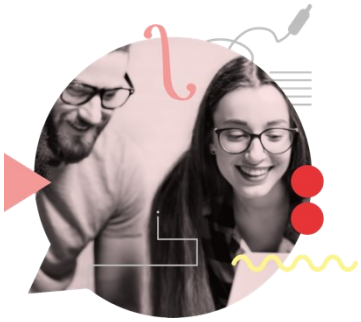
Toward a Telepresence of Sound

Video Conferencing in Spatial Audio

Jackson Goode



Master's program in
Music, Communication, and Technology

Department of Music
Norwegian University of
Science and Technology

Department of Musicology
University
of Oslo

May 2021

© 2021 Jackson Goode

Abstract

Digital communications technologies have developed at an increasingly rapid pace, with the COVID-19 pandemic accelerating its recent adoption. This shift over the last few decades has seen a mass migration online, where utilities like video conferencing software have become essential to entire industries and institutions. This thesis proposes the integration of binaural spatialized audio within a web-based video conferencing platform for distributed conversations. The proposed system builds upon findings on the benefits of spatial audio in video conferencing platforms and is guided by the tenets of telepresence. The developed implementation is based on Jitsi Meet, a robust open-source conferencing system. It localizes participant's voices through sound spatialization methods provided by the Web Audio API. . This project treads new ground in exploring how localized audio can be conceptualized within an accessible telecommunications platform, proposing a novel integration of HRTF-based binaural spatialization within a standard video conferencing layout. System design and experimental questions used in a technical evaluation and user study are informed from a review of audio and video conference systems found in the literature and commercial market. The system evaluation suggests its viability from a compatibility and performance perspective. Perceptual metrics of cognitive load, social presence, and intelligibility are further investigated by a user study where four remote subjects were asked to engage in a short group discussion on a live deployment of the system. Results find support for improvements across all defined metrics as well as increased opinion scores regarding the preference of conferencing with a spatial audio system.

Acknowledgments

Completing a master's program is no walk in the park. It is even more challenging when it takes place in a country you have never visited. Add the "once in a lifetime" pandemic to the mix and the remote learning experience we were studying from a pedagogical perspective became a lot more real than any of us had anticipated. Yet in a way, this was a climate that we had prepared for in theory and practice. The program in Music, Communication, and Technology (MCT) levied the intersection of music technology, communication theory, and the critical reflection of the humanities onto a foreseeable future where physical distance might be easily overcome by digital presence. It is a future that seems to be arriving sooner rather than later.

Adjusting to my brief stay in a foreign nation wouldn't have been possible without the tremendous support from my classmates, many of whom were navigating the same challenges in communications and distance on a daily basis. The researchers and academic staff at UiO were welcoming and helpful in every fashion they could be. I want to thank my professor, advisor, and colleague Stefano Fasciani who has helped me at every stage of the thesis and has taken such care in shaping the MCT program into one that aligns with the interests and reflections of its students. Completing this work and program would have not been possible without the constant support and inspiration from my parents, friends, and partner who encouraged me to take a risk and pursue an education in the fields I love.

Table of Contents

1. Introduction

1.1. Overview

Innovations in telecommunications have dramatically transformed the way humans interact over the last century, alongside the development and accessibility of the internet. This revolution has shifted nearly every industry that exists. The changes that have followed have allowed businesses and organizations to expand beyond their immediate locale, interact with remote collaborators, and reduced the need for individual travel simply for communication and as a result, the resources that accompany it. For rural communities, it has allowed them to stay in touch with the happenings in their societies. Within educational institutions, widespread access to the internet has led to the distribution of knowledge and has generally provided unprecedented access to information. And for each individual, the social networks afforded by the progress in telecommunications are far greater than at any point in history. It is hard to imagine daily life without the ability to communicate in near real-time with any person on earth.

Since the outbreak of the COVID epidemic beginning in early 2020, there has been a striking rise in the use of telecommunication platforms. The extensive closure of businesses, schools, and entertainment venues has forced those affected to migrate their practices and work online. Indeed, many industries have been upended by this migration to digital platforms in their attempt to continue to provide services and enable communication between clients and fellow teammates. Many businesses and institutions, much to their surprise, have been able to adapt to the exclusively digital environments. Individuals have also reclaimed as much of their lost social interactions through telecommunications. Even still, there has been major disruption from our standard means of physical communication and the behavioral nuance that we often take for granted in face-to-face interactions. While comprehensive data has yet to be compiled regarding the full effects of COVID-19 on the landscape of digital communication, it will regardless stand as a major trial for telecommunication platforms in adapting to a socially distanced world.

Given the recent dependence upon video-conferencing tools, many users of these services often struggle with issues across the board in adapting to modern communication tools facilitated by computers. It is likely that digital illiteracy among older demographics and those without access to technology has hobbled the transition to online communications. This is compounded with the fact that current internet infrastructure in most countries can barely accommodate the requirements needed for high-quality teleconferencing. But even within a user's control there are a myriad of devices that may contribute to the success or failure of a digital experience: a user's router, microphone, speakers, display, and processing capabilities of the device used. In most cases, none of the components are ideal for clear, networked conversations. While developments in hardware and networking infrastructure are not commonly adjustable from a user level, there may be techniques in the design and processing of participant's audio that may lead to more intelligible and engaging conversations and ones that lead to less "Zoom fatigue" (Ramachandran, 2021).

This work offers a viable technique to enable audio spatialization for distributed conversations within an open-source, video-conferencing application. The addition of a spatialized audio appears to offer several perceptible benefits in addition to being generally preferred by users of these systems. At the present moment, there are native, features present in all modern browsers that enable a high-quality binaural rendering of sound sources. This will be employed and evaluated against academic and commercial video conferencing systems in the hope of a more realistic conferencing experience.

1.2. Concepts

The scope of this project is transdisciplinary and brings together the topics of teleconferencing, telepresence, spatial audio, and the respective user experience. Each has its own history and technical successes, and ought to be considered as a part of the genesis from which this project arises. Their relevance will be discussed in context to how a spatially distributed audio system contributes to the goal of telepresence or the experience of feeling located in an environment through the assistance of technology. As a corollary, this work will investigate how spatial audio may contribute to reducing the cognitive effort, perception of social presence, and intelligibility through a number of evaluative queries.

This work will begin with an overview of how video conferencing has become a critical technology in the modern day, over the COVID-19 pandemic, and why this current project is motivated by failures of standard video conferencing applications. Afterward, literature is reviewed to address the themes of telepresence and how early experiments with telecommunication establish its terms for evaluation. These studies, in the evaluation of telecommunication platforms, provide tenets and methods that remain at the seat of interest in the scientific literature. Another topic that follows in the context of simulating a non-local environment, is spatial audio; specifically, those tools and methods which enable the experience of synthetic, three-dimensional sound. Investigations of spatial audio will be explored and how it might compare to systems with spatial video.

These topics will serve as the canvas upon which one can examine how similar systems work in practice, from experimental to commercial products, and how these models can shed light on the experience that a supposed system might enable today. Systems that implement spatialized audio will be examined from both within and outside of the literature. Examples in the literature motivate the idea that a spatially coherent system would provide tangible benefits to users in metrics of performance and ease of use. Moving to related platforms that are publicly available, a brief survey of the field will highlight the current absence of a spatial audio system within a conventional video conferencing design.

1.3. Contribution and research question

This thesis discusses spatial audio within research on telecommunications, its evaluation as framed within the goal of telepresence, and the context of modern teleconferencing systems as a preface to establishing a novel, spatial audio, teleconferencing system. Specifically, this thesis contributes with:

- A conceptual design of the sonic and visual features of a spatialized video conferencing system against a foundation of related literature.
- A method to integrate spatialized audio within an existing WebRTC-based conferencing system.
- An implementation and integration of the design and methods within the platform, scalable up to five concurrent participants.
- An evaluation of the system's performance and design choices from a computational perspective and in the context of modern web technologies.
- A user study validating the proposed system and providing feedback on how the integration of spatial audio affects the experience of video conferencing.

This project will contribute to the growing literature on novel configurations of video conferencing applications with a valid prototype of such a spatial audio system. A technical evaluation of this system as well as a user study will examine the experience of the system from several critical perspectives. This builds upon existing academic evaluations of experimental systems but within an open-source, web-based application. A significant outcome of this work is an open-source project that anyone can access as a conferencing solution or fork for development. Previous attempts to integrate spatial audio into teleconferencing applications failed to publish their platforms openly to the research community, hindering replication of the studies and comparisons. It also prevents the general public from benefiting from original contributions to software development. In addition, an evaluation of the current web technologies that enable spatial audio provides a basis for future development of audio on web-based conferencing.

1.4. Motivation

Something that cannot yet be estimated, is how many industries, reliant on communication, have been disrupted by COVID-19 imposed lockdowns. These lockdowns were mandated by governments for the sake of the health and safety of their citizens but left a void in many industries, education, and communities. One solution taken by many who were now faced with the impossibility of interacting in a social environment was to move to available digital solutions. While teleconferencing platforms have been freely available over the last decade at both a retail and commercial level for users and businesses, these solutions have never been tested to this scale and extent. Notably, industries in healthcare (Wosik et al., 2020), education (Crawford et al., 2020; Chen et al., 2020), and music (Rendell, 2020), to name just a few, were mobilized to adopt these new video communications platforms. What Ryan described as the "University of Tomorrow", in his sweeping discussion of telematics within education, was one that many students were both theoretically and practically thrown into in this period (Ryan, 1981). Though they imagine this future as a bright, interconnected, network of learners, this last year has shown that there is room for improvement. Furthermore, the fallout from COVID-19 related shutdowns has left not only working environments in limbo but also social circles as well.

These workspaces would have deeply struggled without access to these technologies, yet there are clear failings when it comes to the unique spaces of each of these industries. Many industries simply require multi-modal engagement in these shared spaces, such as education. The consensus is clear, that even the highest quality video telecommunication pales in

comparison to face-to-face instruction (Buxton, 1992). In industries like healthcare, video conferencing serves as a weak alternative, without the ability for physical assessment. As a replacement for casual socialization, our current video conferencing solutions struggle to represent the subtle, visual cues that exist in a physical space. Standard video conferencing platforms struggle with the inability to quickly discern facial expressions, eye contact, bodily gestures, and other subtle gestures. In the performing arts, however, even the state-of-the-art telecommunication systems struggle to meet the high-fidelity and low-latency requirements.

It is clear that new standards of work, education, and socialization are emerging on a global scale as a result of this pandemic. Statistics from Cloudflare, a US-based web infrastructure, provides evidence that internet activity has blossomed from 10-40% in some regions after the first international shutdowns in March of 2020 (Poinsignon, 2020; Graham-Cumming, 2020). This evidence is mirrored by Nokia's Deepfield in their intelligence report for 2020 with a 20-30% increase in average traffic (Nokia, 2021). In addition, Cloudflare's data can provide a heatmap detailing the change in web traffic from a working day in mid-February compared to a day in March. It shows a decrease in average internet traffic in cities and an increase in surrounding suburban and urban areas (Asturiano, 2020). This suggests a migration of internet activity from populated business districts to residential areas, from workplaces to homes. While the larger architecture of the internet and its distributors appear able to handle this new burden (Estes, 2020), the increased presence on the net has directed considerable attention to how these services perform in daily applications.

Another dramatic change throughout the virus-imposed lockdowns is the recent reduction in CO2 emissions compared to its anticipated climb (Friedlingstein et al, 2020; Le Quéré et al., 2020; Liu & Ciaias, 2020). This reduction chiefly comes as a result of the travel and contract restrictions imposed by governments and companies and the aforementioned shift to telework. As vaccines begin their mass distribution in many countries, travel both near and far will begin to pick up. This period of crisis, while taking an enormous toil of life, has also offered a vision of how societies and markets can feasibly operate without the extensive carbon-producing activities involved in communication.

1.5. Shortcomings of telecommunications

Over the last year, many people have personally grappled with the struggles of integrating video conferencing solutions into their daily routine. This collective frustration can often be seen in how major media distributions have reported on the topic (Murphy, 2020). In the pipeline that enables video conferencing, much of the technical foundation that supports video communication is outside of a user's reach. Most of these are dependent on the inherent capacities of an Internet Service Provider's (ISP) infrastructure in a given area. This is the difference between being able to video chat in high quality with a group of friends and struggling to download a song due to the wide differences in speed and bandwidth across a country. In response to the massive influx of users, user hardware like routers and switches have struggled with accommodating more users. And servers themselves strain under the weight of an unanticipated number of users engaged on these platforms. While the infrastructure of the internet was built to deal with these high-capacity scenarios, artifacts of this effort can be experienced daily.

These issues lie beyond what the user is typically in control of but determine the nature of audio. Bandwidth, throughput, latency, jitter, and synchronicity for telecommunication platforms can pose issues for conventional telecommunications. Bandwidth is the theoretical maximum capacity of a given network, while throughput is the realistic amount of data that can be transmitted in a given window of time. Both vary widely by telecommunications infrastructure available in one's area and the router and physical obstructions that exist. Also affected is latency is the time it takes for one packet of data to the destination. This is experienced in the often perceptible delay in communication experienced in quick conversational exchanges in telecommunications systems. Jitter is the irregularity in transmitting packets, often a result of network congestion, and can lead to intermittent delays in transmission with short drops in telecommunications. From the infrastructure level to the consumer hardware and software available in a user's market, there are many hurdles on the road to ideal conditions for digital interactions.

