Trends in Hearing
Volume 27: 1–9
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165231152356
journals.sagepub.com/home/tia
SAGE

An Effect of Gaze Direction in Cocktail Party Listening

Virginia Best[1], Alex D. Boyd[2] and Kamal Sen[2]

Abstract

It is well established that gaze direction can influence auditory spatial perception, but the implications of this interaction for performance in complex listening tasks is unclear. In the current study, we investigated whether there is a measurable effect of gaze direction on speech intelligibility in a "cocktail party" listening situation. We presented sequences of digits from five loudspeakers positioned at 0°, ± 15°, and ± 30° azimuth, and asked participants to repeat back the digits presented from a designated target loudspeaker. In different blocks of trials, the participant visually fixated on a cue presented at the target location or at a nontarget location. Eye position was tracked continuously to monitor compliance. Performance was best when fixation was on-target (vs. off-target) and the size of this effect depended on the specific configuration. This result demonstrates an influence of gaze direction in multitalker mixtures, even in the absence of visual speech information.

Keywords

multitalker speech, spatial attention, sound localization, crossmodal attention, eye tracking

Received 5 August 2022; Revised 18 November 2022; accepted 4 January 2023

Introduction

It is well established that there are interactions between gaze direction and auditory spatial perception. For example, gaze direction has been shown to influence sound localization (e.g., Getzmann, 2002; Jones & Kabanoff, 1975; Lewald, 1998), and auditory spatial resolution is improved in the vicinity of visual fixation under certain conditions (Maddox et al., 2014; Rorden & Driver, 1999), although this effect has not always been replicated (Wood & Bizley, 2015). Effects of gaze direction have also been observed in various contexts involving competing sounds. For example, Pomper and Chait (2017) used a task involving competing streams of tones at different horizontal locations, and reported reduced reaction times for detecting deviant tones in a target stream when the participant fixated on the target loudspeaker rather than a distractor loudspeaker.

Visual behavior plays a special role in speech communication. Eye movements to a talkers face and lips provide access to redundant speech information that can improve intelligibility especially under difficult listening conditions (Hadley et al., 2019; Hendrikse et al., 2019; Sumby & Pollack, 1954). However, even in the absence of lip-reading cues, people tend to move their eyes toward the location of a talker that they are attending to in the presence of competition (Gopher, 1973). Eye movements are thus presumed to represent a general pattern of orientation that comes into play when the listening task is effortful. However, there is a scarcity of compelling evidence to suggest that gaze direction actually influences speech intelligibility in challenging situations. A handful of studies measured the recall of speech from a target talker in the presence of competing talker, but the results are mixed, with some but not all studies finding a beneficial effect of fixating on the target (e.g., Reisberg et al., 1981; Wolters & Schiano, 1989). Driver and Spence (1994) had listeners shadow speech, presented to either the left or right side (at ± 30° azimuth), in the presence of a distractor on the other side. They found that fixating to the correct side was beneficial when there was relevant visual information presented (lip movements) but not when the visual information was irrelevant (chewing movements), which they interpreted to mean that gaze direction

[1] Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, USA
[2] Department of Biomedical Engineering, Boston University, Boston, MA, USA

Corresponding Author:
Virginia Best, Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA 02215, USA.
Email: ginbest@bu.edu

itself was not important. Recently, this question was revisited by Fleming et al. (2021), who presented competing talkers at ± 15° azimuth, in both audiovisual and audio-only conditions. In the audiovisual condition, a mismatch between the target auditory and visual locations reduced word recall accuracy by about 10 percentage points. In the audio-only condition, there was a smaller decrement in performance when the participant fixated to the opposite side (around 5 percentage points). In that study (and most of the previous studies) spatial attention was defined rather broadly (left vs. right hemifield) and it remains unclear whether spatial alignment at a finer level is also important. Moreover, eye position was not monitored closely, and thus effects may have been modulated by how well participants were able to follow the instructions and maintain fixation.

The aim of the current study was to determine if there is a measurable effect of gaze direction in "cocktail party" listening situations involving multiple competing talkers. The study was motivated by an intuition that gaze direction may be especially relevant in complex listening situations where spatial attention is required to select the source of interest (Shinn-Cunningham et al., 2017). A secondary motivation was our interest in visually guided beamforming (Kidd, 2017) where the gaze direction is used to steer a highly directional hearing device. To optimize the benefits provided by such a device in everyday listening situations, it is critically important to understand the interplay between gaze direction and selective listening in multitalker situations.

Methods

Participants

Participants were 10 young adults with audiometrically normal hearing and self-reported normal (or corrected-to-normal) vision. They ranged in age from 18 to 23 years (mean 21 years) and were recruited from the Boston University community. All procedures were approved by the Boston University Institutional Review Board.

Equipment

The experiments took place in a single-walled sound booth (Industrial Acoustics Company) with interior dimensions of 3.8 × 4.0 × 2.3 m (length × width × height), with perforated metal panels on the ceiling and walls and a carpeted floor. Stimuli were presented via five loudspeakers (Acoustic Research 215PS) located on a horizontal arc of radius 1.5 m at azimuths of 0°, ± 15°, and ± 30°. The loudspeakers were driven by a Lenovo PC, a multichannel soundcard (MOTU 16A) and a bank of power amplifiers (Crown Audio XTi 1002) that were all located outside of the booth. For the presentation of instructions and visual cues, an array of ASUS 16" monitors was situated above and just slightly behind the array of loudspeakers.

