Technical Report

UCAM-CL-TR-952
ISSN 1476-2986

Number 952

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Exploring the effect of spatial faithfulness on group decision-making

David Adeboye

October 2020

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted 2020 by the author for the degree of Master of Philosophy (Advanced Computer Science) to the University of Cambridge, Jesus College.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

Remote working is becoming increasingly popular and many large organisations are asking their employees to work from home. However, several studies have shown that groups who make decisions over videoconferencing take longer to complete tasks, are less effective and are less satisfied with the result. The ability for a communication medium to convey information, cues or symbols to its users has been theorised to influence team/communication performance. Videoconferencing fails to communicate these non-verbal behaviours, which provide complementary information to speech. For example, the inability to use gaze to help indicate the next speaker means that conversations over videoconferencing typically contain more explicit handovers such as names.

This thesis presents *Spatial*, a new spatially faithful videoconferencing application that captures the aspects of face-to-face conversations that are not available on standard systems. Unlike previous work, which requires specialised equipment or setups, *Spatial* focuses on work-from-home environments. *Spatial* aims to replicate the spatial characteristics of face-to-face conversations, using commodity hardware. It builds environments that ensure that both visual and auditory communication can be transmitted directionally and as wholly as possible. Using *Spatial* they can calibrate their working environments to ensure that their experience is free from perspective distortions.

We show that under *Spatial*, groups replicate conversation characteristics of face-to-face interactions. They completed a cooperative decision-making task in a shorter amount of time, took less time for turns and interrupted each other less.

Contents

List of Tables and Figures

Chapter 1

Introduction

The Internet and advances in technology have introduced and promoted new possibilities in working arrangements. Remote working has become increasingly popular [7] and many large companies are switching to virtual teams [8]. In the ongoing COVID-19 pandemic, a report stated that 71% of employers are struggling to adapt to remote work, with three in four employers asking employees to work from home [9]. Large communication firms are reporting significant upticks in their videoconferencing services [10, 11]; however, surveys suggest that teams' productivity has decreased [9]. Furthermore, *shifting communications to meet remote needs* is listed in the top three challenges of moving to virtual teams, and 35% of employers are grappling with changes in employee productivity [9]. The functionality and productivity of virtual teams are profoundly affected by the mediation of these communication and collaboration technologies [12].

Virtual teams typically use online videoconferencing as a replacement for face-to-face meetings, however, research has shown teams take more time to complete tasks over this medium [13]. Group decision-making tasks are particularly affected, with previous work suggesting that videoconferencing leads to a reduction in group effectiveness, increase in task duration and decrease in member satisfaction [13]. Many theories that explain and predict the impact of the communication medium on performance, focus on the amount of information, cues or symbols that the medium conveys. They make the argument that communication media which convey more information lead to individuals who are better able to understand more complex messages, and therefore increased task performance [14]. These non-verbal behaviours, which provide complementary information to speech, help to sustain and regulate the conversation [15]. However, even though some visual cues are used in a similar way to face-to-face communication, this does not translate to a similar task performance or accuracy [16].

In group conversations, visual cues such as head-turning and eye-gaze play a vital role in speaker switching. These two behaviours rely on the notion of *directionality* [17]. However, videoconferencing systems typically use a single camera input and a single video output. This single perspective view of the camera doesn't preserve the spatial characteristics of a face-to-face situation. In a real physical environment, different users do not share the same view of others. In particular, in group videoconferencing, the collapsed viewer effect, nicknamed the *Mona Lisa Effect* occurs (see Figure 1.1) when *observers feel that are looked at when other participants look at the camera* [18]. Selective gaze is the ability to *concentrate on a particular person, and all observers in the conversation are aware who the attention target is.* Selective gaze supports the regulation and syn-

Figure 1.1: *Screenshot of the gallery-view used in videoconferencing applications.*

chronisation of conversation, including providing feedback on how the listener perceives a verbal message [19]. Without selective gaze, conversations are filled with unnecessary conversation negotiations for a person to gain the floor. By building a spatial environment into our videoconferencing applications, we can use these visual cues again and have more efficient conversations.

Previous work on restoring these non-verbal cues, have used idealistic environments. This includes the use of specialised hardware such as video-tunnel translucent displays [20], custom-built hardware [21], or expensive hardware. However, it is important to determine whether gains in virtual team performance can be achieved with the use of commodity hardware, and thereby help companies adapt to new working environments. Work-from-home environments typically include a laptop connected to an external monitor. The goal of this project is to investigate, with the addition of an external webcam, whether we can recreate these spatial environments and whether they improve a team's performance.

Figure 1.2: *Photo of a curved monitor.*

1.1 Objectives and Scope

This project investigates the effect of spatial-faithfulness on work-from-home environments in group decision-making. Unlike previous work, we specifically focus on the growing number of virtual teams, who do not have access to specialised hardware or expensive commodity hardware. We aim to reduce the amount of setup work each participant has

to do to use such a system. Curved monitors (see Figure 1.2) do a better job of emulating a three-dimensional space than flat monitors, creating the illusion of peripheral vision. Their price and low uptake make them unsuitable for this project. Also, because of the time constraints of this project, building a full videoconferencing system from scratch would be infeasible. Therefore, as part of the design work, an open-source project was selected from a desired feature list. This project aims to investigate whether a spatially faithful videoconferencing solution will improve group decision making in a typical work-from-home environment. This could help solve the productivity gap between virtual teams and conventional workplace teams and provide the basis for further innovation in videoconferencing.

1.2 Contributions and Novelty

The contributions within this work as follows:

- I introduce *Spatial*, a spatially faithful videoconferencing application. It is a WebRTC videoconferencing application built on the existing OpenVidu platform which generates and manages spatially faithful video conferences. It also provides calibration functionality and performs media transformations to replicate a face-to-face conversation over video.

- I also design a new evaluation task for analysis of group decision-making behaviour. Evaluation of previous work either focused on structured conversations or other subjective aims such as trust. While important to study, they do not focus on the typical group decision-making conversations that occur over videoconferencing. This new evaluation task is based on the hidden profile paradigm, exhibited in many group decision-making situations.

- Finally, I present an evaluation of *Spatial* using the new evaluation task, to determine whether spatially faithful videoconferencing applications can help group decision-making through conversation and team performance analysis.

Our results show that groups using *Spatial*, compared to a gallery view system, took less time per turn, required fewer turns to complete the task, and interrupted each other less. These results show that introducing spatial cues to work-from-home environments is possible, and they improve our group decision-making process over conventional videoconferencing.

Chapter 2

Background

This chapter contextualises the research presented in this thesis by discussing the theoretical background, model, and previous work. It first discusses the architecture and technologies used by a typical videoconferencing application. Then we introduce the theoretical sociolinguistics background of conversation analysis, showing the use of gaze to present the problem it seeks to address. Finally, we present a model for spatial awareness, defining the problem and use it to compare prior work attempting to achieve spatial awareness.

2.1 Architecture of a videoconferencing application

This section discusses the typical architecture of a videoconferencing application and the technologies that it uses. The architecture shown is based on OpenVidu [22] and BigBlueButton [23], two popular open-source videoconferencing applications.

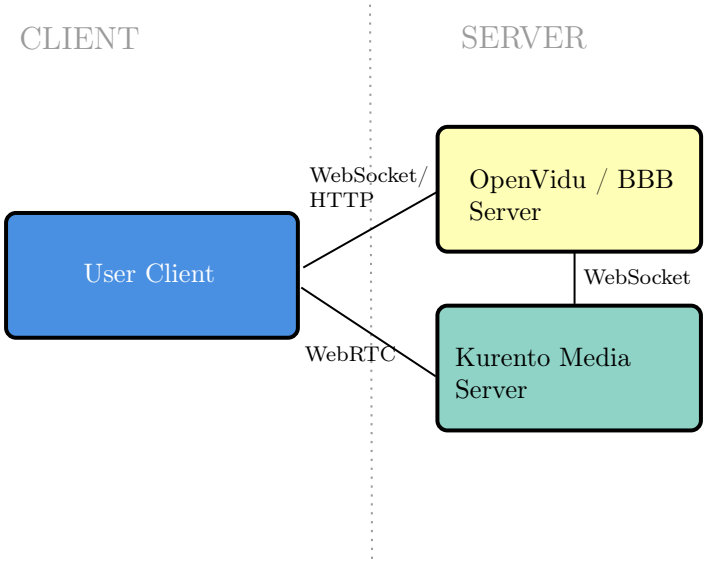

Figure 2.1: *Typical videoconferencing application architecture*

2.1.1 Client-Server model

Video conferencing applications typically use the client-server model. The client-server model partitions tasks between the provider of a service or resource, called *servers* and the

requesters of the service, called *clients*. Typical videoconferencing applications partition their services over two separate types of servers: the application server and the media server (as shown in Figure 2.1).

The main (OpenVidu or BigBlueButton) application server handles the administrative side of the application, including authentication, creating sessions and managing the users. The media server handles the transcoding, recording and multiplexing of the media streams between the clients. The benefit of this separation is that videoconferencing applications need not *reinvent the wheel* and can use an existing media server such as the Kurento Media Server [24].

2.1.2 WebRTC

WebRTC is an open standard which supports video, voice, and data to be sent between peers; typically used for voice and video communication solutions in the browser. The main benefit of WebRTC is that it is supported by all modern browsers, with SDKs available for desktop and mobile applications. This presents WebRTC as the preferred option to a native videoconferencing, as developers are assured their software is supported by the vast majority of platforms.

2.1.3 Kurento media server

While users can communicate directly with each other over WebRTC, in practice, a media server is used for larger-scale deployments. Kurento Media Server (KMS) is a popular media server used by several open-source videoconferencing applications [24]. The main benefits of using a media server over a direct peer-to-peer connection are:

1. **Transcoding:** The media server can dynamically change the quality of a media stream of a peer, based on their internet connection, without affecting other peers' feeds.

2. **Group Communications:** The media server can mix multiple streams together into one composite stream to reduce bandwidth. This is especially important in homes, where the upload speed of a home broadband connection is, on average, an order of magnitude slower than download speed [25]. Therefore, sending multiple media streams can easily create a bottleneck.

3. **Recording:** Since all media streams are routed through a server, it can act as a central recording site.

2.1.4 Broadcast channel

The HTML living standard [26] defines a `BroadcastChannel` interface which supports communication between different browsing contexts of the same origin (same scheme, e.g. `https`, domain, e.g. `google.com` and port, e.g. `80`). Messages are broadcast to all objects listening on the channel. This supports client-side communication between tabs and windows without requiring a server relay. This is easily enabled by creating a `BroadcastChannel` object, provided by the Javascript browser API.

2.2 Turn-taking and feedback

In sociolinguistics, studies have explored a model of turn-taking in conversations and whether a relationship exists between gaze behaviour and turn-taking. This section summarises this work for further understanding of the theoretical background of conversations.

In group discussion and many forms of verbal interaction participants, take *turns* to talk [27], named "one party talks at a time" [1]. This widely accepted model (see Figure 2.2) developed by Schegloff, Sacks and Jefferson (SSJ), provides a simple mechanism by which participants engage in conversation [27]. The current speaker in conversation may select the next speaker, e.g. by addressing them by name. If this does not occur, a speaker can self select once the previous turn breaks down. If not, the current speaker can continue to speak.

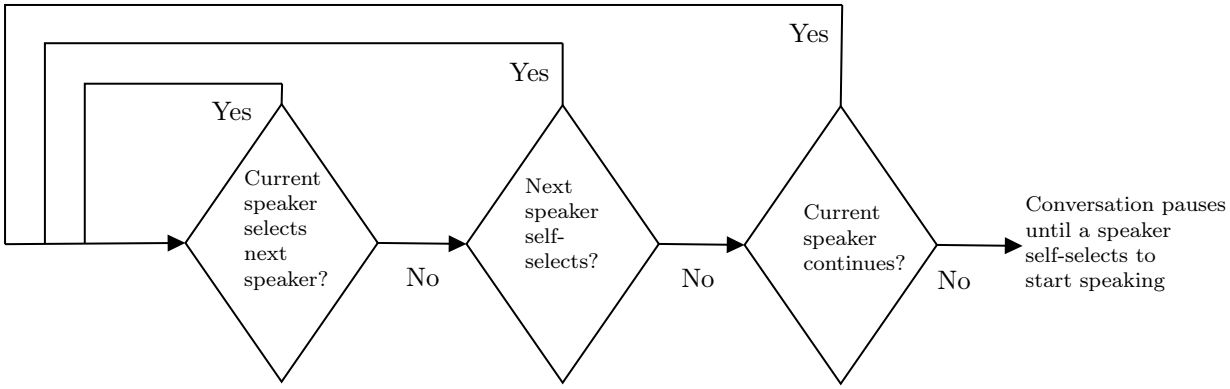

Figure 2.2: *SSJ model of turn-taking in conversation.*

Schegloff et al. [27] describes four ways of *selecting* someone to speak next:

1. *Directly addressing a specific party:* "John, is this right", "What do you think about this?"

2. *Addressed tag questions* attached to the end of an utterance "..., you know?", "..., aren't you?"

3. *Elliptics*, reduced questions that follow or interrupt a turn, interpreted by reference to that turns' talk, thereby automatically addressing its speaker. "How much did you say?", "today?"

4. *Social identities* can make someone immediately selectable without explicit addressing. Schegloff et al. give the example of two couples in conversation, so if someone says "You should go to the movies with us" there is no doubt who "you" and "us" refer, and consequently who is selected to speak.

2.2.1 The role of gaze

While SSJ presents a simple model for conversations, it omits several important aspects of face-to-face conversations. For example, it does not discuss the use of nonverbal signals such as gaze, which is used to regulate turn-taking [19]. Kendon identified it as an

[1] Critics have argued this is not universal or essential for communication, however concurrent communication's turns are typically required to be short and simple for comprehension. [28]

essential signal of both yielding and holding the turn. The paper noted that over 70% of the utterances terminating with the speaker gazing at another participant were followed immediately by talk from the gaze target, in contrast to only 29% ending without gaze. This effect is even more important in group interactions [29], when a speaker may gaze at the desired next speaker to win their attention and thus have an acknowledged speaking turn. Gaze in group interactions serves, not only as a cue for turn-taking, but also engaging someone's attention, which is more important in group conversations, *"a group speaker cannot assume an auditor but must engage one"* [29, p. 168].

