


Practical utility of a head-mounted gaze-directed beamforming system

John F. Culling,^{1,a)}  Emilie F. C. D'Olne,² Bryn D. Davies,¹ Niamh Powell,¹ and Patrick A. Naylor²

¹*School of Psychology, Cardiff University, 70 Park Place, Cardiff CF10 3AT, United Kingdom*

²*Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, United Kingdom*

ABSTRACT:

Assistive auditory devices that enhance signal-to-noise ratio must follow the user's changing attention; errors could lead to the desired source being suppressed as noise. A method for measuring the practical benefit of attention-following speech enhancement is described and used to show a benefit for gaze-directed beamforming over natural binaural hearing. First, participants watched a recorded video conference call between two people with six additional interfering voices in different directions. The directions of the target voices corresponded to the spatial layout of their video streams. A simulated beamformer was yoked to the participant's gaze direction using an eye tracker. For the control condition, all eight voices were spatially distributed in a simulation of unaided binaural hearing. Participants completed questionnaires on the content of the conversation, scoring twice as high in the questionnaires for the beamforming condition. Sentence-by-sentence intelligibility was then measured using new participants who viewed the same audiovisual stimulus for each isolated sentence. Participants recognized twice as many words in the beamforming condition. The results demonstrate the potential practical benefit of gaze-directed beamforming for hearing aids and illustrate how detailed intelligibility data can be retrieved from an experiment that involves behavioral engagement in an ongoing listening task. © 2023 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.1121/10.0023961>

(Received 3 August 2023; revised 13 November 2023; accepted 29 November 2023; published online 15 December 2023)

[Editor: Christopher A. Brown]

Pages: 3760–3768

I. INTRODUCTION

The handicap experienced by hearing individuals with hearing impairment is primarily driven by their difficulty understanding speech in background noise (Kramer *et al.*, 1998). The only existing features of commercial hearing aids that improve intelligibility in noise are adaptive directional microphones, but these tend only to provide attenuation of sources to the rear (Kates, 2008). The adaptive aspect of the microphone is used to direct a null toward the most prominent source of sound in the rear hemifield. Multimicrophone beamforming systems have the potential to provide greater improvements to signal-to-noise ratio than conventional adaptive directional microphones through a more spatially selective beam. Such a system can markedly improve target-speech intelligibility in complex listening situations by attenuating background noise outside of this beam (Soede *et al.*, 1993). A growing number of studies have begun to explore the potential of multimicrophone beamforming systems (Adiloğlu *et al.*, 2015; Best *et al.*, 2017a; Best *et al.*, 2017b; Kidd, 2017; Lindqvist and Sollenberg, 2018; Jennings and Kidd, 2019; Kidd *et al.*, 2020; Yun *et al.*, 2021; Skoglund *et al.*, 2022).

However, a problem with using a very spatially selective beam is that the more spatially selective it is, the more

precisely and consistently it needs to be directed at the target source. The spatial selectivity of a directional beam can have negative as well as positive consequences because sound outside the beam is strongly suppressed. For instance, the system might enhance the intelligibility of one voice in a conversation but lose so much of the interlocutor that the conversation becomes incomprehensible. To evaluate the practical benefit of such a device, therefore, its effectiveness needs to be tested in circumstances in which more than one voice needs to be understood. That is, the user needs to be able to switch the beam between voices in a conversation with sufficient speed and proficiency that the beam is almost always directed to the active talker.

If the beamformer is fixed and the microphone array is head- or body-mounted, then to follow the conversation, the user must always orient the array toward the target's source by making body/head movements, which may be quite slow and awkward in many realistic listening situations. However, it is possible to use signal processing to steer the beam of an array of microphones toward a target source without physically rotating the array. Given this option, the problem becomes one of predicting the appropriate orientation of the beam in a given situation. In other words, from which direction does the user want to hear the sound?

An obvious possibility is to employ the user's gaze because people usually look at the person they are listening

^{a)}Email: cullingj@cf.ac.uk

to. Indeed, for a listener with a hearing impairment, looking at the talker is regarded as an important listening tactic, because it is necessary for gaining the benefit of lip reading; [Hadley et al. \(2019\)](#) showed that individuals with mild hearing loss looked at each other's mouths more during a conversation as the background noise level increased. Using gaze is also likely to be more spatially accurate than relying on head orientation. [Lu et al. \(2021\)](#) found that during a three-way conversation, listeners' overall gaze was directed quite precisely and accurately toward the current speaker, whereas the head orientation, which would determine the direction of a fixed beamformer, consistently undershot the target azimuth and was considerably more variable. In principle, therefore, the user's gaze could be a more accurate and precise indicator of the current talker direction, factors that become more important the narrower the selective beam.

The user's gaze direction could be recovered using an eye tracker (e.g., [Kidd, 2017](#)) or could use electro-oculography collected from electrodes in the user's earpieces (e.g., [Hládek et al., 2018](#)). However, since many individuals with hearing impairment are suffering from age-related hearing loss, they are likely also to be spectacle-wearers, and the spectacle frames might be used as a platform for discretely and conveniently mounting both the eye tracker components and the microphone array (e.g., [Anderson et al., 2018](#)). In this study, we compare user comprehension of a dyadic conversation with a gaze-directed beamformer versus natural, spatialized, and distributed audio.