While many outstanding issues ought to be considered as teleconferencing platforms move towards adapting to natural human interactions, this thesis focuses specifically on the localization of sound and its effects. However, while this work will not directly concern any issues resulting from the infrastructure of the internet or consumer hardware, there are suggestions from the literature that spatial audio may lessen the impact these negative artifacts have on a system. Enabling spatial localization of sound sources involves direct manipulation of the local audio stream of each participant's voice. Yet, even this implementation in audio processing at a local level may require evaluations into capabilities of user hardware for more complex processing, how additional latency may be added and audio/visual desyncing may appear. Evaluations of the system in respect to these measures will take place after describing the implementation. Next, a history of telematics and its intersections with communication, telepresence, and spatial audio will be discussed.

2. Literature Review

To understand and motivate the direction of this thesis, accounts of telematics must be taken from investigations of teleconferencing regarding measures of telepresence, intelligibility, and user experience. Within the literature, it will be useful to examine how evaluations of teleconferencing emerged from the 1970s prior to commercial adoption. These early studies provide perspectives on telecommunication from a conceptual perspective and how audio might serve an essential role within this medium. There will be a critical focus on participants' experiences with spatialized audio from within audio conferencing. Moving onto video conferencing, user's experiences will be explored within the platform, with a critical eye towards sounds. Attempts at introducing spatial audio to video conferencing platforms will be reviewed. These papers discuss the potential benefits that spatial audio conferencing provides to user opinions, cognitive load, social presence, intelligibility, and other metrics of user evaluation. Many of these novel systems designed to integrate spatial features to a conference do so acoustically as well as binaurally. This investigation of the literature will set the stage to look outside in the commercial sphere and eventually help stimulate design decisions for a prototype. To begin, it is useful to describe the origin of the word telepresence and what it means in reference to digital presence.

2.1. Themes and measures

Marvin Minsky, a professor at MIT within cognition and artificial intelligence first coined the term in 1980 in an article submitted to the magazine Omni (Minsky, 1980). Minsky describes a future where a person may operate a remote machine through the motor control and sensory feedback from their hands and a sensor-laden jacket. This future he describes would allow more efficient manufacturing distribution, reduced costs in time and labor, and safety. Though the advances Minsky has in mind were oriented towards the physical mirroring of objects, the idea that a person can impart one's physical or sensory presence, in a shared space, finds strong footing in the ideals of digital communication tools. Telepresence, in this context, will be used to describe the goal of conveying one's physical presence across a video communication system through the integration of spatial audio. Further discussions by Buxton, explore these shared spaces that are enabled with digital tools (Buxton, 1992). They split these locales into the shared person and task spaces, where task spaces allow for interaction and observation of a shared item of interest. In most cases for video conferencing, this takes the place of "screen sharing". But for person spaces, the sensory information we are relayed in group discussions, through cameras and microphones, captures only a fraction of the dynamics and sensory experience one has during in-person interactions.

In the quest for a better video conferencing experience, telepresence serves as a roadmap to bring the experience of communication on a digital platform as close as possible to in-person communication. Researchers take a variety of perspectives on systematizing communication and, as a result, a myriad of unique performance and evaluation metrics will appear in this survey: comprehension, memory, cognitive load, intelligibility, focal assurance, social presence, and mean opinion scores (MOS). It should be noted that there are many factors that will be out of reach in comparisons between the two scenarios. For example, there is no way

of replicating the dynamic depth of a user through a flat display monitor except through techniques in augmented or virtual reality. There is more progress to be made before considering the futuristic world of Minsky's imagination. Even now, video conferencing at scale is a challenging endeavor and is mired by internet infrastructure and consumer hardware. Looking into the academic literature provides a lens into experimental methods that would otherwise be challenging to replicate from a consumer perspective.

In addition to telepresence and telematics generally, social presence provides a reflection on the state of participants in relationship to one another. Outside of digital experiences, social presence has had a multi-tiered history of theories (Argyle and Dean, 1965; Wiener and Mehrabian, 1968; Short et al., 1976). However, only in the two decades has social presence had to be reconsidered in the frame of digital interactions. Gunawardena and Zittle discuss how social presence in digital, text-based conference environments are a predictor of learner satisfaction and enhanced socio-emotional experiences (Gunawardena and Zittle, 1997). This highlights the importance of conveying a sense of presence during computer-mediated interactions yet doesn't clearly take into account how video and audio might further influence one's perception of presence. In a systematic review of the literature, however, richer forms of media like video and audio in communication lead to an enhanced perception of social presence as well as, specifically, increasing the quality of audio (Oh et al., 2018). Social presence ought to be of critical focus in evaluating a teleconferencing system.

The focus in the literature review that follows will concern the role audio plays, both in isolation and interaction with video, in affecting the mentioned metrics of quality of experience. In addition to social presence, of notable interest are cognitive load and intelligibility as they will directly inform the design and evaluation of a novel video conferencing system. Cognitive load or strain is the effort expended in a given task which, in the context of video conferencing, could contribute to mental fatigue. Intelligibility is the ability to clearly understand and comprehend the vocal utterances of an individual. There are potential interaction effects between the two, with lower intelligibility increasing cognitive load, but each examines different facets of the conferencing experience.

It should also be noted that in many of the studies that follow spatial audio is often paired with higher quality audio as a testing condition. Many of these studies group the two conditions together with the aim of widely improving the standards of teleconferencing platforms. There is clear evidence behind the benefit of higher fidelity media. In a revisit of the legacy of the cocktail party effect, Yost notes, "spatialization benefits in discrimination tasks have been shown to increase as signal-to-noise ratios (voice quality) decrease" (Yost, 1997). Or consider research by Ardnt et al., where electroencephalography (EEG) scans of participants watching low bitrate audio and video report higher percentages of alpha waves, which have been correlated to sleepiness, compared to high bitrate media (Arndt et al., 2013). These viewers became more fatigued as a result of low-quality media and rated the low-quality media lower MOS on average (Arndt et al., 2014). While clear evidence supports this movement towards higher fidelity communications, it is not a topic of focus in this thesis.

It will be useful to first discuss spatial audio's origins and the methods that enable the synthetic production of spatial audio over headphones. From here, a discussion of systems that augment audio and video conferencing with various spatial methods will follow, along with their results on a variety of performance and evaluative metrics. These studies will serve to develop the incentive for a novel spatial audio system, explore the range of implementations that have

been tested, and provide a sense of what questions are essential to ask participants who test out such a system.

2.2. Spatial audio

2.2.1. Technical foundations

To understand how it is possible to simulate the location of audio within a virtual space, an explanation of psychoacoustics is useful. The ears make use of multiple cues that appear from the interaction between sound waves and the listener's head. Two essential binaural cues are Interaural Time Difference (ITD) and the Interaural Intensity Difference (IID), described initially as the duplex theory of sound by Lord Rayleigh (Rayleigh, 1907). As sound waves travel to the head they reach the ears at different times, reflecting the difference in distance to each ear. This time difference is prominent at lower frequency ranges, under 1kHz. The intensity of the sound at each ear also varies and this "shadowing" effect, as a result of sound absorption from the head, is more prominent at high-frequency ranges after 2kHz. There is also a monaural spectral cue that is specific to the shape of the listener's outer ear, including the ear canal and pinna. As sound enters a listener's ear, the shape of the cavities, which the pressure waves must travel through, filter the sound in a distinct fashion. These three cues together provide much of the basis for human's ability to localize sound to a high degree of accuracy (Risoud et al., 2018).

There is a wealth of literature that provides the basis for spatialization across various formulations. Physical spatialization resulting from the distance between speakers in an array has been the predominant form of spatialization until work on binaural recordings came into focus in the 1960s (Nordlund, 1962). Then later when head-related transfer functions (HRTF) were devised by and described by Blauert in his book Spatial Hearing (Blauert, 1983). These functions described the filtering effects that can be heard through binaural recordings using a dummy head with microphones placed within replica ears that mimic the monaural and binaural cues that exist (Nordlund, 1963). In hearing this stereo recording with headphones, a listener should hear the recordings in realistic spatial fidelity. However, the synthesis of audio within a virtual space, wherein the audio object can be located through parameters took the development of high-end computing systems to achieve.

In the late '90s, technological progress and interest in spatial sound synthesis made it possible to place sounds in space virtually (Brown and Duda, 1993). This method required binaural recordings of impulse responses, called head-related impulse responses (HRIR) at locations all around the recording head. For ideal coverage, this would be an impulse at every angle around the head. However, this is not physically, nor technically feasible. As a compromise, these impulse responses would then be sent to convolvers for each ear that would process the input signal, in this case a mono signal, within the virtual space that corresponds to the impulse responses from that area. In an ideal world, one would need infinitely many impulse recordings to accurately convolve a signal into that specific location. Instead, a given location takes an interpolation between its nearest HRIR's producing realistic binaural localizations (Vorländer, 2020). This implementation is computationally expensive but even possible within a web browser in mobile devices today.

2.2.2. Cognitive foundations

From a cognitive perspective, the perception of spatial audio is an essential sensory tool that allows us to localize information in space without visual feedback. The well-known Cocktail party effect, described by Collin Cherry, serves as a chief insight into a human's ability to levy their aural perception of space to selectively attend to sonic objects (Cherry, 1953). One prevailing theory of attention that supports this is Kahneman's model of capacity wherein allocation policies determine how one selectively distributes one's available attention (Kahneman, 1973). Considering spatial audio from this angle, one can appreciate how essential this perceptual ability is to daily interactions and especially group conversations.

Indeed, Baldis suggests this phenomenon may happen as a result of the independent processing of working memory between what is known as the Visuo-Spatial Sketch Pad (VSSP) and the phonological loop (Bladis, 2001). The VSSP deals primarily with visual content and its spatial correspondence while the phonological loop processes verbal and auditory information. In this sense, a dry, monaural source produced by headphones relies entirely on the phonological loop as it has no other aural cues that we typically use to locate the sound in space. Sadly, this is what is received at the end of most video conferencing applications and is likely more difficult to process as a result. The reintroduction of spatial information into the source may allow both the VSSP and phonological loop to process the sound as one does with sonic objects in reality. This may be especially useful during double-talk with many participants, the state of concurrent and overlapping speakers.

Binaural lateralization of audio sources with applied noise has been shown to provide significant benefit in speech intelligibility (Ortiz and Orduña-Bustamante, 2015). Ortiz and Orduña-Bustamante's study found that 30-degree angles of lateralization led to an increase of 7% intelligibility as compared to listening at 0 degrees. These listening tests were recorded in a physical space with a binaural model head, but could likely be reproduced with synthetic HRTF processing. This builds upon the claim that, compared to monaural audio, binaural listening offers higher intelligibility at every angle as found in older literature (Nordlund 1962, Nordlund and Lidén 1963, Plomp and Mimpen 1981). This line of research approaches the auditory dimension with multimodality in mind, enabling one's spatial domain to allow information, like speaker identification or word recognition, to be distributed across our sensory capabilities.

2.3. Audio conferencing

Many studies aware of the benefits of spatial audio in intelligibility have implemented this feature within audio conferencing systems with great effect. Studies have found that spatialized audio streams, in many cases paired with high-fidelity audio reproduction, improved subsequent memory and comprehension tasks (Baldis, 2001), increased perceived confidence in remembering topics, and decreased perceived difficulty and attention required in speaker identification (Kilgore et al., 2003), increased audio clarity and social presence (Yankelovich et al., 2006). It has also increased mean opinion scores, improved judgments of speaker recognition, vocal intelligibility, required attention, usefulness of spatial audio (Raake et al., 2010), increased technical quality as well as decreasing the cognitive effort involved in attending to a conversation (Skowronek and Raake, 2015). It will be helpful to describe experiments in which spatial audio was first integrated into audio conferencing environments. These studies offer insights into markers of performance from a cognitive and communication

perspective as well as subjective user evaluations. They provide traction for the exploration of how video conferencing systems might incorporate similar features but with a focus on audio exclusively.

Baldis found that in pre-recorded listening comprehension and speaker identification tasks, a spatial audio setup, using separate loudspeakers, greatly enhanced performance of memory, focal assurance, and perceived comprehension (Baldis, 2001). Participants also preferred spatial audio to mono-aural audio in a follow-up questionnaire. The author suggests that these findings result from an increase in dimensionality, saying the "spatial location provided an additional memory cue that aided in recall, and the presence of spatial audio allowed for more efficient use of working memory." (Baldis, 2001, p. 7) They used two methods of spatialization whereby each of the four voices in the pre-recorded conference was sent through four loudspeakers either with 10 degrees (co-located) of horizontal separation or 40 degrees (scaled). While Baldis hypothesized the scaled condition would lead to greater intelligibility over the co-located condition, through an increase in spatial separation, there was no significant change in comprehension across these conditions, though participants did prefer the scaled over the co-located condition.

One possibility to consider in this study, which the authors mention as well, was that the quality of the recorded speech, with a high signal-to-noise ratio, no environmental noise or compression, was already highly intelligible such that further spatial separation did not provide any benefit. Baldis suggests that these conditions ought to be tested where the audio streams from each speaker may be compressed and vary in audio quality as would occur in a realistic teleconferencing setting. It is interesting, however, that many participants found the co-located condition "unnatural" and significantly preferred a condition where the speakers were spread further apart than would be allowed by a typical computer display.