Gaze has been shown to improve a participant's ability to predict the outcome of an upcoming transfer of the floor [30]. In a three-person (triad) party, the primary addressee looks at the speaker rather than the secondary addressee, contributing to the primary addressee being more likely to take the next turn. Previous research has shown in video-mediated communication, where gaze information is not readily available, participants use more *explicit handovers*, as it is the easiest method to hand over the floor to a specific individual [31].

2.2.2 The role of directionality

Another important factor lost in gallery-view video-mediated conversations is directionality. In face-to-face communication, where participants are in different physical locations, the speaker can use directionality to control the floor. For example, a speaker can acknowledge an interruption, with directional hand-gestures, indicating to the interrupter they intend to yield the floor soon. Spatial auditory signals typically play a role in identifying the source of the simultaneous starters, and gaze plays a role in the negotiation of who will gain control of the floor next.

Research has noted, in audio-conferencing, notably with no visual cues, turn-taking becomes a lot more explicit, and difficulties occur. In dyadic conversations, Beattie et al. [32] noted that verbal cues such as intonation and grammatical junctures would *take over* the function of turn-yielding.

2.3 Conversation Analysis

This section summarises of the conversation characteristics used in coding conversations based on approaches from previous work [31]. We use this specification in the evaluation section to test the different systems.

Auditory Backchannels: Short utterances given in the background, produced by listeners to indicate functions such as attention, support or acceptance. Importantly, backchannels do not compete with the speaker for the floor.

Interruptions: Instances of simultaneous speech when there is no indication by the first speaker that they are about to relinquish the floor. These are deliberate attempts to gain the conversational floor without the prior consent of the current speaker, and they always occur in mid-turn.

Overlaps: Instances of simultaneous speech which follow signals given by the speaker to indicate that they may relinquish the conversation floor. We measure three different types of overlap:

1. *Projection/completion:* This overlap occurs when the next speaker anticipates that the current speaker is about to finish or tries to help with the "forward movement" of an outgoing turn. The next speaker may overlap to complete the current speaker's turn simultaneously. This may occur when the next speaker perceives that the first is having some difficulty in completing their turn.

2. *Floorholding:* This overlap occurs when the next speaker tries to take the floor while the current speaker attempts to hold the floor while producing words that do not contain any information.

3. *Simultaneous starts:* This overlap occurs when two participants concurrently begin a new turn. These occur when a previous speaker has just finished, and two or more speakers (which may include the previous speaker) compete for the floor.

Handovers: When speakers signal that they intend to relinquish the floor using explicit verbal cues. These involve [27]:

1. *Direct addressing:* Using a name or personally identifiable description to indicate whom a question is aimed about.

2. *Addressed tag questions:* These include "..., you know?", "..., aren't you?"

3. *Elliptics*: Reduced questions that follow a turn, thereby automatically addressing its speaker. Note, if the elliptic interrupts rather than follows the speaker, it is marked as a projection/completion.

4. *Social Identities:* Making someone immediately selectable without explicit addressing.

2.4 Group decision-making

Group decision-making occurs under the assumption that no single participant can determine, with certainty, the correct or best choice pre-discussion; the optimal decision can only be found by pooling the unshared information. Such situations are called *hidden profiles* [33].

In workplaces, groups bring together individuals, typically from different departments and levels, with *"unique, relevant and often diverse information sets"* [34], and if pooled efficiently should be able to achieve superior outcomes. However, research has consistently shown that groups fail to pool information, relying on pre-group discussion alternatives, rather than discussing new information that could help [35]. This effect is reinforced over videoconferencing, with video-mediated groups producing worse quality decisions compared to face-to-face [36]. Our goal is to improve the quality of decisions made using videoconferencing.

2.5 Spatial awareness

This section will introduce vocabulary to ease discussion of both the capabilities of this project and prior work. Built on, from widely used gaze awareness definitions [18, 37]:

Mutual Spatial Faithfulness: every observer in a system is simultaneously aware whether they are the attention target.

Partial Spatial Faithfulness: a system in which the apparent direction (up, down, left, or right) of the attention target, as seen by the observer, is the actual direction of the attention target.

Full Spatial Faithfulness: a system in which there is a one-to-one mapping between the apparent attention target and the actual attention target.

From these definitions, most videoconferencing systems do not achieve any form of spatial faithfulness.

2.6 Previous work

Over the past three decades, researchers in the fields of Sociolinguistics and Computer Science have explored whether spatial faithfulness affects group discussions, both for decision-making and other objectives. This section will summarise several key publications and present their contributions and differences. This section will compare the performance measures they used to evaluate their systems, e.g. trust and conversation efficiency, and the tasks types, e.g. debates, cooperative tasks.

Figure 2.3: *User seated in front of Hydra units. [1]*

2.6.1 Conversation efficiency

One method of measuring the task efficiency is by performing conversation analysis to determine whether visual cues are leading to more productive conversations, e.g. fewer interruptions, and fewer simultaneous starts. Sellen et al. [21] investigated the relationship between conversation efficiency and directionality using the Hydra system. They simulated 4-way round-table meetings using phone-like Hydra devices. Each device contained a camera, a small LCD monitor and a speaker, replacing a participants' position in each meeting with a Hydra device (see Figure 2.3). While Hydra used a similar number of explicit handovers as compared to a standard gallery-view solution (see Figure 1.1), they noted fewer conversation breakdowns where participants did not know whom the speaker was addressing. Hydra was successful in facilitating selective listening and selective gaze, achieving full spatial-awareness, however the authors evaluated their system using informal debates. In competitive conversations, holding the floor is equated with winning the argument and thus, competition for the floor would be expected [38]. While in cooperative

conversations, as individuals share the same goal, participants are motivated to cooperate in floor sharing.

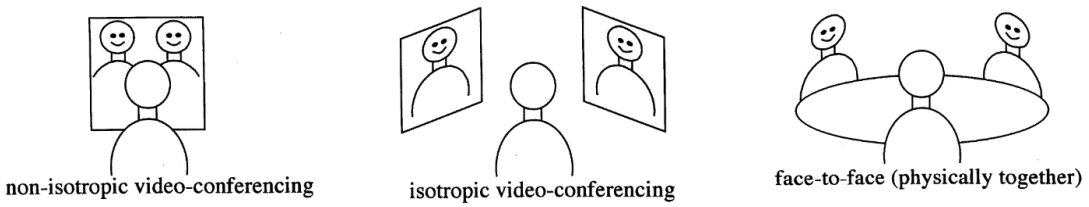

Figure 2.4: *A sketch of the three conversation conditions [2]*

Wekhoven et al. [2] also investigated the relationship of spatiality in video conferences and conversation efficiency. They constructed three identical workplaces which utilised an isotropic layout of recorders. In isotropic layouts each participant has a monitor and video camera pair for each of the remote participants, also achieving full spatial-awareness (see Figure 2.4). They evaluated their system using a problem-solving and decision-making exercise. They concluded both that persuasive force (the ability to change another person's opinion) is significantly stronger under isotropic conditions and participants communicate almost twice as much unshared information under mediated conditions than face-to-face conditions. The project was aimed at replicating a face-to-face situation, as closely as possible, therefore users sat far away (1.25m) from the screen and camera in identical workspaces. While the researchers performed conversation analysis, it only covered the wider categories of conversation artefacts (e.g. handovers, overlaps etc.) and did not present an in-depth discussion on their results. Furthermore, while enough to make out facial expressions, the researchers were limited to low image quality streams, by today's standards.

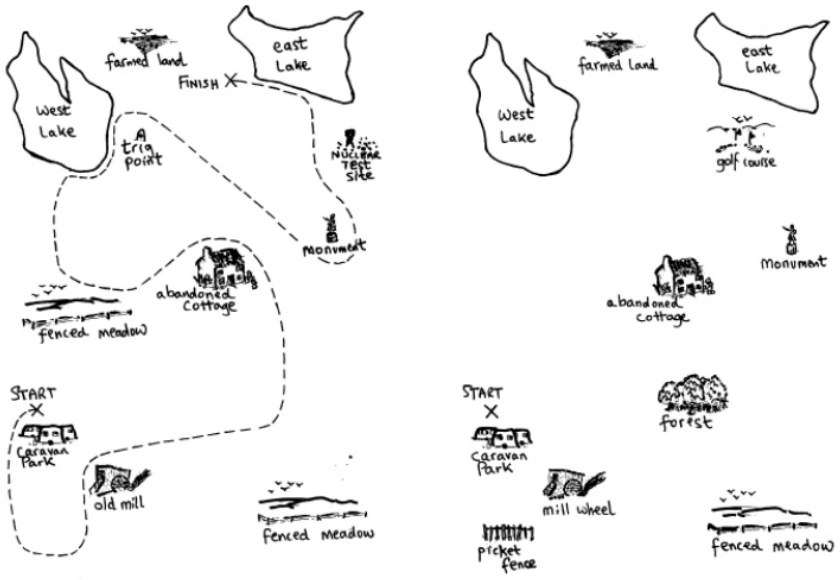

Figure 2.5: *Sample maps from the HCRC Map Task Corpus [3]*

Researchers [6, 39, 40] have used the *map task* as a standard test for videoconferencing analysis because it requires information transfer through spoken dialogue (see Figure 2.5). Introduced by Anderson et al. [3], a pair of participants are given a map scattered with landmarks, but only one person's map contains a route. The goal of the task is for the giver to verbally articulate the trail's path to the receiver, who must replicate the

trail on their own map. Boyle et al. [6] used this task to compare task outcome and conversation efficiency between users who could or could not see each other. Doherty-Sneddeon et al. [40] created the illusion of eye-contact using half-silvered mirrors in front of a videocamera and compared it to an audio-only solution.

They reported that groups using videoconferencing that enabled eye contact took longer to complete the task, than groups with no eye-contact or even no video. However, the map task typically creates structured conversation, as the dialogue is one-sided, from the participant with the filled map to the one with the empty map. This leads to conversations, where turn-taking only occurs for confirmation or clarification. This makes it a poor measure for group decision-making problems where conversation is typically a lot more investigative, and information isn't concentrated with one participant.

2.6.2 Trust

Trust is often cited as an essential prerequisite for effective virtual teams [41]. Research has shown that virtual teams with *high degrees of trust* are more focused on task output and provide more substantive, productive feedback [42]. Nguyen et al. [43] investigated the link between spatial faithfulness and trust, showing that systems that introduce these spatial distortions negatively affect trust formation patterns, while spatial faithfulness eliminates these effects. Full spatial faithfulness for a group was achieved using an array of projectors and cameras. Pan et al. [20] also presented similar results for videoconferencing in which participants sought advice from supposed experts. Comparing traditional flat and spherical display, they demonstrated that spatial distortions negatively affect trust formation patterns. However, other research has shown that the formation of trust is unnecessary for a virtual team to deliver a quality result, and meta-analysis has suggested the existence of common method biases and potential overestimation of correlations [44]. Therefore, this project focuses on objective performance indicators, rather than trust, to combat these biases.

2.6.3 Eye accuracy

(a) *The user is looking into the middle of the display.*

(b) *The user is looking straight into the video camera.*

Figure 2.6: *Images of a user captured by the camera mounted above a 27inch computer display at a normal working distance [4]*

Gaze awareness, seen as a subset of spatial awareness [18], has also been explored in literature. Video conferencing applications rarely allow mutual gaze due to the discrepancy

between the camera's position and the image of the other person's eyes. Figure 2.6 illustrates when the user is looking at their partner's eyes, in the middle of the screen, it appears as if they are not looking at their partner.

Chen and Milton [45] investigated this effect in videoconferencing and formulated requirements for avoiding the eye gaze skew. They stated that the camera needs to be located within 1 degree horizontally and 5 degrees vertically from the on-screen representation of the remote participant to ensure *eye contact*. A popular solution to this problem is the use of videotunnels [37, 40] These use a combination of half-silvered mirrors and translucent displays to remove the camera offset while not obscuring the monitor. Vertegaal et al. [46] achieved gaze awareness using eye trackers and an array of cameras around each subject. These approaches required specialised setups and therefore, unlikely to be used in a work-from-home environment. Recent work has explored the use of neural networks [47] to correct any unwanted eye offset. While these are purely software solutions, there is a lack of user studies or consideration of real-world issues such as when users are intentionally looking away, e.g. to signal thought.

Most studies of eye contact and spatial awareness focus on two-person (dyadic) communication [6, 38, 40]. However, as research has shown [48], the benefits of video-mediated communication are more substantial in larger groups, e.g. visual identification of the momentary speaker. Turn-taking and feedback, become much more important in larger groups since the attention target is not immediately apparent.

2.6.4 Summary

While several parts of this project have been already explored in papers dated 1996 - 2003, many of them were forced to compromise due to the limitations in hardware available. For example, Anderson et al. [49] notes a limitation in only been able to encode and decode 240p video streams. Since the publication of this work in 2000, there has been a substantial increase in the quality of videoconferencing systems. While the ability to transmit additional information, such as visual cues or gestures, does not directly impact the ability to absorb or communicate information [50]. Further research has shown that perceived video quality impacts attitudes such as user satisfaction and acceptance which indirectly impact communication ability [51]. They all describe setups which either used expensive hardware, expert calibration or idealistic environments, such as large working areas. One of the main objectives of this project is to investigate a system that can run on commodity hardware in typical home offices. Another critical change is advancement in network and hardware availability. Isotropic layouts incur a higher network demand because of the increase in media streams sent. However, broadband speed has increased massively, with the average UK residential household internet speed doubling almost four times in a decade [25].

Chapter 3

Model

This chapter provides the reader with a description of *Spatial*, a spatially faithful video-conferencing application. First, the model of the environment is presented, and the desired physical setup of the environment. Secondly, a calibration process is described, showing how we can use classical camera calibration techniques to calibrate our environments.

3.1 Abstract model

This section describes the abstract environment of the spatially faithful system. Our environment is based on an example physical setting, which is then mapped to a virtual environment. Next, constraints are introduced to ensure the environment is spatially faithful.

The model introduced in this section can extend to any number of participants.

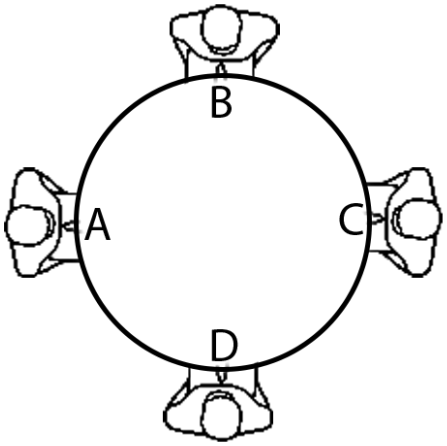

Figure 3.1: *Diagram of four individuals around a round-top table*

The virtual environment we plan to emulate by this system is a circular round-top table, each equidistant away from each other and in view of the rest of the participants (see Figure 3.1). [1]

[1]Seating positions in group contexts can have a significant impact on the group dynamic, interpersonal communication and friendship formations [52], however this is not covered in this work.

