Seli et al., 2018). Yet, empirically testing the studying the frequency and characteristics of attention shifts is operationally difficult since it pertains to participants' internal state that experimenters do not have direct access to. The use of eye-gaze position as a continuous metric for the locus of momentary overt attention in a dynamic scene in the current study contributes to this endeavor.

Here, we found that indeed, in many participants eye-gaze was not maintained on the target speaker throughout the entire trial. Roughly 30% of participants spent over 10% of each trial looking at places in the environment other than the to-be-attended speaker, across all conditions. Interestingly, this proportion is similar to that reported in previous studies for

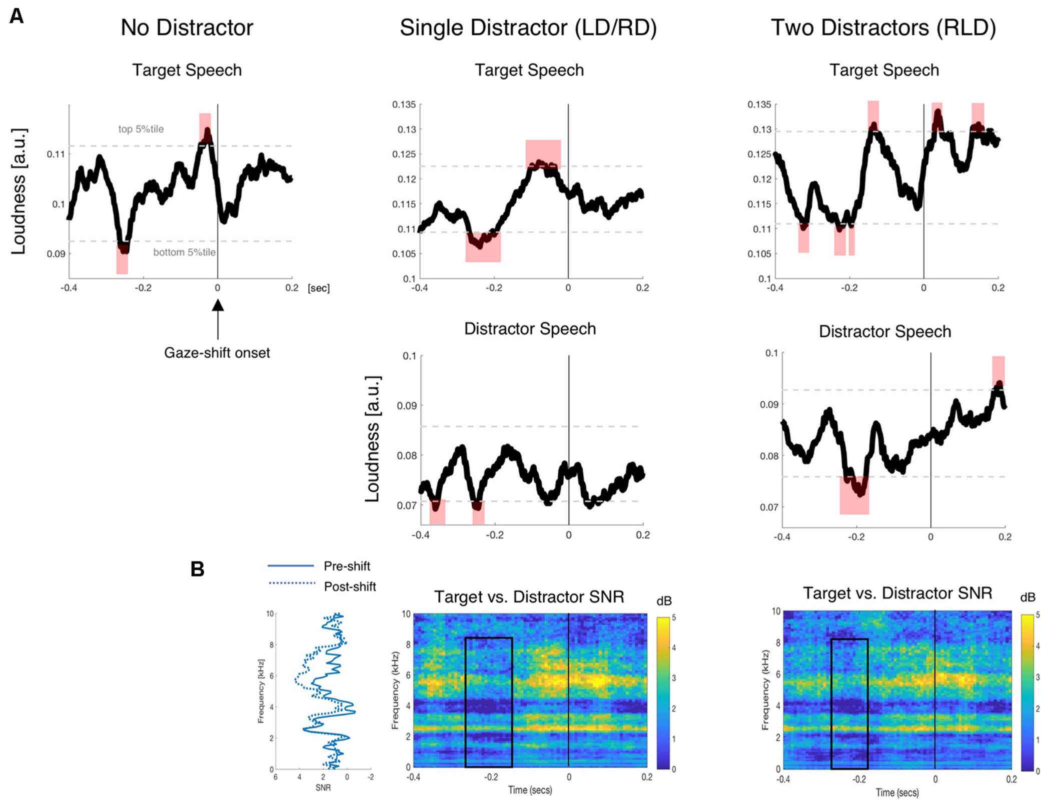

FIGURE 5 | Gaze-shift locked analysis of speech acoustics. **(A)** Average time-course of the target (top) and distractor (bottom) speech envelopes relative to gaze shift onset (*t* = 0). Horizontal dotted gray lines depict the top and bottom 5%tile of loudness values generated through the permutation procedure of non-gaze-locked acoustics segments. The shaded red areas indicate time-periods where the speech sound-level fell within the lower/upper 5% tile of the distribution, respectively. **(B)** Spectrograms depicting the signal-to-noise ratio (SNR) between the target and distractor speaker(s), surrounding the onset of a gaze-shift, in the single and two-distractor conditions. A reduction in SNR is seen in a 200 ms pre-shift time window, primarily in the higher "unvoiced" portion of the spectrogram (4–8 KHz).

the prevalence of detecting ones' own name in a so-called unattended message (Cherry, 1953; Wood and Cowan, 1995), an effect attributed by some to rapid attention shifts (Lachter et al., 2004; Beaman et al., 2007; Lin and Yeh, 2014). Although in the current study we did not test whether these participants also gleaned more information from distractors' speech, we did find that comprehension of the target speaker was reduced as a function of the time spent looking away from the target speaker. Participants were also more likely to miss information from the target-speech during gaze-shifts away, yielding slightly higher mistake-rates. These results emphasize the dynamic nature of attention and attention-shifts, and demonstrate that brief overt attention-shifts can negatively impact speech processing in ecological multi-speaker and multisensory contexts.

