Speech-Augmented Cone-of-Vision for Exploratory Data Analysis

Riccardo Bovo
Imperial College London
London, United Kingdom
rb1619@ic.ac.uk

Daniele Giunchi
University College London
London, United Kingdom
d.giunchi@ucl.ac.uk

Ludwig Sidenmark
Lancaster University
Lancaster, United Kingdom
l.sidenmark@lancaster.ac.uk

Joshua Newn
Lancaster University
Lancaster, United Kingdom
j.newn@lancaster.ac.uk

Hans Gellersen
Lancaster University
Lancaster, United Kingdom
Aarhus University
Aarhus, Denmark
h.gellersen@lancaster.ac.uk

Enrico Costanza
University College London
London, United Kingdom
e.costanza@ucl.ac.uk

Thomas Heinis
t.heinis@ic.ac.uk
Imperial College London
London, United Kingdom

ABSTRACT

Mutual awareness of visual attention is crucial for successful collaboration. Previous research has explored various ways to represent visual attention, such as field-of-view visualizations and cursor visualizations based on eye-tracking, but these methods have limitations. Verbal communication is often utilized as a complementary strategy to overcome such disadvantages. This paper proposes a novel method that combines verbal communication with the Cone of Vision to improve gaze inference and mutual awareness in VR. We conducted a within-group study with pairs of participants who performed a collaborative analysis of data visualizations in VR. We found that our proposed method provides a better approximation of eye gaze than the approximation provided by head direction. Furthermore, we release the first collaborative head, eyes, and verbal behaviour dataset. The results of this study provide a foundation for investigating the potential of verbal communication as a tool for enhancing visual cues for joint attention.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**; **Collaborative interaction**; **Virtual reality**; **Visual analytics**.

KEYWORDS

Field of View, multi-modal visual attention cues, VR collaborative analytics, eye-tracking

CHI'23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581283

ACM Reference Format:
Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Joshua Newn, Hans Gellersen, Enrico Costanza, and Thomas Heinis. 2023. Speech-Augmented Cone-of-Vision for Exploratory Data Analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581283

1 INTRODUCTION

Mutual awareness of visual attention–the ability to identify collaborators' visual attention–is crucial for successful collaboration [22, 26, 77, 99, 111]. As such, prior studies have shown that introducing bi-directional visual attention cues in collaborative VR can improve mutual awareness [53]. Although they offer improvements over having no visual attention cues in virtual collaborative environments, correctly representing users' attention remains an open challenge. Field-of-view-based visualisations only provide an estimate of visual attention [16], while pointer-based using natural pointing modalities, such as the eye gaze [44] and head [115], does not afford the dynamic visual representation of different types of attention (*e.g.* focused and distributed [103]). Moreover, there are inherent limitations to using natural pointing modalities to represent visual attention. For example, attention cues based on gaze input can be distracting for an observer due to natural looking behaviour [119], or 'confusing' when there is a misalignment between a collaborator's verbal references and the depicted eye-gaze location due to eye-tracker calibration issues [26].

In this paper, we explore how combining an existing field-of-view-based visual attention cue ('Cone of Vision' [16]) with verbal communication can improve gaze inference and mutual awareness for exploratory data analysis in VR. The Cone of Vision (CoV) visual attention cue is a novel technique developed by Bovo et al. [16] that leverages head behaviour to allow a more accurate representation of users' attention within their field of view (FoV). Existing FoV-based techniques display the entire area within a user's vision. Though the technique narrows the FoV (based on gaze probability within head coordinates), the visualisation can still contain high

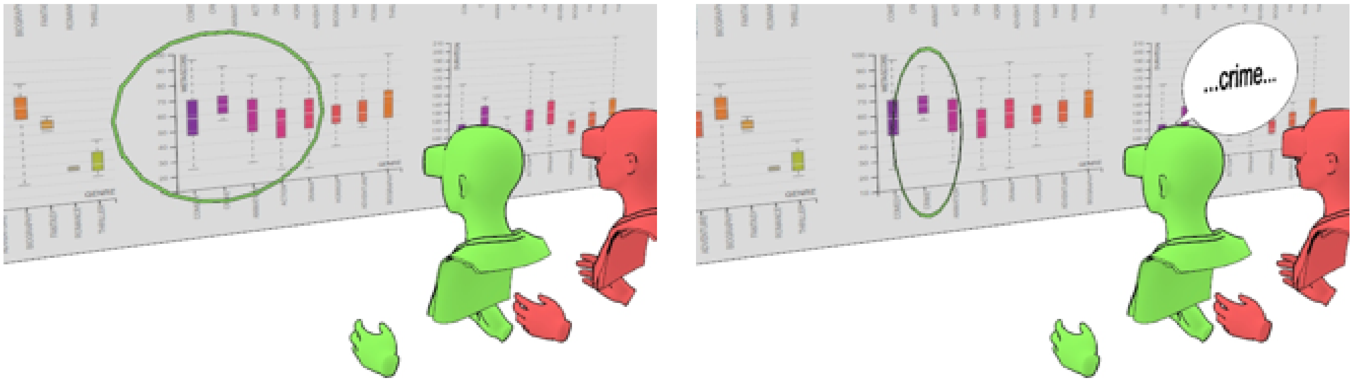


Figure 1: CoV+Speech is a multi-modal visual attention cue that narrows the cone of visual attention around keywords uttered by participants during collaborative verbal communication.

densities of information within the ‘cone’. By using speech to direct the CoV region towards the visual elements mentioned in verbal communication, we can create an adaptive multi-modal approach that continuously refines the focus of visual attention towards such elements. Figure 1 shows how the combination of CoV+Speech can narrow down the CoV visual attention cue based on keywords uttered by a collaborator during exploratory data analysis.

Our proposed approach of combining CoV and verbal communication mirrors how collaborators communicate in face-to-face settings. Research has shown that collaborators often leverage cues from multiple modalities to gauge the visual attention of collaborators, including verbal cues [117]. In particular, they first understand the general orientation of their collaborators (i.e., by evaluating the general direction of their head gaze) and then confirm or refine the location of the visual context using verbal communication [26, 89]. Collaborative verbal communication is also used as a fallback method when visual cues are not accurate enough or when there are calibration errors [117]. Further, our approach is well-suited to cross-virtuality analytics (XVA) context because XR headsets typically have access to head and verbal behaviour, while they do not always have eye-tracking capabilities. This makes our approach widely applicable within the XR device ecosystem. Due to the potential benefits of our proposal technique, we aim to answer the following research question: How does speech in conjunction with head behaviour impact joint attention during collaboration and gaze inference? To address this question, we designed and conducted a within-group study that compares three conditions: *CoV*, *CoV+Speech*, and *Eye-Gaze Cursor*. In the study, ten pairs of participants performed collaborative exploratory data analysis tasks using three different dataset visualisations, testing each of the three conditions. In the *CoV* condition, we used a model to model a cone of vision using the statistical model of gaze probability, which was projected onto VR screens [16]. In the *CoV+Speech* condition, we processed the collaborative verbal communication using speech recognition as input to narrow the CoV around the enunciated elements of the visualisation. Lastly, we added the *Eye-Gaze Cursor* condition as it is a widely used method to represent visual attention, in which we mapped the raw eye-gaze position to a live cursor.

Our results showed that speech recognition did not lead to better joint attention compared to CoV, due to lag and limited speech recognition accuracy. To further investigate the potential of verbal communication to negotiate shared attention, we performed a follow-up analysis using a highly accurate speech-to-text model to transcribe the verbal communication data collected during our study. The transcription allowed us to analyse the types of verbal references used by participants. This analysis validated our assumption that the most common form of communication relies on explicit keyword utterances rather than implicit verbal references or pointing-based communication. In addition to allowing us to perform an offline approximation of eye gaze using speech as an input, the transcription allowed us to analyse the types of verbal references used by participants. Our analysis demonstrated that our proposed method improves eye gaze approximation accuracy by 50 pixels when the CoV regions do not constrain the eye gaze. This suggests that speech has the potential to improve the shared context of visual attention. Therefore, we also present the collected data and accurate transcriptions as a dataset, which is the first dataset of collaborative head, eyes and transcribed speech behaviour to the best of our knowledge. Our findings and dataset contribute to a deeper understanding of verbal communication and gaze during collaboration.

Furthermore, we were able to estimate the impact that CoV and CoV+speech have on individual visual attention by testing the statistical model of gaze on which the CoV cues are based. In the *Eye-Gaze Cursor* condition, the gaze distribution followed the earlier model (70% of gaze samples within the non-displayed CoV). In contrast, the CoV conditions showed that eye-gaze distribution was considerably narrower: more than 85% of gaze samples fell within the displayed CoV. When head-based visual attention cues are visible (i.e. CoV), the head gaze becomes a better predictor of eye gaze than when they are not used. Our study also enabled us to compare bidirectional head-gaze visual attention cues and eye-tracking cues, finding that CoV cues foster joint attention equally or better than eye-tracking cues. We measured joint attention as the fraction of concurrent gaze on the same area of interest (AOI), that is a method used in research to analyze attention to individual objects [84], we do so at two resolutions: the chart level and the screen level. We discuss the implications of this finding in the discussion and conclusion section. The contribution of this paper is threefold:

(1) A novel visual FoV-based cue for collaboration that dynamically changes size based on verbal communication to balance broad and narrow information.
(2) The results of a study compared three visual cues for collaboration during an exploratory data analysis task in VR. Results showed that our proposed approach better approximates the head gaze if compared to the approximation offered by the head gaze alone. Moreover, head-based visual attention cues foster joint attention equally or better than eye-tracking visual attention cues.
(3) A dataset [1] containing the verbal, head, and eye behaviour of ten pairs of participants collaborating in VR.

2 RELATED WORK

2.1 Cross-virtuality Analytics in Immersive Environments

Cross-virtuality Analytics (XVA) can support users simultaneously via collaborative interfaces encompassing the reality–virtuality continuum [35] and has been adopted by a large set of works [20, 21, 60, 70, 79, 91, 96, 97]. XVA has recently gained interest from researchers due to the COVID pandemic due to the increase in demand for remote working and collaboration. Within this scenario, VR enables portable personal bespoke working environments that can be used at home on the go or in hybrid modes with normal screens [14, 32, 83]. While some of the XVAs are tailored for specific 3D data applications [20, 21, 70], there is a growing trend to support 2D content, which is cross-compatible with the standard multi-purpose applications available on desktop PC [30, 46, 60, 71, 87, 97]. Such an approach is also followed by commercial applications [5, 6, 27, 47, 74, 85], which either enable the display of standard 2D documents (e.g. web pages, calendar layout, office documents) or capture arbitrary 2D windows from a desktop PC and display them as a 2D surface in the virtual environment. Such software allows users to set up their own layout of 2D windows in the 3D environment (Figure 2 a) and b)) which is a problem that previous XVA studies have explored (Figure 2 c) and d)). For example, Lee et al. [60] illustrates how users behave when solving an analysis task in a squared-room scenario with the freedom to position 2D surfaces. The experiment results show that 2D screens are often placed on the walls to present the content efficiently to others.

Qualitative analysis of Satriadi et al. [97] focused on determining the optimal shape of 2D screens around a VR user, exploring different layouts such as spherical, spherical cap, planar, and unconstrained (i.e. users are free to arrange in any form in the space). The results highlight how users prefer constrained layouts in curved topologies, such as spherical or cylindrical. Both studies suggest that such layouts guide users in setting up the panel configuration, constraining them to the edges of the virtual environment. Such findings show convergence for VR 2D screen layouts toward a convex and egocentric layout, as seen in commercial VR applications such as Meta Infinite Office [74] or XVA research ([71, 87]).

2.2 Mutual Awareness of Visual attention in Collaborative VR

Awareness of other people's visual attention is a crucial component of collaboration. Several works have shown that visualising collaborators' visual attention can be an effective tool to enhance collaboration by allowing users to predict others' intentions and awareness [26, 29, 77, 99, 111] or desire to communicate [26, 99]. In AR and VR settings, such visualisations are commonly displayed through gaze cursors and have been extensively investigated and proven to improve collaboration [10, 53, 58, 61, 67]. However, modern VR equipment commonly does not incorporate eye-tracking. Therefore, several studies propose head orientation as an approximation of gaze [7, 24, 34, 45, 66, 90, 91] using FoV visualisations or exploiting attention models based on head movements [18]. These works have shown that FoV visualisations can help collaborators establish mutual awareness and attention [10, 24, 34, 45, 91, 91, 92]. Piumsomboon et al. [91] further propose that these visualisations could be displayed or hidden accordingly to the context of collaboration to minimise visual clutter while maximising collaboration. Based on this insight, we introduce a visualisation technique that adapts to participants' speech during collaboration. Because FoV-based vizualisations help maintain mutual awareness among collaborators, we adopt the same metrics used in previous CSCW studies to evaluate visual coordination [16, 26, 82, 99].

Moreover, we determine metrics based on head-tracked movements embraced by previous work [15, 88, 115]. In many VR/AR studies, visual cues are not self-visible but only shown to collaborators; as such, they are called uni-directional visual cues [44, 91, 116]. The justification for mono-directionality is that users already know where they are looking; therefore, they do need such redundant information. Nevertheless, recent HCI studies started to explore bi-directional cues, which can be seen by collaborators and the producer of eye and head behaviour [16, 54]. Such studies highlight that the feedback loop of self-visible visual attention cues is beneficial for collaborative work, leading to less effort (i.e. drop in task physical demand [54] and increase in visual coordination [16]. However, previous studies have yet to compare bi-directional head-based and eye-based cues. In our work, we fill this gap to see which supports collaboration better and to understand whether eye-tracking-less VR headsets can support joint attention during collaboration. In line with the findings described, we developed a layout for our experiment (Figure 2(e), Section 3.2.1), which consists of a convex egocentric layout of 2D VR screens.