To follow this study, Kilgore et al. tested similar hypotheses with their system, Vocal Village, using low fidelity (11k 8bit) audio and spatial audio implemented ITD and IID filters over listening tests with four participants equally spaced apart (Kilgore et al., 2003). While not a full binaural simulation with HRTFs, the authors managed to find that spatialization returned higher favorability from participants over mono audio, increased the perceived confidence in remembering conferee viewpoints, and decreased the perceived difficulty and attention needed to identify speakers. However, they found that the participant's actual memory of who said what was not significantly impacted by spatialization in contrast to the study by Baldis. Their study also implemented a condition where participants were able to place conferees along a horizontal axis in a graphical user interface. This led to higher perceived benefits across all metrics.

Yankelovich et al. followed a similar structure with their system, including a similarly configurable GUI layout for conferees (Yankelovich et al., 2006). They measured the impact of the addition of high-fidelity audio and stereo audio on measures of audio clarity, presence within a conference room, and social presence. They found that high-fidelity stereo audio had potential benefits for both clarity and intelligibility of speech and a sense of social presence for conditions mono at 8k, stereo at 8k, and stereo 44.1k. The authors are unable to decisively conclude whether stereo or higher bitrate was the greatest contributor to these improvements as there was no condition for mono at 44.1k. Thus, it is possible that an interaction effect took place between high-fidelity audio and stereo.

Both studies from Kilgore et al. and Yankelovich et al. introduce the idea of coherence between a visual element and the auditory stream coming from a participant. Participants could

have control over the degree to which other users are spatialized and where in space they would be located. This visual representation of space might have facilitated a better coherence between the conference's voices and their perceived origin in space. This would suggest that video conferencing might provide this visuospatial representation upon which spatial audio could be coherent.

Raake et al. reached a similar conclusion with their extensive testing of both listening and group interaction over bandwidth (narrow, wide, full-band) and spatial presentation conditions (diotic, spatial, head-tracking) in a conferencing environment (Raake et al., 2010). Users were able to recognize the benefits from spatial audio and full range bandwidth in both a listening and conversation task. Metrics of MOS, speaker recognition, intelligibility, required attention, and usefulness were significantly improved with the reproduction method. They concluded that spatial reproduction offered more benefits than bandwidth improvements in the listening task with a higher number of speakers, while this was inverted in the conversation task. Additionally, they found that head tracking offered no distinction across any measures compared to the spatial condition.

In two studies, Skowronek and Raake followed up on their findings to test how spatial audio, audio quality, and number of participants influence both cognitive effort and perceived quality of speech while listening to a pre-recorded audio conference. These studies made use of spatial audio via headphones with head tracking that dynamically adjusted audio sources relative to the direction of the head. In the first study, Skowronek and Raake found that conditions with spatial audio and high-quality audio compared to mono, limited bandwidth audio, improved one's perception of the technical quality and reduced cognitive effort when tested on greater numbers of interlocutors (Skowronek and Raake, 2011). The change in the number of interlocutors did present greater effects across all conditions compared to the change in audio condition. One conclusion is that spatial audio may mitigate the effects of the increased cognitive load due to the difficulty in tracking. The following study further supported significantly decreased self-reported measures over all three evaluative metrics, speech communication quality, cognitive load, and general quality of experience (Skowronek and Raake, 2015). This suggests that a spatialized audio environment can dramatically improve the general perception of quality. The work done by Skowronek and Raake provides a take on how audio spatialization might exist in a conferencing call. However, their work examines only audio as a communication stream and within a listening-only context.

Spatial audio also appears to mask issues that may be inherent within the conferencing medium such as packet loss, or the phenomenon of double-talk. These two major issues in audio teleconferencing, packet loss, and double-talk, were investigated by Spur et al. in the context of spatial audio (Spur et al., 2016). Just like in mono-audio environments, packet loss experienced by a single connection was infectious in the quality ratings of other participants. However, during experiments where double-talk was present in the context of binaural spatial audio, mean opinion scores of participants were higher, appearing to mitigate the negative effects of packet loss. This suggests that environments in which participants' audio has been spatialized may provide a higher-quality listening experience during group conversations.

There appear to be clear benefits to various kinds of acoustic and binaural spatial audio in intelligibility, comprehension, and attention within audio conferencing. While a lot can be gleaned from the literature of the treatment of audio in its evaluation over spatial and non-spatial environments, this collection of studies are a starting point for the later integration of video within these systems. Moving onto literature that integrates both spatial audio and video

will be essential in building a history of past experimental platforms to explore the interaction effects of video and audio within both visual and auditory space.

2.4. Video conferencing

Compared to audio telecommunications, real-time video communication requires far more bandwidth and processing power from video stream's inherent bandwidth requirements and the codecs used to compress and decompress video streams from server to receiver respectively. Even with advances in highly efficient video codecs, video is a far more challenging front to advance the aim of telepresence. As a result, systems integrating both audio and video did not meet the commercial market until much later in the mid-2000s with the arrival of commercial software for the end-user. Today, only under ideal connections with fiber connections on both ends can streaming lossless, uncompressed audio and video be attempted. The addition of video provokes new questions about the goal of telepresence within the context of communication such as how we represent multiple participants visually in a group discussion.

Video, on the one hand, can provide key markers of tone and expression from reading another's face, gaze, and the context that surrounds them, but it can also pose significant technical challenges to implement successfully such as audio/video synchronicity and the tradeoff in sharing bandwidth between audio and video. Investigations into video conferencing will be explored with an emphasis on the role that audio and spatial elements play. Studies have found that audio is far more essential than video for task-oriented and pedagogical purposes (Ryan, 1976; Watson and Sasse, 1996) and that video did not provide any additional benefits in conversational fluency compared to audio exclusively though it enhances interpersonal awareness and had higher MOS (Daaly-Jones et al., 1998). Studies that evaluate spatial audio in a video conferencing are few but Sellen et al.'s (1992) Hydra, Nguyen and Canny's (2007) system MultiView, and Inkpen et. al's (2010) system found benefits in double-talk comprehension, trust formation, cooperation frequency, and cooperation resilience. These systems show novel, imaginative approaches to video conferencing and support more enriching communication experiences.

2.4.1. Evaluations of video conferencing systems

In 1976, Ryan looked for subjective differences across the communication modes in aestheticism, evaluation, privacy, potency, and activity by comparing face-to-face video conferencing and audio-only interactions between sets of pairs. They found that both face-to-face interactions and video communication were rated more highly on aestheticism and general evaluation. However, users rated the audio-only channel more potent, suggesting that the audio channel may provide a more capable method of communicating ideas, especially when the conversation is task-oriented. This being one of the first studies to address the effects of video conferencing, there might be an effect of novelty upon participants testing this system. This early study provides some insight into the relationship between audio and video communication and anticipates how burgeoning video telecommunications might be eventually received by the public.

Later during the rise of the first publicly available systems for teleconferencing, Watson and Sasse studied and attempted to establish standards of evaluation for users in multimedia systems (Watson and Sasse, 1996). Many of their insights discuss the inconsistency of

bandwidth over the internet and how software developers ought to prioritize audio and video transmission to compensate. Following this, their field study shows feasibility in teaching a language course over an early video conferencing network. Their study reveals a priority of audio quality over video, without which it would have been impossible to facilitate even the most basic educational instruction over a teleconferencing format. Simultaneous speech was hindered early on in the study, making it the major concern for both students and teachers. Additionally, their studies struggled from the technological standards at the time, which choked video streams at 4-5 frames/second and audio with 10-15% packet loss. Indeed, in working in environments where the repeated repetition of speech is essential, double-talk becomes a major hurdle of intelligibility. Their case study provides real-life cases of the importance of audio from a pedagogical perspective.

In the late nineties, Daly-Jones et al. compiled one of the more comprehensive reviews of the challenges and features of telecommunication at the time, as well as expanding the literature with two studies to find what effects the mode of conversation had on conversational fluency and interpersonal awareness. In two studies, they asked two dyads and two quartets of participants to engage in a fictional application assessment task both over an audio-only stream and audio video stream communication. In anticipation of further advancements in bandwidth and technology, the connections were purely analog, so lossless video and audio reached two local rooms which each set of participants. For these measures, the researchers recorded length of utterance, number of turns speaking, speaking length, vocal and visual backchannels (affirmations), overlapping speech, and explicit questions asked to the other group as a metric of fluency.

In accordance with past evidence, their first experiment with pairs only appeared to increase the number of explicit questions with the addition of a video channel compared to the audio-only condition. In their second experiment with pairs on each side, there was a significant increase in all measures of fluency. In questionnaire responses detailing interpersonal awareness, both experiments did show that participants rated the video condition with enhanced interpersonal awareness, especially when it concerned the attentional state of the conversational recipient. Their research further supports the idea that audio, rather than video, serves as a fundamental channel within telecommunication that enables most communicative interactions to take place between individuals, highlighting how essential audio is in most task-driven conversations. Yet, in groups of more than one speaker, the inclusion of video did improve fluency promoting the idea that engaging with multiple users from one end may take advantage of the visual representation of video to disambiguate communication.

2.4.2. Video conferencing with spatial audio

There have been a handful of prototypes in deploying both spatial audio and video within a conferencing environment that take advantage of the coherence between audio and visual representations of participants, most notably Hydra and Multiview. In 1992, Sellen et al. created Hydra, a 4-way video conferencing system where each of the three conference participants is embodied in a small unit containing a video monitor, microphone, and speaker (Sellen et al., 1992). The Hydra system can be seen in Figure 1. In their preliminary tests, their device allowed users the ability to detect who is paying attention to whom, make eye contact with participants, and benefit in double talk comprehension. In this case, the spatialization was entirely physical, with both the video streams and audio streams appearing from separate physical devices.

Figure 1: An photo of Sellen et al.'s spatial conferencing system Hydra (Sellen et al., 1992)

Later in a review of different mediating technologies, Sellen found that the distributed design of Hydra may better facilitate parallel conversations and assist in following conversational threads even as there were few differences found between audio-only and video conferencing systems (Sellen, 1995). The physical independence of each unit may have added a unique benefit to their spatialization technique to which single display-based representations have no comparison.

Similarly, Nguyen and Canny's system MultiView provided a display that can show projections of the three connected groups each from three angles (taken from three mounted cameras), dependent on a group member's location in respect to their screen (Nguyen and Canny, 2005). A diagram of MultiView can be seen in Figure 2. Spatial faithfulness and realistic gaze representation were of key focus in their system. In an assessment of trust, they found that compared to standard conferencing setups (one camera, one perspective), spatial video conferencing improved trust formation, cooperation frequency, and cooperation resilience (Nguyen and Canny, 2007). Their study additionally provided further support that standard video conferencing layouts hinder this trust formation process compared to face-to-face interactions.

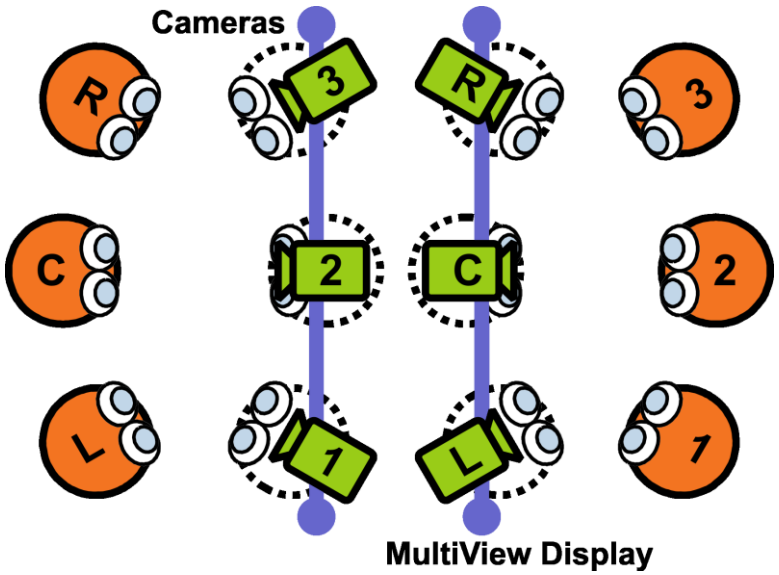


Figure 2: An image of MultiView's gaze preserving conferencing system
(Nguyen and Canny, 2007)

In addition to these systems, Inkpen et al. conducted a study to compare the effects of spatialized audio in group teleconferences with or without spatialized video (Inkpen et al., 2010). Their study involved a split monitor with a participant, a speaker and microphone, and a camera on either side such that participants in a 3-way conversation would be able to perceive one another's gaze. The layout of their system can be seen in Figure 3.