They also highlight the importance of studying individual differences in attentional control. In the current study set, we did not collect additional personal data from participants which may have shed light on the source of the observed variability in gaze-patterns across individuals. However, based on previous literature, individual differences may stem from factors such as susceptibility to distraction (Ellermeier and Zimmer, 1997; Cowan et al., 2005; Avisar and Shalev, 2011; Bourel-Ponchel et al., 2011; Forster and Lavie, 2014; Hughes, 2014), working memory capacity (Conway et al., 2001; Kane and Engle, 2002; Tsuchida et al., 2012; Sörqvist et al., 2013; Hughes, 2014; Naveh-Benjamin et al., 2014; Wiemers and Redick, 2018) or personality traits (Rauthmann et al., 2012; Risko et al., 2012; Baranes et al., 2015; Hoppe et al., 2018). Additional dedicated research is needed to resolve the source of the individual differences observed here.

Is Eye-Gaze a Good Measure for Attention-Shifts Among Concurrent Speech?

One may ask, to what extent do the current results fully capture the prevalence of attention-shifts, since it is known that these can also occur covertly (Posner, 1980; Petersen and Posner, 2012)? This is a valid concern and indeed the current results should be taken as representing a *lower-bound* for the frequency of attention-shifts and we should assume that attention-shifts are probably more prevalent than observed here. This motivates the future development of complementary methods for quantifying covert shifts of attention among concurrent speech, given the current absence of a reliable metrics.

Another concern that may be raised with regard to the current results is that individuals may maintain attention to the target speaker even while looking elsewhere, and hence the gaze-shifts measured here might not reflect true shifts of attention. Although in principle this could be possible, previous research shows that this is probably not the default mode of listening under natural audiovisual conditions. Rather, a wealth of studies demonstrate a tight link between gaze-shifts and attention-shifts (Chelazzi et al., 1995; Deubel and Schneider, 1996; Grosbras et al., 2005; Szinte et al., 2018) and gaze is widely utilized experimentally as a proxy for the locus of visuospatial attention (Gredebäck et al., 2009; Linse et al., 2017). In multi-speaker contexts, it has been shown that participants tend to move their eyes towards the location of attended speech sounds (Gopher and Kahneman, 1971; Gopher, 1973). Similarly, looking towards the location of distractor-speech significantly reduces intelligibility and memory for attended speech and increases intrusions from distractor speech (Reisberg et al., 1981; Spence et al., 2000; Yi et al., 2013). This is in line with the current finding of a negative correlation between the time spent looking at the target speaker and speech comprehension, and higher mistake-rates during gaze-shifts, which further link overt gaze to selective attention to speech. Studies on audiovisual speech processing further indicate that looking at the talking face increases speech intelligibility and neural selectivity for attended speech (Sumby and Pollack, 1954; Zion Golumbic et al., 2013a; Lou et al., 2014; Crosse et al., 2016; Park et al., 2016), even when the video is not informative about the content of speech (Kim and Davis, 2003; Schwartz et al., 2004), and eye-gaze is particularly utilized for focusing attention to speech under adverse listening condition (Yi et al., 2013). Taken together, current findings support the interpretation that gaze-shifts reflect shifts in attention away from the target speaker, in line with the limited resources perspective of attention (Lavie et al., 2004; Esterman et al., 2014), making eye-gaze a useful and reliable metric for studying the dynamics of attention to naturalistic audio-visual speech. Interestingly, this metric has recently been capitalized on for use in assistive listening devices, utilizing eye-gaze direction to indicate the direction of a listener's attention (Favre-Felix et al., 2017; Kidd, 2017). That said, gaze-position is likely only one of several factors in determining successful speech comprehension in multi-speaker environments (e.g., SNR level, audio-visual congruency, engagement in content etc.), as suggested by the significant yet still moderate effect-sizes found here.

Listening Between the Gaps—What Prompts Attention Shifts Among Concurrent Speech?