2.3 Verbal Communication and Mutual Awareness of Visual attention in CSCW

Previous CSCW work by [17, 117, 118] investigated pointing tools used to negotiate visual attention during collaboration. These studies highlight the underlying relationship between pointing tools and the verbal channel (i.e. pointing-based communication). Such dynamics are also explored concerning eye-gaze visualizations by the work of D'Angelo and Begel [26] in remote pair programming tasks. Their study proposes a taxonomy for verbal and gestural spatial references describing different types: explicitly mentioning specific keywords that are displayed on the screen, utterance plus hand pointing, referring directly to the gaze visualizations ("..where

[1] https://github.com/Collaborative-Immersive-Visual-Toolkit/Speech-Gaze-Head-Datatset

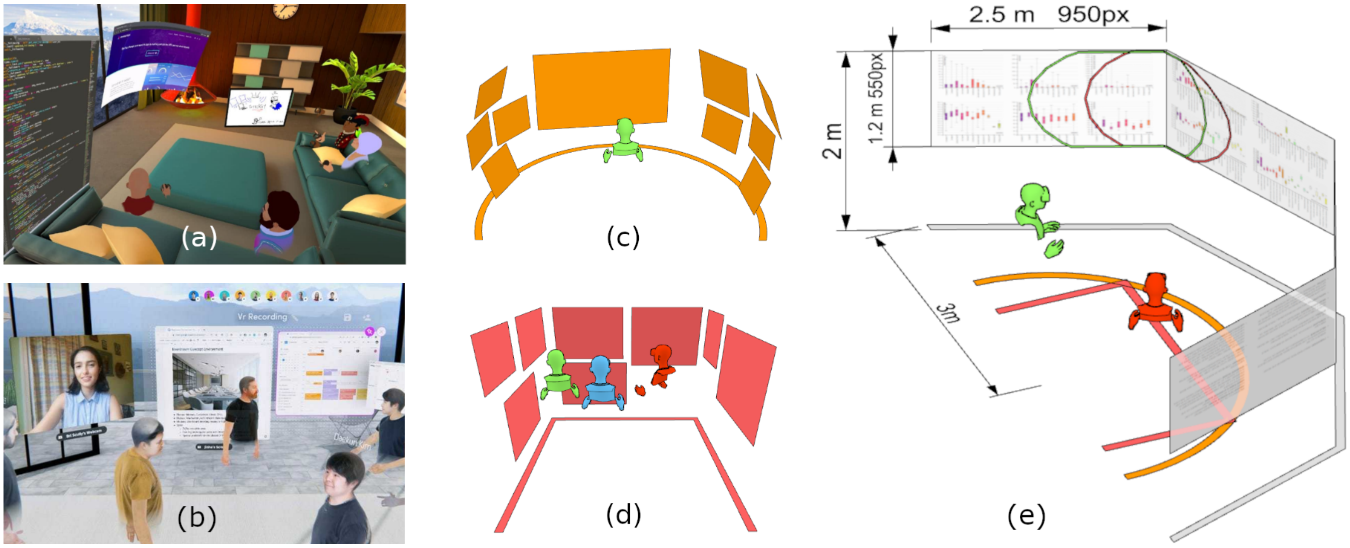


Figure 2: This figure shows on the left two collaborative environments: (a) Immersed [51], and (b) Spatial.io [5], which enable users to generate their configurations of 2D VR screens layouts. Figure (c) highlights circular VR scenario in cross-virtual analytics studies of Satriadi et al. [97] and (d) a squared configuration from the work of Lee et al. [60]. Our immersed collaborative virtual environment is depicted in (e). It consists of 4 VR screens with HTML web pages displayed in a cylindrical layout.

I am looking...") or interacting with the data (i.e. typing and selecting text). Further work by Pettersson et al. [89] explores how verbal communication is used to negotiate shared visual attention in the context of collaborative maps analysis on tabletop displays. Their analysis found three ways of referencing the visualized data: colour statements, size statements, and pointing gestures. Such studies highlight how verbal communication is used by uttering visualized labels or explicitly referring to the characteristics of the objects visualized. Thus, multiple strategies based on these different types of references can be used to improve visual cues through the verbal channel. In this study, we focus specifically on the explicit naming of visualized labels as this solution is a simple and efficient method (by calculating word similarity [68]) to refine visual attention.

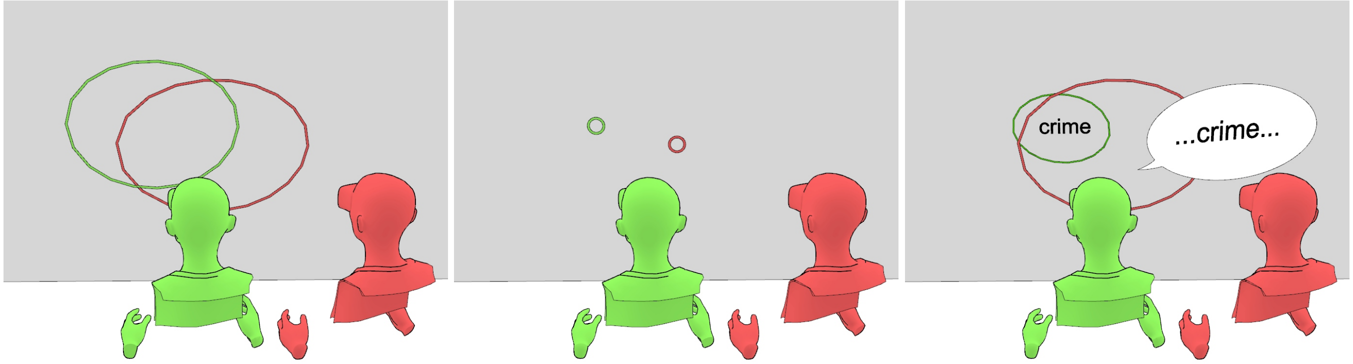
2.4 Speech in VR interaction

Speech interfaces have been widely adopted in a wide variety of interactive contexts [13, 41, 52, 57, 64, 65, 69, 109, 114, 120]. Speech interface involves various challenges, such as speech recognition, phrase interpretation, and interaction. Speech interaction has been used in numerous works, and how the user interacts is highly dependent on the task and the functionalities of the various released systems [2, 37, 40, 42, 100]. Speech interaction and VR met decades ago [28, 72, 73, 76] with the implementation of multi-modal systems with two possible approaches: fully interactive speech or "command and control". The first type was speaker-dependent because the variety of words and sentences forced the user into a training phase called enrollment [81]. Before the recent revolution of the natural language process (NLP), researchers adopted the paradigm of the 'Wizard of Oz' [9, 36, 63] to avoid technical limitations of the free-speech interfaces for both the recognition and the process phase.

However, the command and control system was speaker-independent as the limited number of words to be converted into commands did not require an enrolment stage. In particular, this approach presents advantages over keyboard input or gestures input [104] as the last ones necessitate practice. Furthermore, the interaction style derived from these two approaches originate from studies illustrating that vocabulary size can impact interaction [8, 105] as well as the awareness that a machine or human interprets speech [106]. Given these advantages, speech interaction was experimented with in medicine to treat social phobia [108], civil engineering to help with architectural design [25], and in dealing with the digital twin of complex machines such as airplanes [102].

With the advent of more powerful deep learning models for recognition and NLP, speaker-dependent systems became obsolete, as various services [4, 38, 39, 49, 75] can receive and process audio streaming that provides transcription almost in real-time. We use these capabilities in VR to alter visual cues according to the semantics of spoken sentences during collaboration. To our knowledge, such a multi-modal interface is applied and studied for the first time in exploratory data analysis to understand the head and gaze behaviour. We present a novel method for visual cues that are modified by collaborative verbal communication.

Within the collaborative scenario of 2D VR screens in cross-virtually analytics (Section 2.1) and low-cost eye-tracker-less VR headsets, we address the problem of conveying visual attention to achieve mutual awareness and support collaboration. Such a task is challenging, as the most obvious way of conveying visual attention is to depict eye-tracking information, which is unavailable on low-cost VR headsets. Therefore, we compare bi-directional head-based and eye-based cues eye-tracking-less VR headsets can support joint attention during collaboration. We explore an orthogonal approach to address the same problem by designing and implementing a novel



(a) CoV (b) Eye-tracking (c) CoV + speech

Figure 3: Experiment one conditions: (a) CoV: participants can see the CoV [16] (c) Eye-tracking: a visual cursor is displayed for each participant in the raw gaze location of the eye-gaze. (c) CoV+Speech: when the participant enunciate a word which is displayed on the VR screen the application search within the HTML context and if the word is within the CoV it shrinks the cone to around the bounding box of the matching word

multi-modal interface which leverages natural language processing (NLP) and head behaviour. NLP interprets natural collaborative verbal communication to improve gaze inference, thus conveying visual focus and supporting joint attention. Finally, we extend previously existing spatial verbal reference taxonomy by merging two different domain taxonomies, one from remote pair programming [26] and one from a collaborative analysis of Maps on tabletop displays [89] (Section 2.3). To our knowledge, no dataset merges information from the head, eye and verbal behaviour, so we fill this gap by creating a dataset that includes this information.

3 STUDY

We designed a within-group study that compares three conditions: Cone of Vision (CoV), Cone of Vision+Speech (CoV+Speech) and Eye-Gaze Cursor (Figure 3). Participants were embodied in an avatar ReadyPlayerMe[2] and could also use a hand pointer to reference the observed dataset. We used a Latin Square approach with 3 conditions. However, due to the number of participants, one order had one more pair of participants than the other[3].

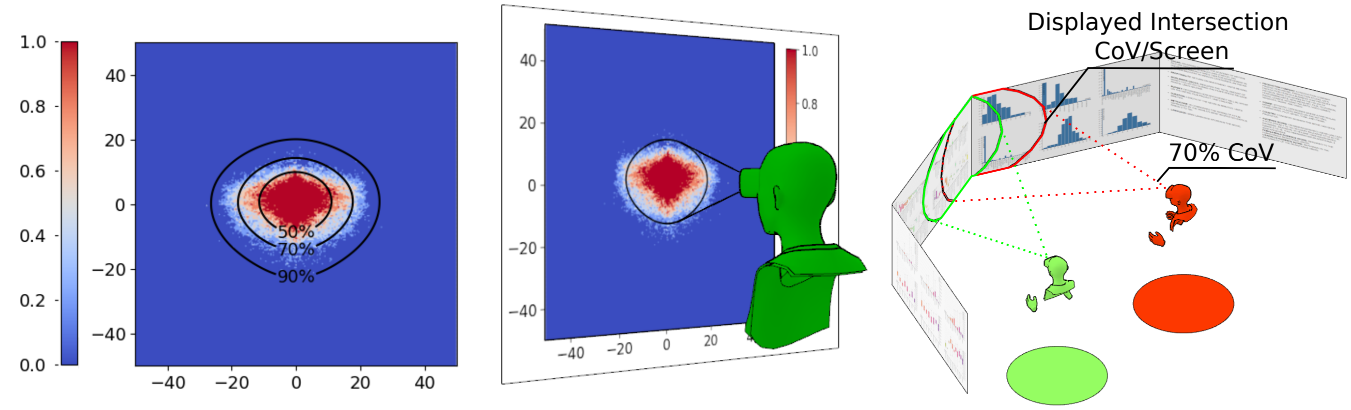
3.1 Conditions

3.1.1 Cone of Vision (CoV). Inspired by Bovo et al. [16], we use a different graphic element from the classic FoV frustum, called the Cone of Vision (CoV) (Figure 4). Geometrically, this visualisation is obtained by intersecting the cone that has the vertex in the centre of the head and the direction parallel to the head direction with the observed 2D surface. This depiction is designed to work with data displayed on 2D surfaces (such as panels or VR screens) but immersed in a 3D scenario (Figure 3). The main difference between the FoV frustum and the CoV is their spatial dimensionality, that is, 3D for the first and 2D for the second. However, both convey probabilistic information about the gaze location since they are aligned with the head. The CoV is displayed using the contour surrounding the area with the 70% probability of containing the users' fixations, achieved using the dataset of Agtzidis et al. [1].

3.1.2 Cone of Vision+Speech (CoV+Speech). The second visual cue is a combination of CoV and the effects of the user's verbal interaction with the system. Although we use the same CoV calculation as in the CoV condition, we designed a novel algorithm that modifies the CoV after processing the user's speech. We describe in Section 3.2.1 the dataset contained in the HTML pages rendered by the virtual screens in the 3D office. To extract the semantics of the speech, we first capture the audio of the user talking to his collaborator. Therefore, we stream this audio to an online speech service that transcribes the speech and returns a string to parse and process with NLP algorithms. Such information is then searched in the HTML context for those elements that contain keywords isolated by NLP. Then, we return their bounding box coordinates within the browser page and convert the local coordinates into world coordinates and add them to a list. The next phase is the modification of the current CoV. To reduce the CoV size, we determine the principal component of the coordinates by doing a linear regression. Then we calculate the standard deviation of the points along the principal component and along its orthogonal direction. Subsequently, we draw the ellipse using the coordinates of the centre of mass with the two standard deviations are the two radii of the ellipse. Ultimately, we interpolate between the CoV points and the ellipse points by a factor of 0.5 (Figure 5 (d)). The visual cue is displayed as shown in Figure 5 (e). Such a condition includes two different input channels: head-based position and orientation coming from the hardware of the HMD, and an analysis that converts the speech to a morphing function of the CoV. While the CoV calculation relies on internal code, a part of speech processing relies on an external service.