The participant was seated on a chair in the center of the array, with their head stabilized by a neck rest attached to the back of the chair. A wearable eye tracker provided a continuous estimate of eye position. For the first five participants, we used a Pupil Core headset (Pupil Labs GmbH), which was configured with one camera per eye and a 60 Hz sampling rate. For the second five participants, we used Tobii Pro Glasses 3 (Tobii Technology, Inc.), which have two sensors per eye and a sampling rate of 100 Hz. We switched to the Tobii system as it gave a more stable readout, was easier to calibrate, and had additional advantages for our purposes including simultaneous headtracking and the ability to add corrective lenses. Headtracking made use of an outward facing scene camera and a detectable landmark placed in the room at 0° azimuth. Because the scene camera operated at 25 Hz, the eye tracker readout was downsampled to 25 Hz (by averaging every four samples) to produce coincident estimates of head and eye position. This downsampling had the additional advantage of smoothing the eye tracker data and reducing the impact of occasional dopped samples. Only the horizontal co-ordinates of the head and eye position estimates were considered.

Participants provided responses via a handheld backlit keypad. MATLAB software (MathWorks) was used for stimulus generation, stimulus presentation, data acquisition, and analysis.

Stimuli and Task

We modified an experimental setup that has been used in several previous studies to investigate spatial attention in multitalker speech mixtures (Best et al., 2008, Best et al., 2018). In this setup, five sequences of digits are presented from five different azimuths (Figure 1A). Each sequence is comprised of four digits (from the set 1–9). In the current study, the digits were spoken by 12 male talkers, and had an average duration of 593 ms. They were drawn at random, with the constraint that for each of the four temporal positions, the five simultaneous digits (and their talkers) were different. The digits in each temporal position were time aligned at their onsets and zero padded such that the duration of each set of digits was

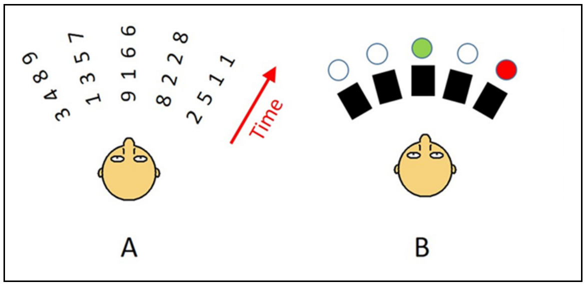

Figure 1. The experimental setup. (A) Stimuli were comprised of five competing sequences of four digits. (B) Visual cues indicated where to listen (red) and where to look (green).

determined by the longest digit for that position. Each digit was presented at a level of 60 dB sound pressure level.

On each trial, two visual cues were provided along with the auditory stimulus (Figure 1B). One cue, which was red in color, indicated which loudspeaker to attend to (i.e., "where to listen"). The participant's task was to report the four digits presented from the target loudspeaker, in order. The other visual cue was green in color and indicated which loudspeaker the participant should fixate on (i.e., "where to look") for the duration of the stimulus. When the "listen" and "look" positions were the same, only the green fixation cue was visible. Trials were organized into blocks of 25 and the target and fixation positions were fixed throughout a block. An instruction screen at the start of each block provided written information about the target position and the fixation position for that block. Across blocks, each combination of three target positions (−30°, 0°, +30° azimuth) and three fixation positions (−30°, 0°, +30° azimuth) was tested. Six blocks per combination were completed by each participant, in a random order. Responses were scored as the percentage of target digits correctly reported across all trials in all blocks. Before the experiment, a brief training block was conducted to familiarize participants with the stimuli and to make sure they understood the two visual cues.

Results

Speech Scores

Inspection of the individual data revealed a small number of blocks (six out of the 540 total blocks) in which performance was at or below chance. These blocks were clear outliers when compared to the other five blocks for that subject in that condition, which were substantially above chance. They were thus excluded from the following analyses. Figure 2 shows individual and mean percent correct scores for the different combinations of target and fixation position. Performance was better for lateral targets than central targets (compare panels), which was expected given that the lateral targets are acoustically more favorable in this configuration (i.e., the signal-to-noise ratio (SNR), at the better ear is higher). Performance was also better and considerably more variable on the right side than on the left. We have no definitive explanation for this asymmetry, but we speculate that it is related to the "right-ear advantage" for speech that is known to be highly variable across people and conditions (Westerhausen & Kompus, 2018). Critically, for a given target position, performance was better when the gaze was directed to that position (filled symbols) than when the gaze was directed elsewhere (open symbols). These observations were confirmed by a repeated-measures analysis of variance (ANOVA), which found significant main effects of target position ($F(2,18) = 79.8$, $p < .001$) and fixation position ($F(2,18) = 10.2, p = .001$), and a significant interaction ($F(4,36) = 7.7$, $p < .001$). Planned comparisons indicated that for the 0° target, on-target fixation was superior to fixation positions of −30° ($t(9) = 3.6$, $p = .006$) and + 30° ($t(9) = 3.4$, $p = .008$). For the −30° target, on-target fixation was superior relative to the fixation position of +30° ($t(9) = 4.7, p = .001$) but not 0° ($t(9) = 1.8, p = .1$). For the +30° target, on-target fixation was superior to the fixation position of −30° ($t(9) = 2.5, p = .03$) but not 0° ($t(9) = 2.0, p = .07$).