Besides characterizing the prevalence and behavioral consequences of attention-shifts in audio-visual multi-talker contexts, it is also critical to understand what prompts these shifts. Here we tested whether there are aspects of the scene acoustics that can be associated with attention-shifts away from the target speaker. We specifically tested two hypotheses: (1) that attention is captured exogenously by highly salient sensory events in distracting speech (Wood and Cowan, 1995; Itti and Koch, 2000; Kayser et al., 2005); and (2) that attention-shifts occur during brief pauses in speech acoustics that momentarily unmask the competing sounds (Lavie et al., 2004; Cooke, 2006).

Regarding the first hypothesis, the current data suggest that distractor saliency is not a primary factor in prompting gaze-shifts. Since gaze-shifts were just as prevalent in the NoD condition as in conditions that contained distractors and since no consistent increase in distractor loudness was observed near gaze-shifts, we conclude that the gaze-shifts performed by participants do not necessarily reflect exogenous attentional capture by distractor saliency. This is in line with previous studies suggesting that sensory saliency is less effective in drawing exogenous attention in dynamic scenarios relative to the stationary contexts typically used in laboratory experiments (Smith et al., 2013).

Rather, our current results seem to support the latter hypothesis that attention-shifts are prompted by momentary acoustic release-from-masking. We find that gaze-shifts occurred more consistently ∼200–250 ms after instances of low acoustic intensity in both target and distractor sounds and low SNR. This time-scale is on-par with the initiation time for saccades (Gilchrist, 2011), and suggests that momentary reduction in masking provide an opportunity for the system to shift attention between speakers. This pattern fits with accounts for comprehension of speech-in-noise, suggesting that listeners utilize brief periods of unmasking or low SNR to glean and piece together information for deciphering speech content ("acoustic glimpsing"; Cooke, 2006; Li and Loizou, 2007; Vestergaard et al., 2011; Rosen et al., 2013). Although this acoustic-glimpsing framework is often used to describe how listeners maintain intelligibility of target-speech in noise, it has not been extensively applied to studying *shifts* of attention among concurrent speech. The current results suggest that brief gaps in the audio or periods of low SNR may serve as triggers for momentary attention shifts, which can manifest overtly (as demonstrated here), and perhaps also covertly. Interestingly, a previous study found that eye-blinks also tend to occur more often around pauses when viewing and listening to audio-visual speech (Nakano and Kitazawa, 2010), pointing to a possible link between acoustic glimpsing and a reset in the oculomotor system, creating optimal conditions for momentary attention-shifts.

CONCLUSION

There is growing understanding that in order to really understand the human cognitive system, it needs to be studied in contexts relevant for real-life behavior, and that tightly constrained artificial laboratory paradigms do not always generalize to real-life (Kingstone et al., 2008; Marius't Hart et al., 2009; Foulsham et al., 2011; Risko et al., 2016; Rochais et al., 2017; Hoppe et al., 2018). The current study represents the attempt to bridge this gap between the laboratory and real-life, by studying how individuals spontaneously deploy overt attention in a naturalistic virtual-reality environment. Using this approach, the current study highlights the characteristics and individual differences in selective attention to speech under naturalistic listening conditions. This pioneering work opens up new horizons for studying how attention operates in real-life and understanding the factors contributing to success as well as the difficulties in paying attention to speech in noisy environments.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study was approved by the Institutional Ethics Committee at Bar-Ilan University, and the research was conducted according to the guidelines of the committee. Signed informed consent was obtained from each participant prior to the experiment.

AUTHOR CONTRIBUTIONS

EZG designed the study, oversaw data collection and analysis. KS-C collected and analyzed the data. Both authors wrote the article.

FUNDING

This work was supported by the Israel Science Foundation I-Core Center for Excellence 51/11, and by the United States–Israel Binational Science Foundation grant #2015385.