3.1.3 Eye-Gaze Cursor. The Eye-gaze cursor condition displays a graphical cursor at the gaze location on the VR screen. The position is calculated by calculating the intersection between the gaze direction and the VR screen. The cursor is visualised as a ring (Figure 3 (b)) and has a radius of $\frac{1}{3}$ of the radius of the fovea region

[2] https://readyplayer.me/
[3] We performed additional statistical analysis to verify that no ordering or learning effects were present, detailed in Section 4.



(a) Gaze 360 dataset AFM [1]

(b) 70% PBC to CoV

(c) Intersection CoV/Screens

Figure 4: (a) shows the eye distribution in the longitudinal and latitude direction for the dataset of Agtzidis et al. [1] and the number of eye samples captured by the percentile base contours (PBC). (b) illustrates the cone of vision in VR with the contour that contains 70% of the dataset (Figure taken from [16]). (c) depicts how the intersection between 70% CoV with the VR screen is achieved in a VR environment indicating the area of visual attention of a user.

determined by pilot testing to ensure that the cursor is noticeable. The eye-gaze cursor is subjected to noise from the eye-tracker.

3.2 Apparatus

For our study, we provide each participant with a PicoNeo 2 HMD with a resolution of 4K (3840 × 2160) at a refresh rate of 75Hz. The HMD has an embedded Tobii eye-tracker that works at 90Hz and a declared accuracy of 0.5 degrees. For verbal communication, we set up a Microsoft Teams connection using Bluetooth headphones and a microphone for the participants' communication, while we used PicoNeo 2 microphone to capture the audio streaming for the transcription. We designed and implemented our collaborative VR application made with Unity2020.3.34.f1, where two users shared the same digital space, but not physical, with three different ways of exchanging visual cues during the exploratory data analysis task. The visual cues of each user are displayed in two different colours: red for the local visual cue and green for the remote visual cue. All cues are refreshed at each Unity loop with a fixed rate of 50Hz.

3.2.1 VR Environment. We designed the 3D scene with the participants positioned in two locations close to each other in front of four panels positioned as in Figure 2. We developed a convex egocentric layout in line with the findings of [60, 97] described in Section 2.1; however, since we are not limited in space by physical constraints as [60], and we have more participants than [97] we set our environment to have a radius of 3m compared to the 2m setup in [97] (Figure 2). The participants were positioned in the initial locations for all the sessions without the possibility of translating their avatars to avoid obfuscating the other's participant view. The avatar was created and imported from ReadyPlayerMe. We used the torso version of a custom avatar and implemented lip synchronisation and eye synchronisation. The four VR screens contain charts rendered by an internal browser embedded in such panels, decoding information from HTML/javascript files where the datasets are contained. The information related to the keywords' position in such HTML is extracted to be used during the CoV+speech condition.

3.2.2 Data Visualisations. The three datasets used during the experiment are the "Hollywood movie gender bias" based on The Bechdel Test [12], the success of Hollywood movies with information taken from IMDB [50], and the insurance risk for cars taken from the UCI machine learning repository [98]. These datasets are also used in collaborative analysis tasks by Bovo et al. [16]. Each test includes 38 views on seven screens, one of which contains instructions. The visualisations contain scatter plots, stacked bar plots, histograms, and box and whisker plots. The dataset is stored in the GitHub repository [4], and the charts at the following link [5].

Speech to Text. Real-time captioning services provide transcriptions for spoken information from audio streams. Google Speech-To-Text[39], Microsoft Cognitive Services[75], Dialogflow [38], IBM Watson[49], Amazon Transcribe [4] are the most used services that allow integrating a real-time API transcription in a system. We choose Google Speech-To-Text as the one compatible with our requirements of the platform (Android) and framework (Unity). The Google Speech-to-Text service can be configured with different parameters such as language, sample rate, automatic punctuation, context adaptation, etc. We ran several pilots to optimise the accuracy and reduce the latency of the Google speech service. We used audio with a sample rate of 16000 Hz, determined the language as British English and consequentially hired participants that were native British speakers and filtered out punctuation.

[4] https://github.com/Collaborative-Immersive-Visual-Toolkit/ConeOfVision
[5] https://graphs-for-collaborative-vr.web.app/

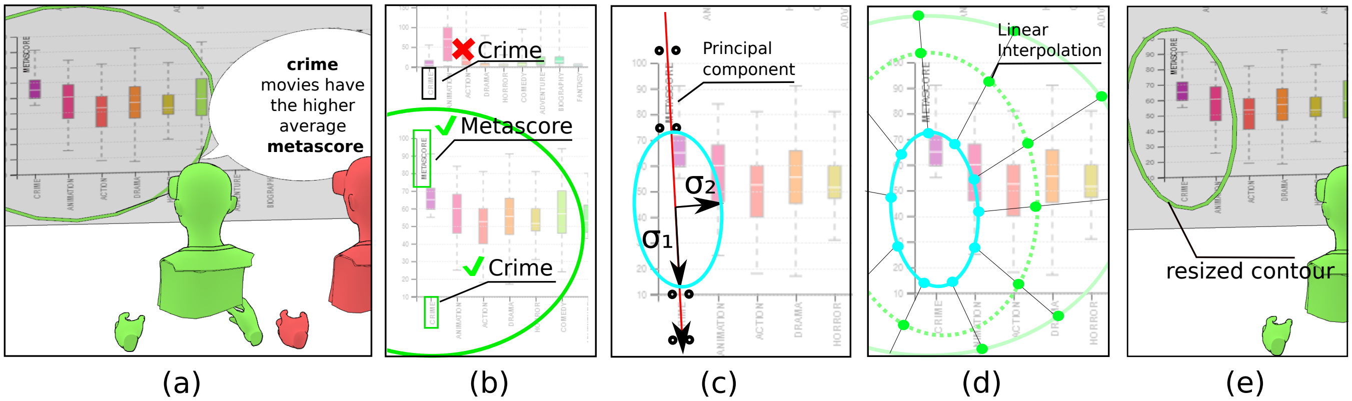


Figure 5: This sequence of figures shows the algorithm's steps to produce the visual cue from the combined action of CoV and speech inputs. From the left (a), CoV is visible on the VR screen, and the user speaks to the collaborator. After speech recognition, (b) the keywords present on the VR screen are found and filtered according to their positions. We accept the keywords inside the CoV and discard the others, and we fit an ellipse of such keywords distribution with the axis corresponding to the standard deviation of their components, X and Y (c). In (d), we interpolate the CoV with such an ellipse with a balanced ratio of 0.5. In (e), the resulting visual cue is proposed to the users.

3.3 Participants

We recruited 20 participants in two weeks(13 women, 7 men, M_{Age} = 29.4, SD_{Age} = 9.1) through an online platform managed by University College London [6]. We applied several inclusion criteria when performing a screening: being a native English speaker, having normal vision, and having a minimum education degree in high school. In particular, the latter criterion was to ensure that participants had sufficient knowledge to interpret the graphs of the visualization. In addition, we required participants to be confident in interpreting the charts we included in the study. Such charts consisted of bar or candlestick plots, histograms, and scatter plots. Each participant was self-assessed with a questionnaire, and we summarised the characteristics of such plots during the task presentation. One participant declared to be an expert VR user, five with average experience, six as occasional users, eleven with low experience and four with no experience. Participants received compensation of £15 each for a 90 min study. We incentivised participants to perform at their best by introducing of an additional reward of £15 each if they reported the highest number of valid insights among all the pairs. We recommended participants to collaborate instead of splitting their attention into different visualisation areas.

3.4 Procedure

Upon arrival, participants were asked to read the information sheet and sign the consent form. We carried out the experiment in the lab using two separate offices, one for each participant. Next, we explained the duration of the task and the three experimental conditions, allowing participants to test each for 1 min. We then asked participants to perform an exploratory data analysis task, extracting insights from the displayed visualisations. We provided participants with examples of valid insights. In our context, we describe a valid insight as a recorded speech where is conveyed a precise and deep understanding of two or more measures displayed on a graph or a series of graphs [80]. Next, we explained how to record insights and use the hand pointer. Once instructions were clear, participants were asked to wear the Pico Neo eye 2, perform an eye-tracking calibration process, connect to the virtual environment, and start the collaborative task. After all VR trials (i.e. experimental conditions), we asked participants to complete the questionnaire (Section 4.3).

At the end of the experiment, we conducted semi-structured interviews with each participant individually. Participants were asked to report cases in which each experimental condition helped with the assigned task and cases that did not. The study lasted between 75–90 minutes (M = 80, $S.D.$ = 12.7), and the duration of the trial lasted approximately 10 min (M = 13 m, SD = 3m). The stop condition was reached when the time was up (13 minutes).

3.5 Offline Analysis Methods

To understand the role of verbal communication concerning negotiating shared visual attention, we transcribed the recorded audio to achieve high-accuracy transcriptions (Section 3.2.2). We analysed the transcribed data to quantify how much participants utter displayed keywords to reference the data and how much they use alternative methods to reference (Section 3.5.2). Furthermore, we evaluated whether participants' utterances can be used in conjunction with the head direction to refine eye-gaze inference (Section 3.5.3).

3.5.1 Speech to Text. We used an offline speech recognition system to analyse the audio recordings with higher accuracy than the real-time system used in the study. This framework, released in the second part of September 2022, is the open-source project Whisper [93], developed by OpenAI. Such a system is trained with many hours of multilingual spoken language. Its end-to-end architecture is based on an encoder-decoder transformer [110] and produces very accurate text captions. We used Whisper with Python 3.8.3 and PyTorch 1.12.3 [86]. The manual analysis described in the following Section 3.5.2 ensured the transcription quality.

[6] https://uclpsychology.sona-systems.com/

Table 1: The summary of the verbal communication taxonomy we developed to analyse the video recordings.

(1) Keyword When a participant explicitly references an element by naming a specific word or name displayed in the data visualization. For example, “I would say that a movie’s budget is ...”.

(2) Sequential When a participant references an element by naming a number representing the element, for example, "shall we move to the third panel" or "look at the second element". An alternative expression might consist of the participant suggesting moving the focus of the collaboration to the next or previous graph/element/page, for example: "the one next to it".

(3) Color When a participant references an element by naming its colour: "..the green one.."

(4) Context reference When a participant references to an element based on its location within the page, for example, saying "..top left corner..“ or "..the graph below ... ". Such references can also be related to data o data, for example:" but the actual value is much lower".

(5) Pointer When a participant performs an implicit verbal reference by using the laser pointer to highlight an element directly and utter words like "..this..", "...over here...", ".. the graph we are looking at.. " or directly mentioning "...where I’m pointing..."

(6) User relative When a participant reference an element in relation to the frame of reference of the other user, such as: on your right/left/side, close to/far from you, above/below you ... for example, "..the graph on your right...".

(7) Temporal When a participant reference an element previously envisioned, such as before, after, or earlier... for example, "Let me check if I can see something else from the previous one"

(8) Visual cue When a participant uses a deictic reference such as this, that or here "..this graph over here.." or when a participant directly refers to the gaze visualization, for example, "Right where I am looking."

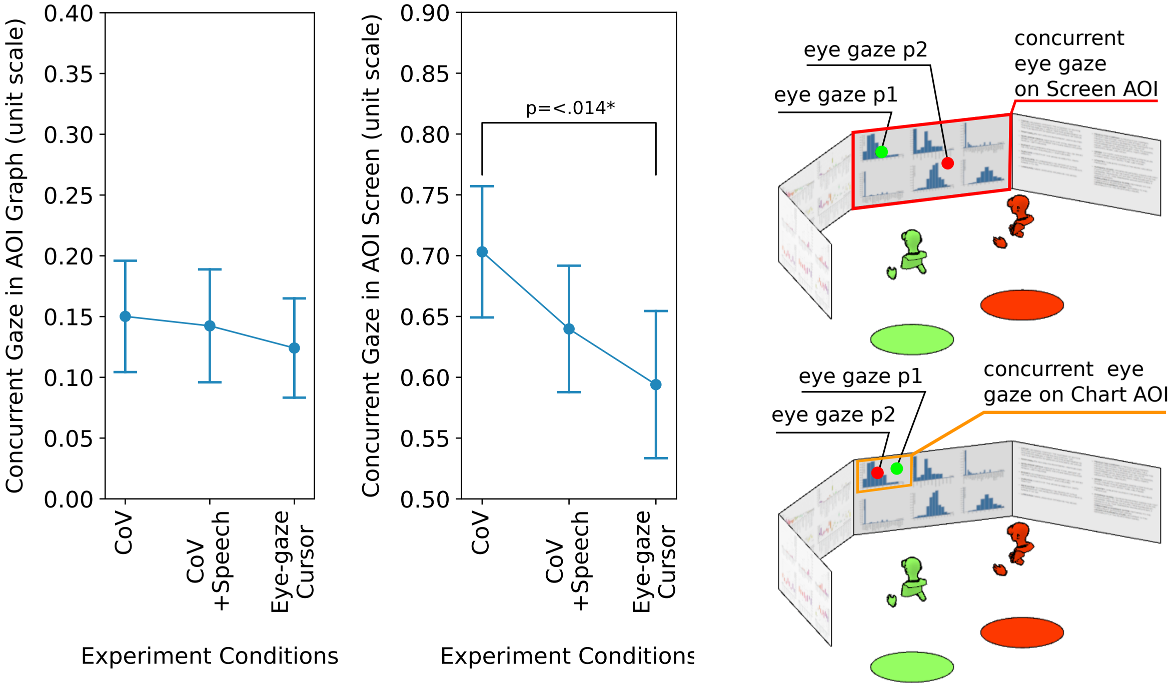
3.5.2 Classification of verbal references. We quantify how much participants utter displayed keywords to reference the data and how much they use alternative methods such as pointing gestures or implicitly referring to visual cues. We start by merging verbal reference taxonomies from D’Angelo and Begel [26] (i.e., remote pair programming; Table 1 (1,3,8)) and Pettersson et al. [89] (i.e., collaboration over tabletop maps visualizations; Table 1 (3, 5)) to include both text and visual element classes in the same context. We expand the resulting taxonomy by considering novel verbal references such as sequential statements (Table 1 (2)) that rely on implicit directional bias left-to-right (LTR) [33]. Furthermore, we add Context/User relative references (Table 1 (4, 6)), and temporal references (Table 1 (7)). The transcripts were analysed alongside video and audio recordings to gather the context of non-verbal communication (i.e., pairs being mutually aligned or orientated in opposite directions, performing pointing gestures with a laser pointer, etc.). Three coders performed the analysis: each transcribed trial was analysed by one coder and then reviewed by a second one; the third coder resolved any disparity between the first and second coders. The roles between coders rotate for each trial. For each transcribed sentence, we identify if it contains a verbal element aimed at identifying or changing the focus of the collaborative exploratory data analysis task concerning the visualized data. If the sentence contains a visual context negotiation, it is classified (i.e. using the aforementioned classes), and we identify which areas of interest the verbal communication was aiming for (i.e., data, chart, page). After classifying all transcriptions, we counted the number of occurrences each pair of participants performed in each experimental condition. The difference between the “Keyword” class and all other classes was immediately apparent, as the Keyword class was more prevalent than all other classes combined.