We define the math operator $<$ as *to the positional left of*. For example, from observers A and D, $B < C$.

Next, we define a *local spatial order*, i.e. from an observers point-of-view, the order in which the other participants are apparent. For example, from observer A, the spatial order is $B < C < D$

Finally, we can define a *global spatial order*, taken in a clockwise direction, the order in which the other participants are sat. This spatial view importantly is circular. For example, the global spatial order is $A < B < C < D$. Each *local spatial order* for every participant is a subset of the *global spatial order*.

We define a **spatial connection** as a virtual connection between two participants. Each spatial connection can be thought of as a *virtual window*, in which directional sound/video from the one individual passes to the other. If a participant looks at one of these windows, the other participant is aware that they are the attention target. These *virtual windows* satisfy the requirement for mutual spatial faithfulness, since for a given attention source, each observer in the system, is aware whether they are the attention target.

As each *local spatial order* is a subset of the *global spatial order*, the order of participants is shared between observers. Therefore, they are aware of the direction of a potential attention source since the order is shared between all observers. This satisfies the requirement for partial spatial faithfulness. As our environment has mutual and partial spatial faithfulness, our system is assured to be fully spatially faithful.

3.2 Physical design

This section describes the physical design of the spatially faithful videoconferencing system, which obeys the constraints of the model introduced in the previous section. This physical design is then used to influence our design and implementation choices.

Literature reviews have highlighted inconsistencies between the positive and negative relations between the "virtualness" (teams that are more or less virtual) and team performance [53]. However, research on memory has found that individuals are better able to remember items when items are *naturally presented* [54]. Therefore mapping our virtual environment into a familiar physical space, should result in higher levels of participant satisfaction.

3.2.1 Individual site design

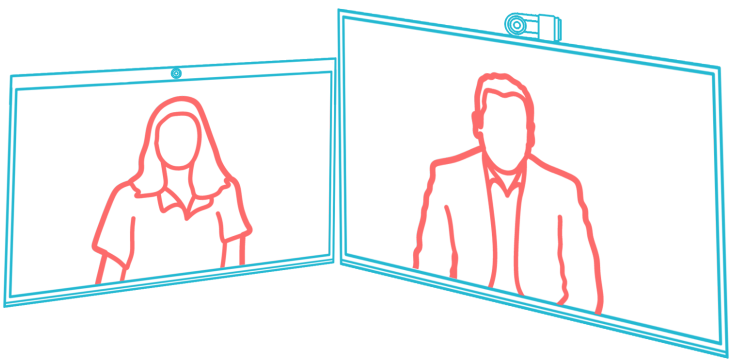

Figure 3.2: *Example of a dual flat monitor setup.*

Our defined work-from-home environment is a participant in their own site using a laptop and an external monitor. Each participant will sit in the centre of the two monitors, with a full view of both screens (see Figure 3.3). This two-screen setup will enable a three-person conference. The aim is to use these monitors to act as a *window* to a remote participant. Each screen will display a different participant (see Figure 3.2). Unlike previous studies, we use webcams instead of video cameras [2,18]. The main benefit is that webcams typically use wide-angle lenses, as they are designed for close usage. Therefore participants can sit closer to their screens, rather than larger distances that had to be employed by previous work. For the laptop screen, the built-in webcam located above the screen is used, and for the external monitor, an external webcam is used. Additionally, each participant will wear headphones or use stereo speakers to take advantage of the positional sound.

3.2.2 Full conference design

Figure 3.3 shows an example three participant conference, each remotely located, each with two monitors and two cameras. Camera A1 is displayed on screen B2, and Camera B2 is displayed on screen A1. Camera B1 is displayed on screen C2, and Camera C2 is displayed on screen B1. Camera C1 is displayed on screen A2, and Camera A2 is displayed on screen C1. This preserves mutual, partial and full spatial awareness across all locations - person A would be able to determine the attention target of person B, even if the target is person C.

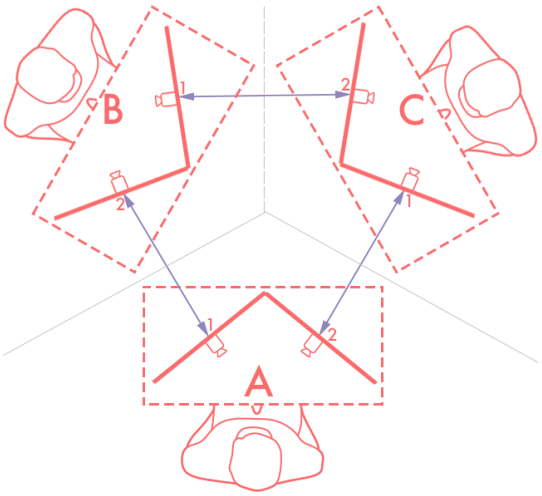

Figure 3.3: *Physical design of a three-party setup*

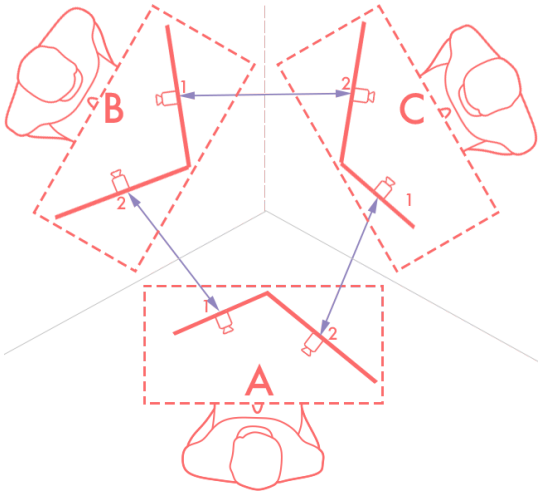

Figure 3.4: *Physical design of a three-party setup, with participants A and C has different sized screens*

As we are targeting a work-from-home environment, each participant is likely to have two screens of different sizes; therefore, the angle between the screens and the desk is likely to be incongruous. The angles will be computationally determined by the relative sizes of the screens and the location of the participant. Figure 3.4 provides an alternative design where participants have screens of different dimensions.

Audio design

We also want to ensure that our environment is audibly spatially faithful, as these play an essential role in identifying a new speaker and turn-taking negotiation. We can achieve this by applying 3D audio effects to our remote participant audio streams. 3D audio

allows you to specify the location of an audio source in virtual space. In Figure 3.3, participant A will hear B's voice from their left side, and C's voice from their right side.

3.2.3 Gaze awareness

A parallax effect occurs when the local participant does not perceive eye contact with the remote participant due to the physical distance between the camera and the image (see Figure 2.6). While not explicitly covered by this work, considerations were made in the physical design of this system to minimise this effect. Chen [45] presented a requirement for eye contact: the video camera needs to be located within 1 degree horizontally and 5 degrees vertically from the on-screen representation of the on-screen representation. Assuming the remote participant's eyes are located 20% of the screen's height, from the top edge and a distance of 1cm between the screen edge and webcam lens. For a 13-inch (33.0cm) laptop screen, this translates into a distance of 87cm from screen edge to eyes, and for a 23-inch (58.4cm) monitor, a sitting distance of 142cm. This is unfeasible for a standard working environment as the majority of people will be sitting at a desk, a lot closer than 142cm.

Several researchers have explored solutions to this parallax problem. Giger et al. [55] generated a 3D face model that matched the user's face and inserted it into the original's image. He et al. [56] trained a generative adversarial neural network to generate different gaze directions, which was also inserted into the original image. However, the lack of literature on user studies or implementations into existing solutions is concerning, furthermore, neither provided solutions for determining whether the user is actually looking at the screen or not. The GAZE-2 [46] study noted that large angular eye shifts were actually considered distracting by the participants, which could prove detrimental to performance and user satisfaction.

Therefore, we decided against including gaze correction techniques, as it would require a large amount of work to address problems that current literature has not solved. A benefit, however, of the physical design, is participants will be centre aligned in each screen, so while eye contact may not strictly be achieved, participants should be able to determine whether they are the attention target easily.

3.3 Calibration model

To ensure that each participant's individual site, correctly follows the physical design, we must design a mechanism for individual setups to be calibrated. It is worth noting that people do not sit still during video conferences so any precise calibration effort would be futile as they would require to be continually updated. Instead, we aim for a best-effort approach.

Humans have been shown to be less sensitive to small spatial distortions in natural scences [57], and observers' expectations of familiar shapes, such as faces, allow them to tolerate noticeable distortions [58]. This allows us to consider less accurate but more user-friendly solutions.

3.3.1 Calibration measurements

We need to ensure that each participant's visual and audio setup does not lead to distortions or misaligned perspectives that look incorrect to other participants or worse confuse

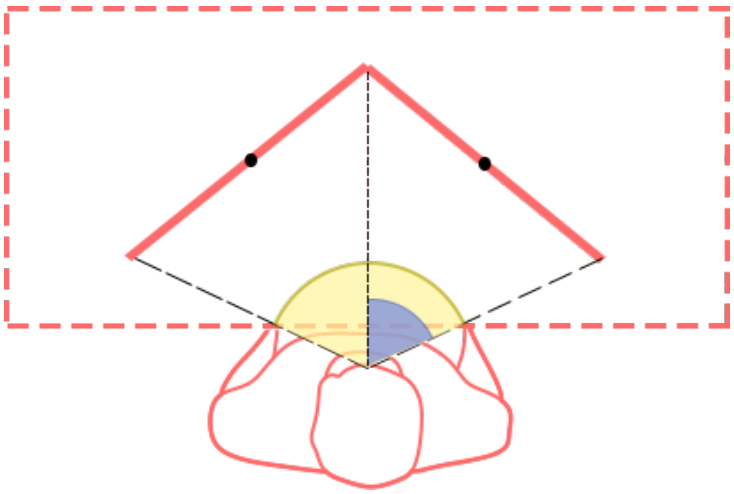

Figure 3.5: *Diagram showing two calibration measurements. Yellow: Total field-of-view angle, Blue: Screen field-of-view*

them. Previous work avoided this problem by providing specialised laboratory workspaces and hardware [18,21], which required experts to set up and specialised equipment installed; however, we are targeting commodity setups. We are concerned about changing the angle of our monitors (and thereby webcams) to ensure that the field-of-view of each monitor is correct.

1. **Total Field-of-View**: The angle of a participant's field of view that is obscured by the monitors. This is the amount of physical space dedicated to the virtual environment.

2. **Screen Field-of-View:** The angle of a participant's field of view that is obscured by a particular monitor. This is the amount of physical space dedicated to a remote participant.

From Figure 3.5 it is clear to see that the total field-of-view is simply the sum of each of the screen's field-of-views. Therefore, we can work on calibrating each screen separately, ensuring each remote participant is given the correct proportion of the participant's physical space.

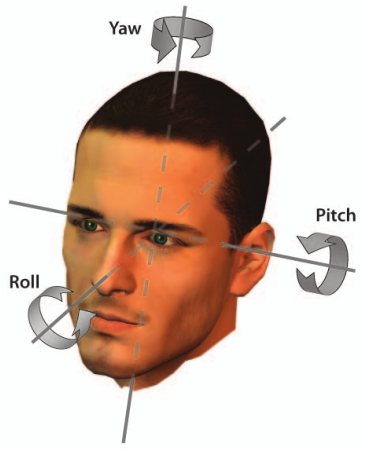

Figure 3.6: *The three degrees of freedom of a human head can be described by the egocentric rotation angles pitch, roll, and yaw. [5]*

3.3.2 Single screen calibration

Typical camera calibration algorithms provide rotational information such as head pose (see Figure 3.6); this is typically denoted using Tait-Bryan angles, which split into yaw, pitch and roll. We only care about the **yaw**, shown as the (green angle on Figure 3.7).

However, we want to calculate the angle between the participant's neutral position (looking in the centre of their setup) and the far edge of the monitor. That is, we want to calculate the amount of space that is obscured in the person's view (blue angle on Figure 3.7).

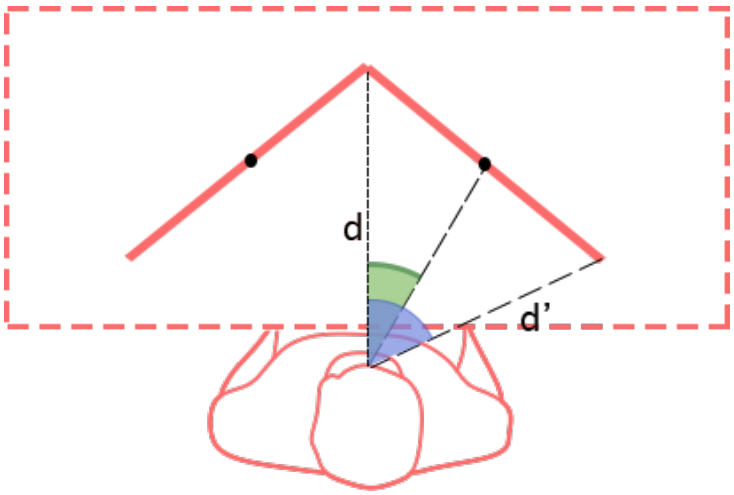

Figure 3.7: *Diagram comparing the angle of webcams with true field-of-view.*

Therefore we must devise a mathematical relationship between these two angles.

3.3.3 Calculating field-of-view

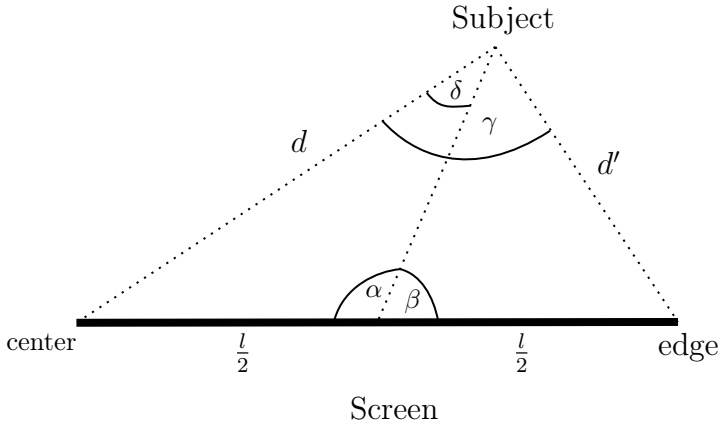

Figure 3.8: *Diagram of the lengths and angles created between the subject's head, their screen and webcam. The bottom plane represents the screen, and the top point represents the subject's eyes.*

Measured values (inputs)

Screen width l: The width length of the selected screen in cm.