REFERENCES

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., and Donk, M. (2015). It depends on *when* you look at it: salience influences eye movements in natural scene viewing and search early in time. *J. Vis.* 15:9. doi: 10.1167/15.5.9

Atal, B. S., and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.* 50, 637–655. doi: 10.1121/1.1912679

Avisar, A., and Shalev, L. (2011). Sustained attention and behavioral characteristics associated with ADHD in adults. *Appl. Neuropsychol.* 18, 107–116. doi: 10.1080/09084282.2010.547777

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Baranes, A., Oudeyer, P.-Y., and Gottlieb, J. (2015). Eye movements reveal epistemic curiosity in human observers. *Vision Res.* 117, 81–90. doi: 10.1016/j.visres.2015.10.009

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Beaman, C. P., Bridges, A. M., and Scott, S. K. (2007). From dichotic listening to the irrelevant sound effect: a behavioural and neuroimaging analysis of the processing of unattended speech. *Cortex* 43, 124–134. doi: 10.1016/s0010-9452(08)70450-7

Bompas, A., and Sumner, P. (2009). Temporal dynamics of saccadic distraction. *J. Vis.* 9, 17–17. doi: 10.1167/9.9.17

Boudewyn, M. A., and Carter, C. S. (2018). I must have missed that: α-band oscillations track attention to spoken language. *Neuropsychologia* 117, 148–155. doi: 10.1016/j.neuropsychologia.2018.05.024

Bourel-Ponchel, E., Querné, L., Le Moing, A. G., Delignières, A., de Broca, A., and Berquin, P. (2011). Maturation of response time and attentional control in ADHD: evidence from an attentional capture paradigm. *Eur. J. Paediatr. Neurol.* 15, 123–130. doi: 10.1016/j.ejpn.2010.08.008

Broadbent, D. E. (1958). "Selective listening to speech," in *Perception and Communication*, ed. D. E. Broadbent (Elmsford, NY: Pergamon Press), 11–35.

Calandruccio, L., Buss, E., Bencheck, P., and Jett, B. (2018). Does the semantic content or syntactic regularity of masker speech affect speech-on-speech recognition? *J. Acoust. Soc. Am.* 144, 3289–3302. doi: 10.1121/1.5081679

Chelazzi, L., Biscaldi, M., Corbetta, M., Peru, A., Tassinari, G., and Berlucchi, G. (1995). Oculomotor activity and visual spatial attention. *Behav. Brain Res.* 71, 81–88. doi: 10.1016/0166-4328(95)00134-4

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi: 10.1121/1.1907229

Conway, A. R. A., Cowan, N., and Bunting, M. F. (2001). The cocktail party phenomenon revisited: the importance of working memory capacity. *Psychon. Bull. Rev.* 8, 331–335. doi: 10.3758/bf03196169

Cooke, M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* 119, 1562–1573. doi: 10.1121/1.2166600

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cogn. Psychol.* 51, 42–100. doi: 10.1016/j.cogpsych.2004.12.001

Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (2016). Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *J. Neurosci.* 36, 9888–9895. doi: 10.1523/jneurosci.1396-16.2016

Deubel, H., and Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Res.* 36, 1827–1837. doi: 10.1016/0042-6989(95)00294-4

Deutsch, J. A., and Deutsch, D. (1963). Attention: some theoretical considerations. *Psychol. Rev.* 70, 80–90. doi: 10.1037/h0039515

Ding, N., and Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U S A* 109, 11854–11859. doi: 10.1073/pnas.1205381109

Ding, N., and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* 107, 78–89. doi: 10.1152/jn.00297.2011

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychol. Rev.* 87, 272–300. doi: 10.1037/0033-295x.87.3.272

Ellermeier, W., and Zimmer, K. (1997). Individual differences in susceptibility to the "irrelevant speech effect". *J. Acoust. Soc. Am.* 102, 2191–2199. doi: 10.1121/1.419596

Escera, C., Yago, E., Corral, M.-J., Corbera, S., and Nuñez, M. I. (2003). Attention capture by auditory significant stimuli: semantic analysis follows attention switching. *Eur. J. Neurosci.* 18, 2408–2412. doi: 10.1046/j.1460-9568.2003.02937.x

Esterman, M., Noonan, S. K., Rosenberg, M., and Degutis, J. (2013). In the zone or zoning out? Tracking behavioral and neural fluctuations during sustained attention. *Cereb. Cortex* 23, 2712–2723. doi: 10.1093/cercor/bhs261

Esterman, M., Rosenberg, M. D., and Noonan, S. K. (2014). Intrinsic fluctuations in sustained attention and distractor processing. *J. Neurosci.* 34, 1724–1730. doi: 10.1523/JNEUROSCI.2658-13.2014

Favre-Felix, A., Graversen, C., Dau, T., and Lunner, T. (2017). "Real-time estimation of eye gaze by in-ear electrodes. in," in *Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Seogwipo: IEEE), 4086–4089.