3.5.3 Head+Speech Gaze inference . We evaluated whether the utterances of keywords by the participants can be used in conjunction with the head direction to refine eye-gaze inference. We consider the data segments in which verbal communication is used to perform a fairer analysis. As shown in Figure 9b, we describe the steps we used to calculate the accuracy of Head/Gaze and Head/Gaze+Speech methods in our analysis with respect to the ground truth, the gaze information. Firstly, our model focuses on bi-grams, the last two spoken words by the user at any point in time. Such a number of words is optimal among the other n-grams. Secondly, we ran the well-established text similarity metric Recall-Oriented Understudy for Gisting Evaluation Lin [68] for longest common subsequences (ROUGE-L) through all the possible keywords located inside the CoV. Such similarity metrics range between 0 to 1, and we kept only positive scores. We determined the bounding boxes of the accepted keywords and evaluated which box is closest by calculating its Euclidean distance with the Head/Gaze. Finally, we calculate the RMSE for Head-Gaze and Head/Gaze+Speech with the eye gaze.

4 RESULTS

We structured the results into three sections. Section 4.1 reports our analysis of how the different visual cues affected concurrent (Section 4.1.1) and individual (Section 4.1.2) visual attention. Section 4.2 reports our analysis of how participants use verbal communication to negotiate the shared context of visual attention (Section 4.2.1) and how effective speech is an input to infer gaze during collaboration (Section 4.2.2). Section 4.3 reports our analysis to evaluate participants’ experience using different visual attention cues. Unless otherwise stated, the analysis was performed with a one-way repeated measures (RM) ANOVA (α=.05) with Condition (*COV*, *EyeGazeCursor*, *CoV+Speech*) as an independent variable. When the assumption of sphericity was violated, as tested with Mauchly’s test, Greenhouse-Geisser corrected values were used in the analysis. QQ-plots were used to validate the assumption of normality. Holm-corrected post-hoc tests were used when applicable.

As mentioned in Section 3, our Latin square was not fully balanced: due to the number of participants, three participant pairs started with the CoV condition, three with the CoV+Speech condition, and four with the Eye-Gaze Cursor condition. To understand whether our configuration led to potential biases and confounding factors, we performed a statistical analysis to see if there were any ordering or learning effects. We searched for an ordering effect in our joint attention analysis, "concurrent AOI", and individual visual attention, "gaze on own cone", by running an ANOVA analysis and found no evidence of such an effect. Furthermore, we checked



(a) Chart AOI

(b) Screen AOI

(c) AOI depiction

Figure 6: (a) on the y-axis the percentage of time in unit scale that users spent concurrently looking at a graph together (b) on the y-axis the percentage of time that users spent concurrently looking at the same page together. (a)(b) On the x-axis of both graphs, there are the experimental conditions. The error bars displayed represents 95% of the confidence interval. (c) show the granularity with which we measured the joint attention. On the top (c), the screen granularity, when collaborators are focused on the same screen. On the bottom (c), the chart granularity when collaborators are focused on the same chart.

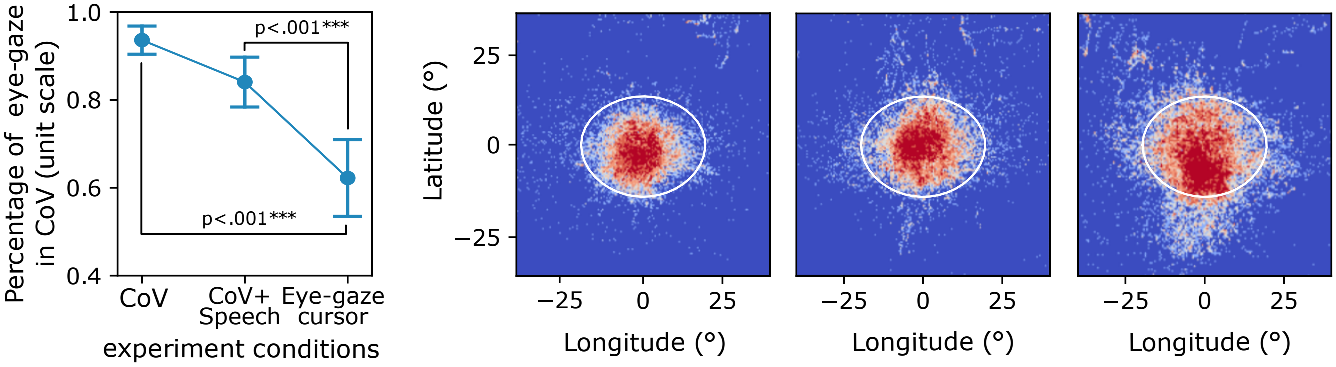
whether there was any learning effect by performing an ANOVA analysis on the number of insights generated across the conditions, and we found no learning effect. This suggests that the order in which the conditions were presented did not significantly impact the participants' behaviour or performance in the task, thus suggesting that the approach effectively improves collaboration and communication in data analysis. All statistical analyses are included in the supplementary materials.

4.1 Visual Attention

We segmented the recorded visual attention data by dividing it for each reported insight. Concurrent (Section 4.1.1) and individual (Section 4.1.2) visual attention behaviour were averaged for each insight segment. Participants reported 179 insights. Each pair reported a mean number of 17.6 insights (SD=5.25). Statistical analysis did not show significant differences in the number of insights. We then investigated the duration of the insight by segmenting each trial into the time between each insight. Participants spent, on average, 127.24s (SD=63.57s). We found no statistical difference between the conditions.

4.1.1 Concurrent Visual Attention in Areas Of Interest. We measured the amount of concurrent visual attention within pairs using a semantic segmentation of the visualization. This was achieved by calculating the percentage of time participants spent concurrently looking at the same AOI for each insight. We did this for two AOI levels: *screens* and *charts*. The screen AOI is defined as a full page, 981px wide, and 551px high (Figure 6c top). The chart AOI is defined as an individual chart, varying between 300px and 500px wide and 250px high (Figure 6c bottom). For the screen AOI level, the RM ANOVA analysis revealed a statistically significant difference ($F(2,116)$=4.191, p=.017). Post-hoc comparisons only revealed a significant difference between CoV ($M = 0.703, SD = 0.211$) and Eye-tracking ($M = 0.583, SD = 0.238$) conditions ($t = 2.895, p = .014$). The effect size test ($Cohen'sd = 0.377$) indicated a small to medium-sized effect. The results showed that the presence of the CoV led to an increase in the time participants spent concurrently looking at the same screen by 20%. For the chart AOI level, ANOVA analysis showed no significance. The participants spent on average 13% of each insight looking at the same graph, but no differences emerged from the different experimental conditions.

4.1.2 Individual visual attention behaviour. We measured how the experimental conditions affected the participants' eye-gaze behaviour by calculating the percentage of gaze samples that were within the visual cone (Figure 7a). This measure gives us an understanding of how much the gaze diverges from the head's direction. RM ANOVA showed a significant difference ($F(1,106)$=29.007,



(a) Eye-gaze inside CoV **(b) Gaze in CoV** **(c) Gaze in CoV+Speech** **(d) Gaze in eye-cursor**

Figure 7: (a) Eye-gaze within the projected cone: on the y-axis the percentage of time in unit scale that the eye-gaze was within the projected cone of vision, on the x-axis the experimental conditions. (b) (c) (d) Gaze sample distributions in head angular coordinates of respectively the experimental condition of (b) CoV (c) the CoV+speech and (d) eye-cursor. On the x-axis, the longitude in degrees; on the y-axis, the latitude in degrees. In each of (a)(b)(c), the CoV is depicted in white.

p<.001). Post hoc comparisons showed significant differences between the CoV (M = 0.929, SD = 0.151) and Eye-tracking (M = 0.541, SD = 0.383) conditions (t = 7.337, p < .001); and CoV+Speech (M = 0.829, SD = 0.232) and Eye-tracking conditions (t = 5.441, p < 0.001). All significant pairwise differences showed Cohen's d > 0.8. These results show that when the CoV's cone intersection is visualized, the participant's gaze is significantly less likely to be outside the depicted region. Therefore, the region visualization acts as a container. This effect is also present in the CoV+speech condition, where the region is dynamically modified by voice input. To further explore how these changes affect gaze behaviour, we plot gaze distributions for each of the conditions (Figure 7b, Figure 7c, Figure 7d). By comparing each distribution, two insights become apparent. First, the gaze distribution in the Eye-tracking condition is more sparse than in other conditions. Second, the Eye-tracking gaze distribution is most spread in the downward direction. This indicates that users moved their gaze further away from the head in the downward direction than in other directions, as previously reported [101] and that other conditions may lead to more head movement as the gaze remains within a smaller area.

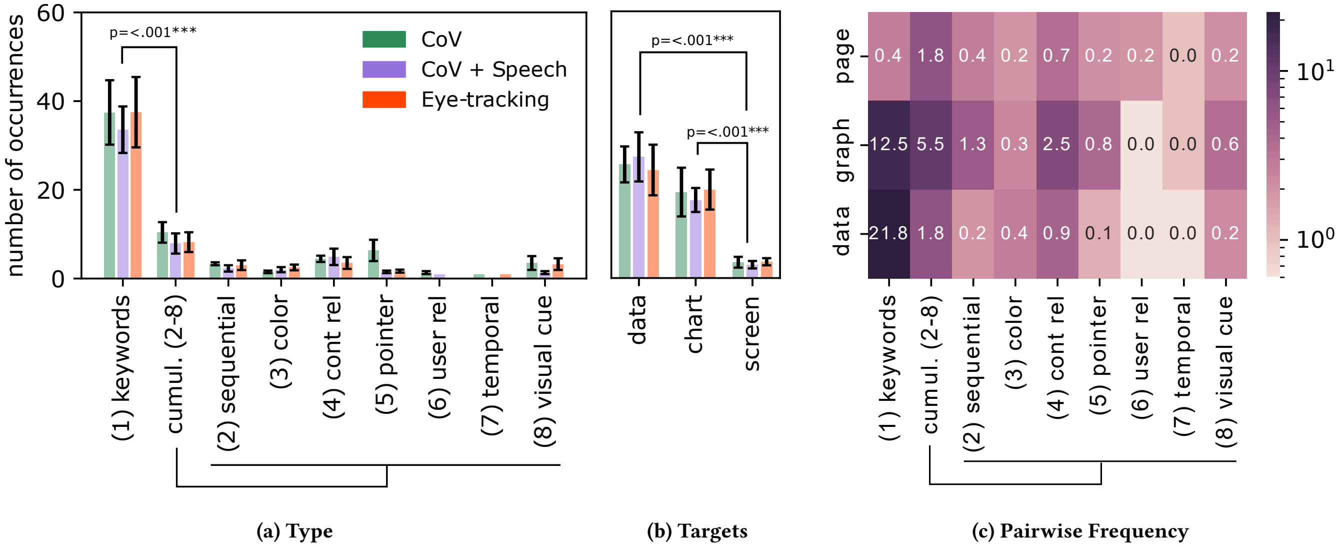
4.2 Speech and Visual Attention

4.2.1 Classification of Verbal References. We performed a classification aiming to quantify the types (Section 4.2.1) and the targets (Section 4.2.1) of verbal references used to negotiate the shared context of visual attention (Section 3.5.2). This classification helps us calculate the frequency of use of various verbal references (Table 1) during collaboration. Also, understanding what targets require more frequent referencing (Section 4.2.1) or if the typologies of verbal references frequency change when the target changes.

Verbal References Types. We performed an RM ANOVA analysis to see if there is any significant difference in the type of verbal reference ("(1) keywords" and "cumulative (2-8)") under experimental conditions (Figure 8a). There was a significant difference in the number of verbal references consisting of participants enunciating visualization keywords compared to the cumulative sum of all other types of verbal references ($F(1, 8)$=50.190, p<.001). These results highlight that our proposed approach, which intersects meanings extracted from verbal communication with keywords on the VR display, improves visual attention inference. RM ANOVA analysis showed no significant differences between the experimental conditions ($F(2, 16)$=0.211, p=0.812) nor any interactions between the reference type and the experimental conditions ($F(2, 16)$=0.013, p=.987). These results imply that participants did not change their verbal communication depending on the type of visual attention cue, and suggest that using verbal communication as an input does not impact the verbal behaviour of users.