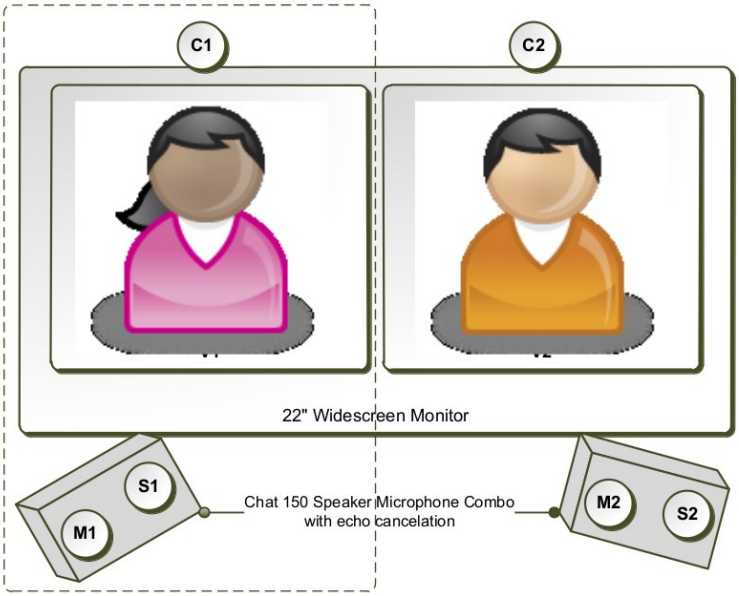


Figure 3: A diagram from Inkpen et al.'s in-house, video conferencing software
(Inkpen et al., 2010)

They found that spatialized video led to higher quality and more engaging conversations as well as a participant's ability to track a conversation. However, the audio conditions, mono versus spatial, did not appear to significantly impact any of the conditions they measured, such as perceived audio quality, conversation quality or engagement, awareness of who is listening, and tracking the conversation, except for improving conversation tracking over mono audio

without video. This study contradicts a number of findings showing that spatial audio leads to conversational benefits. The authors' reason that this may be due to the spatial video system overshadowing any effects of spatial audio. The participants were also familiar with each others' voices and frequently engaged in group meetings which may have diminished the effect of spatial audio. Indeed, the influence of video may overpower the effects found in standard investigations of audio conferencing.

2.5. Summary

There is a wide gap between the richness of in-person interactions and modern conferencing systems as expressed by evaluations of audio and video platforms employing novel techniques in aural spatialization. Concepts of telepresence, telematics, and evaluative metrics of telecommunication systems can provide some guidelines needed to pursue a more natural experience with telecommunications. For many researchers, audio appears to be the essential medium of communication within our faculties and thus one critical to which a more realistic treatment of sound ought to be applied. High-quality synthesis of spatial audio has become mature enough to experimentally test within audio and video conferencing systems with marked success across metrics of cognitive load, comprehension, intelligibility, and social presence.

First, the themes of telepresence, social presence, and the varied metrics of evaluation were described. Spatial audio was then addressed both a technical and psychoacoustic perspective to discuss the possibilities in simulating binaural listening and the ways in which the ear accurately decodes audio information into a spatial coordinate. Finally, a review of the literature in both audio conferencing and video conferencing systems outlines past discoveries of the benefits and effects of communicating within video conferencing systems. Academic literature can offer critical insights into how users communicate with experimental technologies and reveal the space mapped by user evaluations.

However, in the scope of providing a novel implementation, this can only provide a sliver of the platforms that exist to approach the integration of spatial audio within video conferencing applications. Establishing a well-informed methodology that provides the public with a viable application, integrating best practices of current video conferencing platforms with spatial audio, means closely examining what exists today for users in and outside of the market. Therefore, academic literature is not enough to paint an accurate picture of the landscape of video conferencing platforms. Platforms similar to those discussed in this section are further evaluated.

3. Related Works

3.1. Commercial landscape

In the modern day, there are several platforms that offer paid video conferencing solutions at scale. Many of these platforms are developed by large technology companies who have the resources and spare capacity to host these services for free as an incentive to either integrate a user into their ecosystem, harvest user data or have paid plans. As of the time of writing, the largest five such services for desktop appear to be Zoom, Microsoft Teams, Google Meet, Cisco WebEx (Statista, 2020; EmailToolTester, 2021). These commercial platforms share many of the same features and user interfaces such as a gridded display of speakers, noise suppression, echo cancellation, and can host a large number of participants. All of these services require signing up for an account prior to their usage; this is both a barrier to entry and a means of tracking user behavior. As an essential point to the current work, all are closed-source, meaning that their source code is proprietary and is unable to be viewed, modified, or forked. This licensing further prevents any disclosure of which digital languages or technologies are being employed. With no opportunity to modify the application directly, one would be forced to rely upon external programs for routing sound if one was to implement spatial audio. Thus, all commercial programs were out of the scope of development within this project.

However, it is important to consider that these major platforms have established standards for web-based video conferencing in terms of both the backend engineering as well as the frontend user interface and experience. Because of this, it is natural to assume that most of the commercially available products will share many of the same features and design language to remain competitive and relevant as a product for consumers. Consequently, there is no spatial treatment of audio in any of these major commercial platforms. This type of market relationship makes it fiscally risky for the companies developing this software to make changes that might displease customers.

Unfortunately, this further incentivizes other projects to keep to the standards established. As mentioned, some novel projects do exist that attempt to reinvent how spatial teleconferencing might exist, but there is a clear distinction between platforms that intend to serve as general teleconferencing and those that offer a completely novel experience. The gulf between these groups of platforms is wide and there is no clear platform aimed at introducing spatial audio within small, conferencing scenarios. Even still, established and well-funded companies who develop these commonly used platforms are invested in innovation and actively test new features, albeit at a slower pace.

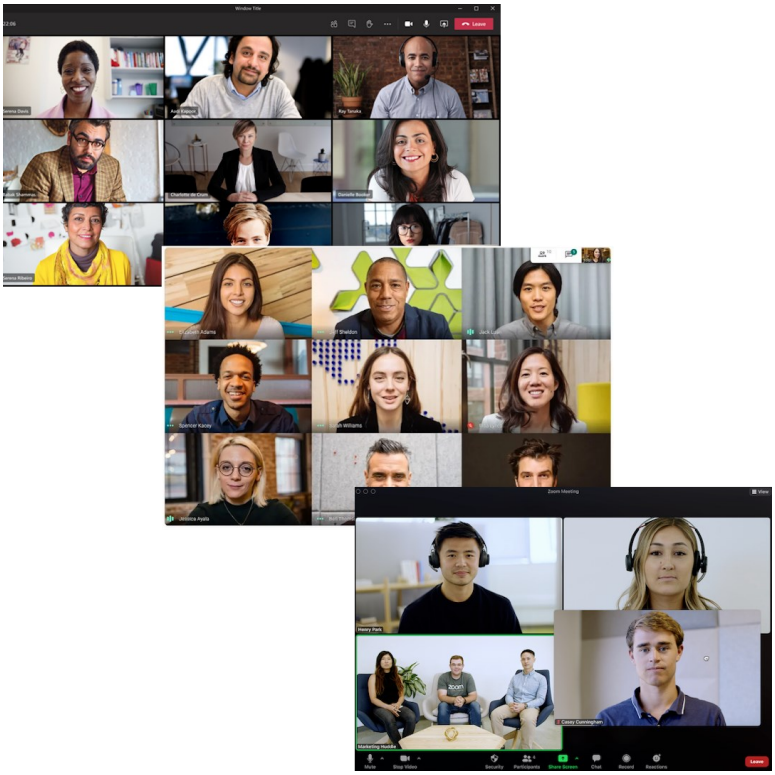


Figure 5: From top to bottom, Microsoft Teams, Google Meet, and Zoom, each following a fairly standardized UI/UX strategy

3.2. Spatial platforms

Commercial and academic software-based projects benefit from institutional resources and capital incentivized markets. In addition to the commonly used, well-funded, platforms, there are a host of notable projects that do attempt to integrate spatialization to an extent and are useful cases to explore for the sake of delineating the space of what is currently being attempted outside of the academic world. There are three major styles of platforms that are implementing spatial audio within their proposed system, 1) platforms incorporating augmented or virtual reality, 2) platforms taking a two-dimensional, gamified environment, and 3) platforms serving spatial audio as a service through an API. While these may not have the same reach and recognition as other well-known products, they represent a market interest in bringing greater realism to telecommunications systems.

The first, and most publicized to some extent, is the development of systems that incorporate Augmented Reality (AR) or Virtual Reality (VR) through the use of dedicated headsets that connect. The most popular platforms with working products dedicated to conferencing are Spatial[1], Glue[2], and vSpatial[3]. These platforms virtualize a user as an avatar within a space with other collaborators, wherein audio is spatialized through a user's headphones via the directional and physical relationship between the users. This approach is incredibly innovative and experiments suggest numerous benefits in social presence and trust (Donsik et al., 2017; Pazour et al., 2018). but struggles today with the issue of expensive, and

[1] https://spatial.io/
[2] https://glue.work/
[3] https://www.vspatial.com/

even often experimental, hardware requirements (AR/VR devices) as well as graphics processing requirements similar to a modern video game.

The second group of developers approached a spatial platform by thinking about how space could be represented without leaving the monitor. The direction that companies like HighFidelity, GroupRoom, SpatialWeb, and TeamFlow take is one that places users in a two-dimensional room and gives control to users to turn and move along the plane to allow separate, breakout groups, to form naturally. Some of these systems include the user's webcam as their 2D avatar while others exclude video streams completely. All of the platforms in this category utilize spatial audio in respect to the direction the avatar is facing on the 2D map. Of interest in this category, Calla is an open-source project based upon Jitsi Meet's source.

Finally, the final group of projects is API-based, meaning they act as routing for audio during conferences for spatialization. Due to the high processing load of spatializing many individual streams, companies like Dolby, with Dolby.io Interactivity[4], DIRAC,[5] Immersitech[6], and HighFidelity[7] offer the ability for developers to hand off the processing to their own servers for spatialization and return the streams to each user. These services may be integrated into existing video conferencing solutions, and as of now, Verizon's conferencing service BlueJeans[8] does implement Dolby's API for spatial audio. This may be the closest commercial implementation to the current project in effect but, again, there is no way to examine the source code or test an implementation of this product without purchasing a subscription.

All of the groups mentioned are novel approaches to teleconferencing that integrate spatialized audio into a new platform for virtual/video spatialization or serve as an endpoint to send audio from an application to be spatialized. Each product makes a strong attempt at challenging the current standards set by its commercial predecessors but often does so through a kind of gamification of a user's presence. These platforms are far from conventional video conferencing and there are no solutions that appear in between. Furthermore, there are no easily accessible applications that would provide a visual experience with the augmentation of spatial audio. This is to say that of the few services that might offer this experience, none are non-commercial, can be publicly demoed, or used in this scope, and none that are non-commercial. As a result, this current work must turn to alternative platforms that satisfy the criteria of accessibility, both from the perspectives of a developer and end-user.

3.3. Open-source platforms

Among services that offer a comparable feature set, there are few alternatives built as free and open-source software (FOSS). For development purposes, only open-source platforms would serve as a framework for this integration. However, not all open-source projects are free. Again, the decision to choose a free alternative was motivated by accessibility and the interest of allowing the project to be freely forked without licensing concerns. If this platform is to be adopted for experimentation or use, the barrier for use, development, and deployment ought to be as low as possible. Additionally, the open-source project must have enough of a community

[4] https://dolby.io/products/interactivity-apis
[5] https://www.dirac.com/spatialaudio
[6] https://immersitech.io/spatial-audio-conferencing/
[7] https://www.highfidelity.com/
[8] https://www.bluejeans.com/

behind it to provide support during the development process. Thus, the criteria for platform selection were development accessibility and available resources.

The only two platforms widely used and developed at scale are Jitsi Meet and BigBlueButton[9]. From a feature-rich perspective, BigBlueButton has many tools such as breakout rooms and presentation features that gear the platform strictly towards education. In contrast, Jitsi Meet[10], developed by 8x8[11], is designed as a more general purpose platform for telecommunication and offers basic features like screen sharing, chat, and recording. Another point between these two projects is their licensing. BigBlueButton is licensed under the GNU Lesser General Public License v3.0 which is a copy-left license and requires that forks of the project carry the same license. Jitsi Meet is under the Apache-2.0 license, a permissive license, which is freer in the sense that it allows the author to choose a different license for forked works. Another FOSS platform that is supported by the EU's Horizon 2020 research and innovation program is eduMEET[12]. It offers feature parity with Jitsi Meet as mentioned in their technical overview[13], but unfortunately does not share the kind of community or reputation as Jitsi Meet.

In this project, the major factor in adopting Jitsi Meet over other open-source alternatives was its vibrant community of developers and the resources they offer on their community forums. Though it is not a direct measure of concurrent development, at the time of writing, Jitsi Meet repository has 15.5k stars on GitHub compared to 6.4k on BigBlueButton's repository. A star is given when a user of GitHub favorites a repository that provides some rough estimate of interest. This may be more of a direct measure of how well known the application is compared to its active development but it is worthwhile in noting its popularity as a platform with significant public interest. Compared to BigBlueButton, Jitsi Meet also has community forums that serve as a resource to ask developers and community members questions if one is forking the project. For these reasons, Jitsi Meet was chosen as the platform from which to integrate a dynamic spatial audio system.

3.4. Summary

Given the scope of this project, the academic research can only tell one side of the story of the latest developments in spatial audio, video conferencing platforms. Commercial platforms can offer insights into the current best practices of video conferencing applications. In the same light, they can also provide evidence for the lack of treatment of audio as a critical medium in communication. Exploring deeper into the market, there are a number of solutions that attempt to address this want of more immersive audio in telecommunications. Many of these platforms depart from the standards found in mainstream applications but are innovative in their approach to representing visual presence. Only a few attempt to integrate spatial audio coherently within an established, accessible video layout, and within this group there are no clear solutions that are open-source. Jitsi Meet appears to be a robust video conferencing application from which to implement spatial audio and one that is accessible from any web browser.