Conventional speech-intelligibility tests are not suitable for evaluating the benefit of a gaze-directed beamforming system, because they focus on the understanding of isolated sentences from a single individual. In contrast, the ability to reorient a directional beam is useful for situations in which the target voice is changing dynamically. For a gaze-directed beamforming system to be effective, users will also need to direct their gaze to the right person sufficiently promptly to follow a conversation better than they would without the system. Some studies have created a spatially dynamic situation by presenting target materials from random directions (e.g., [Best et al., 2017b](#); [Hládek et al., 2018](#)). Such studies have not tended to report benefits for a directional beam. Other studies have added artificial indicators of the target direction, such as warning lights ([Favre-Félix et al., 2018](#)). However, real conversations naturally contain visual and verbal cues to upcoming changes in the conversational floor, which may enable listeners to reorientate efficiently. For instance, [Skoglund et al. \(2022\)](#) demonstrated modest benefits of an artificial directional effect (a 6-dB gain in the look direction) using comprehension of a natural conversation. Comprehension was evaluated using multiple-choice questions. The present study takes a novel approach, which enables the conventional measurement of intelligibility as a percentage of words correctly understood, but within a fluent, connected conversation. Moreover, the beamforming effect is based on a real spectacle-mounted microphone array.

The test is performed in two phases. In the first phase, participants listened to an extended audiovisual conversation either through a simulated gaze-directed beamformer or, as a control condition, using natural binaural cues. Since the listeners were engaged in following the conversation, they were unable to concurrently report its contents, so instead they were asked questions after the video about the content of the conversation to assess their understanding. However, questionnaires can have very variable sensitivity, often caused by floor and ceiling effects. In this case, the questionnaires also relied on accurate recollection of the material after a delay of several minutes. To address these problems, a second phase was included, in which formal sentence-by-sentence intelligibility measurement was performed using the audiovisual input from the first phase. To provide the same audiovisual input, participants in the second phase were yoked in both conditions to the eye movements made by participants in the first phase and to the consequent beamforming effect in the beamforming condition. By measuring sentence-by-sentence intelligibility, no significant memory load was involved, and sensitivity was determined only by the speech material and not by question difficulty. In the following, the two phases of the overall experiment will be described as "experiment 1" and "experiment 2." Experiment 1 tests whether overall comprehension of the conversation is improved through the use of the beamformer, while experiment 2 assesses any intelligibility improvement using standard measurement of percent words correct.

II. EXPERIMENT 1

A. Participants

The participants were ten students and staff from Cardiff University whose first language was English. None of the participants reported any existing hearing impairment.

B. Materials

A video call was recorded between two of the authors (N.P. and B.D.D.) with British English accents using the Zoom™ platform (version 5.11.11). The audio settings were used to record a separate audio file for each participant, and video settings were used to select "gallery view," which placed the two video images side by side (Fig. 1). The video was recorded at 25 frames/s at a resolution of 1920 × 1080, and the two audio recordings were at 32 kHz sampling frequency. The recording was cut into six equal sections of 2 min and 45 s. The video recording was cut using ffmpeg (ffmpeg.org, version git-2020-06-20-29ea4e1) and the audio recording using MATLAB™ (R2021b). The audio recordings were upsampled using MATLAB to 44.1 kHz. To normalize the speech levels between the talkers, one voice was boosted in level by 10 dB. The recordings of six interfering voices at 44.1 kHz were readings gathered from LibriVox (librivox.org) in British English accents (three male and three



FIG. 1. Screenshot of video conference call between N.P. and B.D.D.

female). Their levels were approximately equal to the two target voices before directional processing.

The conversation was unscripted and spontaneous, ranging over topics about sport and music. Since the ability to switch visual attention between talkers was a key determinant of success with the beamformer, relatively long gaps between one voice and the next could make the task easier. The transitions were consequently analyzed for gap duration. Figure 2 shows a histogram of all the gap durations from the six video segments. The modal gap duration was 400–600 ms, while the mean was 1.04 s. The negative gaps reflect overlapping dialogue.

For both conditions, the acoustic simulation placed the two target talkers at $\pm 15^\circ$, matching their visual angles with respect to the viewer. The competing talkers were at 0° , $\pm 60^\circ$, $\pm 135^\circ$, and 180° . During presentation of the beamforming condition, the signal from the eye tracker (EyeLink 1000, SR Research, Ottawa, Canada) was used to select filters from a lookup table for each video frame. Filters for the current gaze orientation were selected for each of the eight source directions and applied to the corresponding audio segments for the current video frame. The eight sources were summed before presentation with the corresponding video frame, and continuous audio output was maintained using the overlap-and-add method. The signals to each ear were identical. During presentation of the binaural condition, the eight sources were filtered separately for each ear

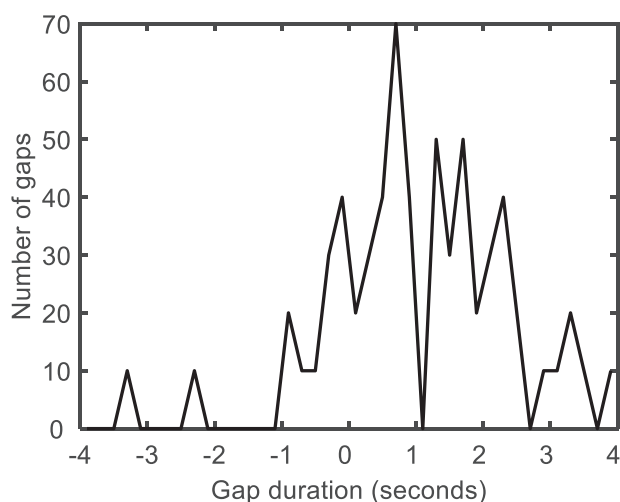


FIG. 2. Histogram of the duration of gaps between voice activity of the two talkers.

using corresponding head-related impulse responses recorded at MIT at 44.1 kHz from Gardner and Martin (1995).