Head pose γ: The angle between the centre of the screen and the subject can be measured using camera calibration algorithms.

Head to screen length d: The distance between the user's head and the centre of the screen.

Now we can apply multiple iterations of the sine and cosine rule to calculate γ.

To calculate i:

$$(\frac{l}{2})^2 = i^2 + d^2 - 2id\cos(\delta) \tag{3.1}$$

Next we can calculate α and β:

$$\frac{\sin\alpha}{d} = \frac{\sin\delta}{\frac{l}{2}} \tag{3.2}$$

$$\beta = \pi - \alpha \tag{3.3}$$

We can use this information to calculate the inside distance d':

$$d'^2 = i^2 + (\frac{l}{2})^2 - 2i\frac{l}{2}\cos(\beta) \tag{3.4}$$

Now finally, we can calculate γ:

$$\frac{\sin\beta}{d'} = \frac{\sin(\gamma - \delta)}{\frac{l}{2}} \tag{3.5}$$

Once we have a value for our screen field-of-view, we can instruct users to adjust their screen and webcam angle as necessary. If the centre-to-subject distance is kept constant, users can simply adjust the angle of each monitor without recalculating the distance between their head and the inside edge.

For flat monitors, users are told to adjust the angle of their monitors, and thereby their webcam, pivoting around the centre point.

Chapter 4

Preparation and Design

This Chapter presents the design of the implementation of *Spatial*. First, we discuss the implementation design of the videoconferencing system, explaining any design choices that have been made. Secondly, we provide a design for the evaluation, identifying the limitations with previous solutions and introducing an alternative evaluation task.

4.1 Implementation design

This section presents the architecture design of the implementation. We conducted a small survey of open-source videoconferencing applications that we could begin to build on for this project.

4.1.1 Requirements

Open-Source: An open-source system would allow us to make any required changes to the underlying system.

Well documented: We wished to use a system that provided clear documentation for use and extension of their system.

Language familiarity: We wished to use a language we were familiar in as it would speed up development time.

Ability to record sessions: This functionality will allow us to record evaluation sessions for further analysis.

Web application: A web application means the final system is platform-independent, which does not limit our choice of evaluation hardware.

Requirements	Apache Open Meetings [59]	Open-Vidu [22]	Jitsi Meet [60]	BigBlue-Button [23]
Open source:	✓	✓	✓	✓
Well documented:	✓	✓		
Language familiarity:	✓	✓		
Ability to record sessions:	✓	✓	✓	✓
Web application:		✓	✓	✓
Ability to easily manipulate video:	✓	✓		✓

Table 4.1: *Comparison of a sample of videoconferencing tools*

BigBlueButton provided excellent documentation and a well-built system; however its architecture of more than six individual modules compared to OpenVidu's four meant we opted to use the later due to more simplicity.

4.1.2 OpenVidu

This section provides a brief overview of the OpenVidu framework.

OpenVidu architecture

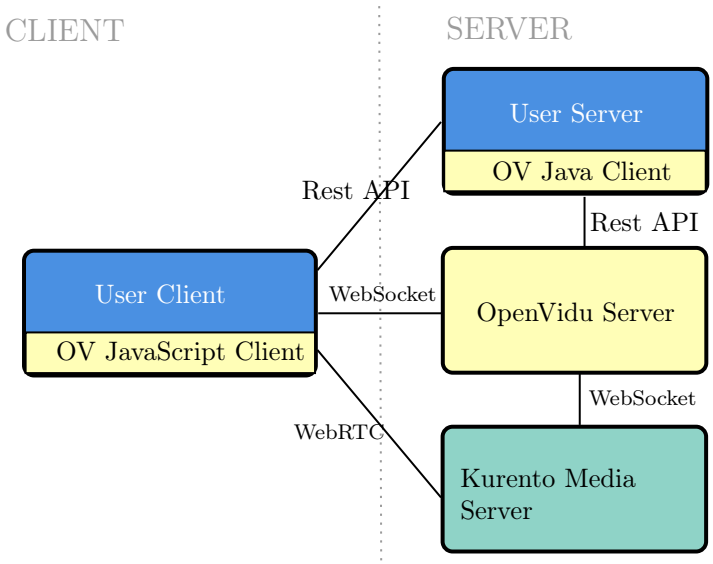

Figure 4.2: *Typical architecture of an OpenVidu-based application.*

OpenVidu [22] is an open-source, videoconferencing framework built in Java. It provides a simple API, enabling developers to use high-level abstractions, such as sessions and publishers, without handling the media streams directly. OpenVidu applications typically only require four components for usable applications, as presented in Figure 4.2. Developers can then use the OpenVidu server and client libraries to handle the video streams, while focusing on the other functionality of their applications. The user server has full

access to the OpenVidu API, while the user clients only have access to a particular video conference session.

Below, I summarise the role of each component in the architecture:

1. **OpenVidu Client and Server:** These provide an easy-to-use abstraction over the media streams and provide the ability to create rooms for video conferences.

2. **Kurento Media Server:** This is a popular, open-source, WebRTC media server which handles the transcoding, recording, mixing, routing and broadcasting of audiovisual flows [24].

3. **User Server:** The user server can control the creation of video conferences using the OpenVidu Java client.

4. **User Client:** The user client receives information from the user server about the video conference and interacts with the media streams using the OpenVidu Javascript clients.

OpenVidu model

OpenVidu provides three primary abstractions, which developers interact with to create their videoconferencing sessions.

1. **Session:** These represent video conference calls, all publishers and subscribers in a single session have access to each other's streams. Only the user servers can create sessions servers. Once a session is created, the user server typically requests for each of its publishers and subscribers and issues them to the user clients.

2. **Publisher:** Publisher tokens give user clients the ability to join a session. Once joined, they can publish a single stream comprising of video, audio or both to the session. They will also be able to subscribe to other publisheBelow, I summarize the role of each component in the architecture:rs in the same session.

3. **Subscriber:** Subscribers are publishers except they do not have the permission to publish their own media stream.

A typical session involves the user server setting up the video sessions, where options like recording and image filters can be enabled. Next, the user server creates tokens for all of the user clients: a typical videoconferencing application would create one token per user. Finally, the user clients connect to each session and subscribe to the incoming streams from the OpenVidu server.

OpenVidu includes two useful features that proved helpful for this project. Firstly, custom metadata can be sent with streams; this allows *Spatial* to send spatial information alongside the stream, rather than maintaining an additional data source to match both up. Secondly, users can selectively subscribe to different streams based on the metadata sent with each stream. Both these features meant that clients could read the metadata included with each stream and to decide whether if they require the incoming stream. This reduces the complexity of the implementation.

4.2 Evaluation design

Previous videoconferencing analyses have relied on either tasks such as the map task [39], focused on task efficiency or on a trust-based task such as the cooperative investment task [43].

As introduced in Chapter 2, the map task, typically performed in pairs, involves two similar maps distributed to each participant. However, only one map has a route marked on it, and also, landmarks found on both maps did not necessarily match (see Figure 2.5). The collective goal is to draw the path using the mode of communication.

This task is a poor choice for measuring group decision-making for several reasons:
1. Focus on paired conversation: The task was originally designed for two participants. A three-participant model can be achieved either by using two experts, each with a partial view of the map or three participants, each with a different section of the map. However, this reduces the amount of cross-study analysis that can be performed.
2. Structured conversation: The task, as shown in the evaluation, exhibits reduced turn-taking and fewer interruptions. Participants typically alternate based on which participant has the relevant part of the route.
Interruptions only take place for clarification and confirmation; they are dependent mostly on the quality of the communication, rather the quality of the medium.
3. Focus on paper: Since the task is to complete the route, most of each subject's focus is on the sheet, rather than the other participant. Therefore, any improvements to the video communication would only be used sparsely and therefore, its effects muted. Subjects would typically only look up at each other when they clarified a particular instruction.

These reaons present explanations why the aforementioned studies did not find a statistically significant difference in task performance and efficiency between the individual modes of communication.

Alternatives to the map task rely on task-specific knowledge [20]; or the subjects' self-reporting on subjective measures [43], which cannot be standardised; or the group self-reporting on measures such as *"confidence in the decision reached by the group"* [53]. Therefore, it is important to design a task that can objectively measure team performance that does not require task-specific knowledge. Instead, it should comprise of reading and communicating information to other team members, skills that underlay multiple tasks.

4.2.1 Evaluation task

As a result, I have designed a new evaluation task, focused on exploring the dynamics of the hidden profile problem in a videoconferencing chat. Each participant is given their own character profile, our *unshared information*, with a brief description of who they are, and their own personal objectives from this task. Character profiles were kept short, to minimise the tim that participants were taken away from the screen. In addition, they are each given another document, our *shared information*, which outlines the options for the group's objectives. Finally, they are given the group's objective.

Unlike other hidden profile problems, the search space of the group objective was large, reducing the chance of participants relying on pre-discussion biases [35]. Interdependent connections were included between different pieces of unshared information, this required participants to interact with each other to gain a broader understanding of both the

group's objective and their own personal objectives. Connections among pieces of information are seen as critical in problem-solving research [61] and increasing the cognitive load of the group [62].

Two similar problems were designed to compare *Spatial* to alternative methods.

Evaluation task 1

Task 1 required the three participants, each with their own character profiles, to organise a dinner and drinks together the following week. They were each given a food and drink menu of the desired location as the shared information. The food and drink choices changed each day. The information for this task can be found in Appendix A.1

Evaluation task 2

Task 2 required the three participants, each with their own character profiles, to organise a board games night and drinks together the following week. They were each given a board games and drinks menu of the desired location as the shared information. The board games and drink choices changed each day. The information for this task can be found in Appendix A.2

Chapter 5

Implementation

This chapter describes the implementation of *Spatial*, a spatially faithful videoconferencing platform. The platform is split into two components: `Spatial-Server`, the server application, which creates and controls the environment and the users and `Spatial-Client`, the client-side website, used by the participants. Due to the close-coupling of their roles, the chapter does not emphasise functionality implemented in each component, unless stated otherwise.

`Spatial` has three primary functionalities and responsibilities:

1. Users can calibrate their individual sites to ensure there aren't any spatial distortions
2. Users can setup spatially-faithful video conferences
3. It provides a management layer over OpenVidu to manage spatially faithful videoconferences.

This chapter presents the architecture of the system, and then describes of how these three functionalities are implemented.

COVID Notice: *Unfortunately, because of restrictions imposed by the COVID-19 pandemic, I was forced to complete my evaluation early due to the early closure of the University. Therefore, I was unable to implement all of the planned functionality and have deferred this to future work.*

5.1 Application Overview

The architecture of `Spatial`, shown in Figure 5.1, is the same organisation as the basic OpenVidu architecture introduced in Section 4.1.2.

`Spatial` is a Java Spring application. Java Spring is a modular framework focused on creating *enterprise-ready* web applications [63]. Its large support base and my familiarity made it an obvious choice for this project. The implementation was built using Spring MVC, which follows the popular Model-View-Controller design [64].

The benefit of using the MVC design is that we can define a model of our spatial video conference, named *SpatialSession*, which is shared by `Spatial-Server` and `Spatial-Client`. This *SpatialSession* directly manages the data about the users in this call, the spatial re-

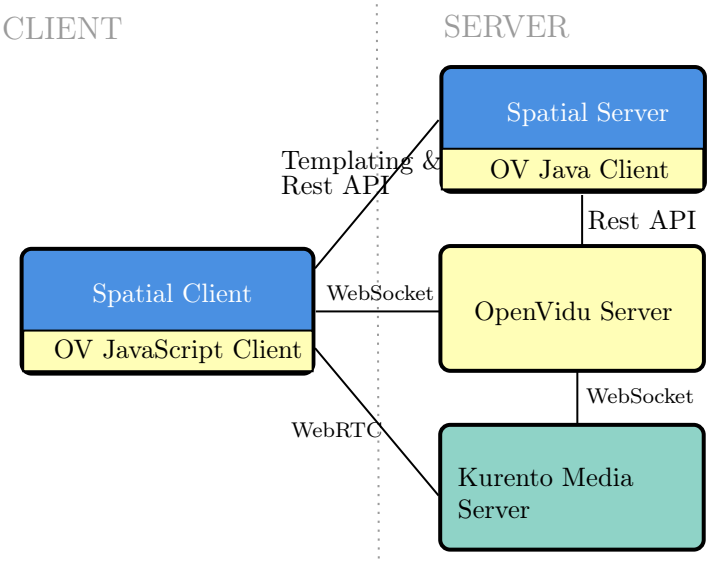

Figure 5.1: *Architecture of Spatial with communication links with OpenVidu and Kurento Media Server*

lationships between the users and the information about their media streams. We give a UML diagram of this model in Appendix B.

5.1.1 Communication between server and client applications

`Spatial` uses two methods to communicate between the server and client application.

1. **Templating:** `Spatial` uses the Thymeleaf [65] template engine to display the client-side webpages. `Spatial-Server` injects the model directly into the source code for `Spatial-Client`. This increases the prototyping speed as we do not need to design API endpoints for each part of the model.

2. **API:** `Spatial-Server` also exposes REST endpoints for dynamic loading and changes of content. The API is used by the client to submit model changes back to the server for it, as well as receive live updates without a page reload.

⁂

The following three sections discuss the main responsibilities of `Spatial`.

5.2 Calibration

To ensure each participant's individual workspace is correctly set up, `Spatial` provides a process for users to calibrate their setups. This section describes how the calibration model (introduced in Section 3.3) is implemented in `Spatial`. Figure 5.3 shows the calibration process. For head pose estimation, we use FSA-NET, a neural network-based approach.