[9] https://bigbluebutton.org/
[10] https://meet.jit.si/
[11] https://www.8x8.com/
[12] https://edumeet.org/
[13] https://edumeet.org/technical-overview/

4. Methods and Implementation

4.1. Framework

Jitsi Meet is a robust and approachable framework to integrate methods for spatial audio as motivated in the prior chapter. There are a number of potential solutions and with them decisions on what may be best for incorporating spatial audio into a platform like Jitsi Meet. Current web browsers are built upon a stack of technology, HTML, CSS, Javascript, that enables a dynamic and unified web browsing experience. Web browsers are capable of quite complex and intensive applications that can deal with media streams provided from a server, or between peers across the Internet with the acceptance of Web Real-Time Communication (WebRTC) standards. WebRTC[14] is an open-source set of API's that enable real-time communication easily from a set of high-level functions. The technology built to accommodate the web acts as one of the most reliable, cross-platform environments to build applications within and thus quite accessible to users across the technical and hardware spectrum.

Unlike applications that have to be downloaded and installed, web applications can be loaded on demand by simply visiting a website. In fact, the ubiquity of applications that web technology is able to provide prompted the development and rise of Chromebooks, a line of affordable computers from Google featuring a slim, Linux-based operating system. These machines are centered around the use of Chrome, a web browser developed by Google. In fact, this idea has become incredibly successful and a huge portion of American schools have adopted these devices (Singer, 2017).

However, many criticisms concerning privacy and data mining have been made to this rollout onto the web (Gebhart, 2017; Petrone, 2018). It shouldn't be understated that, though incredibly accessible, the web is laden with applications that aggressively procure user data, including web-based video conferencing applications. Here again, open-source projects act as a beacon of transparency for web-based software as one can freely audit the source code of a given project. Jitsi Meet collects no analytics in its native deployment from source, avoiding concerns of what might be collected by using a free service. The application also does not require an account, unlike many other major services. From a perspective of accessibility, the web appears to be an environment in which one can develop with maximum outreach and minimal considerations of the user's operating system or hardware. Thus, building upon an open-source, web-based application like Jitsi Meet is an appropriate choice for the implementation of spatial audio presented in this thesis.

4.2. Web Audio API

Within modern JavaScript, the Web Audio API[15] is a high-level interface that allows complex manipulation of audio beyond simple HTML5 audio controls. Web Audio allows for timing control, analysis, buffers, simultaneous playback of sound, as well as real-time processing

[14] https://webrtc.org/
[15] https://www.w3.org/TR/webaudio/

effects. The API is based on a graph schema wherein a series of nodes can be connected to audio sources and one another to apply effects in serial. In particular, Web Audio has a number of native methods that allow panning and spatialization. Interactive websites, applications and even full games can employ this API to create rich, dynamic, sonic environments. Though powerful, the Web Audio API is quite new to the modern web. It was originally proposed in 2011 by the World Wide Web Consortium (W3C) and has only been supported by all major web browsers since late 2013[16]. Even in its current implementation, some browsers have enforced unique methods that must be used as alternatives and particular limitations with media capturing have yet to be addressed. It is a challenging front to unify standards across many different browser engines. However, the audio system offers essential digital audio transformations that enable a more interactive web. The ease of use from the API facilitates a fast development process and was a major reason to employ Web Audio for this work.

4.3. Methods of spatializing sound

The Web Audio API offers a wealth of methods that enable the rendering of audio in complex digital environments and provides two spatialization methods for rendering audio over headsets and speakers. The system contains a fairly standard panning implementation, *StereoPannerNode*, for panning based on the equal power law. Web Audio API also offers a method to easily spatialize binaural audio using HRTFs with its *PannerNode*[17]. This node allows the processing of a mono or stereo source via a set of convolvers who load a dataset of HRIRs from the IRCAM Listen HRTF Database[18]. This database contains 51 subjects from which each HRIR is averaged into a single set. Upon review by Carpentier, the *PannerNode* does make some modifications to the database, such as truncating the samples to half of their original sample length (Carpentier, 2015). Performance was a clear focus of developers when integrating this method into web standards.

Indeed, there are many considerations in working with this implementation. For example, this averaged set of HRTFs does not approximate each individual's ears to the same degree. The unique shape of an individual's ears, head, and even shoulders can influence the perception of sound and attribute to widely different HRTFs if recorded. In the current Web Audio API it is not possible to configure a custom HRTF dataset, though it is proposed in the next version[19]. With the rise of potentially viable methods of calculating one's own HRTFs, this may be a solution for those with poorly tuned binaural experiences (Lee & Kim, 2018; Huttunen et al., 2014). There is also the performance and latency cost of using the *PannerNode* implementation for binaural sound. To include multiple, dynamic audio objects, each source requires its own *PannerNode*. In the context of a video conference, every speaker must have a *PannerNode* associated with their stream. The implications of this in respect to performance are discussed in the following chapter.

In light of this, there are a few alternatives to the standard implementation, such as Resonance Audio[20]. However, these libraries do not appear to be maintained or have notice of

[16] https://caniuse.com/audio-api
[17] https://developer.mozilla.org/en-US/docs/Web/API/PannerNode
[18] http://recherche.ircam.fr/equipes/salles/listen/
[19] https://github.com/WebAudio/web-audio-api-v2/issues/21
[20] https://resonance-audio.github.io/resonance-audio/

further support if the Web Audio methods change in a new revision. One possibility to consider an alternative binaural framework would be if there was a considerable performance or quality advantage. Figure 4 comes from Resonance Audio's developer guide where they detail the performance between their system for web-based ambisonic reproduction compared against the HRTF mode of the *PannerNode* over a large number of sources[21]. Sadly, the methods used to generate this figure are undisclosed.

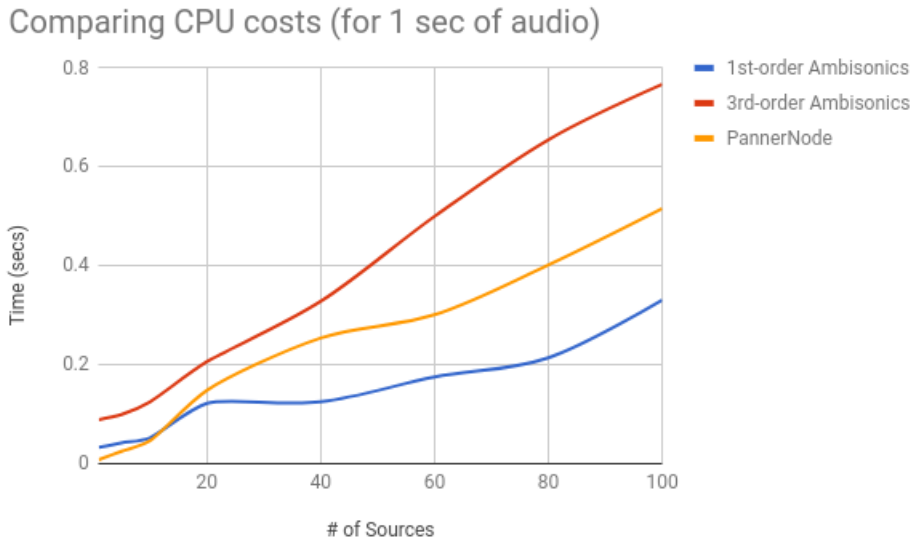


Figure 6: CPU cost against number of sources for Omnitone's ambisonic implementation compared to *PannerNode*

Resonance Audio, with its spatialization engine Omnitone, does not appear to offer any performance benefits at the low number of sonic objects involved in a small distributed conversation. Thus, the standard Web Audio protocol was used in this system both for stability, ease of integration, and the long-term sustainability of the web application. It should be recognized that even Web Audio must make sacrifices for scalability and flexibility in serving devices of various processing capabilities. Hopefully, as both software and hardware continue to improve the API will advance in the fidelity of acoustic reproduction and performance.

4.4. Implementation

4.4.1. Browser compatibility

In developing applications for the web, there are unified standards across all web browsers as to what features ought to be supported and how they ought to be implemented. While there is considerable compatibility across modern web browsers today, there are still several unresolved issues with new libraries like Web Audio. The web-platform-tests (WPT)[22] are a series of automated tests created by the W3C that are run with each new version of the most popular web browsers. These tests evaluate the adherence of browsers to the

[21] https://resonance-audio.github.io/resonance-audio/develop/web/developer-guide
[22] https://wpt.fyi/

specifications of the W3C standards. In the current context, this tool provides useful information on the compatibility of web browsers to current Web Audio standards. Figure X shows a list of compatibility tests between the four major browsers.

Path	Chrome 91 Linux 20.04 e53fe3b Apr 19, 2021	Edge 91 Windows 10.0 e53fe3b Apr 19, 2021	Firefox 89 Linux 20.04 e53fe3b Apr 19, 2021	Safari 123 preview macOS 10.15 e53fe3b Apr 19, 2021
processing-model/	3 / 6	3 / 6	6 / 6	3 / 6
the-analysernode-interface/	177 / 177	177 / 177	177 / 177	173 / 177
the-audiobuffer-interface/	153 / 160	160 / 160	160 / 160	153 / 160
the-audiobuffersourcenode-interface/	647 / 652	647 / 652	613 / 648	647 / 652
the-audiocontext-interface/	95 / 115	95 / 115	93 / 103	98 / 111
the-audionode-interface/	353 / 353	353 / 353	343 / 345	353 / 353
the-audioparam-interface/	1770 / 1770	1770 / 1770	1370 / 1585	1767 / 1770
the-audioworklet-interface/	342 / 353	349 / 357	289 / 321	304 / 336
the-biquadfilternode-interface/	378 / 378	378 / 378	336 / 355	378 / 378
the-channelmergernode-interface/	161 / 164	161 / 164	156 / 160	156 / 159
the-channelsplitternode-interface/	68 / 68	68 / 68	68 / 68	68 / 68
the-constantsourcenode-interface/	111 / 111	111 / 111	111 / 111	111 / 111
the-convolvernode-interface/	346 / 349	346 / 349	341 / 345	335 / 341
the-delaynode-interface/	153 / 154	153 / 154	141 / 144	153 / 154
the-destinationnode-interface/	2 / 2	2 / 2	1 / 2	2 / 2
the-dynamicscompressornode-interface/	78 / 78	78 / 78	78 / 78	78 / 78
the-gainnode-interface/	98 / 98	98 / 98	86 / 89	98 / 98
the-iirfilternode-interface/	199 / 199	199 / 199	195 / 199	199 / 199
the-mediaelementaudiosourcenode-interface/	32 / 32	31 / 32	26 / 32	10 / 18
the-mediastreamaudiodestinationnode-interface/	47 / 47	47 / 47	47 / 47	47 / 47
the-mediastreamaudiosourcenode-interface/	7 / 7	7 / 7	7 / 7	5 / 7
the-offlineaudiocontext-interface/	54 / 57	54 / 57	57 / 57	57 / 57
the-oscillatornode-interface/	122 / 122	122 / 122	105 / 122	122 / 122
the-pannernode-interface/	849 / 849	849 / 849	636 / 670	849 / 849
the-periodicwave-interface/	32 / 32	32 / 32	32 / 32	32 / 32
the-scriptprocessornode-interface/	10 / 10	10 / 10	6 / 10	10 / 10
the-stereopanner-interface/	125 / 125	125 / 125	91 / 100	125 / 125
the-waveshapernode-interface/	145 / 145	145 / 145	134 / 142	145 / 145

Figure 7: Comparison of supported Web Audio functionality within popular browsers

It's useful to note that Microsoft's Edge browser, like many others, is based on the open-source browser Chromium, the same browser that Chrome is flavored from. It is clear that Firefox fails to complete more tests compared to either Chrome or Safari as a general metric of compatibility. Additionally, Spotify's Web Audio Bench[23] is a utility designed to test the performance of each component of the Web Audio API. It provides clear evidence that for the majority of tests, Chrome performs better across any OS and especially in the *PannerNode's* HRTF mode. It also reveals that the *PannerNode* is one of the most computationally taxing components in the API. As a result, Chrome was selected as the browser that development would take place on so as to minimize issues with compatibility. From the perspective of an eventual user study, Chrome also takes the lead in popularity in desktop browser usage[24] making it likely more convenient for users to participate without requiring an extraneous install.

4.4.2. Localization design

As reviewed in the literature, spatialization through lateralization improves intelligibility and spatial representation of both sound and video so design choices were made to reflect this. While not acoustic, binaural spatialization can realistically represent sound coming from different origins in space. Because of human's higher native acuity in lateral sound detection compared to vertical, it was decided to use only lateralization to differentiate the speaker's audio streams. This also involved changing the layout of Jitsi Meet's default speaker

[23] https://github.com/spotify/web-audio-bench
[24] https://en.wikipedia.org/wiki/Usage_share_of_web_browsers

tiling to force a horizontal array of speakers rather than its default 2x2 tile arrangement. As a result, the visual size of the participant's video stream is reduced to fit the horizontal span of the browser window. Given that the ratio of most computer monitors are 16:9, this trade-off makes sense. Incorporating elevation to the visual and aural representation of the participants would cluster the participants closer along the azimuth dimension and likely make it far more difficult to differentiate between the participant's voices. A screenshot from a demo of three phony participants can be seen in Figure 8. Clicking on the figure will link to the full video.