Verbal References Targets. We performed an RM ANOVA to compare the effect of the experimental condition on the target AOI of the verbal reference performed (Figure 8b) and found a significant difference ($F(2, 14)$=30.368, p<.001). Post hoc tests showed a significant difference (p < .001) between the data (M = 25.333, SD = 11.313) and screen (M = 2.125, SD = 2.853) targets, as well as a significant difference (p < .001) between chart (M = 20.875, SD = 15.376) and screen. These differences showed that participants equally negotiated visual attention via verbal communication for both the "chart" and "data" levels, while for the "screen" level such negotiation is less necessary. The analysis did not show any main effect on experimental conditions ($F(2, 14)$=0.211, p=.934), nor an interaction between the target factors and the experimental conditions ($F(4, 14)$=0.211, p=0.696) meaning that participants did not change the way they verbally communicated depending on the type of visual attention cue displayed. These results imply that the experimental conditions did not alter verbal communication.

Pairwise comparison of Types and Targets of verbal references. We further explored the relationship between the target AOI and the verbal reference method by generating a pairwise frequency matrix



(a) Type

(b) Targets

(c) Pairwise Frequency

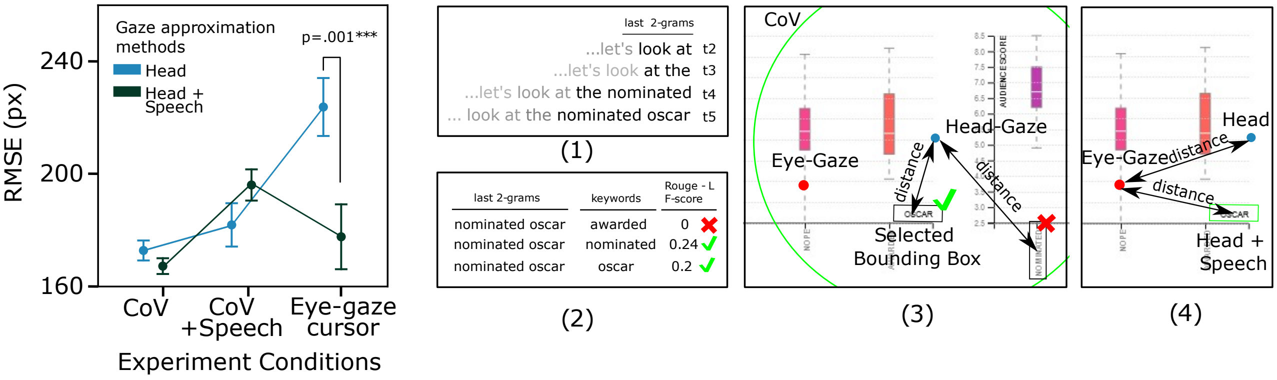
Figure 8: Analysis of verbal references to the dataset. (a) Bar charts representing the mean count of verbal reference occurrence per trial grouped by type (types description can be seen in Table 1). In the x-axis, the classes representing the different types of verbal communication references to the visualized data, and on the y-axis, the number of occurrences that each pair of participants did during a single 15 min trial. Error bars represent unbiased standard error of the mean Normalized by N-1. (b) Bar charts representing the mean count of verbal reference occurrence per trial grouped by a target area. On the x-axis, the classes representing the different AOI targets of the verbal communication references the visualized data, and on the y-axis, the number of occurrences that each pair of participants did during a single 15 min trial. Error bars represent unbiased standard error of the mean Normalized by N-1. (c) Verbal communication pairwise (type/AOI) frequencies, each cell shows the mean count per trial of each verbal references combination type/AOI. The Colour bar shows a logarithmic scale palette.

(Figure 8c) to further characterise the references. For example, it is clear that when participants refer to the chart AOI they tend to use a larger array of methods, which is visible by comparing the chart row and data row. This effect is even stronger for the page AOI where (1) keywords are no longer the most frequent method, as in data and chart AOIs.

4.2.2 Speech and Head-Gaze as Approximation of Eye Gaze. To evaluate whether speech helps to approximate eye gaze, we conducted an offline simulation across the three experimental conditions using the method described in Figure 9b. As the results of Questionnaire Q10 (Figure 10) highlighted that the implementation of speech-to-text used during the experiment had accuracy and latency problems, we transcribed the recorded audio of the experiment with a novel speech-to-text algorithm as described in Section 3.2.2. As we aim to evaluate verbal communication as a supplementary input for gaze inference, we performed such a comparison only for the time segments where participants were speaking. We calculated the Euclidean distance of head-gaze and our method (Figure 9b) from the eye gaze (ground truth) and the root-mean-square error (RMSE) of the distances. The RMSE was calculated in screen space; therefore, the results are pixels. The RM ANOVA results (Figure 9a) showed a main effect for the gaze approximation method ($F(1, 45)$=7.065, p=.011) and an interaction between the study condition and gaze approximation method ($F(1.501, 63.210)$=8.420, p=.002). Post hoc comparisons highlighted a difference between Head+Speech ($M = 178.327, SD = 142.517$) and Head ($M = 174.407, SD = 40.401$) in the Eye-tracking condition ($t = 4.703, p < .001$). This difference highlights how our method outperforms the head gaze as an approximation of eye gaze in the eye-tracking condition. An interesting insight is that this difference is only present in the eye-tracking condition where the CoV is not present; in the CoV and CoV+Speech conditions, the presence of the contours keeps the vision closer to the head-gaze, therefore, hindering the eye-gaze from spreading wider.

4.3 Questionnaire results

As part of the evaluation, we conducted a series of questionnaires with 11 questions answered on five-point Likert scales (Figure 10). The first two questions (Q1 and Q2) come from the System Usability Scale (SUS) questionnaire [11], and the second pair of questions (Q3 and Q4) come from the NASA Task Load Index (NASA TLX) questionnaire [43]. Q5 and Q6 aim to understand how visual attention cues affect attention allocation. Q5 question comes from an attention allocation questionnaire and is intended to measure how visual cues attract attention [55] and Q6 is intended to measure how much visual cues act as distraction [55]. We also introduced two questions (Q7 and Q8) related to communication to understand if, across conditions, participants altered their verbal communication behaviour (Q8) or the pointing-based communication behaviour (Q9). Q10 is specifically related to how accurately participants perceived the underlying technologies (i.e., eye-tracking, head-tracking, and



(a) RMSE gaze approximation methods

(b) Head+Speech method

Figure 9: In this sequence of images, we illustrate how we measure the accuracy of Head/Gaze and Head/Gaze+Speech method. (a) we first create a queue of the last 2-grams (b) we then run the ROUGE-L metric Lin [68] for each keyword within the CoV (which is a score from 0 to 1 indicating how similar two sequences are), we keep only the keywords which score above 0. (c) we calculate the euclidean distance between the head gaze and each of the bounding box and we determine the closest. (d) we measure the RMSE for both the head gaze and the Head+Speech using as ground truth the eye gaze.

speech-to-text). Finally, Q11 relates to how users actively alter their natural behaviour to improve feedback passed to collaborators via the visual attention cue.

To analyse Likert items, we used the Wilcoxon test, with a Kruskal–Wallis pairwise comparison for post-hoc testing where relevant. We only report the statistical results where significant differences were found. Tests revealed statistically significant differences for Q2 ($W = 5.0, p = 0.002$). A pairwise comparison revealed differences ($H = 9.26, p = .0097$) between CoV + speech ($M = 2.40, SD = 1.02$) and Eye-tracking conditions ($M = 3.35, SD = 0.57$). We also found a significant Wilcoxon test for Q10 ($W = 0.0, p < 0.001$). Pairwise comparisons revealed differences between the CoV+speech ($M = 1.65, SD = 1.11$) and Eye-tracking ($M = 3.15, SD = 0.73$) conditions ($H = 24.08, p < 0.001$), and between CoV+speech and CoV ($M = 3.30, SD = 0.78$) conditions ($H = 3.0, p < 0.001$). Responses to these questions highlight well-known problems when dealing with speech recognition technologies of latency during real-time use and difficulty in accent recognition [95]. We found no significant differences for Nasa TLX, attention, allocation, distraction, verbal communication behaviour, pointing-based communication behaviour, and active engagement questions.

5 QUALITATIVE ANALYSIS

Post-experiment semi-structured interviews were audio-recorded, fully transcribed and analysed through thematic analysis [19]. Our research questions focused on verbal communication as input for visual attention cues and, more broadly, the role of verbal communication in collaborative exploratory data analysis. The codes for the analysis were initially based on our research questions. Therefore, we focused on capturing aspects relative to the perception of the cues, comments about verbal communication, and the impact of the CoV. However, we also included codes from the interviews, such as lag and accuracy issues with speech-to-text technology, workarounds when the visual attention cues lacked precision, or CoV helping individuals to focus. The resulting 30 codes were grouped into three themes reported in the following subsections.

5.1 Comparing the different visual cues

Although most participants reported the eye-tracking condition as their favourite due to its precision, some noted that it did not allow them to focus on the charts and, for this reason, preferred the larger contoured region of the cones. For example, we heard:

"Eye-tracking was helpful If I was trying to say something specific. But then if either of us were talking about something broader then it would not be helpful because you missed the bigger picture. [P12]", or: *"...so sometimes during a discussion you are not talking about the specific data point but more about the broader and the specific cursor led me to focus on one thing at the detriment of other facts. [P16]"*

Some participants reported that gaze movements were hectic and distracting. For example, P13 mentioned: *"...it was like really distracting as I have ADHD, so it's hard to focus its hard to concentrate on the task, so I could not focus on my collaborator's visual cue because it was very confusing"*. Similar P3 said: *"...the eye-tracking visual cue it felt like he's pulling me away from the where I need to focus..."*.

Most participants reported the CoV to be most useful during the initial phase of mutual alignment. For example: *"it was very helpful to get aligned initially and so just for the moment what he needs to align and maybe the moment when the other one goes away [P4]"* or: *"it was helpful when the person I was collaborating with was talking about something, but I didn't know where she was looking for pictures so I could see the different colour area and turns towards it [P7]"*. P15 said that the CoV was useful to confirm the two participants were looking at the same thing: *"I found it helpful because I knew where she was looking, so we were able to basically be on the same page"*.

Several participants reported that the CoV contours allowed them to focus on the encircled data: *"...the good part of the cone was that it helped me focus on where I was looking at, so I won't look in under directions. [P9]"*. P17 explicitly stated that they 'liked' the

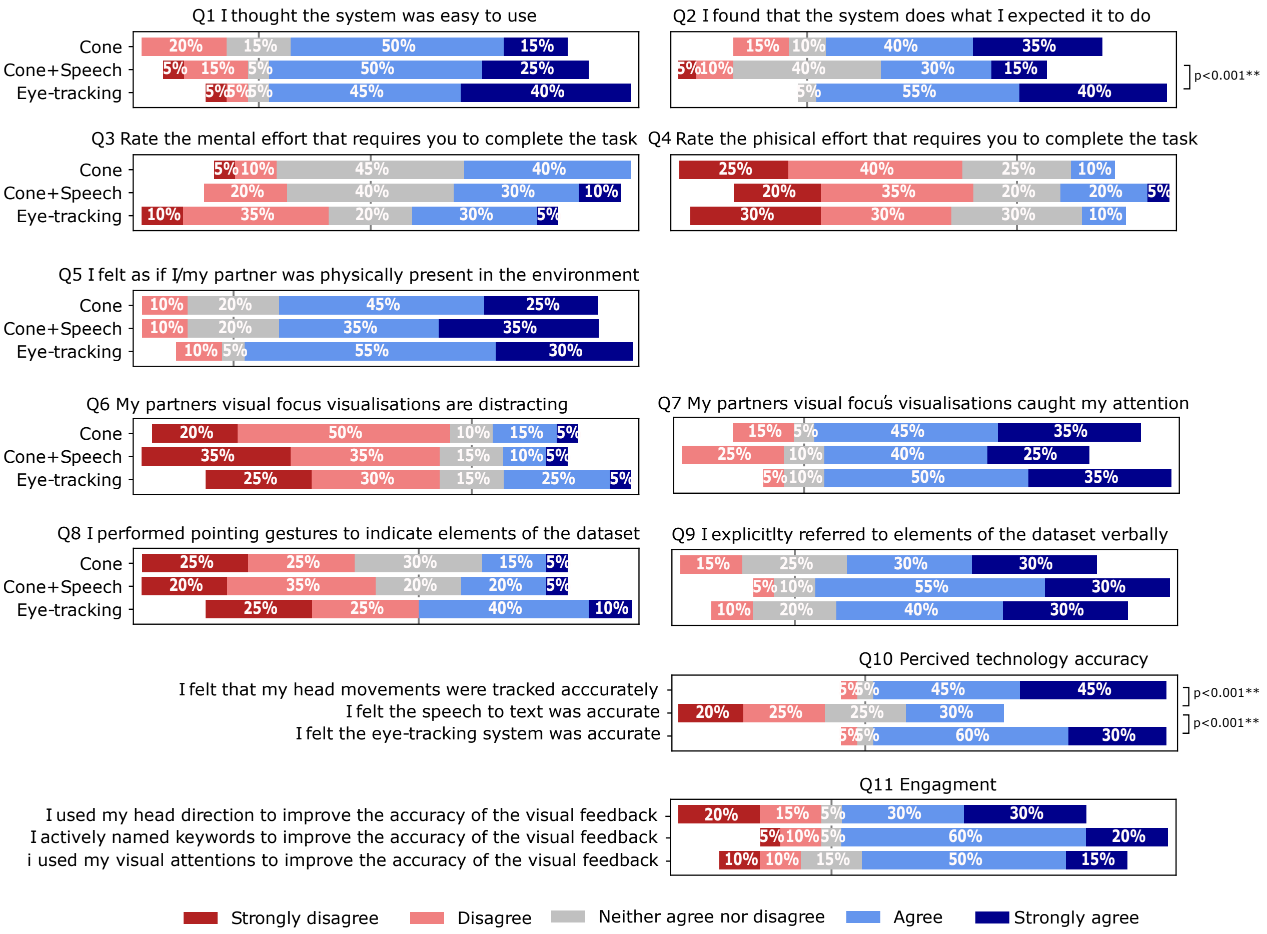


Figure 10: Questionnaire results

circle produced by the cones: *"We were both focussing on the same things. Like having a line around, you know what you are focussing your attention on, that kind of helps the kind of block out everything else"*. Similarly, P2 stated: *"The cone was helpful as it just helps me concentrate. I don't really feel like it was getting on the way"*

Participants commented that the bidirectional visual attention cues made them feel more coupled and accountable and not wandering around but staying on the same page. For example, P13 reported: *"The cone was helpful in that it kind of kept me in the room so I need to look at the same thing that she was describing, so she would see that I was looking at it was very helpful"*.