Error Patterns

In previous studies involving competing speech stimuli like those used here, a useful distinction has been made between "masker errors," where the response corresponds to a masker word that was presented simultaneously with the target, and "random errors," where the response does

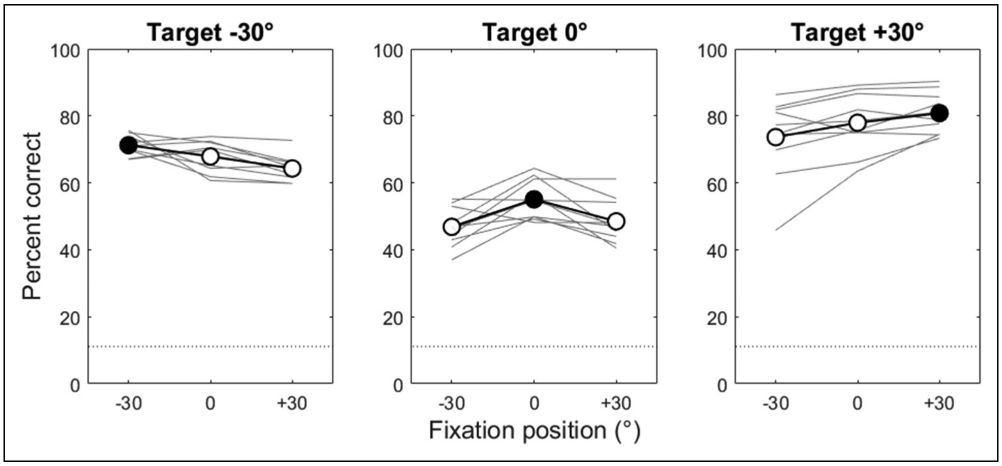

Figure 2. Individual and mean speech scores for different combinations of target and fixation position. Thin gray lines show individual participants and the symbols show across-subject means. The filled symbol in each panel identifies the combination in which the target and fixation position coincided. The dashed line shows chance performance for this task (11%).

not match any of the presented words (e.g., Best et al., 2008, 2018; Brungart, 2001). Masker errors often reflect confusion between competing talkers or attention to the incorrect talker. When the current dataset was analyzed in this way, we found that around 90% of errors could be classified as masker errors. We then looked into the spatial distribution of errors, by sorting responses (each digit at a time) according to the loudspeaker that presented that digit. Figure 3 (top row) shows this distribution for each target position (different panels). The small proportion of responses that did not match a digit in the stimulus (marked "R") are also shown on each panel and can be categorized as random errors. Responses showed a characteristic error pattern overall, with a tendency to report digits arising from loudspeakers adjacent to the target loudspeaker on incorrect trials. This pattern is consistent with previous studies using a similar arrangement of stimuli (Best et al., 2008, 2018). A comparison of the different colored bars in this figure shows how errors differed as a function of fixation position. Because we were specifically interested in how off-target fixation disrupted performance, the bottom row of Figure 3 shows the change in response rates relative to on-target fixation. In this display, we can see that off-target fixation caused a decrease in responses to the target loudspeaker (negative values at −30°, 0°, and +30° in the left middle, and right panels, respectively), with larger decreases for lateral fixation (white and black bars) than for central fixation (gray bars). These effects mirror the score reductions seen in Figure 2. Otherwise, the changes appear to be rather nonsystematic. Small increases in responses to nontarget loudspeakers are visible (i.e., there are more masker errors), as well as small increases in responses that were not present in the stimulus (i.e., there are more random errors). There was no strong evidence for a specific tendency to report digits arising from the fixation position when there was a mismatch.

Eye Position Data

Interpretation of the behavioral data in terms of gaze direction requires confirmation that the participants were in