Forster, S., and Lavie, N. (2014). Distracted by your mind? Individual differences in distractibility predict mind wandering. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 251–260. doi: 10.1037/a0034108

Foulsham, T., Walker, E., and Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Res.* 51, 1920–1931. doi: 10.1016/j.visres.2011.07.002

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *J. Acoust. Soc. Am.* 115, 2246–2256. doi: 10.1121/1.1689343

Gilchrist, I. D. (2011). "Saccades," in *The Oxford Handbook of Eye Movements*, eds S. P. Liversedge, I. D. Gilchrist and S. Everling (Oxford, UK: Oxford University Press), 85–94.

Gopher, D. (1973). Eye-movement patterns in selective listening tasks of focused attention. *Percept. Psychophys.* 14, 259–264. doi: 10.3758/bf03212387

Gopher, D., and Kahneman, D. (1971). Individual differences in attention and the prediction of flight criteria. *Percept. Mot. Skills* 33, 1335–1342. doi: 10.2466/pms.1971.33.3f.1335

Gredebäck, G., Johnson, S., and von Hofsten, C. (2009). Eye tracking in infancy research. *Dev. Neuropsychol.* 35, 1–19. doi: 10.1080/875656409033 25758

Grosbras, M. H., Laird, A. R., and Paus, T. (2005). Cortical regions involved in eye movements, shifts of attention and gaze perception. *Hum. Brain Mapp.* 25, 140–154. doi: 10.1002/hbm.20145

Hoppe, S., Loetscher, T., Morey, S. A., and Bulling, A. (2018). Eye movements during everyday behavior predict personality traits. *Front. Hum. Neurosci.* 12:105.doi: 10.3389/fnhum.2018.00105

Hughes, R. W. (2014). Auditory distraction: a duplex-mechanism account. *Psych J.* 3, 30–41. doi: 10.1002/pchj.44

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40, 1489–1506. doi: 10.1016/s0042-6989(99)00163-7

Kane, M. J., and Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention and general fluid intelligence: an individual-differences perspective. *Psychon. Bull. Rev.* 9, 637–671. doi: 10.3758/bf03196323

Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Curr. Biol.* 15, 1943–1947. doi: 10.1016/j.cub.2005.09.040

Kidd, G.Jr. (2017). Enhancing auditory selective attention using a visually guided hearing aid. *J. Speech Lang. Hear. Res.* 60, 3027–3038. doi: 10.1044/2017_JSLHR-H-17-0071

Killingsworth, M. A., and Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science* 330:932.doi: 10.1126/science.1192439

Kim, J., and Davis, C. (2003). Hearing foreign voices: does knowing what is said affect visual-masked-speech detection? *Perception* 32, 111–120. doi: 10.1068/p3466

Kingstone, A., Smilek, D., and Eastwood, J. D. (2008). Cognitive ethology: a new approach for studying human cognition. *Br. J. Psychol.* 99, 317–340. doi: 10.1348/000712607X251243

Koller, M. (2016). robustlmm: an R package for robust estimation of linear mixed-effects models. *J. Stat. Softw.* 75, 1–24. doi: 10.18637/jss.v075.i06

Lachter, J., Forster, K. I., and Ruthruff, E. (2004). Forty-five years after broadbent (1958): still no identification without attention. *Psychol. Rev.* 111, 880–913. doi: 10.1037/0033-295X.111.4.880

Lavie, N., Hirst, A., de Fockert, J. W., and Viding, E. (2004). Load theory of selective attention and cognitive control. *J. Exp. Psychol. Gen.* 133, 339–354. doi: 10.1037/0096-3445.133.3.339

Li, N., and Loizou, P. C. (2007). Factors influencing glimpsing of speech in noise. *J. Acoust. Soc. Am.* 122, 1165–1172. doi: 10.1121/1.2749454

Lin, S.-H., and Yeh, Y.-Y. (2014). Attentional load and the consciousness of one's own name. *Conscious. Cogn.* 26, 197–203. doi: 10.1016/j.concog.2014.03.008

Linse, K., Rüger, W., Joos, M., Schmitz-Peiffer, H., Storch, A., and Hermann, A. (2017). Eye-tracking-based assessment suggests preserved well-being in locked-in patients. *Ann. Neurol.* 81, 310–315. doi: 10.1002/ana.24871