The beamformer simulation was based on impulse responses recorded from a circular array of 48 loudspeakers (i.e., at 7.5° intervals) within a sound-treated room at eight MEMS microphones using a sampling frequency of 32 kHz. The microphones were glued to a pair of safety glasses and mounted on a Brüel & Kjaer (Naerum, Denmark) acoustic manikin (Fig. 3, left panel). The diffuse-noise directivity pattern for a minimum-variance distortionless response beamformer was calculated for each angular direction from these impulse responses in MATLABTM for frequencies from 100 to 16 000 Hz in 100-Hz steps. A diffuse-noise response was selected on the basis that it is independent of the spatial configuration of the interferers. It simulates a system that would be relatively straightforward to implement and provides a baseline for eight-channel beamformer performance. This set of beam patterns was used to derive 512-point linear-phase finite impulse response (FIR) filters using the host-window method (Abed and Cain, 1984) for each source direction and each beam orientation. The filters were upsampled to 704 points at 44.1 kHz and formed into a lookup table for use during the experiment.

The beam patterns were also used to create speech-weighted beam shapes (Fig. 3, right panel) for five example beam directions. These shapes were derived from a weighted sum over frequency of the narrowband beam patterns using weights from the speech intelligibility index (ANSI, 1997; Table I). The effective beam widths are quite spatially narrow with attenuation of approximately 6 dB for sources more than 30° away from the center of the beam, increasing to about 10 dB at the extremes.

In both conditions, a small yellow square was added to the image, which tracked the current eye position. This addition enabled the experimenter to verify that the eye tracker was functioning accurately throughout the experiment and would also have maintained awareness in the minds of the participants that their eyes were being tracked.

C. Procedure

Participants attended to each of the six videos and completed a questionnaire about the conversational content at the end of each. For each video, they sat with their head in the frame of the eye tracker and looked at an LCD monitor (52 cm wide \times 30 cm high), whose distance (48 cm) was

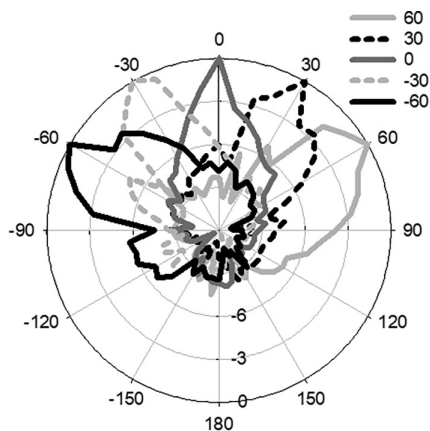
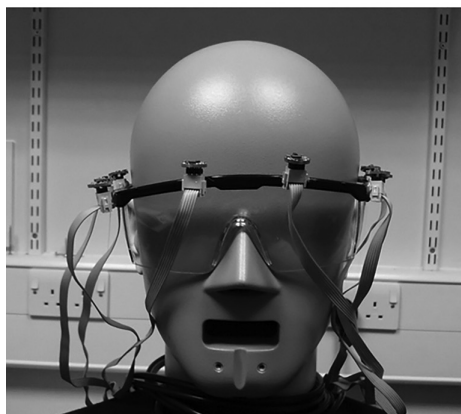


FIG. 3. (Left) Eight-microphone array mounted on a pair of safety glasses and worn by an acoustic manikin, ready for head-related-impulse-response measurement. (Right) Speech-weighted beam directivity for five different target azimuths, calculated from the measured impulse responses.

calculated to place the faces of the two target talkers at $\pm 15^\circ$. The eye tracker was calibrated, and the calibration was validated before the presentation of each video segment using five points in a vertical cross-formation. The speech mixture was presented over Sennheiser (Wedemark, Germany) HD650 headphones. The two conditions were presented in the first three or last three videos, so that participants had optimal opportunity to acclimatise to the different conditions. The order of the conditions was alternated between successive participants, so that any differences between the intrinsic intelligibility of the two halves of the conversation were counterbalanced. Participants were advised that in some cases, direction of their gaze would influence what they heard but that they should ignore this effect and concentrate on the task of following the conversation.

The questionnaire consisted of five questions for each of the six video segments, which asked for specific information content from the video, rather than being yes/no responses (e.g., question, “What sport does Bryn play with the University”; answer, “Tennis”). Marking was strict with no marks awarded for partially correct answers.

D. Results

The primary research question was whether the beamforming simulation improved intelligibility when following a conversation. Figure 4 shows the mean score on the questionnaires was more than twice as high in the beamforming condition as in the binaural control condition [$t(9) = 5.2$, $p < 0.001$]. The result shows that listeners’ overall performance at following the conversational content was better when the beamformer simulation was activated.

In addition to testing the main hypothesis, we were interested to see whether the beamforming induced any changes in behavior or whether participants were simply watching the conversation as they normally would. Figure 5 shows an example track. One question was whether the beamformer affected the fidelity with which the user could redirect the beam before a new talker has started. Aside from the method for extracting eye-saccade timing, the method was identical to that used by Hadley and Culling (2022) to analyze head movements during a live

conversation. In brief, the timings of speech onsets and offsets were extracted from the temporal envelopes of the microphone recordings, while the timings of eye saccades were extracted by thresholding the first differential of the eye-track recordings. The threshold for accepting a peak as evidence of a saccade was chosen from visual inspection of example eye tracks with the first differential superimposed (see Fig. 5, bottom panel). Some participants made saccades only at the point at which there was an exchange of the conversational floor, that is, when one talker stopped and the other started. Others made saccades much more frequently, but the extra movements were generally the result of only brief glances at the inactive talker.