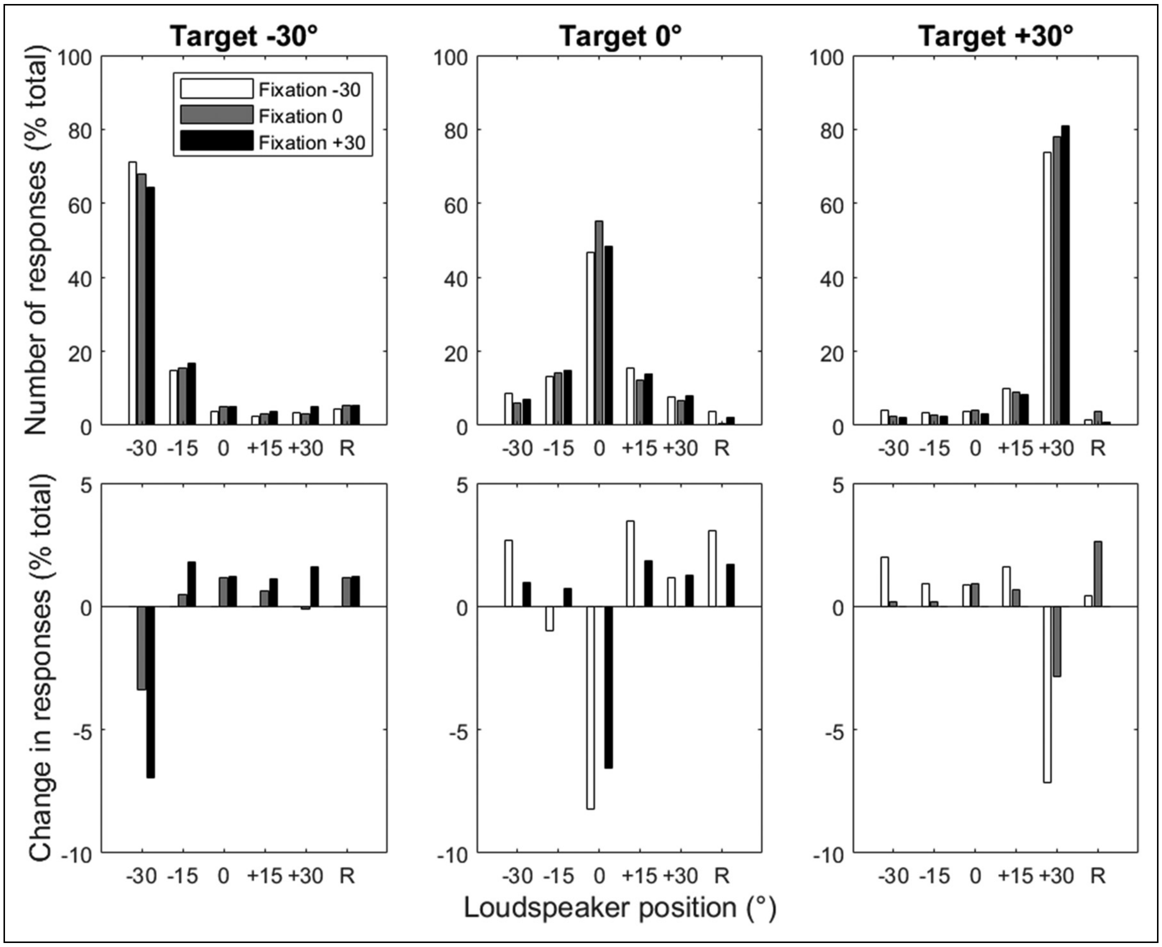

Figure 3. Top row: Distribution of responses according to the loudspeaker presenting each digit in each response. R refers to random responses (i.e., digits that were not present in the stimulus). Each panel shows one target position, and the different bars in each cluster correspond to the three fixation positions. Bottom row: Change in response distributions relative to the case in which the target and fixation positions coincided. One bar in each cluster is (by definition) at zero.

fact fixating on the cued location, and doing so reliably throughout each block of trials. Examination of the data obtained using both the Pupil Core headset and the Tobii Pro Glasses 3 provide such a confirmation.

In Figure 4 we show data for the five participants who were fitted with the Tobii system, as that system provided a more detailed set of outputs. For each participant, median head and eye position estimates were calculated for every trial based on all samples that fell within the stimulus window for that trial. The top row of Figure 4 shows histograms of head position estimates, pooled across trials and participants. The fact that the distributions are centered around 0° confirms that participants were generally able to maintain a forward-facing head position (per the instructions, and with the help of the neck rest). The distributions are slightly skewed for the lateral fixation conditions (left and right panels), which is consistent with our observation that the head would sometimes drift to the side of fixation in those blocks.

The middle and bottom rows of Figure 4 show histograms of eye position estimates, plotted in a similar way to the head position estimates. The middle row shows estimates based on the raw eye position data obtained from the eye tracker (i.e., relative to the head), while the bottom row shows those same data corrected for head position, so that the azimuth corresponds to the actual loudspeaker positions in the room. These figures show that eye position distributions were centered close to the correct azimuths. The distributions were tighter for central fixation than for lateral fixation, which we attribute to both listener factors (it is more strenuous to fixate laterally than centrally) and technical factors (eye tracking is less reliable for lateral eye positions). There was no strong evidence for systematic eye movements to the target loudspeaker (i.e., histograms within a panel are