Lou, Y., Yoon, J. W., and Huh, H. (2014). Modeling of shear ductile fracture considering a changeable cut-off value for stress triaxiality. *Int. J. Plast.* 54, 56–80. doi: 10.1016/j.ijplas.2013.08.006

Marius't Hart, B. M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., König, P., et al. (2009). Gaze allocation in natural stimuli: comparing free exploration to head-fixed viewing conditions. *Vis. cogn.* 17, 1132–1158. doi: 10.1080/13506280902812304

McDermott, J. H. (2009). The cocktail party problem. *Curr. Biol.* 19, R1024–R1027. doi: 10.1016/j.cub.2009.09.005.

Mesgarani, N., and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485, 233–236. doi: 10.1038/nature11020

Nakano, T., and Kitazawa, S. (2010). Eyeblink entrainment at breakpoints of speech. *Exp. Brain Res.* 205, 577–581. doi: 10.1007/s00221-010-2387-z

Naveh-Benjamin, M., Kilb, A., Maddox, G. B., Thomas, J., Fine, H. C., Chen, T., et al. (2014). Older adults do not notice their names: a new twist to a classic attention task. *J. Exp. Psychol. Learn. Mem. Cogn.* 40, 1540–1550. doi: 10.1037/xlm0000020

Neely, C., and LeCompte, D. (1999). The importance of semantic similarity to the irrelevant speech effect. *Mem. Cogn.* 27, 37–44. doi: 10.3758/bf032 01211

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Park, H., Kayser, C., Thut, G., and Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *Elife* 5:e14521. doi: 10.7554/elife.14521

Parmentier, F. B. R., Pacheco-Unguetti, A. P., and Valero, S. (2018). Food words distract the hungry: evidence of involuntary semantic processing of task-irrelevant but biologically-relevant unexpected auditory words. *PLoS One* 13:e0190644. doi: 10.1371/journal.pone.0190644

Petersen, S. E., and Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annu. Rev. Neurosci.* 35, 73–89. doi: 10.1146/annurev-neuro-062111-150525

Posner, M. I. (1980). Orienting of attention. *Q. J. Exp. Psychol.* 32, 3–25. doi: 10.1080/00335558008248231

Rauthmann, J. F., Seubert, C. T., Sachse, P., and Furtner, M. R. (2012). Eyes as windows to the soul: gazing behavior is related to personality. *J. Res. Pers.* 46, 147–156. doi: 10.1016/j.jrp.2011.12.010

Raveh, D., and Lavie, N. (2015). Load-induced inattentional deafness. *Atten. Percept. Psychophys.* 77, 483–492. doi: 10.3758/s13414-014-0776-2

Reisberg, D., Scheiber, R., and Potemken, L. (1981). Eye position and the control of auditory attention. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 318–323. doi: 10.1037/0096-1523.7.2.318

Risko, E. F., Anderson, N. C., Lanthier, S., and Kingstone, A. (2012). Curious eyes: individual differences in personality predict eye movement behavior in scene-viewing. *Cognition* 122, 86–90. doi: 10.1016/j.cognition.2011.08.014

Risko, E. F., Richardson, D. C., and Kingstone, A. (2016). Breaking the fourth wall of cognitive science. *Curr. Dir. Psychol. Sci.* 25, 70–74. doi: 10.1177/0963721415617806

Rochais, C., Henry, S., and Hausberger, M. (2017). Spontaneous attention-capture by auditory distractors as predictor of distractibility: a study of domestic horses *(Equus caballus)*. *Sci. Rep.* 7:15283. doi: 10.1038/s41598-017-15654-5

Röer, J. P., Körner, U., Buchner, A., and Bell, R. (2017). Attentional capture by taboo words: a functional view of auditory distraction. *Emotion* 17, 740–750. doi: 10.1037/emo0000274

Rosen, S., Souza, P., Ekelund, C., and Majeed, A. A. (2013). Listening to speech in a background of other talkers: effects of talker number and noise vocoding. *J. Acoust. Soc. Am.* 133, 2431–2443. doi: 10.1121/1.4794379

R Development Core Team. (2012). R: A Language and Environment for Statistical Computing. Vienna: R foundation for Statistical Computing. Available online at: http://www.R-project.org/.