We implemented the face detection, and head pose measurement functionality in C++, as the maturity and stability of the desired libraries (OpenCV and Tensorflow) were much

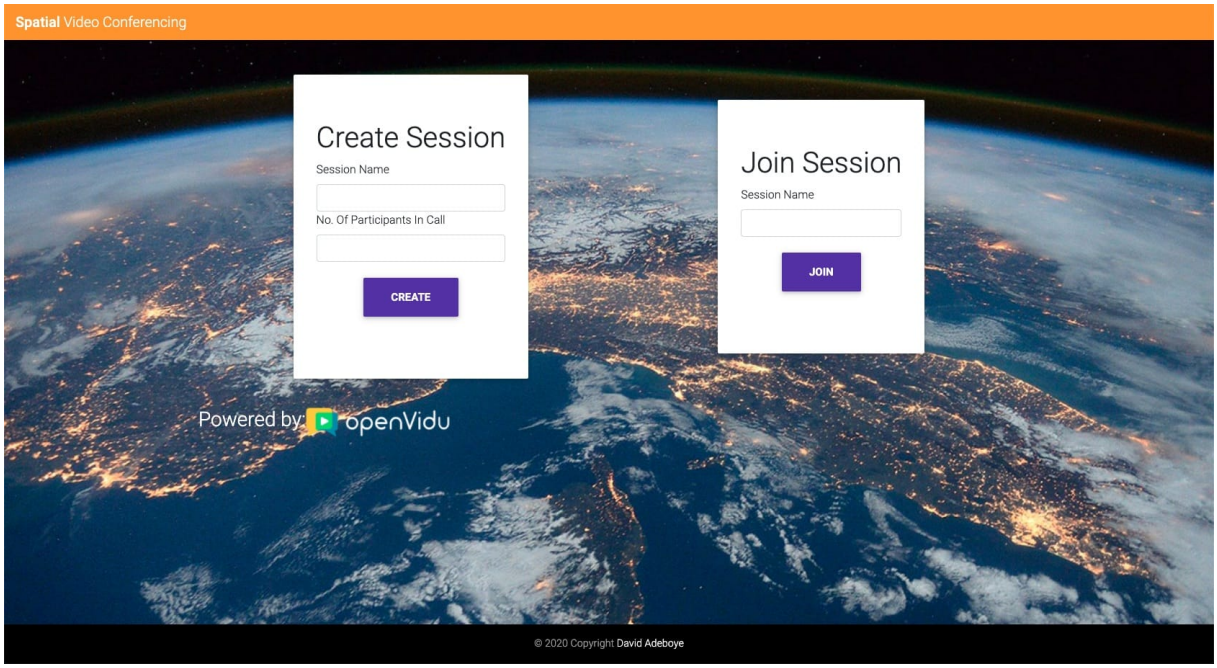

Figure 5.2: *Screenshot of Spatial's welcome page*

higher, compared to Java. Therefore, once the `Spatial-Server` has received the data, we use the Java Native interface to communicate with the face detection and head pose measurement library. The Java Native interface [66] allows Java code to call and be called by native libraries and applications written in other languages.

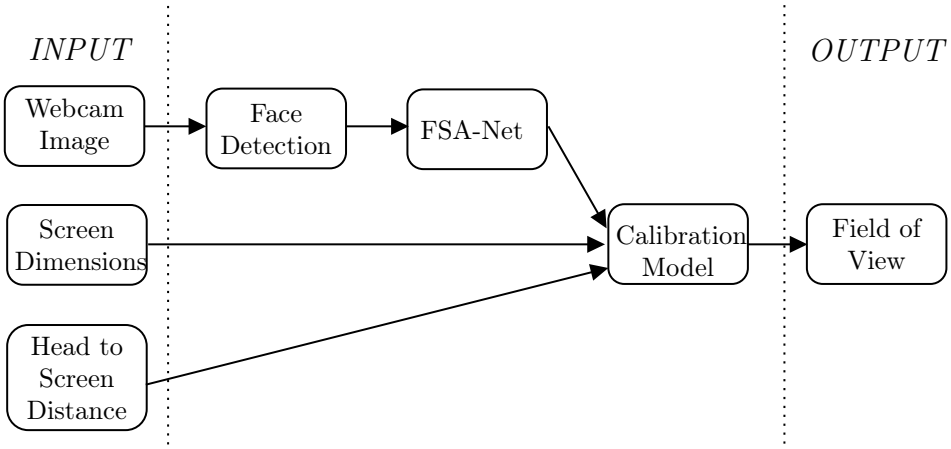

Figure 5.3: *A dataflow diagram showing the functions required to calculate the field-of-view for each screen.*

5.2.1 Obtaining input data

We obtain the input data using a web form in *Spatial-Client*, which sends a Web `POST` request to an API endpoint in `Spatial-Server`.

Spatial-Client captures a webcam image, and draws it onto a Javascript `canvas`. The canvas is converted to an image and `base64` encoded.

The monitor's physical horizontal dimension is calculated by drawing an HTML element on the screen of a set physical size (e.g. 2cm). Then, we use its pixel width to calculate

the pixels per inch ratio; we can then use the browser width to calculate the actual screen dimensions.

Finally, we ask the user to estimate and enter the distance between their head and the centre of the screen. We give advice on the size of common household objects, e.g. a sheet of paper. In a three-person setup, each participant uses two monitors, whose inner edge is adjacent to each other and therefore this distance (shown on Figure 3.7 and 3.8) should only need to be measured once.

5.2.2 Head detection

Our head pose measurement algorithm requires a cropped image of the individual's head. Therefore, we use standard face detection algorithms to identify the individual's head and crop it. OpenCV provides in-built classifiers for object detection. `Spatial` uses the local binary patterns histogram (LBPH) method, implemented in OpenCV.

If we cannot detect a face or detect more than one face, the calibration process is halted. To ensure that the entire head was captured, the cropping was performed on a box 10% larger than what was returned by the algorithm.

5.2.3 Head pose measurement

Next, we must determine the head pose angles before we can provide them to our calibration model. We opted to use a machine learning solution as it reports accurate results, without the use of specialised hardware or additional materials. Yang et al. [67] presented a method for head pose estimation through the use of their neural network: FSA-Net. Results published in the paper show that FSA-Net achieves the lowest average error for yaw and mean average than any of the other state-of-the-art methods. While previous methods typically ignore the spatial relationships between the features of a subject's face, FSA-Net instead groups pixel-level features together to form more powerful region-level features.

While there are other publications with similar results, FSA-Net was chosen for the following reasons. First, the authors provided the source code of the work, making their results easily-reproducible and I didn't have to waste time re-creating and re-training a similar model. Second, the system is implemented in Tensorflow, a framework, I am familiar with, reducing the amount of learning I required to adapt the model for this project. Third, they boast a small model size, *around 100x smaller than those of previous methods*, which is especially helpful, as memory on video-processing servers is typically scarce and therefore appreciated.

The FSA-Net Tensorflow model was exported as a Tensorflow Lite model for use in `Spatial`. Tensorflow Lite provides tools to export a model's entire graph easily, leaving developers to only focus on the handling of the model's inputs and outputs [68].

The first step was resizing the OpenCV image output from the face detection to 64x64. Secondly, the image channels of the image output were cycled since OpenCV stores images in BGR (Blue, Green, Red), instead of RGB (Red, Blue, Green), used by FSA-Net. Finally, we could convert the underlying image matrix into a Tensor, ready for FSA-Net.

Once the model has finished executing, it outputs an array with the three pose angles.

5.2.4 Calibration measurement

Finally, with head pose information, screen dimensions and the distance between the user and the edge of the screen, we can calculate the field-of-view. The calibration model presented in Section 3.3, is used to calculate the field-of-view and return it to the user. The user can then make adjustments of the angle of their monitor until the current field-of-view matches the desired one Users are told to adjust the angle of their monitors, and thereby their webcam, pivoting around the centre point of their setup.

5.3 Creating spatially faithful environments

Unlike standard videoconferencing platforms which allow users to join, as they please, `Spatial` requires the video conference to remain in a spatially faithful state. Therefore, all participants information must join before the video conference can begin for any user. This section describes of the steps involved in creating the spatial environment.

Step 1: Waiting room

First, `Spatial` needs to be told the number of participants to create a new spatial video conference. The current version requires all participants to be active on the website before the rest of the setup can continue. Currently, co-located participants are required to use the same client to connect to the video conference.

Step 2: Generate environment plan

To create a spatial plan, we assign a random global order to all of the participants and then create a spatial connection between all non-colocated users.

After creating this spatially faithful environment, the server also determines the angles between each pair of participants. This information is used by the spatial audio processing.

Step 3: Setting up input and output devices

The generated environment plan is shown as a simple diagram to participants, so they have a visual idea of the environment they are virtually stepping into (see Figure 5.4).

Next, participants are required to set up their individual environments. Each participant was asked to select their desired audio input and output device. Finally, for each spatial connection, they selected the order of their webcams, from leftmost to right, to ensure that remote participants was recieving the correct feed. A screenshot of this panel is shown in Figure 5.5.

We realise this approach is very rigid. Users must all synchronously join the system, so that the system can create spatial connections. A better approach would be to adopt a dynamic setup, in which users can come and go as they please, and the system regenerates the spatial connection on change. We defer this to future work.

5.4 Managing spatially faithful videoconferencing

This section describes the implementation of the spatial video conference.

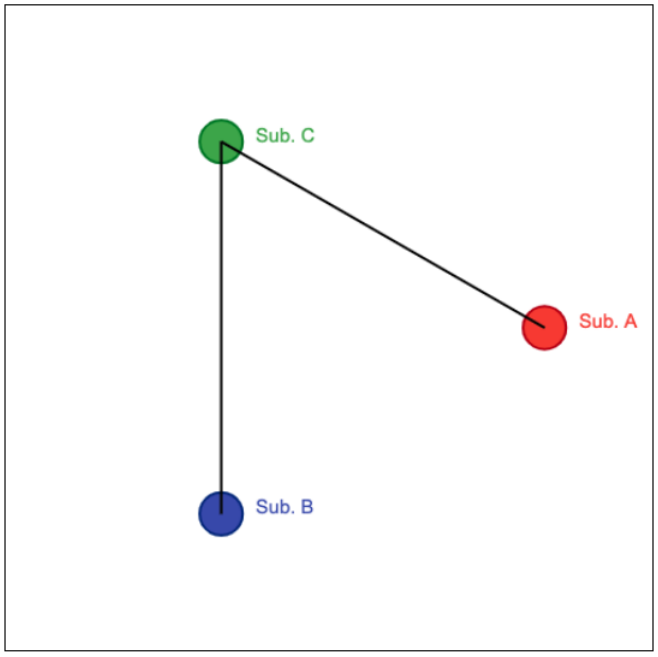

Figure 5.4: *Screenshot of environment diagram shown by Spatial. Diagram shows the connections from* ***Sub. C*** *view, therefore the connection between* ***Sub. A*** *and* ***Sub. B*** *is omitted.*

In *Spatial* video and audio inputs are treated separately, and each are assigned a *Publisher* entity. `Spatial-Server` sets up the OpenVidu videoconferencing sessions and creates publisher tokens for each of the participants using the OpenVidu Rest API.

Each browser window is responsible for a *SpatialConnection* with a remote participant, therefore it receives the video and audio streams from a remote participant (see Figure 5.6). Additionally, one of the browser windows sends the audio stream of the local participant.

5.4.1 Spatial metadata

OpenVidu allows `JSON` metadata alongside the stream to their clients. `Spatial-Server` embeds a list of the media stream's recipients in this metadata. If the browsers' ID is found in the recipient list (see Listing 5.7), then the `Spatial-Client` subscribes to the stream. If not, the stream is discarded, and no further media data is sent. The metadata also contains information about audio translation (see Section 5.4.3), dictating the position, for a given participant, where the audio stream should be located in 3D spatial space.

5.4.2 Browser Communication

Each monitor will have an open browser window, therefore, there are multiple `Spatial-Clients` running on a user's machine. `Spatial-Client` uses Javascript's BroadcastChannel interface to communicate between windows and tabs. This provides some useful features. For example, a user can use the mute button on any of the browser windows, even if the current window isn't the window publishing. Additionally, if another browser context is not active, it can inform the user of any problems it detects.

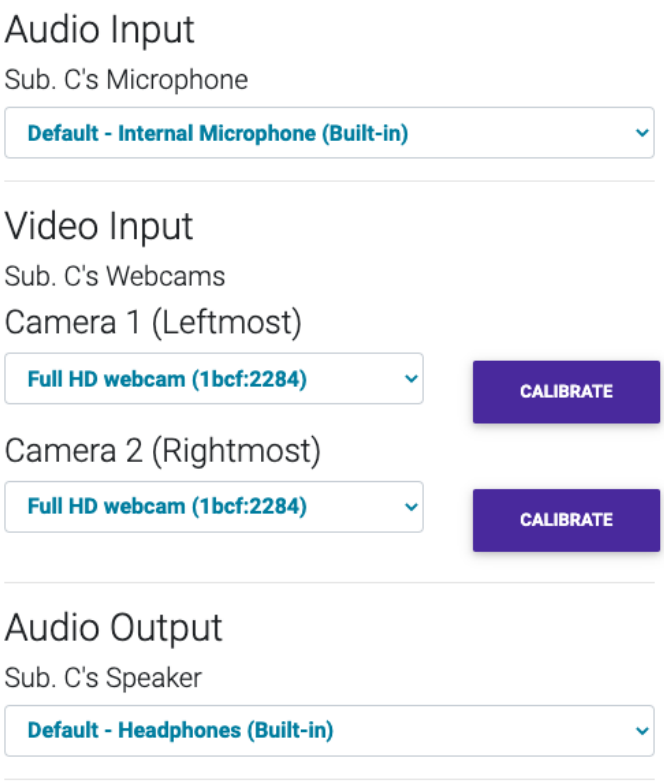

Figure 5.5: *Screenshot of the environment setup screen for* ***Sub. C****. The participant is invited to set up their microphone, webcams and speaker.*

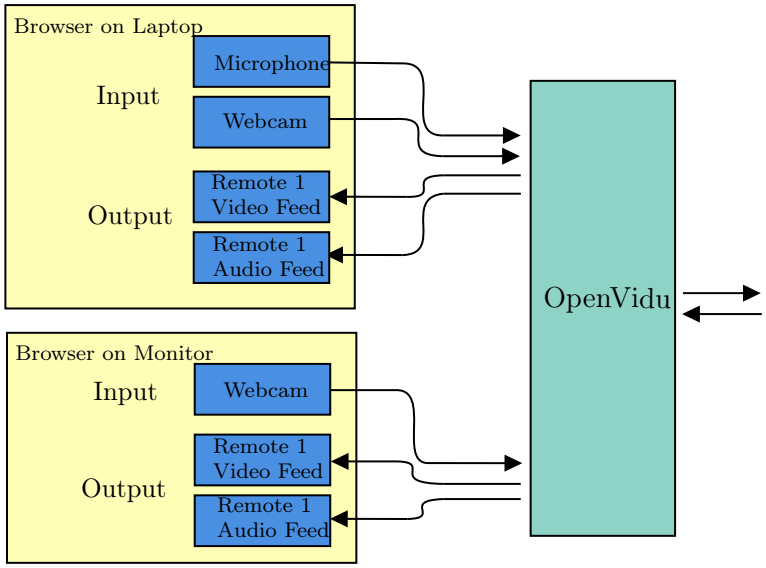

Figure 5.6: *A diagram showing the input and output streams of a typical Spatial participant's browser windows.*

5.4.3 Audio processing

This subsection discusses the audio processing functionality to ensure our audio streams are consistent with the visual spatial conference. Since each audio environment is location dependent, all processing is performed in `Spatial-Client`.

```
{
        "sendingUser": "Sending User Id",
        "sendingUserName": "David",
        "videoStream": true,
        "targetUsers": [
                {
                "id": "Recieving Screen Id",
                "angle": -60,
                "order": 1
                },
                {
                "id": "Recieving Screen Id",
                "angle": 60,
                "order": 2
                }
        ]
}
```

Listing 5.7: *Example JSON of the Spatial Metadata sent alongside media streams.*

Input processing

Multi-channel (stereo) microphones tend to produce unwanted 3D audio effects, since they provide their own positional characteristics. This can create confusing and inconsistent spatial effects as the audio is translated again before it's played for remote participants. To solve this we down-mix all audio channels and send a single channel stream to OpenVidu, this discards the channels, keeping only the frequencies. Specifically, `Spatial-Client` sends the microphone source to `ChannelMergerNode`, which combines all incoming channels into one output channel, before it is sent to OpenVidu.