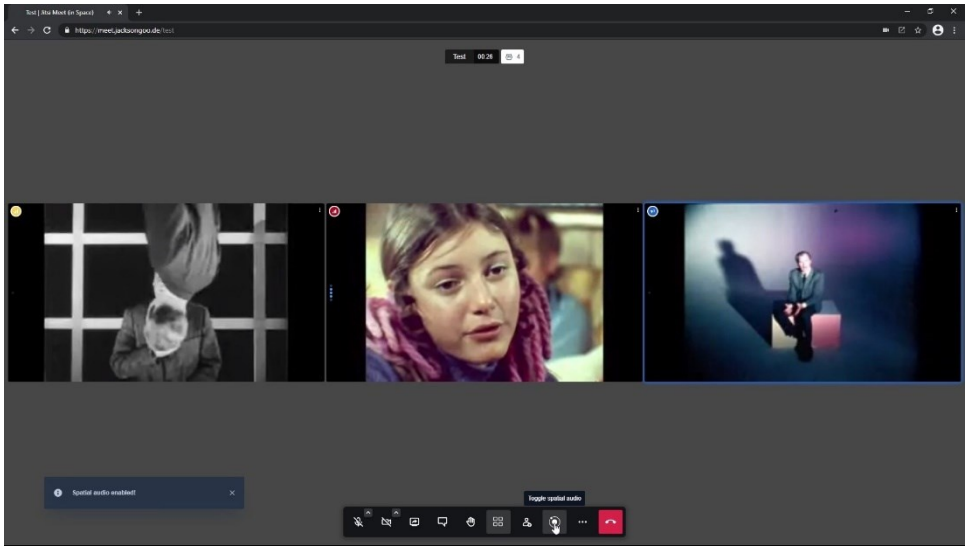


Figure 8: Screenshot of a video demo of the system[25]

In addition, a decision had to be made as to the degree of separation between each speaker in space. Following Kahneman's theory of capacity, a highly complex attentional space captures more attentional resources. Disentangling the sonic sources then suggests a less taxing local environment for each voice, allowing a higher number of attentional resources to be dedicated to each voice selectively. Furthermore, work by Divenyi and Oliver suggests that a minimum angle of 60° ought to exist between two sound sources for them to be localized reliably (Divenyi and Oliver, 1989). These considerations as well as results from Baldis, where participants preferred voices spaced further than the corresponding visual representation on screen, led to the development of a simple algorithm that equally distributed distance between speakers. Figure 9 shows the layout for arrangements of 3, 4, and 5 total participants in respect to the user or listener.

[25] https://vimeo.com/548286337

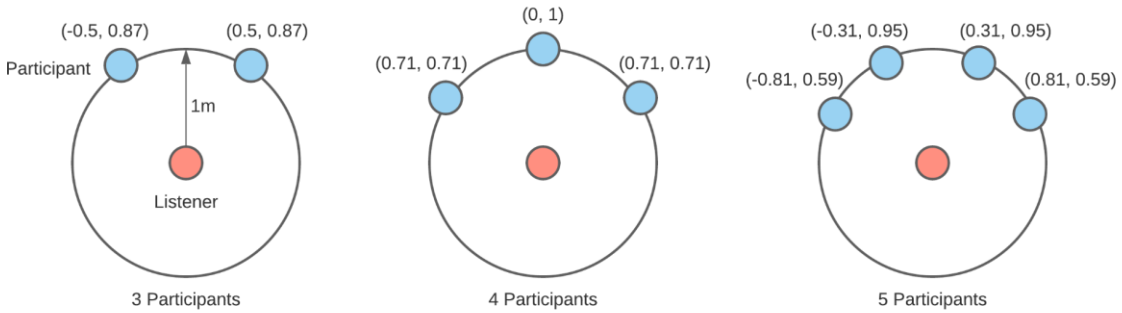


Figure 9: Arrangement of participants in virtual space with respect to user (coordinates in meters)

Arrangements in virtual space were created such that each participant is equidistant from the listening user in virtual space. While the video grid does not incorporate the simulated depth of the audio streams, it is a tradeoff to provide full visibility to the other participants' streams. These arrangements respond dynamically to the entrance or exit of participants where the location specified by the *PannerNode* for each participant's voice adjusts accordingly.

The technical operation of passing vocal streams and linking them within a Web Audio graph is straightforward and can be seen in Figure 10.

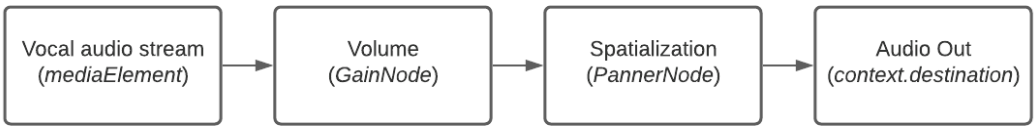


Figure 10: Signal chain for a single audio stream

Audio streams that spawn when a participant connects are first converted into sources via *createMediaStreamSource* or *createMediaElementSource*, depending on browser. These sources can then be connected to a *GainNode* which is in turn connected to a configured *PannerNode* and sent to the audio output. Upon each audio track's appearance, when a participant connects, an initial routine configures the required parameters of the *PannerNode*. This is followed by an update procedure that considers both the number of participants and the relative position of the participant's video as an index. The general schema is described in Figure 11.

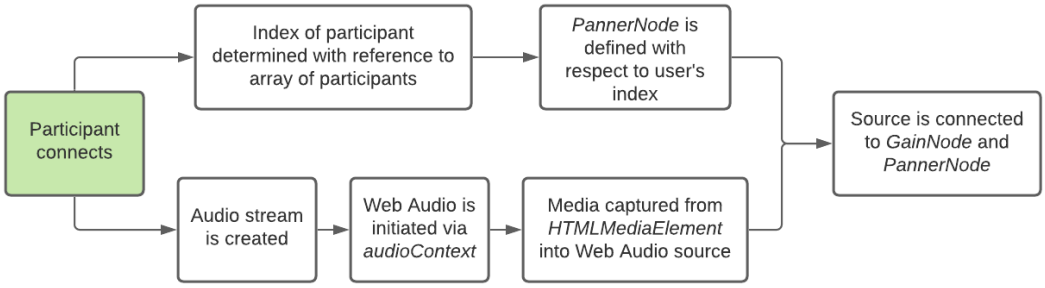


Figure 11: Overview of processes when a user arrives

This update procedure also happens whenever there is a change to the number of participants or the index of the video stream. When a participant disconnects from the conference room, for example, all remaining audio tracks will update their *PannerNode*'s as a result of the number of participants changing. An outline of this is presented in Figure 12. In effect, the audio locations will be properly brought in accordance with the new visual layout. The resources allocated to the audio graph are also released when this occurs.

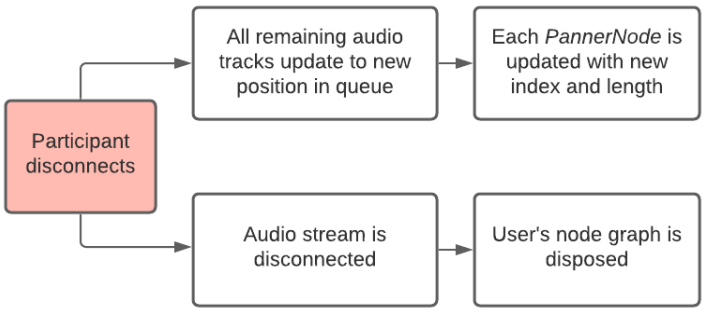


Figure 12: Overview of the processes that occur when a participant disconnects

While most modern computers can handle small quantities of these *PannerNodes*, the computation of each additional object scales linearly, meaning spatializing ten or more participants may be a significant burden on a machines' CPU utilization. Some metrics of Web Audio performance have been recorded by development teams at Spotify, Mozilla, and the Google-developed SDK, Resonance Audio, which includes their ambisonics implementation Omnitone as previously mentioned. The interpolation between HRTFs through two convolvers is computationally expensive and will inherently add some latency via the convolvers' audio block size compared to the original audio stream. Thus, this system was designed and tested to only accommodate up to five users.

4.5. System evaluation

From a theoretical point of view, the integration of real-time, sonic renderings within Jitsi Meet will increase the CPU utilization of the web application and add a degree of latency, inherently from the requirements of spatialization. Still, there may be issues that are unaccounted for. Synchronization and performance tests were made with the limits of the application's current implementation in mind. All tests were performed on the latest Chromium browser (90.0). The server was located quite a distance away (Oslo, Norway) from the testing location (San Diego, California, USA) likely leading to greater latency and higher potential interruption during the connection. The server running the instance of Jitsi Meet (Ubuntu 20.04.2, Intel (Haswell, no TSX) (4) @ 2.399GHz, Cirrus Logic GD 5446, 16GB memory) was granted by the Norwegian Research and Education Cloud (NREC)[26], appropriate for serving the application under modest conditions and never appeared to struggle with any of the benchmarking tests. Tests were run across the proposed system in spatial audio, non-spatial

[26] https://www.nrec.no/

audio, as well as the standard Jitsi Meet instance[27]. This was to compare performance across all systems.

A synchronization test was performed to check whether the additional latency added by the spatialization technique offset the audio from the video to a significant degree. One computer broadcasts a video file used to test audio-video synchronization as a virtual webcam and microphone to a conference room on the web application. A simple tone was played with a flashing visual cue every second. Another computer connected to the same meeting room on the browser within the same wireless network. Once a sample of the sender's audio and video were recorded from the receiving computer's screen and audio device, the resulting video file was brought into a video editor to log the delay between the visual cue and the auditory tone frame by frame. In a sample recording where a tone was played every second, an average of 30 latency samples presented the signal arriving early in every case. Results can be seen in Table 1.

System	Average early arrival
Proposed system (spatial audio)	0.085ms
Proposed system (non-spatial audio)	0.099ms
meet.jit.si	0.066ms

Table 1: Comparisons of the synchronicity of audio and video between the proposed system and Jitsi Meet's official instance

This same test was then performed on the original audio engine and the delay was nearly identical with a ~14ms difference that is likely to account for the latency inherent in the methods of rendering a sound object from a virtual distance for binaural listening. In making a comparison with the standard Jitsi Meet deployment, it should be noted that the servers hosting the Jitsi Meet instance are based in the US, so there may be other unrecognized issues with the network and bandwidth exchange in comparison to serving from Norway. All audio was seen to arrive earlier than the video stream which appears arbitrary from the variance seen over the development period.

The second test analyzed the impact that the spatialization contributes to the performance of the web application and whether it might pose noticeable performance impairments in analyzing up to three spatialized participants. Each computer streamed a video clip of a presentation as a virtual web camera through OBS Studio[28] with the audio of the clip piped as the virtual microphone. A single computer (Ubuntu 20.04.2, Intel i5-10210U (8) @ 1.600GHz, Intel UHD Graphics, 16GB memory) ran the Unix command *top* to capture the CPU and percentage of memory used by the Chromium tab and audio service processes. In Chromium, tasks are spread across multiple processes for network activity, storage, audio, and each tab. It is therefore possible to get some rough estimation of how these components are utilizing the CPU individually. 200 samples of CPU utilization were taken at a 500ms interval. The results for both tab and audio processes are detailed in Table 2 and Table 3.

[27] https://meet.jit.si/
[28] https://obsproject.com/

Tab process	Proposed system		meet.jit.si	
Number of participants	CPU (%)	Memory (%)	CPU (%)	Memory (%)
0	57.354	1.99	67.571	1.634
1	201.808	2.411	294.142	2.246
2	206.034	2.374	268.385	2.083
3	207.032	2.378	318.083	2.121

Table 2: Tab process (Chromium) utilization compared between the proposed system and the standard meet.ji.si instance across number of participants

Audio service	Proposed system		meet.jit.si	
Number of participants	CPU (%)	Memory (%)	CPU (%)	Memory (%)
0	3.298	0.7	7.174	0.7
1	5.453	0.704	9.754	0.704
2	5.238	0.704	9.692	0.704
3	5.25	0.704	9.505	0.704

Table 3: Audio service process (Chromium) utilization compared between the proposed system and the standard meet.ji.si instance across number of participants

It appears that the proposed implementation did not increase the CPU utilization to any significant degree with an increase in the number of participants even as the CPU readings themselves suffer from very high variance. Indeed, it is difficult to record CPU readings, let alone plausibly attribute their readings to the tasks performed in a tab. While all other factors were isolated, it is challenging to make a clear comparison to the generic Jitsi Meet instance as there may be logging or other modifications that are causing the increase in CPU utilization compared to the proposed system. The deployment at *meet.jit.si* is likely to have some considerable customizations for their product, but it is unclear why the CPU increase was so significant across both increases in participants in tab process recordings and baseline audio service recordings.

One other metric to consider is what is called the Render Capacity in the DevTool's Web Audio tab in Chromium. This is the time spent in the rendering of audio divided by instantaneous callback interval times 100. This percentage can be viewed per audio context and appears to linearly increase by around 10% for every participant that joins the meeting. Unfortunately, there is no method of logging and collecting this data to get an accurate average. Comparatively, the standard Jitsi Meet application only employs Web Audio for a voice level meter, so Render Capacity is negligible (under 1%). On the computer used for recording tests, the render capacity was seen to occasionally spike above 80% with 4 participants' audio streams being rendered. This does indicate how taxing this kind of spatialization is on even modern consumer hardware and software and how the capacity of the audio buffer might need to be potentially increased if performance issues occur or lower-end hardware or higher numbers of participants.