Schomaker, J., Walper, D., Wittmann, B. C., and Einhäuser, W. (2017). Attention in natural scenes: affective-motivational factors guide gaze independently of visual salience. *Vision Res.* 133, 161–175. doi: 10.1016/j.visres.2017.02.003

Schwartz, J.-L., Berthommier, F., and Savariaux, C. (2004). Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition* 93, B69–B78. doi: 10.1016/s0010-0277(04)00054-x

Schweizer, K., and Moosbrugger, H. (2004). Attention and working memory as predictors of intelligence. *Intelligence* 32, 329–347. doi: 10.1016/j.intell.2004.06.006

Seli, P., Beaty, R. E., Cheyne, J. A., Smilek, D., Oakman, J., and Schacter, D. L. (2018). How pervasive is mind wandering, really? *Conscious. Cogn.* 66, 74–78. doi: 10.1016/j.concog.2018.10.002

Simpson, S. A., and Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *J. Acoust. Soc. Am.* 118, 2775–2778. doi: 10.1121/1.2062650

Smith, T. J., Lamont, P., and Henderson, J. M. (2013). Change blindness in a dynamic scene due to endogenous override of exogenous attentional cues. *Perception* 42, 884–886. doi: 10.1068/p7377

Sörqvist, P., Marsh, J. E., and Nöstl, A. (2013). High working memory capacity does not always attenuate distraction: bayesian evidence in support of the null hypothesis. *Psychon. Bull. Rev.* 20, 897–904. doi: 10.3758/s13423-013-0419-y

Spence, C., Ranson, J., and Driver, J. (2000). Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention. *Percept. Psychophys.* 62, 410–424. doi: 10.3758/bf03205560

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Szinte, M., Jonikaitis, D., Rangelov, D., and Deubel, H. (2018). Pre-saccadic remapping relies on dynamics of spatial attention. *Elife* 7:e37598. doi: 10.7554/elife.37598

Treisman, A. M. (1964). The effect of irrelevant material on the efficiency of selective listening. *Am. J. Psychol.* 77, 533–546. doi: 10.2307/1420765

Tsuchida, Y., Katayama, J., and Murohashi, H. (2012). Working memory capacity affects the interference control of distractors at auditory gating. *Neurosci. Lett.* 516, 62–66. doi: 10.1016/j.neulet.2012.03.057

Vergauwe, E., Barrouillet, P., and Camos, V. (2010). Do mental processes share a domain-general resource? *Psychol. Sci.* 21, 384–390. doi: 10.1177/0956797610361340

Vestergaard, M. D., Fyson, N. R. C., and Patterson, R. D. (2011). The mutual roles of temporal glimpsing and vocal characteristics in cocktail-party listening. *J. Acoust. Soc. Am.* 130, 429–439. doi: 10.1121/1.3596462

Walker, F., Bucker, B., Anderson, N. C., Schreij, D., and Theeuwes, J. (2017). Looking at paintings in the vincent van gogh museum: eye movement patterns of children and adults. *PLoS One* 12:e0178912. doi: 10.1371/journal.pone.0178912

Warm, J. S., Parasuraman, R., and Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Hum. Factors J. Hum. Factors Ergon. Soc.* 50, 433–441. doi: 10.1518/001872008X312152

Weissman, D. H., Roberts, K. C., Visscher, K. M., and Woldorff, M. G. (2006). The neural bases of momentary lapses in attention. *Nat. Neurosci.* 9, 971–978. doi: 10.1038/nn1727

Wiemers, E. A., and Redick, T. S. (2018). Working memory capacity and intra-individual variability of proactive control. *Acta Psychol.* 182, 21–31. doi: 10.1016/j.actpsy.2017.11.002

Wood, N., and Cowan, N. (1995). The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? *J. Exp. Psychol. Learn. Mem. Cogn.* 21, 255–260. doi: 10.1037/0278-7393.21.1.255

Yi, A., Wong, W., and Eizenman, M. (2013). Gaze patterns and audiovisual speech enhancement. *J. Speech Lang. Hear. Res.* 56, 471–480. doi: 10.1044/1092-4388(2012/10-0288)

Zion Golumbic, E. M., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013a). Visual input enhances selective speech envelope tracking in auditory cortex at a "Cocktail Party". *J. Neurosci.* 33, 1417–1426. doi: 10.1523/JNEUROSCI.3675-12.2013

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013b). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* 77, 980–991. doi: 10.1016/j.neuron.2012.12.037

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.