To assess how promptly participants were able to use natural conversational cues to redirect their gaze, the timing of the saccades was compared with the timing of the nearest speech onsets and offsets for cases where there was an exchange of floor. Since these relationships are potentially coincidental, the chance rate was estimated by mismatching the voice recordings and eye tracks and conducting the same comparison on these unrelated datasets. The difference between the two results (thick versus thin lines in Fig. 6) is thus an estimate of the distribution of eye saccades that can be causally attributed to the exchange of floor.

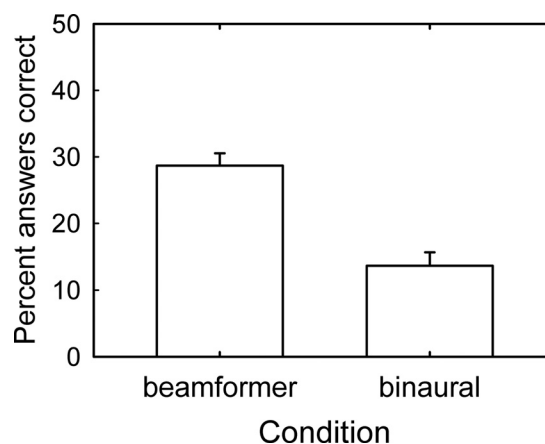


FIG. 4. Scores in the questionnaire about the content of the segments of conversation in each condition. Error bars, one standard error.

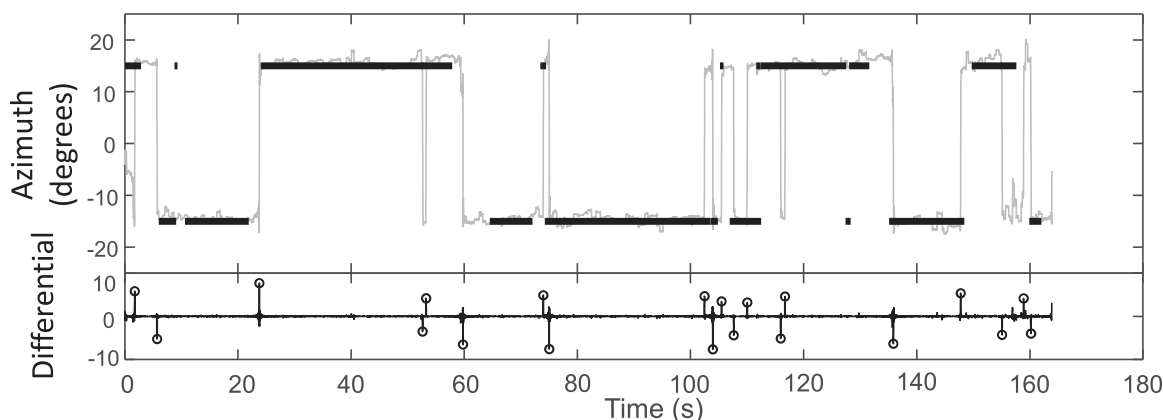


FIG. 5. Example eye-track recording. The top panel shows the voice activity of the two speakers as thick horizontal lines at their respective azimuthal locations. The thin gray line shows the recorded eye track. The bottom panel shows the first differential of the eye track with open symbols marking the points at which the differential exceeded the saccade-detection threshold.

Figure 6 shows the distribution of the timing of eye saccades with respect to onsets (left panels) and offsets (right panels) for the binaural and beamforming conditions. It can be seen that the timing of saccades is almost always within ± 1 s of the onset of the associated exchange of floor, regardless of condition. Based on the area between the two curves, the saccade anticipates the onset of the new speaker 56% of the time. These results are similar to those of Hadley and Culling (2022) for the initiation of head movements during an in-person conversation, in that the movements appear timed to catch the start of a new talker’s intervention. It should be noted, however, that since eye saccades are much more rapid than head movements, the current data effectively plot the timing of both the initiation and completion of the eye movement, rather than only the initiation. In contrast to the situation with onsets, the vast majority of the

saccades follow the offset of the preceding voice (right panels), suggesting that many of the saccades could be a reaction to the previous speaker stopping, rather than requiring the listener to predict an upcoming exchange of floor.

III. EXPERIMENT 2

In experiment 2, a second crew of ten listeners provided direct measurements of speech intelligibility for each of the participants in experiment 1, based upon exactly what those individual participants saw and heard during experiment 1.

A. Participants

The participants were ten different undergraduates and staff at Cardiff University whose first language was English and who had no known hearing impairment.