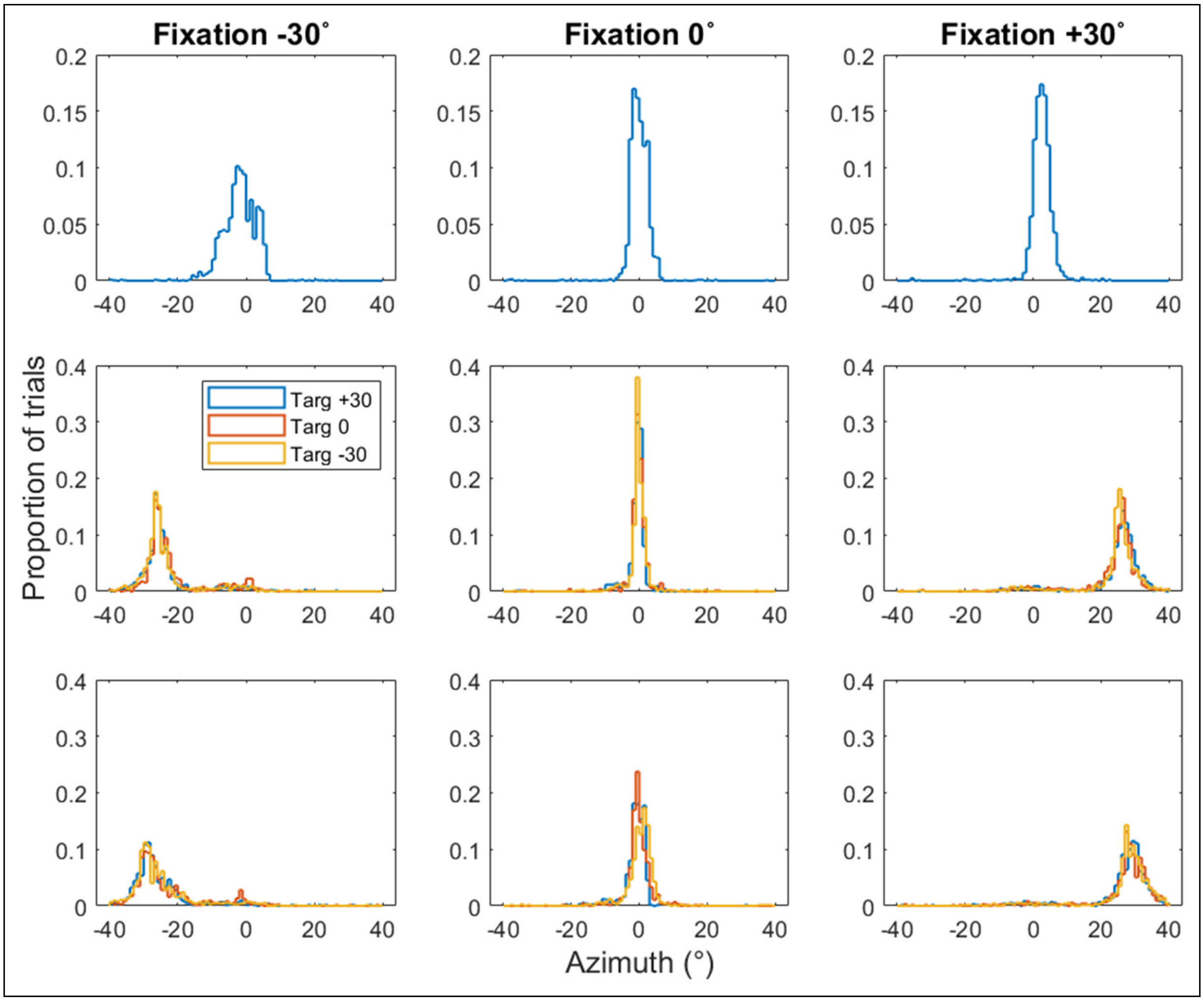

Figure 4. Pooled head and eye position distributions for five participants obtained using the Tobii Pro Glasses 3. Top row: head position estimates. Middle row: eye position estimates relative to the head. Bottom row: eye position estimates after correcting for head position. Each panel shows one fixation position and the three histograms within a panel are for different target positions. Median values and interquartile ranges for each histogram are provided in Table 1.

similar). A few outliers can be seen around 0° in the lateral fixation panels, but it is difficult to know if they represent erroneous eye movements to the target loudspeaker, or simply lapses where the participant looked back to the keypad. The latter seems more likely as there are essentially no instances of the opposite case (erroneously looking to the side in central fixation conditions). Summary measures for the histograms plotted in Figure 4 are provided in Table 1, along with relevant summary measures for the eye position data collected with the Pupil Labs system.

To understand the impact of occasional lapses in fixation, we identified all trials in which the eye position deviated by more than 15° from the cued location, as would occur if the participant fixated on the wrong loudspeaker. These trials made up about 10% of all trials collected using the Pupil Labs system, and about 8% of all trials when considering the more accurate head-corrected Tobii data. Importantly, these trials were distributed in a nonsystematic way across conditions, participants, and blocks. Removing these trials before calculating speech scores did not change the pattern of results or the statistical conclusions.

Discussion

These results demonstrate a measurable (but modest) effect of gaze direction on speech intelligibility in a multitalker mixture, extending the findings of previous studies that used simpler arrangements of stimuli. The implication is that fixating on the talker of interest in a cocktail party situation provides a benefit related to the gaze direction, in addition to the benefits afforded by relevant visual information such as that talker's lip movements.