Participants complained that in the CoV + speech condition, the cue lagged considerably due to lag in recognition: *"I feel the speech was picking it up like in 10 seconds I did not find it to be reliable as when it was shirking down it would do so unreliable not in the specific area. [P13]"*. Similarly, P12 stated: *"it felt like it was slow"*.

5.2 Verbal communication as a fallback.

Most participants reported that the CoV worked well for keeping them on the same page, however, it was not very precise. So they reported using verbal communication as the default method to refine the accuracy of the CoV. For example: *"I think it was helpful seeing in general where the other person is looking at and then also aligning myself with that we did without yeah but it wasn't like with the specifics obviously it wasn't as helpful so I think we used more like verbal things to see like which chart each person's actual we read [P4]"*. P6 explicitly compared the CoV to the eye-tracking cues: *"with the fixed cone compared to the eye-tracker there is a lot more to verbalise so you had to find out like oh yeah, I have a look over there and get more details to say to the other collaborator"*.

Some participants reported feeling that the eye-tracking was sometimes not perfectly calibrated. Although initially they tried to compensate for the error by moving their eyes, they ended up using verbal communication to specify the location of the data they were discussing: *"the eye-tracking did get in the way a little bit because I feel like it was sometimes it wasn't calibrated that well, so I'm trying to fix my attention on the specific part of the chart and then the cursor was slightly off in another place so at first I was trying to compensate with my eyes but that wasn't working so I just had to ignore it and communicate the region of interest verbally [P2]"*.

5.3 CoV + Speech condition.

The CoV narrowing down on the region of interest was reported as a welcome confirmation of the shared visual attention: *"[it] was nice to have a confirmation of the cone shrinking as it increases the confidence that we were both looking at something [P5]"*. Participants commented that it rarely focused on the wrong area: *"I felt it rarely narrows down on the wrong area, but there was delay [P19]"*.

Sometimes participants went beyond explicitly looking for labels to refer to, and they attempted to direct the CoV narrowing with spatial voice commands (e.g., top left, bottom right, etc.). The positions were expected to be understood concerning the virtual screen at which the participant was looking. For example, P13 reported feeling disappointed that such a strategy did not work: *"I found also that was limited in the functionality as it would not recognise top left bottom right corners"*. P2 mentioned that the other participant instead quickly reacted to such spatial references: *"There were few charts in which some of the information on the y and x axis were the same; however, with them, I was mentioning top left or top right, and she would very quickly look there"*. Therefore, the future system which uses speech as inputs for visual attention could integrate recognizing this type of verbal, spatial references to inform the visual cue contours without the semantic knowledge of the context.

6 DISCUSSIONS

In this section, we discuss the results subdividing the findings in Joint Attention (in Section 6.1) and Individual Visual Attention (in Section 6.2). The final Section (Section 6.3) highlights the outcomes of the speech and semantic analysis.

6.1 Comparing Joint attention

Quantitative results indicate that concurrent joint attention on VR screens (Figure 6b) was significantly better (increment of 20%) in the CoV condition than in the eye-tracking conditions of Section 4.1.1. Despite the fact that the qualitative results showed that the participants preferred the eye-tracking method, the interviews also revealed the reasons for the success of the CoV method in joint attention on screen AOI. Participants mentioned that the cone was helpful to mutually orient and that they found the wider head-based cone contour much easier to find than the eye-tracking cursor. The size of the contour was not the only reason it was easier to find, it also moved less. Mutual alignment (i.e., orienting themselves along the general direction of collaborators) has been defined by [118] as an essential phase in negotiating a shared visual attention context and is significant for joint attention. Such results extend previous work related to uni-directional visual attention cues of [91], showing that in the context of 2D VR screens, bi-directional visual attention cues based on head direction outperform the overall mutual alignment when compared to eye-tracking cues.

Moreover, such a result extends the work of [54], which focuses on bi-directional eye behaviour-based attention cues, as well as the work of [16], which focuses on bi-directional head behaviour-based attention cues; by comparing bidirectional eye-based to head-based visual attention cues. Moreover, quantitative results indicate that, for concurrent joint attention, the charts area of interest (Figure 6a) all tree conditions perform equally well. This result is consistent with the Q6 question of the questionnaire, which indicates no statistical difference across the experimental conditions. Qualitative analysis of interviews suggests that this result could be due to the effectiveness of verbal communication in refining and specifying the location of the area of interest (see Section 5.2). While eye-tracking seems to be the favourite visual attention cue, not all VR headsets have eye-tracking capabilities, and perhaps future cheap VR headsets will never incorporate such capabilities. Our results show that in the context of exploratory data analysis tasks on VR screens, cheap VR headsets using bi-directional CoV can still lead to the same amount of joint attention, therefore, being effective.

6.2 How visual cues contours affect Individual visual attention

Quantitative results indicate that contour-based visual cues significantly alter individual visual attention by focussing it within the depicted visual cone with an increase of 20% of gaze sample within the contour when the contour is displayed compared to a condition in which the contour is not displayed (Section 4.1.2). Furthermore, qualitative results highlight that people perceive such a change and focus more on the area within the depicted region (Section 5.1). These results are consistent with the data [1] used to generate the CoV (Figure 4) and validated by [16] with two separate datasets [48, 59]. While the eye-tracking condition gaze samples are comparable in percentage with the sample of the original dataset, the CoV conditions sample shows the reported 20% increase. This could be because the participants were aware (consciously or unconsciously) that the CoV was a signal of their visual attention to the other person. In other words, they kept their visual attention within the highlighted region, where the other person would expect it to be. An alternative explanation is that the contour generates an attention tunnelling effect similar to the one in the small field of view AR/VR devices [56]. Such results could be interesting in relation to managing the attention of participants who have trouble concentrating, for example, because they are affected by ADHD [62]. A key implication of such an effect is that displaying the CoV can be useful when eye-tracking is not available. For example, metrics about the success of collaboration based on eye-tracking [112, 113] could be more accurately approximated using head-tracking. Furthermore, it might be possible to more effectively adapt interaction techniques that support eye-tracking [107] to only rely on head-tracking.

6.3 Speech as input for collaboration support

Our results indicate that participants utter keywords present in visualization was the primary way to perform verbal references (Section 4.2.1, Figure 8a), independently of verbal communication being used as input for visual attention cues (Figure 8b). Such results indicate that there is potential for speech recognition to be used to refine gaze inference in two ways: first, searching the spoken keywords in the users' visual field (i.e. CoV) and exploiting their location to refine the region of the visual cue is the best strategy for interpreting verbal attention if compared to the other strategies aimed at interpreting different verbal references (Table 1) because its frequency of usage (Table 1). Second, such verbal exchange occurs naturally, so our method does not require users to communicate differently. This suggests that when correctly implemented, such a method could lead to no learning/usage costs for the users. Such results are also consistent with the outcomes of the qualitative analysis Section 5.2: participants reported using verbal communication during collaboration as a method to overcome the lack of precision of the CoV (due to headset slippage [78]).

However, our results indicate that further technological advances in automatic speech recognition are needed for this approach to become viable since real-time state-of-the-art speech-to-text services (Section 3.2.2) still suffer from accuracy and lag issues. Such problems emerged from our qualitative analysis (Section 5.1) and the questionnaire responses (Section 4.3, Figure 10). Nevertheless, we obtained a high-accuracy transcription using the audio recording from the experiment with an offline state-of-the-art speech-to-text model (Section 3.5.1). We use such accurate transcription for an additional analysis where we test speech as an input for gaze inference across all conditions (Section 4.2.2, Figure 9a). These results illustrate that our method better approximates eye-gaze than the head gaze alone when the CoV is not pres(Section Section 4.2.2, Figure 9a). We show that in the eye-gaze cursor condition, speech improves the accuracy of the head gaze by about 50px on average (with statistical significance). Results from the same analysis also show that in the conditions in which the CoV or CoV+Speech is used, our method does not show improvements, most probably because the gaze is constrained by the visual cues (as discussed in Section 6.2). We release the dataset related to head, gaze and accurate transcript speech behaviour, hoping that this will foster research in this direction. Gaze inference models alternative to eye-tracking can be beneficial for those low-cost eye-tracker-less headsets or for offline analysis which lacks gaze data but has speech and head direction information.

7 FUTURE WORK AND LIMITATIONS

We recognise several limitations of our work. First, subjective responses highlighted voice detection issues during the CoV+Speech condition. The post hoc analysis addressed this issue with a different speech recognition engine, but it is possible that speech behaviour during the study was affected. Second, our qualitative analysis (Section 5.3) showed that participants often made spatial references (that is, "on my left" or "top right corner"). These references were not used by our technique. Further work could explore spatial references as explicit control of visual cues. In addition, other verbal references could be exploited to infer areas with specific colours, shapes, images, or synonyms of visible keywords (Section 3.5.2). Third, our current speech-based system is limited to HTML-based VR screens that must contain tags useful for verbal referencing. This aspect could be expanded to be viable in other environments, for example, by leveraging meta-information of 3D environments or the real-time segmentation of videos to provide a layer of meta-information to be queried for collaborative communication [94].

We also envision several avenues for future work. First, our analysis highlighted how individual visual attention is affected by the CoV; in future work, we could explore how CoV size affects this phenomenon. Second, the qualitative analysis showed different qualities of head-based and eye-tracking visual cues. The CoV is easier to find because it is wider and more stable, and the participants found it to be the best for mutual alignment. However, it lacks precision once mutual alignment is performed (Section 5.1). Meanwhile, the eye-tracker is precise but moves erratically, distracting users, and the cursor can be difficult to find. Future work could investigate a hybrid version that combines CoV and eye-tracking cues to gather their advantages. Finally, our dataset of human behaviour can be used for multiple purposes, such as evaluating leadership [3], competence skill [23, 31]), the success of collaboration [112, 113], and other behavioural analyses. However, the dataset is limited to 2-dimensional data, and future work can explore 3D data. Future challenges for 3D data include occlusions, illumination, and different approaches to generating visual cues.

8 CONCLUSIONS

In this paper, we investigate how using verbal communication with the Cone of Vision (CoV) can improve gaze inference and mutual awareness for exploratory data analysis in VR. We proposed a novel method named Speech-Augmented Cone-of-Vision which aims to dynamically balance the broadness of the cone of vision with the pinpoint abilities of verbal communication. We conducted a within-group study where ten pairs of participants performed collaborative data analysis tasks under three conditions. We used quantitative and qualitative methods, including participants' head and eye gaze behaviour, post-task questionnaires, and semi-structured interviews. Our findings suggest that visual attention cues based on head gaze (i.e. CoV and CoV + speech) are equally, if not more effective, in fostering joint attention than those based on eye-tracking. This leads to an increase of about 20% in concurrent gaze on the same VR screen. The questionnaire results and the analysis of the interviews suggest that the CoV+Speech condition was affected by the lag and limited accuracy of the real-time speech recognition implementation we used. To overcome this limitation, we used recorded audio to transcribe verbal communication using an offline high-accuracy speech-to-text model. Accurate transcription allowed us to classify the type of verbal references and validate our assumption that participants used keywords to negotiate shared visual attention. This approach allowed us to perform a non-real-time approximation of eye gaze using speech as input. The results of this analysis show that our proposed method improved the accuracy of gaze by 50px when it was not constrained by CoV regions. Therefore, we demonstrate that speech has the potential to be used as input to dynamically alter CoV cues by narrowing the focus of visual attention. To support further research in this area, we release the data collected in our study as a public research dataset. To the best of our knowledge, this is the first dataset on collaborative head, eye, and transcribed speech behaviour made publicly available.

ACKNOWLEDGMENTS

This work was supported in part by the EPSRC Industrial CASE student award No. EP/T517690/1 and the ERC Horizon 2020 programme (Grant No. 101021229 GEMINI: Gaze and Eye Movement in Interaction, and Grant No. 739578 RISE).

REFERENCES

[1] Ioannis Agtzidis, Mikhail Startsev, and Michael Dorr. 2019. 360-Degree Video Gaze Behaviour: A Ground-Truth Data Set and a Classification Algorithm for Eye Movements. In *Proceedings of the 27th ACM International Conference on*

Multimedia (Nice, France) *(MM '19)*. ACM, New York, NY, USA, 1007–1015. https://doi.org/10.1145/3343031.3350947

[2] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan. 2009. An audio indexing system for election video material. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Taipei, Taiwan, 4873–4876. https://doi.org/10.1109/ICASSP.2009.4960723

[3] Mariano Alcañiz, Elena Parra, and Irene Alice Chicchi Giglioli. 2018. Virtual reality as an emerging methodology for leadership assessment and training. *Frontiers in Psychology* 9 (2018). https://doi.org/10.3389/fpsyg.2018.01658

[4] Amazon. 2022. AWS Transcribe. https://aws.amazon.com/transcribe/

[5] Anand Agarwala; Jinha Lee. 2021. Spatial.io. https://spatial.io/

[6] Chris Anton and Rasmus Larsen. 2021. meetinvr. https://www.meetinvr.com/

[7] Rowel Atienza, Ryan Blonna, Maria Isabel Saludares, Joel Casimiro, and Vivencio Fuentes. 2016. Interaction techniques using head gaze for virtual reality. In *2016 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 110–114. https://doi.org/10.1109/TENCONSpring.2016.7519387

[8] C. Baber, B. Mellor, R. Graham, J.M. Noyes, and C. Tunley. 1996. Workload and the use of automatic speech recognition: The effects of time and resource demands. *Speech Communication* 20, 1 (1996), 37–53. https://doi.org/10.1016/S0167-6393(96)00043-X

[9] C Baber and RB Stammers. 1989. Is it natural to talk to computers: an experiment using the Wizard of Oz technique. *ED Megaw,(Ed.), Contemporary Ergonomics 1989* (1989).