B. Materials

The materials for experiment 2 were derived from those of experiment 1. Each participant in experiment 2 was twinned at random with one of the participants from experiment 1 and was presented with materials designed to match the audiovisual input experienced by that participant. Each participant in experiment 2 thus inherited the allocation of materials to conditions that were experienced by their twin, maintaining the same counterbalancing across the group. To match the auditory experience, the audio output to each participant in experiment 1 was digitally stored on the computer during that experiment. To match the visual experience, the eye tracking record was stored and used to process the corresponding video using MATLAB. New videos were created that cut back and forth between the faces of the two talkers as a function of time in accordance with the eye movements of the participant in experiment 1. The processing used the hemifield toward which the experiment 1 participant directed his/her gaze to determine which half of the corresponding video frame to conserve and which to discard. This approach was adopted because following the exact movement of the eye, which makes many

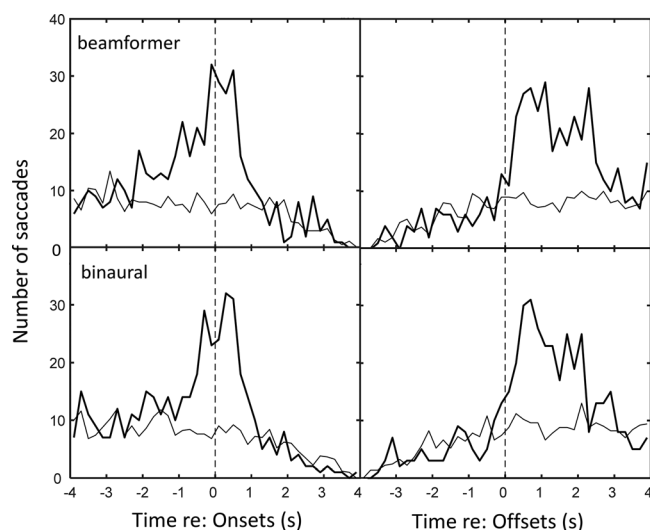


FIG. 6. Histograms of the timing of eye saccades with respect to the onset of a new talker (left panels) and the offset of the preceding talker (right panels) for the beamforming and binaural conditions (top and bottom panels, respectively). Thick lines are for the correctly associated eye and voice recordings, while the thin lines are the distributions for unrelated recordings (i.e., chance levels).

microsaccades when viewing the image of a face, was likely to result in a shaky image. Where the experiment 1 participants had made repeated saccades back and forth between the two interlocutors, the editing could still seem erratic, but not in a distracting way. Importantly, it reflected all the audiovisual information that the corresponding experiment 1 participant had gathered. This editing initially reduced the resolution to 960×1080 . However, since there was a vertical offset between the faces of the two talkers (see Fig. 1), the image was also reframed to improve this alignment by removing the top 200 lines of one stream and the bottom 200 lines of the other stream. The resulting video thus had a resolution of 960×880 . Examples of these video segments for the binaural and beamforming conditions are included (Mm. 1 and Mm. 2).

Mm. 1. Example of the binaural simulation (13 MB).

Mm. 2. Example of the beamforming simulation (13 MB).

The audio and video files were recombined and cut into short segments using ffmpeg (<https://ffmpeg.org/>). These short segments were each timed to encompass a complete sentence from one of the two talkers. There were a total of 117 sentences, and on 53 occasions, there was an exchange of floor between the two talkers between one sentence and the next.

C. Procedure

The isolated audiovisual sentences were presented one at a time to the participants following the same order as the original conversation so that contextual information was conserved. The participants were instructed to type in each sentence following the presentation of the corresponding video clip. Once they submitted their response, the actual transcript of the sentence was presented on the screen, and participants were instructed to score their own transcript with a score of up to 12 words correct using a single key-press on labelled keys. All words in the sentence were scored. Listener transcripts and self-marked scores were all stored in the computer to verify accurate scoring. Since the participants were shown the transcript of each sentence during this marking process, the preceding conversational context was always available, regardless of the intelligibility of the audio.

The experiment was controlled by a purpose-written MATLAB program. The video material was presented on a standard 22-in. LCD computer monitor, and the audio was presented over another pair of Sennheiser HD650 headphones in a single-walled IAC (Chandlers Ford, UK) audiometric booth.

D. Results

Figure 7 shows that the number of words correct in the beamforming condition was nearly double that in the

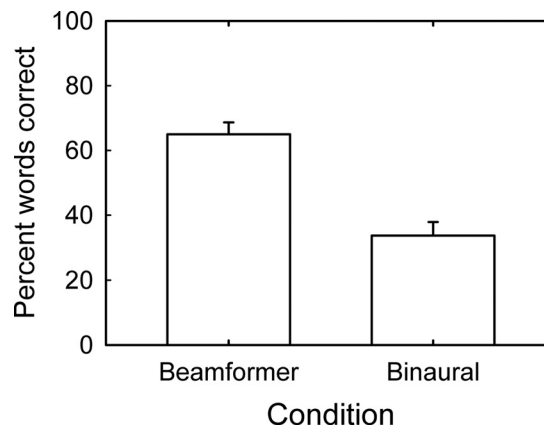


FIG. 7. Mean percentage of words correctly reported in each condition using separately presented sentences. Error bars, one standard error.

binaural condition. This difference was significant in a two-tailed t -test [$t(9)=14.67, p < 0.0001$].

The two conditions could also be compared for each individual listener. Although the audiovisual materials in the two halves of the experiment were not matched and could therefore have differed substantially in their intrinsic intelligibility, the results could still be compared by using an unpaired t -test with sentence scores as the sampled population. This analysis carries the caveat that the sentences were not drawn at random but came from the two halves of the experiment. However, as can be seen from Table I, any differences in the materials were overwhelmed by the effect of condition, since there was a significant effect of condition for every participant regardless of the allocation of materials to the two conditions that was inherited from experiment 1.

IV. DISCUSSION

A. Intelligibility benefit

Both experiments demonstrated an advantage for gaze-directed beamforming over binaural listening in a complex listening situation for following a conversation between two interlocutors. In experiment 1, listeners were able to recall twice as much information after listening to a few minutes of conversation in the beamforming condition as in the

TABLE I. Statistics for each individual participant in experiment 2.