Output processing

When a recipient receives an audio stream from a remote participant, `Spatial-Client` applies a 3D sound effect to move the audio to its correct position. `Spatial` uses the `PannerNode` API found in browsers to perform this transformation client-side. The API requires a 3D translation of the audio source. This is calculated using a smooth curve around the user based on the angle. The y parameter is kept 0: all participants are at the same 3D spatial height:

For a given angle θ, in radians:

$$x = sin(\theta) \tag{5.1}$$

$$z = \left\{ \begin{array}{ll} sin(\frac{\pi}{2} + \theta), & \text{for } \theta < \frac{\pi}{2} \\ sin(\frac{\pi}{2} - \theta), & \text{for } \theta > \frac{\pi}{2} \end{array} \right\}) \tag{5.2}$$

where θ is the relative angle between a pair of participants in the spatial environment.

Limitations

Audio in `Spatial` is not strictly spatially consistent because of the use of a single microphone. Unlike previous work, we used one microphone per individual rather than one

microphone per spatial connection. While laptop webcams are typically found at the top centre of the screens, microphones have no standard placement, making any calibration efforts complicated. Subjects are unlikely to have a secondary microphone at home.

The main limitation of this approach is the inability of the speaker to regulate the volume of their speech for different participants purposely. This removes the possibility for side conversations since participants cannot use the direction of their head to change the amount of sound that is sent to particular participants, e.g. leaning in. Future work could enable this ability through explicit user muting or simulating directionality through the use of head pose detection from the webcams.

5.4.4 Network engineering

Isotropic videoconferencing is network-intensive by nature. For every remote participant, a user must both send a new video feed and receive a video and audio feed from them. We identified two key factors that improve user satisfaction, by changing bandwidth allocation, from previous literature [69]. Firstly, ensuring a minimum quality of service for audio, then improving video quality as much as possible.

Prioritisation of audio streams

OpenVidu enables adaptive bitrate for all media streams, indiscriminately. However, we want to ensure guaranteed flow for audio streams as users place a minimum service of audio much higher than video.

To implement this functionality, we edited the source of OpenVidu. For all audio streams, we disabled Kurento's variable transcoding of the stream. We also set a minimum send bandwidth of 28 kilobits per second for audio streams and 0 for video streams, so ensure that audio streams are given a dedicated bandwidth part. [1].

Increasing the bandwidth of the browser

Browsers, by default, apply a limit on the amount of bandwidth by a WebRTC connection to avoid wasting resources in the user endpoints and save money reducing bandwidth usage on the servers. On Google Chrome's browser, this is 2Mbps.

Spatial's upload bandwidth is typically larger than typical videoconferencing applications which only transmit one video and one audio stream. Therefore, `Spatial-Client` overrode the setup of the WebRTC peers and increased their `AS:BITRATE` and `TIAS:BITRATE` attributes which dictate the max bandwidth for Chrome and Firefox and increased them to 10 Mbps.

5.5 Package overview

Spatial uses the Maven package manager to manage library dependencies. To streamline the development process, both *Spatial*'s Java and C++ code are compiled during Maven's build process. *Spatial* includes an example self-signed certificate as access to webcams and microphones requires an SSL certificate. *Spatial* also includes the modified versions of OpenVidu client and server used for network performance benefits.

[1] According to Opus documentation, the audio codec used by all major browsers, 28 kbps provides full-band audio for VoIP [70].

To help the future deployment and the evaluation setup, we created a setup script, which installed system dependencies, such as Kurento Media Server and related modules. Java MVC requires a minimum amount of entropy in `/dev/random` for session management, which is also checked as part of the setup script. We tested the system to run both on macOS (10.15) and Ubuntu (18.04).

Chapter 6

Evaluation

This section introduces and presents the results of a user study to assess whether team performance and conversation analysis are affected through the use of a spatially faithful videoconferencing system.

COVID Notice: *Unfortunately, because of self-isolation due to COVID-19 symptoms, shortly followed by the UK Government's Stay at Home advice, I was unable to test the desired number of participants. Additionally, some groups could only test the control system, not Spatial, the target system. Finally, we would have liked to compare both systems with a face-to-face interaction, however the social distancing guideline prohibited this. Due to this, the evaluation was necessarily limited.*

6.1 Hypothesis

Chapter 2 presented evidence, from prior work, that the addition of spatial faithfulness to videoconferencing, increased participants' awareness of visual cues resulting in more efficient conversations and increased task performance. While we cannot directly calculate the effectiveness of *group decision-making* using quantitative measures, previous work [2, 6, 40] has shown that with specialised setups, groups complete tasks faster and have more efficient conversations. In this chapter, we evaluate whether these same gains can be obtained using *Spatial*, which relies only on off-the-shelf components.

Therefore, we hypothesise that groups using *Spatial* will exhibit these same characteristics: groups will complete tasks faster and have more efficient conversations. By testing these two hypotheses, we can evaluate whether using spatially faithful videoconferencing in work-from-home environments will help group decision-making.

Unfortunately, there is not a *set* definition of the characteristics that make a conversation efficient. Following previous work, we use face-to-face conversations as our ideal situation since this provides the maximum amount of visual and verbal information. We can then evaluate the effectiveness of our system by comparing the conversations, using *Spatial* replicate the characteristics of a face-to-face conversation.

Unfortunately, we were unable to carry out face-to-face comparisons of our own, due to the ongoing COVID-19 pandemic. Therefore, we rely on Boyle and Henderson's [6] study which compares the map task between different conversation contexts. We could not use any other studies because they either performed a different evaluation type, used different

conversation types [21, 31] or did not provide quantitative data [2]. While the map task contains some drawbacks, as outlined in our Evaluation design (see Section 4.2), it is the closest and most documented cooperative task in research, hence our decision.

6.2 Method

6.2.1 Participants

Participants were recruited from the University of Cambridge's Department of Computer Science and Technology and from Jesus College, Cambridge. All participants signed a consent form for their conversations to be recorded. Nine individuals took part in this study, (three groups). One group completed a within-subjects design, the other two groups only performed the control experiment. All participants reported using well-known videoconferencing systems, such as Zoom, Skype and Jitsi *multiple times* before the experiment. Participants comprised of six males and three females, ranging in age from 21 to 23.

We conducted a statistical power analysis for an F-test using G*Power. The power analysis allows us to calculate the minimum number of subjects needed in a study. We planned a repeated measures, within-between interaction ANOVA study. Similar to previous literature [71], we assumed a small medium effect (f = .25) and power set to .80. Using the different conversation analysis types, as the outcome variables (10 measurements) the analysis determined that 14 triads (42 individuals) would need to take part in this study. Unfortunately, a University-wide closure of buildings meant that we were unable to gather enough participants for 14 trials.

6.2.2 Control system

The **Control System** is the non-isotropic videoconferencing system used as the comparison of "standard videoconferencing applications". We used the OpenVidu Basic Webinar application from the OpenVidu Demos repository [72].

By using the same OpenVidu application as *Spatial*, we reduce the potential number of covariates in our study, to ensure our results are because of a difference in system quality. This also simplified testing, since we only required one videoconferencing server installed.

6.2.3 Target system

The **Target System** is the isotopic videoconferencing system introduced in the previous chapter: *Spatial*.

6.2.4 Experiment design

Server setup

All experiments used a DigitalOcean virtual machine, with 2 cores and 2GB of memory, which OpenVidu's load testing documentation indicated was more than sufficient [73]. We verified that during our tests, our virtual machine never reached over 80% CPU & memory utilisation. All experiments were recorded using OpenVidu's inbuilt recording functionality, which created individual video and audio files for each participant.

Experiment setup

The experiments were performed remotely in separate rooms, in which each participant used laptop devices (see Table 6.1). For the spatially faithful system, external monitors were used in addition to the laptop setups. These monitors and webcams were described in Table 6.2. The webcam quality was set to 720p HD in all experiments, and all internet connections were tested pre-experiment to ensure that their speed was sufficient. Participants in each group were familiar with each other pre-experiment and had the opportunity to discuss any questions or concerns about the task with the moderator privately, before starting.

Make & Model	Quantity	Screen Size	Webcam Quality
Macbook Air 13"	2	13 inches (33.02cm)	720p HD
Macbook Pro 13"	3	13 inches (33.02cm)	720p HD
Macbook Pro 15"	1	15 inches (38.10cm)	720p HD
Dell XPS 13	2	13 inches (33.02cm)	720p HD

Table 6.1: *Breakdown of laptop devices used in experiments*

Monitor Make	Screen Size	Webcam Make	Webcam Quality
Fujitsu	23 inches (58.4cm)	Sandstrom	1080p HD
Philips	25 inches (63.5cm)	Sandstrom	1080p HD
Samsung	23 inches (58.4cm)	Logitech	1080p HD

Table 6.2: *Breakdown of external monitors and webcam combinations used.*

Task setup

Participants were assigned to groups of three. Each group was given one of the two hidden profile tasks, described in Section 4.2.1. All participants were given the shared information for the hidden profile task and made aware of the difference between the shared and unshared information. Each participant was randomly assigned a role and given the unshared information for that role. Each participant was provided with a notepad for any brief notes they have during the discussion. The experiment took about 20 minutes, split between pre-discussion reading time and the actual discussion.

6.2.5 Coding method

The first and last five turns and all interactions with the moderator were excluded from the analysis to avoid bias in opening and closing sequences and conversation type. The audio files from each conversation were then transcribed, noting the occurrences of backchannels, interruptions, overlapping speech and handovers. Each turn was assigned a start and end time, and the number of words per turn were noted as well. We did not code phonetic information (e.g. pitch level or a drawl on the final syllable); however, questions were used to denote the type of explicit handover. Sentences were transcribed as they were spoken, including any syntactical errors and speech disfluencies (e.g. “huh”, “erm”, “so”). In the small number of cases where it could not be ascertained if the utterance was intelligible or not, the turn’s word count was excluded from the wider analysis. Since we were not able to carry out significance tests, we opted not to use multiple judges to code the recordings.

As audio was recorded at all locations separately, assessments of simultaneous speech were analysed at source rather than the destination. While the audio transmission lag

was negligible [1] , we do not rule out the possibility that both parties would unknowingly be speaking at the same time.

Transcript key

The transcriptions were created using widely-used conventions from Atkinson and Heritage (1984) [74]. We provide a relevant overview for the reader:

[	A single left square bracket indicates the point of overlap
=	Equal signs at the end of one line and the beginning of the next, indicate no gap between the two lines
( )	Empty parentheses indicate the transcriber's inability to hear that was said.
(word)	Parenthesized words are especially dubious hearings.

Coding turns

Coding a *turn* in practice is difficult since it typically requires inference of what a subject is attempting to do. For example, a verbal backchannel may occur at the same time a new participant is attempting to gain the floor, e.g.

A: . . . is that right?
B: [yeah
C: [but I thought..

In this instance, the transcriber must determine whether participant B intended to take the floor or rather, it was merely a verbal backchannel. Turns can overlap. If another participant interrupts and gains control of the floor, forcing the first participant to stop talking, then their turns will overlap.

A: So if we choose Wednesday for [drinks then
B: [wait no what happens if. . .

Turns were given to participants who successfully gain full control of the floor. For example, if there is a simultaneous start, the participant who subsequently keeps the floor is given the turn. If simultaneous conversation occurs for longer than four words, then the turn is given only once the participant has gained control, i.e. long stretches of people talking one another are classified as no turn.

6.2.6 Dependent variables

The hypothesis under test concern the dependence of the existence of spatial cues on task performance and conversation efficiency. Apart from task duration, a range of other variables were measured to characterise the differences between the conversations on the control and target system. As each test group varied in the number of turns for completion (see Table 6.3), the conversation analysis features are given as rate per 100 turns.

Task Performance: Determined by the duration it took for each group to complete the task. A shorter duration indicated a more efficient group. Task duration was defined as the time elapsed between the start of the first turn and the end of the last turn.

Auditory backchannels rate

[1] This was verified by collecting the WebRTC statistics from the Kurento Media Server

Interruptions rate

Overlaps rate: Overlaps were split into their different types: Simultaneous Starts, Floorholding and Projection/Completion

Handovers rate: Handovers were split into their different types: Direct Addressing, Tag Questions, Elliptics and Social Identities

6.3 Results

This section discusses the results of the user study. As fewer number of groups performed the study than was required by the power analysis, we do not perform any significance tests.

6.3.1 Overview

While three groups took part in our study, only one group (Group 3) could test both systems.