In the current experiment, the benefit of gaze alignment was not observed consistently across all spatial combinations. For the cross-hemifield conditions, in which the target was on one side and the participant fixated on that same location or on a symmetric location on the other side, we found a significant effect of gaze alignment (7 percentage points on average). Previous studies that tested similar conditions for a simpler two-talker mixture found similarly modest effects. For example, Fleming et al. (2021) found an effect of around 5 percentage points (estimated from their Figure 3A) for their particular stimuli, configuration, and participants. In the current study, when fixation was not to the opposite hemifield, but rather to the midline, the effect of gaze alignment was smaller (3 percentage points) and nonsignificant. For the condition in which the target was located at the midline, we found a robust advantage (7 percentage points on average) when fixation was also to the midline rather than to either side. Another interesting feature of these data is the asymmetry of the gaze alignment effect. Specifically, a mismatch between the target and fixation position was more detrimental for a target at 0° and a fixation position of ±30° than vice versa.

While determining the neural basis of this phenomenon is clearly beyond the scope of the current investigation, we briefly consider two general mechanisms that have been described in the literature. First, there is ample evidence to suggest that eye movements influence *spatial representations* in cortical and subcortical auditory regions (Groh et al., 2001; Werner-Reiss et al., 2003; Winowski & Knudsen, 2006). It is possible that the benefit we observed for fixating on the location of an auditory target relates to an improvement in spatial resolution in that region (as proposed by Maddox et al., 2014) and an associated improvement in the ability to spatially segregate the talker of interest from its neighbors. The error patterns in Figure 3, which show a rather global distribution of incorrect responses, do not provide strong support for this idea, and the observed asymmetries are also hard to explain. A more parsimonious explanation

Table 1. Median and Interquartile Ranges of Head Position, Eye Position, and Head-Corrected Eye Position Estimates.

		Tobii (head position)		Tobii (eye position)		Tobii (head-corrected eye position)		Pupil Labs (eye position)	
Fixation position	Target position	Median	IQR	Median	IQR	Median	IQR	Median	IQR
−30	−30	−1.74	6.08	−24.04	8.88	−25.90	3.99	−27.88	7.20
	0			−23.09	8.73	−25.06	4.65	−27.32	9.29
	+30			−23.49	10.45	−25.71	3.68	−27.67	6.73
0	−30	−0.13	3.30	0.07	3.75	−0.18	1.43	0.94	3.42
	0			1.99	4.65	−0.12	1.69	−0.35	2.59
	+30			1.89	6.25	−0.21	1.59	−0.59	2.91
+30	−30	2.66	3.02	25.65	5.25	25.62	3.56	28.95	4.84
	0			26.66	5.75	26.03	4.67	28.78	5.31
	+30			28.36	8.56	26.54	4.40	28.87	5.23

Note. Values are pooled across the five participants who were fitted with each system. All values are in degrees. IQR=interquartile range.

relates to the fact that eccentric (but not forward) eye positions cause a dissociation of eye-centered and head-centered co-ordinate systems, and require a transformation of auditory spatial co-ordinates in order to maintain perceptual alignment (Zimmer et al., 2004). This transformation of the auditory space map may interfere with the ability to listen to a specific location when the gaze is directed laterally.

An alternative way of understanding the current results is in terms of *crossmodal spatial attention*. An attention-based explanation is appealing given the demanding nature of multitalker listening tasks, and the particular importance of spatial attention in our task. While we did not manipulate visual attention, gaze direction can be considered a surrogate for visual spatial attention (Henderson, 2003), in which case the alignment of the "look" and "listen" positions corresponds to the alignment of auditory and visual spatial attention. Within this view, our results suggest that dividing crossmodal spatial attention comes with a cost, just as dividing auditory attention across locations does (e.g., Best et al., 2006). This result is quite surprising given our experimental task, where there was no relevant visual stimulus or task. Other studies suggest that when auditory and visual stimuli must be actively integrated, the costs could be larger (Fleming et al., 2020; Fleming et al., 2021). An attentional explanation might explain the particular pattern of results we observed. For example, because the central location is acoustically the most unfavorable, it may be the case that spatially aligned visual attention is especially important in that case. It is also possible that lateral eye movements strongly engage the visual spatial attention system, leading to interference if auditory attention is focused elsewhere, whereas frontal fixation leaves the visual system in a default state (Nakashima & Kumada, 2017) that is less likely to cause interference.

Further insights into the likely source of the eye gaze effects observed here come from a related study by Pomper and Chait (2017) in which participants detected deviant target tones in one of three competing streams of tones presented from three spatial locations. They found that responses to targets were overall slower when participants gazed away from the attended location, compared to when they gazed toward it, and attributed this to the misalignment of the locus of visual and auditory attention. They also reported substantial differences in brain state between "coherent" and "incoherent" conditions. Specifically, the incoherent condition was associated with increased occipital alpha-band power, likely reflecting the suppression of distracting input, and increased central theta-band activity, which was interpreted in terms of cognitive control mechanisms and overall task demands. Given that our task structure shares many similarities with that of Pomper and Chait, it seems reasonable to expect that similar attentional mechanisms likely drove the changes in speech intelligibility we observed as we manipulated gaze direction.