Participant	Beamformer % correct	Binaural % correct	$t(115)$	p (two-tailed)
1	76.2	39.0	6.64	<0.0001
2	56.9	31.9	4.07	<0.001
3	72.4	40.7	6.10	<0.0001
4	45.7	21.1	4.03	<0.001
5	64.0	42.6	3.25	<0.005
6	76.9	48.2	5.27	<0.0001
7	56.5	17.7	7.83	<0.0001
8	77.0	42.8	6.64	<0.0001
9	51.8	10.3	7.99	<0.0001
10	72.8	44	5.43	<0.0001

binaural condition. In experiment 2, the participants were able to report twice as many individual words as in the binaural condition. These results were obtained using fixed levels of the different target and interfering sound sources, all eight of which were approximately equal before applying the binaural or beamforming simulations. The results of the second experiment are thus fully comparable with intelligibility measurements in the literature that were made at a fixed signal-to-noise ratio. It would be straightforward to extend the paradigm to multiple signal-to-noise ratios and thus to collect psychometric functions and to derive speech reception thresholds for each condition.

As noted in the Introduction, previous attempts to demonstrate advantages for directional beamformers in dynamic listening situations have had limited success. There have been a number of studies that show that prototype beamformers are effective at improving the intelligibility of a static source of speech against competing noise (Soede *et al.*, 1993; Best *et al.*, 2017a; Kidd *et al.*, 2015), but the essence of a gaze-directed system is to gain an intelligibility benefit while making appropriate shifts in the orientation of the beam.

Best *et al.* (2017b) used a task in which the user experienced a simulation of a gaze-directed beamforming array (Kidd *et al.*, 2013; Kidd, 2017). Questions were delivered from one of three different locations at random; one-word answers then came from one of the other two locations at random, and listeners had to report whether these answers were correct or incorrect. The results showed some benefit of the beamformer over binaural listening when all the questions and answers were in front, but the randomly located stimuli eliminated this benefit. The limited benefit seems likely to be associated with the degree of uncertainty in the paradigm, combined with the brevity of the answers that participants were evaluating; there was insufficient time for participants to identify, and shift their gaze to, the talker giving the answer.

Favre-Félix *et al.* (2018) used a simple 12-dB directional gain to crudely simulate a gaze-directed beamformer. The beamforming was thus equally powerful at all frequencies, rather than limited to higher frequencies. The eye movements were tracked using electro-oculography or with an eye tracker. Participants were required to report one of three simultaneous sentences presented from different directions. The target sentence was indicated by an LED, which indicated the target location 2 s before the stimulus began. The results confirmed a benefit of beamforming of similar magnitude to that reported here, although it was considerably smaller if the eye tracking relied on electro-oculography. However, the use of a 2-s warning light and a simple 12-dB gain lacked ecological validity.

Like the present experiment, Skoglund *et al.* (2022) used a video of a free-flowing conversation between two people, so that normal conversational cues indicated the direction in which to look. As in Favre-Félix *et al.*, the beamforming simulation was implemented as a simple power boost, but this time of only 6 dB. Interfering speech

came from eight loudspeakers, all positioned in front of the participant. Although this made the frontal hemifield very crowded, the fact that the gain was only applied to the loudspeaker closest to the gaze direction ensured that a signal-to-noise ratio advantage was always available, provided that gaze was closely directed toward the current talker. Performance was evaluated using multiple-choice questions after short segments of the conversation, lasting 10–39 s. However, the results showed the advantage from the beamformer to be only a 15%–20% increase in these scores. The reasons for the small observed benefit are not obvious but may be related to the use of multiple-choice questions, which can be difficult to optimise. The classic issue with some multiple-choice questions is lack of discrimination. If some individual questions are either too easy or too hard, it dilutes the overall discrimination of the test.

The current evaluation could be argued to be a little too predictable, because only two talkers were involved. An obvious extension will be to use more than two talkers in the conversation. Another potential limitation is that the recorded conversation was conducted at a fairly leisurely pace, as reflected by the relatively generous durations of silence across each exchange of floor. In contrast, Stivers *et al.* (2009) reported a mean silent gap of only 236 ms for English conversational speech. The robustness of gaze control for more rapid-fire interactions is yet to be demonstrated. However, talkers often hold the floor for more than one sentence, so a degree of tardiness in redirecting the beam during an exchange is unlikely to undo the overall intelligibility benefit.

Overall, the present results are the first to show a benefit of gaze-directed beamforming in terms of sentence-by-sentence speech intelligibility using a task that requires redirection of the beam. It also only provided the listener with ecologically valid cues to the direction of incoming target speech and a realistic beamforming effect. Moreover, the observed benefit is quite large.

B. Secondary effects

To gain benefit from the beamformer, listeners must direct their gaze toward the target source, and this behavior is contingent upon accurate judgment of which person is the current talker. When watching others in a conversation, there are many cues, visual, auditory, and linguistic, that enable listeners to anticipate exchanges of conversational floor (Holler and Kendrick, 2015; Hadley and Culling, 2022). For those actively engaged in conversation, anticipating the end of the other party's contribution is also key to a fluent conversation. It is thought that such anticipation is important in enabling talkers to begin their sentence just 200 ms or so after their interlocutor has finished (Stivers *et al.*, 2009). Experiments on such conversational turn-taking indicate that the informational content of the speech is one of the most important cues (Pickering and Gambi, 2018).