Group	Type	No. of Turns	Duration (s)
Group 1	Control	61	407
Group 2	Control	142	564
Group 3	Control	148	645
Group 3	Target	51	257

Table 6.3: *Overview of the tests conducted and the duration of each.*

6.3.2 Turn analysis

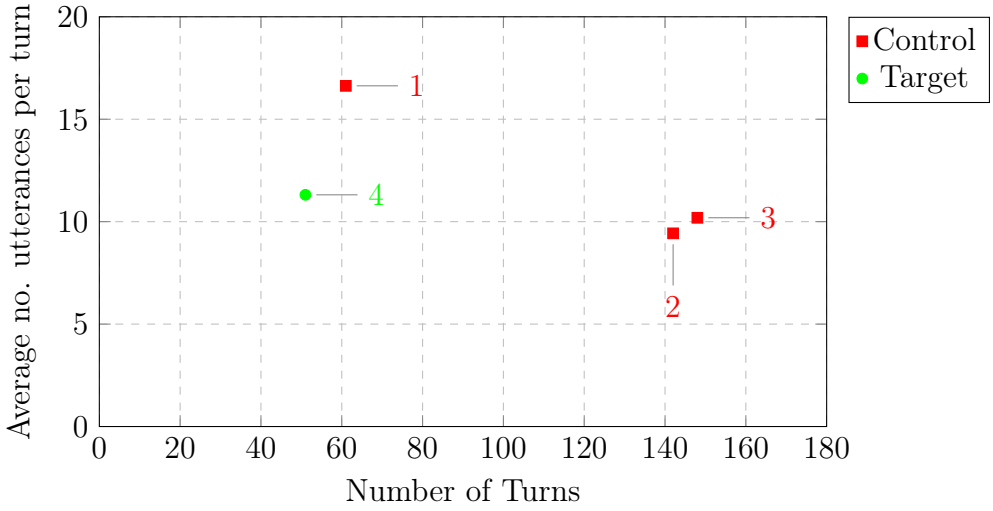

Figure 6.4: *A plot comparing the number of turns in the conversation with the average duration of each turn.*

This subsection provides a short overview of the conversation styles used by the different groups. Figure 6.4 shows the differences between the turns in each experiment. Except for Experiment 1, we can see there is negligible difference between the number of words

per turn between all experiments. This highlights an increased efficiency of conversation using *Spatial*, as while the number of words per turn was similar, the number of turns required to reach the correct answer was less.

Compared to the two other control experiments, experiment 1 produced a low number of turns. This is because the participants opted to share all of their constraints upfront rather than an investigative approach taken in the other experiments. This yielded turns with far more words per turn than the other experiments. The following example shows a typical interaction in Experiment 1 and Experiment 3. The first shows the lecture-like style, where speakers supply large amounts of uninterrupted information (see Excerpt 6.3.1).

Experiment 1	**A:**	I apparently have no car seemingly erm I have a delivery coming this week, erm a new phone it needs to be on either tuesday or sunday so I need to keep one of those two days free but it doesn't matter which of them
	B:	[tuesday or sunday okay
	A:	yeah and erm if its thursday or friday then I will be bringing a plus one
Experiment 3	**A:**	is there? B? B do you have a partner?
	B:	yeah my partner is gonna come on friday if we do it on friday
	A:	on what day does your partner come? C
	C:	my partner is visiting over the weekend, so if dinner dinner or drinks is on friday or sunday they will come along

Excerpt 6.3.1

6.3.3 Task performance

The average duration was 539 seconds for the three control experiments and 257 for the single target experiment. This result provides some evidence for the increased task efficiency of group decision-making under spatial conditions, though this would need to be verified with more experiments.

6.3.4 Auditory backchannels

Since videoconferencing applications typically incur a minor delay, backchannels typically drop when fewer visual cues are communicated by a channel [31]. Late-arriving backchannels are typically detrimental as its communicative impact is reduced and may disrupt the speaker at the remote location by its late arrival. Table 6.5 shows that the rate of auditory backchanneling is higher under the spatially faithful conditions (19.80 compared to 17.13).

Dependent Variable	Measure	Control	Target
Task Duration	Seconds	539	257
Auditory Backchannels	Rate	19.80	17.78
Interruptions	Rate	12.76	8.89
Overlaps			
Simultaneous Starts	Rate	5.68	2.22
Floorholding		1.94	0.00
Projection		9.51	22.22
All		17.13	24.44
Handovers			
Social Identities	Rate	0.00	0.00
Elliptics		6.36	8.89
Tag Questions		5.79	6.67
Direct Addressing		7.33	4.44
All		19.48	20.00

Table 6.5: *Overview of the dependent variables. Rates given per 100 turns. Control is an average of the three groups, while Target is the results of Group 3's test.*

6.3.5 Interruptions

A common finding in previous literature [75, 76] is that face-to-face conversations increase the number of interruptions, compared to video-mediated conversation. However, these works focus on fairly open-ended discussions and debates. One interpretation suggests that systems which convey more visual cues lead to fewer interruptions as participants can read others more accurately, and therefore time their interruptions to ensure success [16]. These interjections may turn into a projection of the end of turn rather than an interruption mid-turn. In either case, a decrease in the number of interruptions leads to more fluid conversations. The target system exhibits a lower rate of interruptions than the control.

6.3.6 Overlaps

While the spatially faithful system reduced the number of simultaneous starts and floorholding, the number of projections increased. This reduction in simultaneous starts and floorholding can be attributed to gaze. In the spatially faithful experiment, subjects often used gaze to hand the floor over to others, which reduced the number of simultaneous starts.

Excerpt 6.3.2 compares a similar conversation from Experiment 2 and 4.

	A:	Do you guys have partners out of interest?	
Experiment 2	**B:**		I do have a partner
	C:		[I do

Experiment 4	**A:**	(*Gaze at B*) sunday my partner would come, would anyone else's partner come on sunday?	
	B:		No
	A:		(*Gaze at C*)
	C:		No, I don't have a partner

Excerpt 6.3.2

Here we can see that the use of gaze in experiment 4, removes the simultaneous speech, leading to a more successful interaction. The reduction in simultaneous starts and floorholding indicates a more efficient conversation. Simultaneous starts are regarded as speaker switching breakdowns.

Subjects under *Spatial* also used projection and gaze to help with the turn negotiations. While projection occurred in all experiments, the use of the spatially faithful system meant participants would often use projection to take over the turn (see Excerpt 6.3.3). The subjects would use gaze to aid tagged questions which did not have an obvious recipient from the dialogue alone.

Experiment 4	**A:**	(*Gaze at B*)	
		. . . drinks you can only	[do tuesday?
	B:		[drinks come before games. . .

Excerpt 6.3.3

This overall increase in the number of overlaps is consistent with some previous work. Werkhoven et al. [2] reported an increase in the number of overlaps of their isotropic system over face to face and their non-isotropic system, however, they provide no explanation for this occurrence or breakdown into types of overlaps.

6.3.7 Handovers

Overall, we can see no difference in the total number of handovers between the spatially faithful system and the control system (see Table 6.5). However, the decomposition of handover types is not similar for both systems. The spatially faithful system shows increases in elliptics and tagged questions, but a decrease in direct addressing.

We can attribute this decrease in direct addressing; participants would instead use elliptics or tagged questions to handover the floor and use their gaze to indicate the addressee. This effect is also shown in Excerpt 6.3.2. This contributed to more efficient conversations as participants in *Spatial* used fewer names in turn negotiation.

6.4 Discussion

We compare the results of our study with the documented comparison of face-to-face conversation and video-mediated discussion presented in literature [6] in Table 6.6. As we could not perform significance testing, we present these comparisons at face-value. Unfortunately, the study did not cover all of the measured values in our study, so we are constrained to only five comparison data points.

Our first hypothesis is that groups using the spatially faithful system will complete the task faster. We can see that the results in Table 6.5, provide supporting evidence for this conclusion.

Our second hypothesis is that conversations under the spatially faithful system will be more efficient due to the increase of visual cues made available to them. We align our results to the comparison of the conversation characteristics between a face-to-face conversation and standard videoconferencing. Since we cannot directly compare our results, as the task type heavily influences the conversation characteristics, we compare our relative relationships since both studies compared to a standard videoconferencing application.

Characteristic	**Our Results**	**Boyle et al.'s Results**	**Supports Hypothesis?**
No of Turns	SC < VC	FTF < VC	✓
Words per Turn	SC > VC	FTF > VC	✓
Auditory Backchannels	SC < VC	FTF < VC	✓
Interruptions	SC < VC	FTF < VC	✓
Overlaps	SC > VC	FTF < VC	✗

Table 6.6: *Observed differences in conversation characteristics and channel properties of our user study, and of Boyle et al. [6] study, comparing videoconferencing and face-to-face conversation for cooperative problem-solving tasks. Final column states whether results support our hypothesis. Key: SC = Spatial videoconferencing, VC = Standard videoconferencing, FTF = Face-to-face*

For four out of five of our comparisons, we see that our results support our hypothesis.

As previously discussed, increase in overlaps under *Spatial* videoconferencing can be attributed to the large increase in the rate of projections: subjects used projection and gaze to help with the turn negotiations. While projection occurred in all experiments, the use of the spatially faithful system meant participants would often use projection to take over the turn (see Excerpt 6.3.3).

Chapter 7

Conclusion

This thesis studied whether the addition of spatial cues to commodity videoconferencing application improves group decision-making. We presented a new videoconferencing application, Spatial, with a description of the design choices and implementation details. Additionally, a comparison of the previous evaluation techniques were critically evaluated and new tasks were designed and evaluated based on group decision-making research.

The system was evaluated with a series of experiments, with triad sessions. The results showed that, compared to a gallery-view videoconferencing application, our system yielded better conversation patterns. Groups using *Spatial*, compared to the standard system, took less time per turn, required fewer turns to complete the task, and interrupted each other less.

While all participants reported using videoconferencing systems, indicating some familiarity; *Spatial*'s users only less than two minutes with the system before starting the exercise. This could indicate some bias to the status quo, where users aren't familiar with using spatial cues in videoconferencing. Long-term studies have shown that familiarity with a video-communication system can increase the efficiency of the participants over a system [77].

While our results were not conclusive, the study indicates that the successes of spatially faithful videoconferencing, reported in previous studies [2] using expensive hardware, could be achieved in a work-from-home environment. As more and more individuals and organisations look to working from home permanently [78], it is important that we ensure that our tools allow us to do so effectively.

7.1 Future Work

When designing and developing *Spatial*, there were several potential extensions we considered however due to limitations was not able to implement or evaluate.

- **Gaze redirection:** Instead of correcting all misaligned gaze back at the webcam, one could apply an offset to their eyes, so they could create eye-contact with remote participant. One could determine, using calibration, the angle between the eyes and the remote participant's position on the screen using eye gaze algorithms, then use the calculated offset to adjust their eyes. Kurento Media Server can apply OpenCV filters to media streams. A filter could then determine whether a face is shown in a video stream and apply the correct gaze offset, server-side. Although a solution for this functionality was designed and developed as part of this project, due to time and hardware constraints we could not test this functionality as part of our evaluation.

- **Media Priority:** A new WebRTC Priority Control API is being drafted, which will allow clients to manipulate the queueing priority of outgoing packets [79]. If used in conjunction with a form of gaze detection, we could dynamically prioritise the video stream that a participant is looking at, saving both bandwidth and increasing user satisfaction.

- **Curved Monitors:** While curved monitors currently have a low take-up, their falling prices could make them an attractive option in the future. *Spatial* could easily support these monitors by displaying multiple participants in one browser window.

- **Dynamic Spatial Environments:** Currently, environments are statically created and users cannot join once one has started. A more user-friendly option is a dynamic environment, similar to existing videoconferencing applications, where users can leave and join as they please. This would require a large change to the OpenVidu system, since currently clients cannot subscribe to a stream once they reject it.

- **Reducing the number of webcams:** The exponential nature of isotropic videoconferencing means that large conferences require lots of equipment. While, an additional monitor, up to a point, will increase the productivity of an individual, an additional webcam would typically be only used for isotropic videoconferencing.

  Here I present two future extensions of reducing the number of webcams for each participant:

  1. **View Morphing** is a technique to generate images from a different viewpoint given existing viewpoints [80]. However, classical methods that rely exclusively on image information, are sensitive to changes in visibility & light, and typically require precise calibration for accurate results. Methods using deep neural networks have been designed, which are a lot more forgiving, however, still cannot properly deal with different illumination and colour characteristics [81]. This can be a major problem in home office environments which rely more on natural light, than commercial offices, creating uneven light distributions.

  2. **Generative Query Networks** take images taken from different viewpoints and create an abstract description of the scene [82]. While in their infancy, they could prove an effective method of re-creating *lost information* and dealing with different illumination characteristics that view morphing methods may fail at.

We hope that these directions can be studied in the future.

Bibliography

[1] Abigail Sellen, Bill Buxton, and John Arnott. Using spatial cues to improve videoconferencing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 651–652, 1992.

[2] Peter J Werkhoven, Jan Maarten Schraagen, and Patrick AJ Punte. Seeing is believing: communication performance under isotropic teleconferencing conditions. *Displays*, 22(4):137–149, 2001.

[3] Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. The hcrc map task corpus. *Language and speech*, 34(4):351–366, 1991.

[4] Aleš Jaklič, Franc Solina, and Luka Šajn. User interface for a better eye contact in videoconferencing. *Displays*, 46:25–36, 2017.

[5] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2008.

[6] Elizabeth A Boyle, Anne H Anderson, and Alison Newlands. The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and speech*, 37(1):1–20, 1994.

[7] Gerald S Oettinger. The incidence and wage consequences of home-based work in the united states, 1980–2000. *Journal of Human Resources*, 46(2):237–260, 2011.

[8] Society for Human Resource Management. Shrm survey findings: Virtual teams. July 2012.

[9] Society for Human Resource Management. Covid-19 research: How the pandemic is challenging and changing employers. April 2020.

[10] Jordan Novet. A message to our users - zoom blog. `https://blog.zoom.us/wordpress/2020/04/01/a-message-to-our-users/`, April 2020.

[11] Eric S. Yuan. Cisco says webex video-calling service is seeing record usage too, even as competitor zoom draws all the attention. `https://www.cnbc.com/2020/03/17/cisco-webex-sees-record-usage-during-coronavirus-expansion-like-zoom.html`, March 2020.

[12] Nishant Bordia. *Role of Technology Selection in Supporting Collaboration and Communication in Globally Distributed Virtual Teams*. PhD thesis, 2017.

[13] Boris B Baltes, Marcus W Dickson, Michael P Sherman, Cara C Bauer, and Jacqueline S LaGanke. Computer-mediated communication and group decision making: A meta-analysis. *Organizational behavior and human decision processes*, 87(1):156–179, 2002.

[14] John Short, Ederyn Williams, and Bruce Christie. *The social psychology of telecommunications*. John Wiley & Sons, 1976.

[15] Judee K Burgoon and Thomas P Saine. *The unspoken dialogue: An introduction to nonverbal communication*. Houghton Mifflin Harcourt (HMH), 1978.