As a final note, the results presented here are potentially relevant for hearing device applications that capitalize on eye movements. For example, we and others have been exploring the concept of a visually guided hearing aid, in which the gaze direction is tracked to estimate the sound source of interest, and in real time used to steer the acoustic look direction of a highly directional microphone array (Anderson et al., 2018; Kidd, 2017; Kidd et al., 2013; Kidd et al., 2015). Beamforming improves the effective SNR and can provide impressive speech intelligibility benefits under simple listening conditions with frontal targets. However, benefits have been harder to demonstrate under dynamic conditions where the target changes location frequently and the beamformer is steered by the user's eye movements (Best et al., 2017b; Roverud et al., 2018). The current results, showing that speech intelligibility is sensitive to the alignment of the "listen" and "look" directions, may offer a partial explanation. Specifically, while improving the SNR, beamformers typically also distort the natural binaural cues that are available at the ears (Best et al., 2017a; Wang et al., 2020). Reducing or eliminating these distortions may be critical for preserving the natural alignment of auditory and visual attention and optimizing speech intelligibility benefits afforded by such devices.

Conclusion

In the current study, we investigated whether there is a measurable effect of gaze direction on speech intelligibility in a demanding listening situation involving multiple competing talkers. Performance was best when gaze direction was aligned with the location of the attended talker, despite the absence of any task-relevant visual information. The results suggest that optimal performance in a cocktail party listening depends on the spatial alignment of auditory and visual attention.

Acknowledgments

This work was supported by NIH-NIDCD grant DC013286. Preliminary results were presented at the 179th Meeting of the Acoustical Society of America (7–11 December 2020) and published in Proceedings of Meetings on Acoustics (https://doi.org/10.1121/2.0001344). The authors would like to thank Gerald Kidd, Chris Mason, Todd Jennings, Ross Maddox, Noro Kopčo and Chris Conroy for their input and helpful discussions. The raw data presented here will be made freely available upon request.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Institute on Deafness and Other Communication Disorders (grant number R01 DC013286).

ORCID iD

Virginia Best https://orcid.org/0000-0002-5535-5736

References

Anderson, M. H., Yazel, B. W., Stickle, M. P. F., Espinosa Inguez, F. D., Gutierrez, N. S., Slaney, M., & Miller, L. M. (2018). Towards mobile gaze-directed beamforming: A novel neurotechnology for hearing loss. *Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, *2018*, 5806–5809. https://doi.org/10.1109/EMBC.2018.8513566

Best, V., Gallun, F. J., Ihlefeld, A., & Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *The Journal of the Acoustical Society of America*, *120*(3), 1506–1516. https://doi.org/10.1121/1.2234849

Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(35), 13173–13177. https://doi.org/10.1073/pnas.0803718105

Best, V., Roverud, E., Mason, C. R., & Kidd, G.Jr. (2017a). Examination of a hybrid beamformer that preserves auditory spatial cues. *The Journal of the Acoustical Society of America*, *142*(4), EL369–EL374. https://doi.org/10.1121/1.5007279

Best, V., Roverud, E., Streeter, T., Mason, C. R., & Kidd, G.Jr. (2017b). The benefit of a visually guided beamformer in a dynamic speech task. *Trends in Hearing*, *21*, 2331216517722304. https://doi.org/10.1177/2331216517722304

Best, V., Swaminathan, J., Kopčo, N., Roverud, E., & Shinn-Cunningham, B. G. (2018). A "buildup" of speech intelligibility in listeners with normal hearing and hearing loss. *Trends in Hearing*, *22*, 2331216518807519. https://doi.org/10.1177/2331216518807519

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Driver, J., & Spence, C. (1994). Spatial synergies between auditory and visual attention. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing*. The MIT Press, pp.311–331.

Fleming, J. T., Maddox, R. K., & Shinn-Cunningham, B. G. (2021). Spatial alignment between faces and voices improves selective attention to audio-visual speech. *The Journal of the Acoustical Society of America*, *150*(4), 3085–3100. https://doi.org/10.1121/10.0006415

Fleming, J. T., Noyce, A. L., & Shinn-Cunningham, B. G. (2020). Audio-visual spatial alignment improves integration in the presence of a competing audio-visual stimulus. *Neuropsychologia*, *146*, 107530. https://doi.org/10.1016/j.neuropsychologia.2020.107530

Getzmann, S. (2002). The effect of eye position and background noise on vertical sound localization. *Hearing Research*, *169*(1-2), 130–139. https://doi.org/10.1016/S0378-5955(02)00387-8

Gopher, D. (1973). Eye-movement patterns in selective listening tasks of focused attention. *Perception & Psychophysics*, *14*(2), 259–264. https://doi.org/10.3758/BF03212387

Groh, J. M., Trause, A. S., Underhill, A. M., Clark, K. R., & Inati, S. (2001). Eye position influences auditory responses in primate inferior colliculus. *Neuron*, *29*(2), 509–518. https://doi.org/10.1016/s0896-6273(01)00222-7