Since the beamformer increases the intelligibility of the attended talker, it is possible that a virtuous circle is created, whereby the improved intelligibility facilitates better anticipation of the exchange of conversational floor, which facilitates more timely movements of gaze (and hence beam direction), which further improves the intelligibility. Alternatively, the narrow focus of the beam may create the problem of “tunnel hearing” (Stadler and Rabinowitz, 1993), in which listeners fail to follow changes in conversation, because they cannot hear new contributions that come from outside the beam. These possibilities were assessed by analyzing the gaze data in comparison with the dominant talker during each video frame.

To establish the dominant talker, the short-term power of the original 32-kHz audio recordings was extracted by squaring the waveforms and applying a 1280-point rectangular moving-average filter (equivalent to one video frame). The envelope of the quieter talker was amplified by 10 dB as in the experiment. Frames without speech were defined as having a combined power that did not exceed a threshold. The value of this threshold was found to have little effect on the results over a wide range of settings. The results reported here use a value of threshold that was set at 1% of the long-term root mean square (rms) of the summed power. By that definition, the proportion of frames for which there was speech above the threshold was 59%. The dominant talker for any frame that exceeded the threshold had greater power than the other. If the direction of gaze was in the appropriate hemifield for the dominant talker during this frame, then the participant was deemed to be following the conversation with their gaze during that frame. Table II shows the fidelity with which participants were following the conversation with their gaze in each of the conditions. The means for each condition are both 85%, so these results are not consistent with either possibility; they do not suggest that the enhanced intelligibility provided by the beamformer is enabling the participants to improve their gaze control, and they do not suggest that tunnel hearing prevents the detection of new contributions. It is conceivable that the two effects offset each other to produce a neutral outcome. These conclusions are also supported by the analysis in Fig. 6, which shows a very similar pattern

TABLE II. The proportion of frames with speech for which the participant was following the conversation in each condition.

Participant	Beamformer (%)	Binaural (%)
1	90	88
2	89	86
3	90	88
4	68	78
5	81	83
6	89	85
7	85	86
8	87	83
9	88	89
10	87	87
Mean	85	85

for the timing of saccades with respect to the onsets and offsets of the two voices.

Finally, the 85% proportion of time looking at the target talker is very similar to the 88% reported by Hadley *et al.* (2019), and the timing of the saccades is similar to the initiation of head rotation reported from Hadley *et al.* data by Hadley and Culling (2022). Thus, the reorientation behavior in Hadley *et al.* data, for which the participants were actively engaged in a triadic conversation, has no obvious distinction from the reorientation behavior of participants who are passively following a conversation.

C. Comparison with adaptive directional microphones

Conventional hearing aids usually use adaptive directional microphones in complex listening situations. They consequently represent a more appropriate benchmark against which a gaze-directed beamformer should be compared than unaided binaural hearing. Making such a comparison would be a little more complicated but would be an essential step for establishing any potential improvement in patient benefit.

Directional microphones can null out sound from one direction to the rear of the listener by subtracting the signal at the rear microphone from that at the front microphone after a delay. The direction of the null is adaptively controlled by adjusting the delay. This technique can substantially enhance the signal-to-noise ratio, particularly if the noise is from a point source at the rear, but is less effective in more complex or diffuse sound fields that are typically challenging to hearing-impaired listeners. The subtractive processing has the disadvantage that low frequencies tend to be suppressed, because the phase difference between the front and rear microphones is small. Consequently, it has rather poor low frequency response, which is compensated by low frequency amplification, but this compensating gain tends to amplify the microphone self-noise in the frequency range of this roll-off [Elko (2004), pp. 11–65]. In contrast, beamformers tend to have a weighted sum of the microphone outputs that does not accentuate self-noise, and their “distortionless response” means a flat frequency response in the target direction. A fair comparison will thus require the development of a sophisticated simulation of adaptive directional microphones, which embodies a realistic frequency response and noise level.

D. Application to other noise suppression methods

The two-phase experimental method described here has wider potential application for evaluating any hearing assistive technology that suppresses interfering noise sources. Many real-life listening situations involve multiple potential target talkers, and the listeners will often wish to switch attention between different people. Machine learning methods that are currently in development can produce dramatic increases in signal-to-noise ratio (e.g., Healy *et al.*, 2023), but as such methods improve, the consequences of mistaking target and interfering sounds become more severe, and knowing the intentions of the user becomes increasingly

important. Gaze is only one method of indexing the user's intentions, but the two-phase experimental technique could be used to extract sentence-by-sentence intelligibility whenever listening to continuous conversation is a critical task.

V. CONCLUSIONS

The two-phase measurement method has shown for the first time that a gaze-directed beamformer provides a substantial improvement over binaural hearing in the understanding of a conversation between two other parties. Comparison with conventional adaptive directional microphones would be important to establish the potential patient benefit of a gaze-directed beamformer. The two-phase method has potential for wider application to evaluating any system that seeks to accommodate shifts in the attention of the user.

AUTHOR DECLARATIONS

Conflict of Interest

The authors had no conflicts of interest.

Ethics Approval

The procedure for both experiments in this work was approved by the School of Psychology Ethics Committee at Cardiff University in compliance with the Declaration of Helsinki.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Abed, A. H. M., and Cain, G. D. (1984). "The host windowing technique," *IEEE Trans. Acoust. Speech Signal Process.* **32**(4), 683–694.

Adilođlu, K., Kayser, H., Baumgärtel, R. M., Rennebeck, S., Dietz, M., and Hohmann, V. (2015). "A binaural steering beamformer system for enhancing a moving speech source," *Trends Hear.* **19**, 233121651561890.