[10] Huidong Bai, Prasanth Sasikumar, Jing Yang, and Mark Billinghurst. 2020. A User Study on Mixed Reality Remote Collaboration with Eye Gaze and Hand Gesture Sharing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376550

[11] Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction* 24, 6 (2008), 574–594. https://doi.org/10.1080/10447310802205776

[12] Alison Bechdel. 1985. Bechdel Test. https://en.wikipedia.org/wiki/Bechdel_test.

[13] Niels Ole Bernsen and Laila Dybkjaer. 2001. Exploring Natural Interaction in the Car. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*. University of Southern Denmark, Denmark, 75–79.

[14] Verena Biener, Daniel Schneider, Travis Gesslein, Alexander Otte, Bastian Kuth, Per Ola Kristensson, Eyal Ofek, Michel Pahud, and Jens Grubert. 2020. *Breaking the Screen: Interaction across Touchscreen Boundaries in Virtual Reality for Mobile Knowledge Workers*. Technical Report 12. 3490–3502 pages. https://doi.org/10.1109/TVCG.2020.3023567 arXiv:2008.04559

[15] B. Biguer, M. Jeannerod, and C. Prablanc. 1982. The coordination of eye, head, and arm movements during reaching at a single visual target. *Experimental Brain Research* 46, 2 (5 1982), 301–304. https://doi.org/10.1007/BF00237188

[16] Riccardo Bovo, Daniele Giunchi, Muna Alebri, Anthony Steed, Enrico Costanza, and Thomas Heinis. 2022. Cone of Vision as a Behavioural Cue for VR Collaboration. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 502 (nov 2022), 27 pages. https://doi.org/10.1145/3555615

[17] Riccardo Bovo, Daniele Giunchi, Enrico Costanza, Anthony Steed, and Thomas Heinis. 2022. Shall I describe it or shall I move closer? Verbal references and locomotion in VR collaborative search tasks. In *Proceedings of 20th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).

[18] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Hans Gellersen, Enrico Costanza, and Thomas Heinis. 2022. Real-Time Head-Based Deep-Learning Model for Gaze Probability Regions in Collaborative VR. In *2022 Symposium on Eye Tracking Research and Applications* (Seattle, WA, USA) *(ETRA '22)*. ACM, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3517031.3529642

[19] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[20] Wolfgang Büschel, Anke Lehmann, and Raimund Dachselt. 2021. MIRIA: A Mixed Reality Toolkit for the In-Situ Visualization and Analysis of Spatio-Temporal Interaction Data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 470, 15 pages. https://doi.org/10.1145/3411764.3445651

[21] Simon Butscher, Sebastian Hubenschmid, Jens Müller, Johannes Fuchs, and Harald Reiterer. 2018. Clusters, Trends, and Outliers: How Immersive Technologies Can Facilitate the Collaborative Analysis of Multidimensional Data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173664

[22] Shiwei Cheng, Jialing Wang, Xiaoquan Shen, Yijian Chen, and Anind Dey. 2022. Collaborative eye tracking based code review through real-time shared gaze visualization. *Frontiers of Computer Science* 16, 3 (2022), 1–11.

[23] Nusrat Choudhury, Nicholas Gélinas-Phaneuf, Sébastien Delorme, and Rolando Del Maestro. 2013. Fundamentals of Neurosurgery: Virtual Reality Tasks for Training and Evaluation of Technical Skills. *World Neurosurgery* 80, 5 (2013), e9–e19. https://doi.org/10.1016/j.wneu.2012.08.022

[24] Maxime Cordeil, Tim Dwyer, Karsten Klein, Bireswar Laha, Kim Marriott, and Bruce H. Thomas. 2017. Immersive Collaborative Analysis of Network Connectivity: CAVE-style or Head-Mounted Display? *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 441–450. https://doi.org/10.1109/TVCG.2016.2599107

[25] Luís Coroado, Tiago Pedro, Jorge D'Alpuim, Sara Eloy, and MiguelSales Dias. 2015. Viarmodes: visualization and interaction in immersive virtual reality for architectural design process. (2015).

[26] Sarah D'Angelo and Andrew Begel. 2017. Improving Communication Between Pair Programmers Using Shared Gaze Awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 6245–6290. https://doi.org/10.1145/3025453.3025573

[27] David Whelan. 2021. engage vr. https://engagevr.io/

[28] Denis V. Dorozhkin and Judy M. Vance. 2002. Implementing Speech Recognition in Virtual Reality *(International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. Volume 1: 22nd Computers and Information in Engineering Conference)*. 61–65. https://doi.org/10.1115/DETC2002/CIE-34390

[29] Jay S. Efran. 1968. Looking for approval: Effects on visual behavior of approbation from persons differing in importance. *Journal of Personality and Social Psychology* 10, 1 (1968), 21–25. https://doi.org/10.1037/h0026383

[30] Barrett Ens, Juan David Hincapié-Ramos, and Pourang Irani. 2014. Ethereal planes: a design framework for 2D information space in 3D mixed reality environments. In *Proceedings of the 2nd ACM symposium on Spatial user interaction*. 2–12.

[31] Charles Faure, Annabelle Limballe, Benoit Bideau, and Richard Kulpa. 2020. Virtual reality to assess and train team ball sports performance: A scoping review. *Journal of Sports Sciences* 38, 2 (2020), 192–205. https://doi.org/10.1080/02640414.2019.1689807

[32] Nadia Fereydooni and Bruce N Walker. 2020. *Virtual Reality as a Remote Workspace Platform : Opportunities and Challenges*. Technical Report.

[33] Meghan E. Flath, Austen K. Smith, and Lorin J. Elias. 2019. Cultural differences in lateral biases on aesthetic judgments: The effect of native reading direction. *Culture and Brain* 7, 1 (2019), 57–66. https://doi.org/10.1007/s40167-018-0062-6

[34] Mike Fraser, Steve Benford, Jon Hindmarsh, and Christian Heath. 1999. Supporting Awareness and Interaction through Collaborative Virtual Interfaces. In *Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology* (Asheville, North Carolina, USA) *(UIST '99)*. ACM, New York, NY, USA, 27–36. https://doi.org/10.1145/320719.322580

[35] B. Fröhler, C. Anthes, F. Pointecker, J. Friedl, D. Schwajda, A. Riegler, S. Tripathi, C. Holzmann, M. Brunner, H. Jodlbauer, H. C. Jetter, and C. Heinzl. 2022. A Survey on Cross-Virtuality Analytics. *Computer Graphics Forum* 41, 1 (2022), 465–494. https://doi.org/10.1111/cgf.14447

[36] Daniele Giunchi, Alejandro Sztrajman, Stuart James, and Anthony Steed. 2021. Mixing Modalities of 3D Sketching and Speech for Interactive Model Retrieval in Virtual Reality. In *ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21)*. ACM, New York, NY, USA, 144–155. https://doi.org/10.1145/3452918.3458806

[37] James R. Glass, Timothy J. Hazen, D. Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay. 2007. Recent progress in the MIT spoken lecture processing project. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*. 2553–2556. http://www.isca-speech.org/archive/interspeech_2007/i07_2553.html

[38] Google. 2022. Dialogflow. https://dialogflow.cloud.google.com

[39] Google. 2022. Speech to Text. https://cloud.google.com/speech-to-text

[40] Masataka Goto, Jun Ogata, and Kouichirou Eto. 2007. Podcastle: a web 2.0 approach to speech recognition research. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*. Curran Associates, Inc., Antwerp, Belgium, 2397–2400. http://www.isca-speech.org/archive/interspeech_2007/i07_2397.html

[41] Alexander Gruenstein, Bo-June Paul Hsu, James Glass, Stephanie Seneff, Lee Hetherington, Scott Cyphers, Ibrahim Badr, Chao Wang, and Sean Liu. 2008. A Multimodal Home Entertainment Interface via a Mobile Device. In *Proceedings of the ACL-08: HLT Workshop on Mobile Language Processing*. Association for Computational Linguistics, Columbus, Ohio, 1–9. https://www.aclweb.org/anthology/W08-0801

[42] J. H. L. Hansen, Rongqing Huang, P. Mangalath, Bowen Zhou, M. Seadle, and J. R. Deller. 2004. SPEECHFIND: spoken document retrieval for a national gallery of the spoken word. In *Proceedings of the 6th Nordic Signal Processing Symposium, 2004. NORSIG 2004.* IEEE, Espoo, Finland, 1–4. https://doi.org/10.1109/TSA.2005.852088

[43] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52, C (1988), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-

9

[44] Keita Higuch, Ryo Yonetani, and Yoichi Sato. 2016. Can Eye Help You? Effects of Visualizing Eye Fixations on Remote Collaboration Scenarios for Physical Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5180–5190. https://doi.org/10.1145/2858036.2858438

[45] Jon Hindmarsh, Mike Fraser, Christian Heath, Steve Benford, and Chris Greenhalgh. 1998. Fragmented Interaction: Establishing Mutual Orientation in Virtual Environments. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '98)*. ACM, New York, NY, USA, 217–226. https://doi.org/10.1145/289444.289496

[46] Adrian H Hoppe, Florian van de Camp, and Rainer Stiefelhagen. 2020. Enabling interaction with arbitrary 2D applications in virtual environments. In *International Conference on Human-Computer Interaction*. Springer, 30–36.

[47] Alex Howland. 2021. Virbela. https://www.virbela.com/

[48] Zhiming Hu, Congyi Zhang, Sheng Li, Guoping Wang, and DInesh Manocha. 2019. SGaze: A Data-Driven Eye-Head Coordination Model for Realtime Gaze Prediction. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (2019), 2002–2010. https://doi.org/10.1109/TVCG.2019.2899187

[49] IBM. 2022. Watson. https://www.ibm.com/watson

[50] IMDb. 1990. IMDb Datasets. https://www.imdb.com/interfaces/.

[51] Immersed Company. 2022. Immersed. https://immersed.com/

[52] Jürgen M. Janas. 1986. The Semantics-Based Natural Language Interface to Relational Databases. In *Cooperative Interfaces to Information Systems*, Leonard Bolc and Matthias Jarke (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 143–188. https://doi.org/10.1007/978-3-642-82815-7_6

[53] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. 2021. EyemR-Vis: Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. ACM, New York, NY, USA, Article 283, 7 pages. https://doi.org/10.1145/3411763.3451844

[54] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. 2021. EyemR-Vis: Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM. https://doi.org/10.1145/3411763.3451844

[55] Allison Jing, Kieran William May, Mahnoor Naeem, Gun Lee, and Mark Billinghurst. 2021. EyemR-Vis: Using Bi-Directional Gaze Behavioural Cues to Improve Mixed Reality Remote Collaboration. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[56] Vinod Karar and Smarajit Ghosh. 2014. Attention tunneling: Effects of limiting field of view due to beam combiner frame of head-up display. *IEEE/OSA Journal of Display Technology* 10, 7 (2014), 582–589. https://doi.org/10.1109/JDT.2014.2311159

[57] Shinya Kikuchi and Partha Chakroborty. 1992. Car-following model based on fuzzy inference system. *Transportation Research Record* (1992), 82–91. http://onlinepubs.trb.org/Onlinepubs/trr/1992/1365/1365-009.pdf

[58] Seungwon Kim, Gun Lee, Mark Billinghurst, and Weidong Huang. 2020. The combination of visual communication cues in mixed reality remote collaboration. *Journal on Multimodal User Interfaces* (7 2020), 1–15. https://doi.org/10.1007/s12193-020-00335-x

[59] Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B. Pelz, and Gabriel J. Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific Reports* 10, 1 (2020), 1–23. https://doi.org/10.1038/s41598-020-59251-5

[60] Benjamin Lee, Xiaoyun Hu, Maxime Cordeil, Arnaud Prouzeau, Bernhard Jenny, and Tim Dwyer. 2021. Shared surfaces and spaces: Collaborative data visualisation in a co-located immersive environment. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1171–1181. https://doi.org/10.1109/TVCG.2020.3030450

[61] Gun A. Lee, Seungwon Kim, Youngho Lee, Arindam Dey, Thammathip Piumsomboon, Mitchell Norman, and Mark Billinghurst. 2017. Improving Collaboration in Augmented Video Conference using Mutually Shared Gaze. In *ICAT-EGVE 2017 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*, Robert W. Lindeman, Gerd Bruder, and Daisuke Iwai (Eds.). The Eurographics Association. https://doi.org/10.2312/egve.20171359

[62] J.M. Lee, B.H. Cho, J.H. Ku, J.S. Kim, J.H. Lee, I.Y. Kim, and S.I. Kim. 2001. A study on the system for treatment of ADHD using virtual reality. In *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 4. 3754–3757 vol.4. https://doi.org/10.1109/IEMBS.2001.1019654