[16] Claire O'Malley, Steve Langton, Anne Anderson, Gwyneth Doherty-Sneddon, and Vicki Bruce. Comparison of face-to-face and video-mediated interaction. *Interacting with computers*, 8(2):177–192, 1996.

[17] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

[18] David Nguyen and John Canny. Multiview: spatially faithful group video conferencing. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 799–808, 2005.

[19] Adam Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63, 1967.

[20] Ye Pan, William Steptoe, and Anthony Steed. Comparing flat and spherical displays in a trust scenario in avatar-mediated interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1397–1406, 2014.

[21] Abigail J Sellen. Remote conversations: The effects of mediating talk with technology. *Human-computer interaction*, 10(4):401–444, 1995.

[22] OpenVidu Homepage. `https://openvidu.io/`, May 2020.

[23] BigBlueBtton Homepage. `https://bigbluebutton.org/`, May 2020.

[24] Luis López, Miguel París, Santiago Carot, Boni García, Micael Gallego, Francisco Gortázar, Raul Benítez, Jose A Santos, David Fernández, Radu Tom Vlad, et al. Kurento: The webrtc modular media server. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1187–1191, 2016.

[25] Ofcom. Uk home broadband performance. December 2018.

[26] HTML Living Standard. `https://html.spec.whatwg.org/multipage/web-messaging.html#broadcastchannel`, May 2020.

[27] Emanuel Schegloff, Gail Jefferson, and Harvey Sacks. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

[28] Elizabeth Spelke, William Hirst, and Ulric Neisser. Skills of divided attention. *Cognition*, 4(3):215–230, 1976.

[29] Jinni A Harrigan and John J Steffen. Gaze as a turn-exchange signal in group conversations. *British Journal of Social Psychology*, page 168, 1983.

[30] Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. Gaze and turn-taking behavior in casual conversational interactions. *ACM Trans. Interact. Intell. Syst.*, 3(2), August 2013.

[31] Brid O'Conaill, Steve Whittaker, and Sylvia Wilbur. Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. *Human-computer interaction*, 8(4):389–428, 1993.

[32] Geoffrey W Beattie. Floor apportionment and gaze in conversational dyads. *British journal of social and clinical psychology*, 17(1):7–15, 1978.

[33] Garold Stasser and William Titus. Hidden profiles: A brief history. *Psychological Inquiry*, 14(3-4):304–313, 2003.

[34] John P Lightle, John H Kagel, and Hal R Arkes. Information exchange in group decision making: The hidden profile problem reconsidered. *Management Science*, 55(4):568–581, 2009.

[35] Garold Stasser and William Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, 48(6):1467, 1985.

[36] Ivan Moser, Sandra Chiquet, Sebastian Kaspar Strahm, Fred Mast, and Per Bergamin. Group decision making in multi-user virtual reality and video conferencing. 2018.

[37] Andrew F Monk and Caroline Gale. A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes*, 33(3):257–278, 2002.

[38] Carmelina Trimboli and Michael B Walker. Switching pauses in cooperative and competitive conversations. *Journal of Experimental Social Psychology*, 20(4):297–311, 1984.

[39] Gillian Brown, Anne Anderson, Richard Shillcock, and George Yule. *Teaching talk: Strategies for production and assessment*. Cambridge University Press, 1985.

[40] Gwyneth Doherty-Sneddon, Anne Anderson, Claire O'malley, Steve Langton, Simon Garrod, and Vicki Bruce. Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied*, 3(2):105, 1997.

[41] Kurt T Dirks and Donald L Ferrin. The role of trust in organizational settings. *Organization science*, 12(4):450–467, 2001.

[42] W Randy Clark, Leigh Anne Clark, and Katie Crossley. Developing multidimensional trust without touch in virtual teams. *Marketing Management Journal*, 20(1):177–193, 2010.

[43] David T Nguyen and John Canny. Multiview: improving trust in group video conferencing through spatial faithfulness. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1465–1474, 2007.

[44] Christina Breuer, Joachim Hüffmeier, and Guido Hertel. Does trust matter more in virtual teams? a meta-analysis of trust and team effectiveness considering virtuality

and documentation as moderators. *Journal of Applied Psychology*, 101(8):1151, 2016.

[45] Milton Chen. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 49–56, 2002.

[46] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. 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*, pages 521–528, 2003.

[47] Takahiko Yamamoto, Masataka Seo, Toshihiko Kitajima, and Yen-Wei Chen. Eye gaze correction using generative adversarial networks. In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pages 276–277. IEEE, 2018.

[48] Lori Foster Thompson and Michael D Coovert. Teamwork online: The effects of computer conferencing on perceived confusion, satisfaction and postdiscussion accuracy. *Group Dynamics: Theory, Research, and Practice*, 7(2):135, 2003.

[49] Anne H Anderson, Lucy Smallwood, Rory MacDonald, Jim Mullin, AnneMarie Fleming, and Clair O'Malley. Video data and video links in mediated communication: what do users value? *International Journal of Human-Computer Studies*, 52(1):165–187, 2000.

[50] Richard L Daft and Robert H Lengel. Organizational information requirements, media richness and structural design. *Management science*, 32(5):554–571, 1986.

[51] Jonathan K Kies, Robert C Williges, and Mary Beth Rosson. Evaluating desktop video conferencing for distance learning. *Computers & Education*, 28(2):79–91, 1997.

[52] Jerald Greenberg. The role of seating position in group interaction: A review, with applications for group trainers. *Group & Organization Studies*, 1(3):310–327, 1976.

[53] Ana Ortiz De Guinea, Jane Webster, and D Sandy Staples. A meta-analysis of the consequences of virtualness on team functioning. *Information & Management*, 49(6):301–308, 2012.

[54] Alan Robinson and Jochen Triesch. Task-specific modulation of memory for object features in natural scenes. *Advances in cognitive psychology*, 4:1, 2008.

[55] Dominik Giger, Jean-Charles Bazin, Claudia Kuster, Tiberiu Popa, and Markus Gross. Gaze correction with a single webcam. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014.

[56] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6932–6941, 2019.

[57] Peter Bex. Sensitivity to spatial distortion in natural scenes. *Journal of Vision*, 8(6):688–688, 2008.

[58] Thomas A Busey, Nuala P Brady, and James E Cutting. Compensation is unnecessary for the perception of faces in slanted pictures. *Perception & Psychophysics*, 48(1):1–11, 1990.

[59] Apache OpenMeetings Project Page. https://openmeetings.apache.org/, May 2020.

[60] Jisti Meet. https://jitsi.org/jitsi-meet/, May 2020.

[61] Nancy Pennington and Reid Hastie. Reasoning in explanation-based decision making. *Cognition*, 49(1-2):123–163, 1993.

[62] Samuel N Fraidin. When is one head better than two? Interdependent information in group decision making. *Organizational Behavior and Human Decision Processes*, 93(2):102–113, 2004.

[63] Spring Homepage. https://spring.io/, May 2020.

[64] Trygve Reenskaug. The model-view-controller (mvc) its past and present. *University of Oslo Draft*, 2003.

[65] Thymeleaf Homepage. https://www.thymeleaf.org/, May 2020.

[66] Oracle. Java native interface specification: 1 - introduction. https://docs.oracle.com/en/java/javase/11/docs/specs/jni/intro.html#java-native-interface-overview.

[67] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1087–1096, 2019.

[68] Tensorflow Lite | ML for Mobile & IoT. https://www.tensorflow.org/lite, May 2020.

[69] J Baraković Husić, S Baraković, and A Veispahić. What factors influence the quality of experience for webrtc video calls? In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 428–433. IEEE, 2017.

[70] Opus Recommended Settings. https://wiki.xiph.org/Opus_Recommended_Settings, May 2020.

[71] Shannon Moore, Michael Geuss, and Joseph Campanelli. Communicating information in virtual reality: Objectively measuring team performance. In *International Conference on Human-Computer Interaction*, pages 473–489. Springer, 2019.

[72] OpenVidu Demos. https://openvidu.io/demos/, May 2020.

[73] OpenVidu. Openvidu load testing: a systematic study of openvidu platform performance. https://medium.com/@openvidu/openvidu-load-testing-a-systematic-study-of-openvidu-platform-performance-b1aa3c4 December 2018.

[74] J Maxwell Atkinson, John Heritage, and Keith Oatley. *Structures of social action*. Cambridge University Press, 1984.

[75] Mark Cook and Mansur G Lalljee. Verbal substitutes for visual signals in interaction. *Semiotica*, 6(3):212–221, 1972.

[76] Derek R Rutter and Geoffrey M Stephenson. The role of visual communication in synchronising conversation. *European Journal of Social Psychology*, 7(1):29–37, 1977.

[77] Rick van der Kleij, Roos M Paashuis, JJ Langefeld, and Jan Maarten C Schraagen. Effects of long-term use of video-communication technologies on the conversational process. *Cognition, Technology & Work*, 6(1):57–59, 2004.

[78] Dynata. Our changing work lives | covid-19 edition. May 2020.

[79] WebRTC Priority Control API. `https://www.w3.org/TR/webrtc-priority/`, May 2020.

[80] Steven M Seitz and Charles R Dyer. View morphing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 21–30, 1996.

[81] Dinghuang Ji, Junghyun Kwon, Max McFarland, and Silvio Savarese. Deep view morphing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2155–2163, 2017.

[82] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

Appendix A

Evaluation Task

A.1 Evaluation Task 1

A.1.1 Shared Information

Background
Three friends, Jessica, Lauren and Kenny are trying to organise a dinner and drinks together at their favourite place: David's next week. However, they each have their own prior commitments, so you must work together to make sure everyone is happy.

Rules

- Dinner and drinks are on different nights
- They will only do one thing a night, i.e. no drinks at dinner, therefore food options do not matter on the drinks night and vice versa

David's Menu
Monday: Food: Risotto (vegan options available) Drinks: Wine, Beer, Cider

Tuesday: Food: Pizza (vegan options available) Drinks: Wine, Beer, Cocktails

Wednesday: Food: Burgers (vegan options available) Drinks: Beer, Wine, Cocktails

Thursday: Food: Steak Drinks: Happy Hour Cocktails!

Friday: Food: Pasta (vegan options available) Drinks: Everything

Saturday: Food: Brunch (vegan options available) Drinks: Wine, Prosecco

Sunday: Food: Roast Drinks: Wine, Cider, Cocktails

A.1.2 Participant 1

Jessica is a Regional Manager at BCG. She has no dietary requirements and no drink requirements. Jessica goes to the gym twice a week in the evenings. However, she can only gym on Monday to Thursday. Jessica can gym after dinner, if it isn't burgers and definitely cannot gym after drinks. Jessica has her car in the shop until Tuesday, however

she can get the tube from work to David's. Jessica states that dinner must come earlier in the week than drinks and can't happen on consecutive days. Jessica is a maid of honour at her best friends' wedding on Saturday and will be unavailable all day. Her partner is coming to visit over the weekend, so if dinner or drinks is on Friday or Sunday, her partner will come along.

Dinner Day:
Drinks Day:
Gym Days (2):

A.1.3 Participant 2

Lauren is a content creator for her own company. Laurent has recently turned vegan but has no restrictions on what she can drink. Lauren works from home and will need a lift from either Jessica or Kenny to dinner or drinks. Lauren also has new equipment arriving this week, she can schedule it for either Tuesday or Sunday evening. If a delivery is arriving, she cannot do anything else that evening. If dinner or drinks is on Thursday or Friday, her partner insists they will come to spend time with Lauren.

Dinner Day:
Drinks Day:
Delivery Day:

A.1.4 Participant 3

Kenny is a salesman for Facebook. Kenny arrives from his business trip on Monday and will be too tired to organise something on that day. Kenny's car is in perfect condition. Kenny will eat/drink anything. Kenny also has a date this week, but will need to organise it for either Thursday or Sunday. Kenny has a client meeting on Friday morning, and cannot be hungover for it, i.e. no drinks on Thursday evening. Kenny is the only single one and will only come if he is not 5th wheeling, i.e. no other partners are attending.

Dinner Day:
Drinks Day:
Date Day:

A.2 Evaluation Task 2

A.2.1 Shared Information

Background
Three friends, Mark, Barnett and Amber are trying to organise a games night and drinks together at their favourite board games cafe: David's next week. However, they each have their own prior commitments, so you must work together to make sure everyone is happy.

Rules

- Games and drinks are on different nights
- They will only do one thing a night, i.e. no drinks at dinner, therefore food options do not matter on the drinks night and vice versa

- If games is arranged on Thursday or Sunday, they will require an even number of people to play (incl. partners).

David Boozy Boardgames's Menu
Monday: Games: Monopoly Drinks: Wine, Beer, Cider

Tuesday: Games: Darts Drinks: Wine, Beer, Cocktails

Wednesday: Games: Mind the Gap, the tube game Drinks: Beer, Wine, Cocktails

Thursday: Games: Chess (even numbers only) Drinks: Happy Hour Cocktails!

Friday: Games: Snakes and Ladders Drinks: Everything

Saturday: Games: Puzzles Drinks: Wine, Prosecco

Sunday: Games: Checkers (even numbers only) Drinks: Wine, Cider, Cocktails

A.2.2 Participant 1

Mark is a Regional Manager at McKinsey. He has no game preferences and no drink requirements. Mark has his car in the shop from Saturday, however she can get the tube from work to David's. Mark plays squash twice a week in the evenings. However, he can only go to squash on weekdays. Mark is going to visit his parents on Friday, so will be out of town for games, but might be able to get back for drinks.

Dinner Day:
Drinks Day:
Squash Days (2):

A.2.3 Participant 2

Barnett is an influencer. Barnett likes any games and has no restrictions on what he can drink. Barnett has a one-day off-site on Wednesday and will be too tired to organise something on that day. Barnett works from home and will need a lift from either Mark or Amber to dinner or drinks. Barnett also has new equipment arriving this week, he can schedule it for either Tuesday or Friday evening. If a delivery is arriving, he cannot do anything else that evening. If dinner or drinks is on Wednesday or Friday, her partner insists they will come to spend time with Barnett.

Dinner Day:
Drinks Day:
Delivery Day:

A.2.4 Participant 3

Amber is a salesman for Facebook. Amber only likes drinking cocktails, she won't drink anything else Amber states that drinks must come earlier in the week than games and can't happen on consecutive days. Amber has agreed to loan a car to a friend on either Monday or Sunday Amber has a client meeting on Friday morning, and cannot be hungover for it, i.e. no drinks on Thursday evening. Amber's partner is coming to visit over the weekend, so if games or drinks is on Friday or Sunday, her partner will come along.

Dinner Day:
Drinks Day:
Loan Day:

Appendix B

Spatial Model