Hadley, L. V., Brimijoin, W. O., & Whitmer, W. M. (2019). Speech, movement, and gaze behaviours during dyadic conversation in noise. *Scientific Reports*, *9*(1), 10451. https://doi.org/10.1038/s41598-019-46416-0

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504. https://doi.org/10.1016/j.tics.2003.09.006

Hendrikse, M. M. E., Llorach, G., Hohmann, V., & Grimm, G. (2019). Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life. *Trends in Hearing*, *23*, 2331216519872362. https://doi.org/10.1177/2331216519872362

Jones, B., & Kabanoff, B. (1975). Eye movements in auditory space perception. *Perception & Psychophysics*, *17*(3), 241–245. https://doi.org/10.3758/BF03203206

Kidd, G.Jr. (2017). Enhancing auditory selective attention using a visually guided hearing aid. *Journal of Speech, Language, and Hearing Research*, *60*(10), 3027–3038. https://doi.org/10.1044/2017_JSLHR-H-17-0071

Kidd, G.Jr., Favrot, S., Desloge, J. G., Streeter, T. M., & Mason, C. R. (2013). Design and preliminary testing of a visually guided hearing aid. *The Journal of the Acoustical Society of America*, *133*(3), EL202–EL207. https://doi.org/10.1121/1.4791710

Kidd, G.Jr., Mason, C. R., Best, V., & Swaminathan, J. (2015). Benefits of acoustic beamforming for solving the cocktail party problem. *Trends in Hearing*, *19*, 2331216515593385. https://doi.org/10.1177/2331216515593385

Lewald, J. (1998). The effect of gaze eccentricity on perceived sound direction and its relation to visual localization. *Hearing Research*, *115*(1-2), 206–216. https://doi.org/10.1016/s0378-5955(97)00190-1

Maddox, R. K., Pospisil, D. A., Stecker, G. C., & Lee, A. K. C. (2014). Directing eye gaze enhances auditory spatial cue discrimination. *Current Biology*, *24*(7), 748–752. https://doi.org/10.1016/j.cub.2014.02.021

Nakashima, R., & Kumada, T. (2017). The whereabouts of visual attention: Involuntary attentional bias toward the default gaze direction. *Attention, Perception, & Psychophysics*, *79*(6), 1666–1673. https://doi.org/10.3758/s13414-017-1332-7

Pomper, U., & Chait, M. (2017). The impact of visual gaze direction on auditory object tracking. *Scientific Reports*, *7*(1), 4640. https://doi.org/10.1038/s41598-017-04475-1

Reisberg, D., Scheiber, R., & Potemken, L. (1981). Eye position and the control of auditory attention. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(2), 318–323. https://doi.org/10.1037//0096-1523.7.2.318

Rorden, C., & Driver, J. (1999). Does auditory attention shift in the direction of an upcoming saccade? *Neuropsychologia*, *37*(3), 357–377. https://doi.org/10.1016/s0028-3932(98)00072-4

Roverud, E., Best, V., Mason, C. R., Streeter, T., & Kidd, G.Jr. (2018). Evaluating the performance of a visually guided hearing aid using a dynamic auditory-visual word congruence task. *Ear and Hearing*, *39*(4), 756–769. https://doi.org/10.1097/AUD.0000000000000532

Shinn-Cunningham, B. G., Best, V., & Lee, A. K. C. (2017). Auditory object formation and selection. In J. C. Middlebrooks,

J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The auditory system at the cocktail party* . Springer, pp. 7–40.

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

Wang, L., Best, V., & Shinn-Cunningham, B. G. (2020). Benefits of beamforming with local spatial-cue preservation for speech localization and segregation. *Trends in Hearing*, *24*, 2331216519896908. https://doi.org/10.1177/2331216519896908

Werner-Reiss, U., Kelly, K. A., Trause, A. S., Underhill, A. M., & Groh, J. M. (2003). Eye position affects activity in primary auditory cortex of primates. *Current Biology*, *13*(7), 554–562. https://doi.org/10.1016/S0960-9822(03)00168-4

Westerhausen, R., & Kompus, K. (2018). How to get a left-ear advantage: A technical review of assessing brain asymmetry with dichotic listening. *Scandinavian Journal of Psychology*, *59*(1), 66–73. https://doi.org/10.1111/sjop.12408

Winowski, D. E., & Knudsen, E. I. (2006). Top-down gain control of the auditory space map by gaze control circuitry in the barn owl. *Nature*, *439*, 336–339. https://doi.org/10.1038/nature04411

Wolters, N. C. W., & Schiano, D. J. (1989). On listening where we look: The fragility of a phenomenon. *Perception & Psychophysics*, *45*(2), 184–186. https://doi.org/10.3758/BF03208053

Wood, K. C., & Bizley, J. K. (2015). Relative sound localisation abilities in human listeners. *The Journal of the Acoustical Society of America*, *138*(2), 674–686. https://doi.org/10.1121/1.4923452

Zimmer, U., Lewald, J., Erb, M., Grodd, W., & Karnath, H. (2004). Is there a role of visual cortex in spatial hearing? *European Journal of Neuroscience*, *20*(11), 3148–3156. https://doi.org/10.1111/j.1460-9568.2004.03766.x