4.6. Summary

Introducing a novel platform for spatial audio requires not only an understanding of previous literature and related commercial applications but also an exploration of the implementation space available, justification for the technical decisions made, and an evaluation of the system. In reviewing the available possibilities for high-fidelity rendering of spatial audio, the Web Audio API can provide a method that works natively, without dependencies, on the web in modern browsers. From the perspective of development, Chrome appears as the more performant and capable browser at the moment for working taxing audio Web Audio systems. The algorithm to lateralize participants places the voices equidistant from the listener with maximum distance between each participant and absolute left and right, inspired by previous research on the benefits of lateralization. The algorithm is dynamic in adjusting to the arrival and departure of participants and has been tested to serve up to 4 participants (not including the user as listener) without audio artifacts or any negative experiences with performance.

The code used to host the deployment of Jitsi Meet with the integrations is hosted publicly on both GitLab[29] and GitHub[30]. The project is licensed under the Apache-2.0 scheme and thereby encourages further open-source publishing. Documentation will also be included to instruct a user on getting the code integrated into their deployment of Jitsi Meet.

[29] https://gitlab.com/jacksongoode/meet-in-space
[30] https://github.com/jacksongoode/meet-in-space

5. User Study

5.1. Methods

5.1.1. Overview

A user study was designed to evaluate the system and explore three concepts that have been critical to the research in spatial teleconferencing: cognitive load, feelings of social presence, and intelligibility in distributed conversations. These themes remain salient both from the perspective of telepresence as well as contemporary struggles with current video conferencing applications. Since COVID-19 has severely prevented the ability for formal, systematic testing in controlled conditions, the experiment will share some features of naturalistic studies and ask participants to use their own machines and headphones with the application. However, compared to most of the research reviewed, this study will provide a stronger impression of how spatial audio might be received "in the wild" as a robust platform deployed on an individual's machine through the web. This will include the possibility of various degrees of quality between the connection of the participants and the server.

To stimulate discussion from the participants, a similar methodology to Inkpen et al.'s discussion-based study was used (Inkpen, 2010). They promote the idea of using free conversations centered around current events or controversial topics. A list of topics was chosen, along with sample questions that were suggested as starting points for discussion. The content of the discussion itself was inconsequential to the study, in as much as it provided a topic to speak of candidly. Survey themes were inspired from studies by Yankelovich et al. (2006), Baldis (2001), and Skowronek (2011) with 3 questions allocated to each cognitive load, social presence, vocal intelligibility, as well as opinions on the quality of the system.

Drawing from the literature, there are four hypotheses that correspond with the three themes asked about on the questionnaire:

> **H1:** In the spatial audio condition, there will be a decrease in the perceived cognitive effort invested during the conference.
>
> **H2:** In the spatial audio condition, there will be an increase in the perceived social presence of other participants.
>
> **H3:** In the spatial audio condition, there will be an increase in the perceived intelligibility and clarity of other participants.
>
> **H4:** In the spatial audio condition, there will be an increase in the perceived quality of the conference.

These hypotheses address the three themes within the literature of similar studies, in addition to evaluating the mean opinions of the quality of the conference.

5.1.2. Participants

Four trials were conducted with each consisting of a group of four participants. Two groups were recruited from students of different intakes of the Music, Communications and Technology master's program[31] and two recruited using snowball sampling, all of whom have been in regular contact using video conferencing applications. Thus, all participants were deeply experienced with video conferencing through their education, employment, or social circles. The participants from the master's program regularly engaged in a cross-campus education over the last year across both high-fidelity and low-latency systems as well as consumer platforms. All participants were familiar with each group member's vocal timbre and could be assumed acquaintances.

5.2. Setup

Due to the COVID-19 epidemic and the remote location of the author there was no viable solution to having participants use similar equipment for this user study. Instead, all participants used computers, cameras, microphones, and headphones within their possession. They were asked to use high-quality, over-the-ear headphones both for sound ambient sound reduction and binaural fidelity and to have their machines connected to a power source for performance optimization and battery concerns. However, two participants were not in possession of over-the-ear headphones and instead used in-ear monitors. For accessing the website, participants were asked to use an up-to-date, web browser based on Chromium for unified compatibility and performance for the web conferencing application. The computer specifications were recorded in the final survey if performance issues appeared in any of the feedback or if there was any trouble loading the website.

5.3. Protocol

A two-factor, repeated measures design was employed with the condition of audio reproduction, either as binaural spatialized audio and monaural audio. Condition order was counterbalanced across the four groups such that two groups received spatial audio first and two groups mono first.

Each group was briefed over in a private group on the telecommunications application Discord[32], with which all participants were familiar. Instructions were posted in a text channel and prior to the study, instructions were reiterated in a voice channel and any questions were received. Then participants were asked to leave the voice channel and visit the meeting room of the deployed conferencing software following a URL posted in the text channel. Participants were instructed to use a Chromium-derived web browser in incognito mode to prevent interference from addons. Upon arriving at the website, the participants were asked to ensure that their video and audio devices were functioning properly. The full instructions provided to participants can be read below:

[31] https://www.uio.no/english/studies/programmes/mct-master/
[32] https://discord.com/

For this evaluation, you will be asked to visit a website from the Chrome browser (or a browser based on Chromium) that hosts a video conferencing application. Please make sure you are wearing headphones (over-ear preferred) and that your computer is plugged into a power source.

Upon arrival, please confirm that your video and audio input devices are functioning. To make you have the latest version of the webpage, please reload the page directly by pressing Ctrl+Shift+R or Command+Shift+R (you may also right click on the Refresh button and select "Hard Reload").

You will meet the researcher to confirm the system is properly running on your computer, be informed of the instructions of the experiment and answer any questions. Once ready, the researcher will turn off their video and audio to observe the discussion to take notes as the conference will not be recorded in any fashion.

You will be asked to have a discussion of a series of current events for 5 minutes after which you will be asked to fill out a short, anonymous survey. The conversation should be casual and candid - there will be no evaluation of the content of the discussion. This will be repeated once more, followed by a second survey. The condition tested will be an implementation of spatial audio for distributed conversations - a method of spatially localizing each participant's voice.

The researcher confirmed that everyone was able to see and hear one another and explained that they would be hidden to take notes on the discussion and that they would have 5 minutes to discuss the topics provided. Three new topics were presented to the participants as possible topics of discussion before each of the two sessions and that they were in no way required to address all of the topics or questions. The topics and sample questions are included in Table 4.

Condition	Topic	Sample question posed
Condition 1	Climate change	Are Western nations responsible for leading an effort to reduce climate change?
	Cryptocurrency	What is the future of cryptocurrencies?
	Universal basic income (UBI)	Is UBI feasible and what would be its benefits or consequences?
Condition 2	Social networks	To what extent should privacy exist in today's social networks?
	COVID-19	How will COVID-19 have changed society in the long term (in a post-pandemic world)?
	Drug legalization	How should countries handle the legalization of marijuana?

Table 4: Topics and sample questions posed to the participants over two conditions

Participants were asked to complete two, anonymous surveys after each short discussion. The surveys asked participants questions concerning the expended effort during the conference, the clarity and intelligibility of each participant's voice, and their sense of social and

physical presence to the other participants, and their opinions on the quality of the conference. Three questions were asked in the second condition's survey for the frequency of video conferencing use in their daily lives, their evaluation of video conferencing, and the machines with which they were using to join the meeting. No audio-video information or other personal data was recorded during the sessions. All questions are presented in Table 5 below.

Metric	**Survey question posed**
Cognitive load	The effort required to determine which participant was speaking was: *Very difficult - Very easy*
	The effort required to follow the speaker was: *Very difficult - Very easy*
	The effort required to follow the overall conference was: *Very difficult - Very easy*
Social presence	To what degree did a participant's voice correspond to their video: *No correspondence - Complete correspondence*
	How close did other participants appear to be to you: *Nearby - Far away*
	How comfortable were you in participating in the conference: *Not comfortable at all - Very comfortable*
Vocal intelligibility	How clear was the speech from other participants: *Not clear at all - Very clear*
	How often did overlapping voices interrupt the conference: *Never - Always**
	How confident were you in understanding what other participants had said: *Not confident at all - Very confident*
Preferences and opinion scores	The locations of the participants' voices was: *Not helpful at all - Very helpful*
	The audio quality of the conference was: *Poor - Excellent*
	The overall quality of the conference experience was: *Poor - Excellent*
Comments	Any additional comments regarding the conference experience?

* This question's scale was incorrectly inverted. It should have read *Always - Never*.

Table 5: Survey questions posed to participants at the end of each discussion period. All non-explicit response options were arranged as 5-point Likert scales.

5.4. Results

A Wilcoxon signed-rank test was performed on each of the survey questions (Siegel, 1956). This test is chosen as the data analysis method for its designed use with ordinal data like the 5-point Likert scales employed for the surveys. Since the sample size tested (n = 16) was less than 30, results may depart from normality which the Wilcoxon test can accommodate while using an exact test. While it serves as a non-parametric alternative for a t-test for paired samples, a two-tailed paired t-test is also included for comparison. All 12 questions are prefixed with their corresponding category, *C* = cognitive load, *P* = social presence, *I* = intelligibility, and *O* = opinion scores. A distribution of the average question score over each topic can be seen in Figure 12. The results of the paired t-test and Wilcoxon Signed Rank test are listed in Table 6.

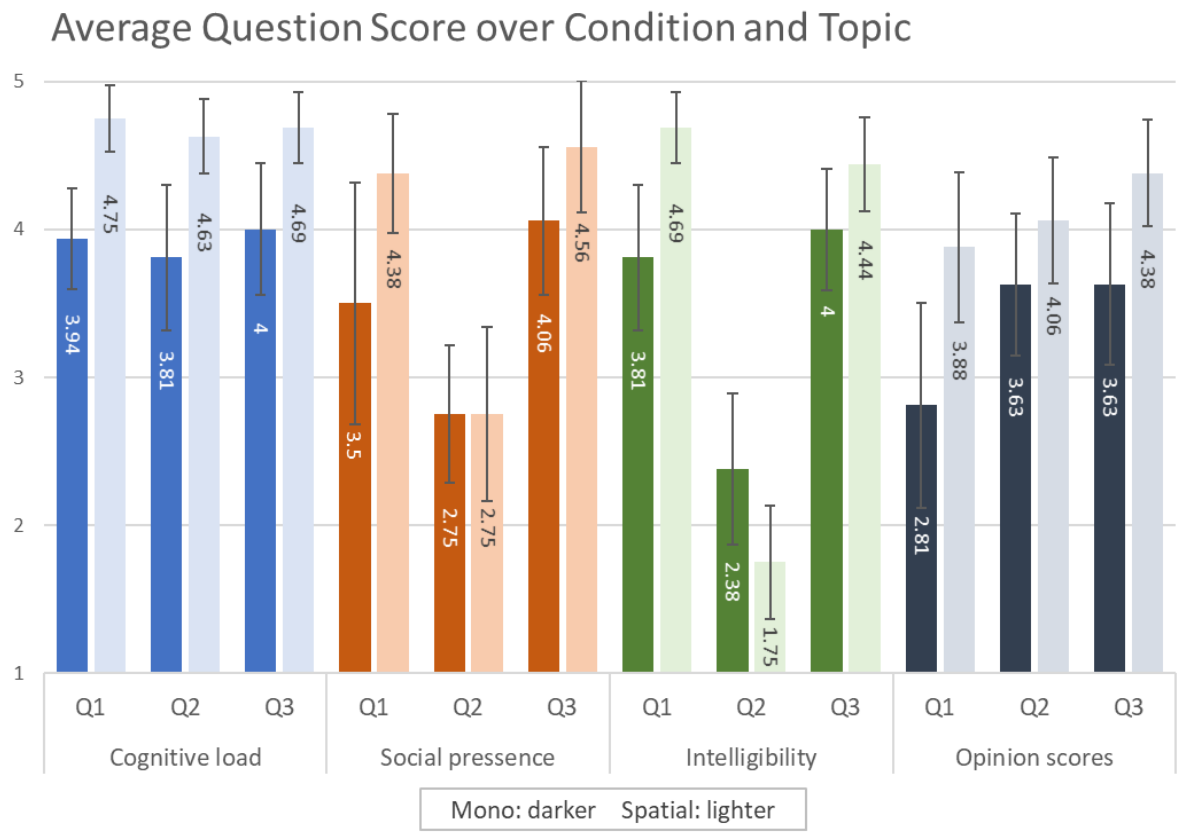


Figure 12: Average question score over audio reproduction method, grouped by topic. Standard deviation bars are included.

Table 6 provides the average scores reported for both mono and spatial conditions as well as their standard deviation, written as SD, for each condition. The p-values of a paired t-test are included. For the Wilcoxon test, both the positive and negative sum of signed ranks (W+ and W-) as well as the *T* statistic as the minimum value between the sums of signed ranks. A critical value of *T* for a given sample size and tail can be looked up in an associated table[33]. For the t-test, rejection of the null hypothesis was set at $p < 0.05$. Wilcoxon significance is determined by checking if *T* is less than a critical value, T_{crit}, for $\alpha = 0.05$, where n = 16 and the hypothesis is two-tailed.