[63] Minkyung Lee and Mark Billinghurst. 2008. A Wizard of Oz Study for an AR Multimodal Interface. In *Proceedings of the 10th International Conference on Multimodal Interfaces* (Chania, Crete, Greece) *(ICMI '08)*. ACM, New York, NY, USA, 249–256. https://doi.org/10.1145/1452392.1452444

[64] Fei Li and Hosagrahar V Jagadish. 2014. NaLIR: An Interactive Natural Language Interface for Querying Relational Databases. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data* (Snowbird, Utah, USA) *(SIGMOD '14)*. ACM, New York, NY, USA, 709–712. https://doi.org/10.1145/2588555.2594519

[65] Sean Li and Xiaojun (Jenny) Yuan. 2018. A Review of the Current Intelligent Personal Agents. In *HCI International 2018 – Posters' Extended Abstracts*, Constantine Stephanidis (Ed.). Springer International Publishing, Cham, 253–257. https://doi.org/10.1007/978-3-319-92270-6_35

[66] Yin Li, Alireza Fathi, and James M. Rehg. 2013. Learning to Predict Gaze in Egocentric Video. In *2013 IEEE International Conference on Computer Vision*. 3216–3223. https://doi.org/10.1109/ICCV.2013.399

[67] Yuan Li, Feiyu Lu, Wallace S Lages, and Doug Bowman. 2019. Gaze Direction Visualization Techniques for Collaborative Wide-Area Model-Free Augmented Reality. In *Symposium on Spatial User Interaction* (New Orleans, LA, USA) *(SUI '19)*. ACM, New York, NY, USA, Article 11, 11 pages. https://doi.org/10.1145/3357251.3357583

[68] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[69] Gustavo López, Luis Quesada, and Luis A. Guerrero. 2018. Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In *Advances in Human Factors and Systems Interaction*, Isabel L. Nunes (Ed.). Springer International Publishing, Cham, 241–250. https://doi.org/10.1007/978-3-319-60366-7_23

[70] Tahir Mahmood, Erik Butler, Nicholas Davis, Jian Huang, and Aidong Lu. 2018. Building Multiple Coordinated Spaces for Effective Immersive Analytics through Distributed Cognition. In *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*. IEEE, 1–11. https://doi.org/10.1109/BDVA.2018.8533893

[71] Mark McGill, Aidan Kehoe, Euan Freeman, and Stephen Brewster. 2020. Expanding the Bounds of Seated Virtual Workspaces. *ACM Transactions on Computer-Human Interaction* 27, 3 (2020). https://doi.org/10.1145/3380959

[72] Scott McGlashan. 1995. Speech interfaces to virtual reality.

[73] Scott McGlashan and Tomas Axling. 1996. A speech interface to virtual environments.

[74] Meta. 2022. Infinite Office. https://www.workplace.com/metaverse-work-infinite-office

[75] Microsoft. 2022. Azure Cognitive Services. https://azure.microsoft.com/

[76] J. Muller, C. Krapichler, Lam Son Nguyen, K. Hans Englmeier, and M. Lang. 1998. Speech interaction in virtual reality. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, Vol. 6. 3757–3760 vol.6. https://doi.org/10.1109/ICASSP.1998.679701

[77] David Nguyen and John Canny. 2005. MultiView: Spatially Faithful Group Video Conferencing. (2005), 799–808. https://doi.org/10.1145/1054972.1055084

[78] Diederick C. Niehorster, Thiago Santini, Roy S. Hessels, Ignace T.C. Hooge, Enkelejda Kasneci, and Marcus Nyström. 2020. The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods* 52, 3 (2020), 1140–1160. https://doi.org/10.3758/s13428-019-01307-0

[79] Arthur Nishimoto and Andrew E Johnson. 2019. Extending Virtual Reality Display Wall Environments Using Augmented Reality. In *Symposium on Spatial User Interaction* (New Orleans, LA, USA) *(SUI '19)*. ACM, New York, NY, USA, Article 7, 5 pages. https://doi.org/10.1145/3357251.3357579

[80] Chris North. 2006. Toward measuring visualization insight. *IEEE Computer Graphics and Applications* 26, 3 (2006), 6–9. https://doi.org/10.1109/MCG.2006.70

[81] JM Noyes. 1993. Speech technology in the future. In *Interactive speech technology: Human factors issues in the application of speech input/output to computers*. Taylor & Francis London, 189–208.

[82] Marc-Antoine Nüssli. 2011. Dual Eye-Tracking Methods for the Study of Remote Collaborative Problem Solving. *PhD Thesis, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE* 5232 (2011). https://doi.org/10.5075/epfl-thesis-5232

[83] Eyal Ofek, Jens Grubert, Michel Pahud, Mark Phillips, and Per Ola Kristensson. 2020. *Towards a Practical Virtual Office for Mobile Knowledge Workers*. Technical Report. arXiv:2009.02947 http://arxiv.org/abs/2009.02947

[84] Jacob L Orquin, Nathaniel JS Ashby, and Alasdair DF Clarke. 2016. Areas of interest as a signal detection problem in behavioural eye-tracking research. *Journal of Behavioral Decision Making* 29, 2-3 (2016), 103–115.

[85] Santtu Parikka and Juha Ruistola. 2021. Glue Collaboration. https://glue.work/

[86] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[87] Leonardo Pavanatto, Chris North, Doug A. Bowman, Carmen Badea, and Richard Stoakley. 2021. Do we still need physical monitors? An evaluation of the usability of AR virtual monitors for productivity work. *Proceedings - 2021 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2021* (2021), 759–767. https://doi.org/10.1109/VR50410.2021.00103

[88] Jeff Pelz, Mary Hayhoe, and Russ Loeber. 2001. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research* 139, 3 (8

2001), 266–277. https://doi.org/10.1007/s002210100745

[89] Lars Winkler Pettersson, Andreas Kjellin, Mats Lind, and Stefan Seipel. 2010. On the role of visual references in collaborative visualization. *Information Visualization* 9, 2 (2010), 98–114. https://doi.org/10.1057/ivs.2009.2

[90] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. 2017. [POSTER] CoVAR: Mixed-Platform Remote Collaborative Augmented and Virtual Realities System with Shared Collaboration Cues. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. 218–219. https://doi.org/10.1109/ISMAR-Adjunct.2017.72

[91] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun Lee, and Mark Billinghurst. 2019. The effects of sharing awareness cues in collaborative mixed reality. *Frontiers Robotics AI* 6, FEB (2019). https://doi.org/10.3389/frobt.2019.00005

[92] Michael Prilla. 2019. "I Simply Watched Where She Was Looking at": Coordination in Short-Term Synchronous Cooperative Mixed Reality. *Proc. ACM Hum.-Comput. Interact.* 3, GROUP, Article 246 (dec 2019), 21 pages. https://doi.org/10.1145/3361127

[93] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *OpenAI Blog* (2022).

[94] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788. https://doi.org/10.1109/CVPR.2016.91

[95] Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, and Neil Sebire. 2022. It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. *ACM Trans. Comput.-Hum. Interact.* 29, 3, Article 25 (jan 2022), 41 pages. https://doi.org/10.1145/3484221

[96] Patrick Reipschlager, Tamara Flemisch, and Raimund Dachselt. 2021. Personal augmented reality for information visualization on large interactive displays. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1182–1192. https://doi.org/10.1109/TVCG.2020.3030460 arXiv:2009.03237

[97] Kadek Ananta Satriadi, Barrett Ens, Maxime Cordeil, Tobias Czauderna, and Bernhard Jenny. 2020. Maps Around Me: 3D Multiview Layouts in Immersive Spaces. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 201 (nov 2020), 20 pages. https://doi.org/10.1145/3427329

[98] Jeffrey Schlimmer. 1985. Automobile data set. http://archive.ics.uci.edu/ml/datasets/Automobile.

[99] Bertrand Schneider and Roy Pea. 2013. Real-time mutual gaze perception enhances collaborative learning and collaboration quality. *International Journal of Computer-Supported Collaborative Learning* 8, 4 (2013), 375–397. https://doi.org/10.1007/s11412-013-9181-4

[100] Lin shan Lee Sheng-yi Kong, Miao ru Wu, Che kuang Lin, Yi sheng Fu, Yungyu Chung, Yu Huang, and Yun-Nung Chen. 2009. NTU Virtual Instructor - A Spoken Language System Offering Services of Learning on Demand Using Video/Audio/Slides of Course Lectures. *ICASSP* (2009).

[101] Ludwig Sidenmark and Hans Gellersen. 2019. Eye, Head and Torso Coordination During Gaze Shifts in Virtual Reality. *ACM Trans. Comput.-Hum. Interact.* 27, 1, Article 4 (dec 2019), 40 pages. https://doi.org/10.1145/3361218

[102] Aziz Siyaev and Geun-Sik Jo. 2021. Towards Aircraft Maintenance Metaverse Using Speech Interactions with Virtual Objects in Mixed Reality. *Sensors* 21, 6 (2021). https://doi.org/10.3390/s21062066

[103] Narayanan Srinivasan, Priyanka Srivastava, Monika Lohani, and Shruti Baijal. 2009. Focused and distributed attention. In *Attention*, Narayanan Srinivasan (Ed.). Progress in Brain Research, Vol. 176. Elsevier, 87–100. https://doi.org/10.1016/S0079-6123(09)17606-9

[104] Alexander Winstan Stedmon. 2005. *Putting Speech in, taking speech out: human factors in the use of speech interfaces*. Ph. D. Dissertation. University of Nottingham.

[105] Alex W Stedmon and Chris Baber. 1999. Evaluating stress in the development of speech interface technology. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I-Volume I*. 545–549.

[106] Alex W. Stedmon, Harshada Patel, Sarah C. Sharples, and John R. Wilson. 2011. Developing speech input for virtual reality applications: A reality based interaction approach. *International Journal of Human-Computer Studies* 69, 1 (2011), 3–8. https://doi.org/10.1016/j.ijhcs.2010.09.002

[107] William Steptoe, Oyewole Oyekoya, Alessio Murgia, Robin Wolff, John Rae, Estefania Guimaraes, David Roberts, and Anthony Steed. 2009. Eye Tracking for Avatar Eye Gaze Control During Object-Focused Multiparty Interaction in Immersive Collaborative Virtual Environments. In *2009 IEEE Virtual Reality Conference*. 83–90. https://doi.org/10.1109/VR.2009.4811003

[108] Niels ter Heijden and Willem-Paul Brinkman. 2011. Design and evaluation of a virtual reality exposure therapy system with automatic free speech interaction. *Journal of CyberTherapy and Rehabilitation* 4, 1 (2011), 41–55.

[109] Amrita S. Tulshan and Sudhir Namdeorao Dhage. 2019. Survey on Virtual Assistant: Google Assistant, Siri, Cortana, Alexa. In *Advances in Signal Processing and Intelligent Recognition Systems*, Sabu M. Thampi, Oge Marques, Sri Krishnan, Kuan-Ching Li, Domenico Ciuonzo, and Maheshkumar H. Kolekar (Eds.). Springer Singapore, Singapore, 190–201. https://doi.org/10.1007/978-981-13-5758-9_17

[110] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[111] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. 2003. GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. ACM, New York, NY, USA, 521–528. https://doi.org/10.1145/642611.642702

[112] Maureen Villamor and Ma. Mercedes Rodrigo. 2018. Predicting Successful Collaboration in a Pair Programming Eye Tracking Experiment. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. ACM, New York, NY, USA, 263–268. https://doi.org/10.1145/3213586.3225234

[113] Hana Vrzakova, Mary Jean Amon, Angela E. B. Stewart, and Sidney K. D'Mello. 2019. Dynamics of Visual Attention in Multiparty Collaborative Problem Solving Using Multidimensional Recurrence Quantification Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300572

[114] David L. Waltz. 1978. An English Language Question Answering System for a Large Relational Database. *Commun. ACM* 21, 7 (July 1978), 526–539. https://doi.org/10.1145/359545.359550

[115] Peng Wang, Shusheng Zhang, Xiaoliang Bai, Mark Billinghurst, Weiping He, Shuxia Wang, Xiaokun Zhang, Jiaxiang Du, and Yongxing Chen. 2019. Head pointer or eye gaze: Which helps more in MR remote collaboration. In *26th IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2019 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 1219–1220. https://doi.org/10.1109/VR.2019.8798024

[116] Peng Wang, Shusheng Zhang, Xiaoliang Bai, Mark Billinghurst, Weiping He, Shuxia Wang, Xiaokun Zhang, Jiaxiang Du, and Yongxing Chen. 2019. Head Pointer or Eye Gaze: Which Helps More in MR Remote Collaboration?. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 1219–1220. https://doi.org/10.1109/VR.2019.8798024

[117] Nelson Wong and Carl Gutwin. 2010. Where are you pointing? The accuracy of deictic pointing in CVEs. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2. ACM Press, New York, New York, USA, 1029–1038. https://doi.org/10.1145/1753326.1753480

[118] Nelson Wong and Carl Gutwin. 2014. Support for Deictic Pointing in CVEs: Still Fragmented after All These Years'. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. ACM, New York, NY, USA, 1377–1387. https://doi.org/10.1145/2531602.2531691

[119] Yanxia Zhang, Ken Pfeuffer, Ming Ki Chong, Jason Alexander, Andreas Bulling, and Hans Gellersen. 2017. Look together: using gaze for assisting co-located collaborative search. *Personal and Ubiquitous Computing* 21, 1 (2017), 173–186. https://doi.org/10.1007/s00779-016-0969-x

[120] Lina Zhou, Mohammedammar Shaikh, and Dongsong Zhang. 2005. Natural Language Interface to Mobile Devices. In *Intelligent Information Processing II*, Zhongzhi Shi and Qing He (Eds.). Springer US, Boston, MA, 283–286. https://doi.org/10.1007/0-387-23152-8_37