Anderson, M. H., Yazel, B. W., Stickle, M. P. F., Espinosa Iñiguez, F. D., Gutierrez, N.-G. S., Slaney, M., Joshi, S. S., and Miller, L. M. (2018). "Towards mobile gaze-directed beamforming: A novel neuro-technology for hearing loss," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2018**, 5806–5809.

ANSI (1997). ANSI S3.5-1997, *Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

Best, V., Roverud, E., Mason, C. R., and Kidd, G. (2017a). "Examination of a hybrid beamformer that preserves auditory spatial cues," *J. Acoust. Soc. Am.* **142**(4), EL369–EL374.

Best, V., Roverud, E., Streeter, T., Mason, C. R., and Kidd, G. (2017b). "The benefit of a visually guided beamformer in a dynamic speech task," *Trends Hear.* **21**, 233121651772230.

Elko, G. (2004). "Differential microphone arrays," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Springer, Boston, MA), pp. 11–65.

Favre-Félix, A., Graversen, C., Hietkamp, R. K., Dau, T., and Lunner, T. (2018). "Improving speech intelligibility by hearing aid eye-gaze steering: Conditions with head fixated in a multitalker environment," *Trends Hear.* **22**, 233121651881438.

Gardner, W. G., and Martin, K. D. (1995). "HRTF measurements of a KEMAR," *J. Acoust. Soc. Am.* **97**(6), 3907–3908.

Hadley, L. V., Brimijoin, W. O., and Whitmer, W. M. (2019). "Speech, movement, and gaze behaviours during dyadic conversation in noise," *Sci. Rep.* **9**, 10451.

Hadley, L. V., and Culling, J. F. (2022). "Timing of head turns to upcoming talkers in triadic conversation: Evidence for prediction of turn ends and interruptions," *Front. Psychol.* **13**(December), 1061582.

Healy, E. W., Johnson, E. M., Pandey, A., and Wang, D. (2023). "Progress made in the efficacy and viability of deep-learning-based noise reduction," *J. Acoust. Soc. Am.* **153**(5), 2751–2768.

Hládek, L., Porr, B., and Brimijoin, O. W. (2018). "Real-time estimation of horizontal gaze angle by saccade integration using in-ear electro-oculography," *PLoS One* **13**(1), e0190420.

Holler, J., and Kendrick, K. H. (2015). "Unaddressed participants' gaze in multi-person interaction: Optimizing reciprocity," *Front. Psychol.* **6**, 98.

Jennings, T. R., and Kidd, G. (2019). "A visually guided beamformer to aid listening in complex acoustic environments," *Proc. Mtgs. Acoust.* **33**, 050005.

Kates, J. M. (2008). *Digital Hearing Aids* (Plural, San Diego, CA).

Kidd, G. (2017). "Enhancing auditory selective attention using a visually guided hearing aid," *J. Speech Lang. Hear. Res.* **60**(10), 3027–3038.

Kidd, G., Favrot, S., Desloge, J. G., Streeter, T. M., and Mason, C. R. (2013). "Design and preliminary testing of a visually guided hearing aid," *J. Acoust. Soc. Am.* **133**(3), EL202–EL207.

Kidd, G., Jennings, T. R., and Byrne, A. J. (2020). "Enhancing the perceptual segregation and localization of sound sources with a triple beamformer" *J. Acoust. Soc. Am.* **148**(6), 3598–3611.

Kidd, G., Mason, C. R., Best, V., and Swaminatha, J. (2015). "Benefits of acoustic beamforming for solving the cocktail party problem," *Trends Hear.* **19**, 233121651559338.

Kramer, S. E., Kapteyn, T. S., and Festen, J. M. (1998). "The self-reported handicapping effect of hearing disabilities," *Audiology* **37**, 302–312.

Lindqvist, J., and Sollenberg, M. (2018). "Real-time multiple audio beamforming system," Ph.D. thesis, Lund University, Lund, Sweden.

Lu, H., McKinney, M. F., Zhang, T., and Oxenham, A. J. (2021). "Investigating age, hearing loss, and background noise effects on speaker-targeted head and eye movements in three-way conversations," *J. Acoust. Soc. Am.* **149**, 1889–1900.

Pickering, M. J., and Gambi, C. (2018). "Predicting while comprehending language: A theory and review," *Psych. Bull.* **144**, 1002–1104.

Skoglund, M. A., Andersen, M., Shiell, M. M., Keidser, G., Rank, M. L., and Rotger-Grifol, S. (2022). "Comparing in-ear EOG for eye-movement estimation with eye-tracking: Accuracy, calibration, and speech comprehension," *Front. Neurosci.* **16**(June), 873201.

Soede, W., Berkhout, J., and Bilsen, F. (1993). "Development of a directional hearing instrument based on array technology," *J. Acoust. Soc. Am.* **94**, 785–798.

Stadler, R. W., and Rabinowitz, W. M. (1993). "On the potential of fixed arrays for hearing aids," *J. Acoust. Soc. Am.* **94**, 1332–1342.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., and Levinson, S. C. (2009). "Universals and cultural variation in turn-taking in conversation," *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10587–10592.

Yun, D., Jennings, T. R., Kidd, G., Jr., and Goupell, M. J. (2021). "Benefits of triple acoustic beamforming during speech-on-speech masking and sound localization for bilateral cochlear-implant users," *J. Acoust. Soc. Am.* **149**(5), 3052–3072.