[33] https://www.oreilly.com/library/view/nonparametric-statistics-a/9781118840429/bapp02.xhtml

	Cognitive load			Social presence			Intelligibility			Opinion scores		
	C-Q1	C-Q2	C-Q3	P-Q1	P-Q2	P-Q3	I-Q1	I-Q2	I-Q3	O-Q1	O-Q2	O-Q3
Mean (Mono)	3.94	3.81	4.00	3.50	2.75	4.06	3.81	2.38	4.00	2.81	3.63	3.63
Mean (Spatial)	4.75	4.63	4.69	4.38	2.75	4.56	4.69	1.75	4.44	3.88	4.06	4.38
SD (Mono)	0.68	0.98	0.89	1.63	0.93	1.00	0.98	1.02	0.82	1.38	0.96	1.09
SD (Spatial)	0.45	0.50	0.48	0.81	1.18	0.89	0.48	0.77	0.63	1.02	0.85	0.72
p-value for paired t-test	0.001*	0.001*	0.011*	0.039*	1.000	0.027*	0.006*	0.076	0.089	0.012*	0.168	0.009*
W+	45	45	42	39	25	15	52	14	51	79	27.5	42.5
W-	0	0	3	6	30	0	3	52	15	12	8.5	2.5
T	0	0	3	6	25	0	3	14	15	12	8.5	2.5
Wilcoxon significance ($T < T_{crit} = 29$)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* $p < 0.05$

Table 6: Statistics related to data collected from survey questions. Results of a paired t-test and Wilcoxon Signed Rank test are included.

There is strong support for **H1** with a significant effect found from both the t-test and Wilcoxon test for each of the three questions related to perceived effort. This suggests that the effort involved in determining which participant was speaking and following the speaker and the conference as a whole decreased during the spatial audio condition.

There is partial support for **H2** with significant effects found for two questions related to social presence through a paired t-test and for all three questions with the Wilcoxon test. One question (P-Q2) within the questions on social presence question did not appear to be interpreted in the same way as expected in the formulation stage. The question investigating the perceived proximity of the other participants in the conference had a mean of 2.75 in both the spatial and mono conditions with high deviations in each (1.18, 0.93). Unfortunately, this question might be confusing for some subjects as non-localized audio might be perceived as louder, "inside the head", and as a result closer compared to a sound with a synthetic spatial source. The questions scale "Nearby - Far away" might also be inappropriate as there will never be a condition that might be perceived as far away.

There is partial support for **H3** with a perceived increase in metrics of vocal intelligibility. The listening condition was significant for all three questions by the Wilcoxon test and near significance for the paired t-test. Questions of vocal clarity, effects of double-talk and comprehension confidence all appeared to improve during the spatial condition. There was a mistake made in question I-Q2 where the scale was inverted, thus the reason for a higher mean for the mono condition (2.38) over the spatial condition (1.75).

Finally, there is support for **H4** in the mean opinion scores of participants regarding the helpfulness of the localization and quality of the conference as a whole. There does seem to be less support for the judgment of higher audio quality as an effect of spatialization (O-Q2, p-value = .168), but this was added in the interest of determining whether spatial audio might lead to an increased perception of audio quality as a whole.

Across the subjects tested there were a variety of monitor sizes, mostly concentrating around smaller screen panels (sub-15 inches) while there were five participants with external monitors or using desktop computers. The distribution of monitor sizes is provided in Figure 13.

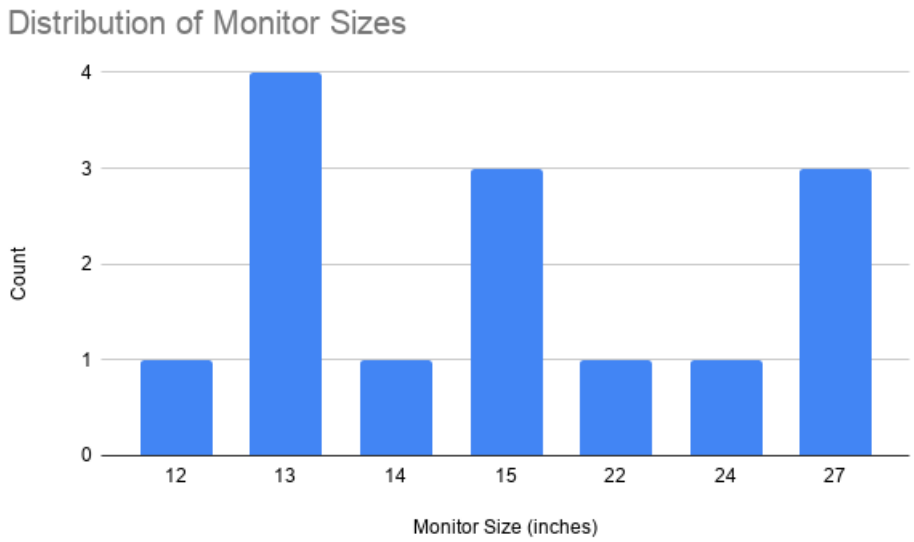


Figure 13: Distribution of user's monitor sizes in inches

While not appropriate in the present survey due to sample size, it would be interesting to study the eventual interaction between display size and spatial perception, investigating whether larger displays might better facilitate the correspondence of spatial audio to wider, distributed visual representations of participants. Perceived benefits from this difference in hardware might then emphasize its importance in media-rich communications where the real estate of the screen is coveted.

The distribution of video conferencing applications primarily used by participants was fairly close to what statistics have suggested of the current market distribution. However, there do appear to be applications like Facebook Messenger and FaceTime which might not be considered proper video conferencing applications as they have emerged from larger communications ecosystems of their companies (Facebook, Apple). Figures 14 and 15 display the primary and total count of video conferencing platforms cited by subjects.

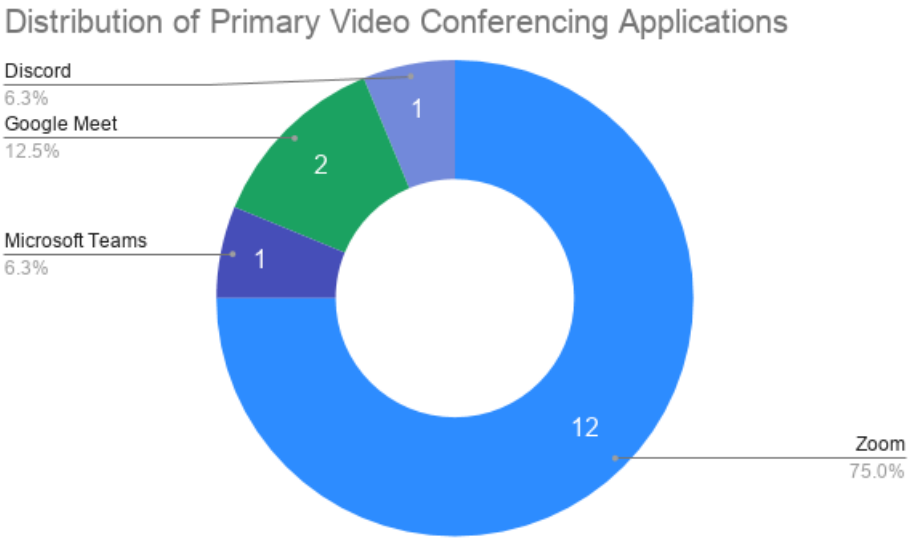


Figure 14: Distribution of user's primary video conferencing applications

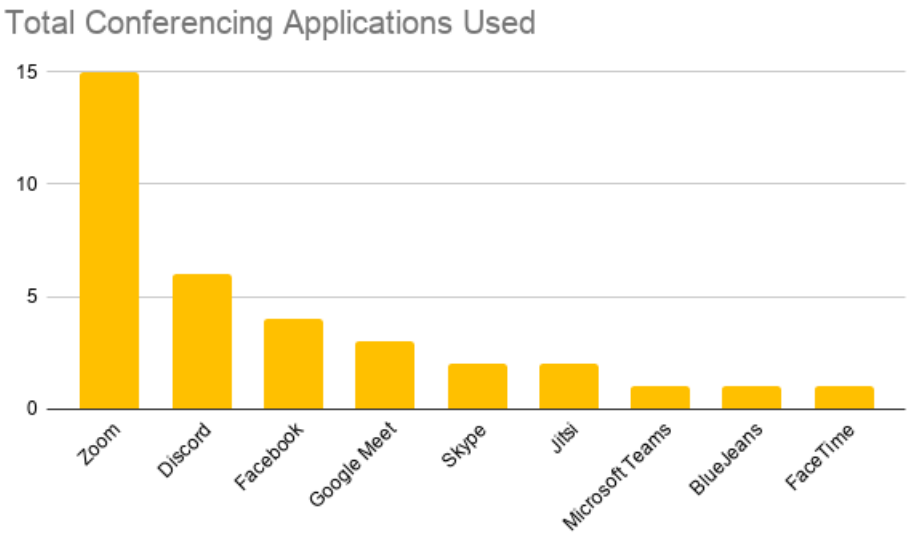


Figure 15: Conferencing applications count as mentioned by subjects

Finally, the frequency of video conferencing use is what might be expected of the population that was sampled with most subjects using some service daily and others more than once a week. That said, it draws to attention the clear bias that might exist within this population of highly skilled conferencers. The frequency of video conferencing across the subjects can be seen in Figure 16.

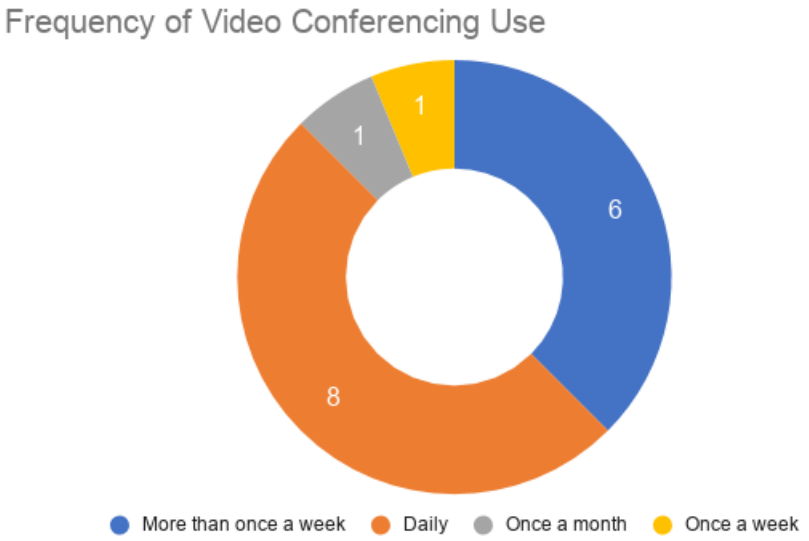


Figure 16: Reported frequency of user video conferencing use

5.5. Discussion

The results of this study are aligned with prior research on the benefits of spatial audio within conferencing environments reviewed in Chapter 2. These findings ought to merit continued attention toward the virtues of realistic aural renderings in telecommunications as a method of achieving interactions that better reflect in-person, group conversations. Each of the four hypotheses posed had some metric of support considering the low statistical power from the small sample size and brief period of exposure to a novel spatial audio system. Thus, these results must be viewed as preliminary and will benefit from further validation with larger sample sizes, refined questions, and a controlled experimental setting. One discussion that ought to be further examined from this work is how to evaluate the quality of binaural localization of sonic streams. Questions asking subjects to estimate indirect measures of social presence begs the question as to whether the subjects themselves associate these measures as well. Topics like intelligibility and cognitive effort or strain may be more readily understood compared to abstract concepts. how social dynamics are shaped by the technologies they pass through and, critically, how telepresence can arise from this exchange.

Due to the COVID epidemic, it was impossible to create an environment that would have allowed a controlled experiment to take place. As a result, development was spent on the application so that a researcher would be able to monitor the discussion and take observational notes without appearing both visually or aurally. Nevertheless, many assumptions on behalf of the participants' experiences have to be made given that there is no method of monitoring a user's conferencing machine remotely. Of course, this environment detracts from the general validity of the assumptions of each participant's experience and whether they directly followed instructions like maintaining focus on the conferencing window of the web browser or switching conditions when requested. Yet, it does stand as an example for remote studies of telecommunication's platforms that utilize the platform itself to observe and take notes.

There are a number of risks and unaccountable factors in testing a live system across a number of differently networked locations each with a different set of hardware. While Jitsi Meet, like other robust video conferencing applications, can mitigate many of these issues there

is always the possibility of disruptions in networked communication that could clearly impact the metrics evaluated. There were no comments made by participants about any failures of the network connection in either US-based or Norway-based groups, though it did seem that connections may have taken longer to establish with US-based groups. This is not to say that no network issues existed, but that if there were, they were likely to be similar to experiences with other video conferencing platforms.

In addition, each participant's computer, operating system, and headphones varied. However, none of these factors appeared to have an effect on the experience of most users and was only collected to track performance-related issues. In the pre-discussion setup period of the spatial condition, one participant did report being uncertain whether spatialization was taking place. Unfortunately, there was no way to determine if this was hardware, software, or perceptual incongruity. Since all users were on one of the three most recent versions of Chrome browser, the application did appear to operate as it had during prototyping.

Some subjects also had mentioned the fact that the spatialization localized the sound sources further than what the screen would provide, with one suggesting that this felt a bit unusual:

> It kind of felt off-putting having the sound panned so far that it was only heard in one ear. There didn't seem to feel like a natural connection automatically made by my brain that the sound was from the left because their video was on the left, it was more of a conscious thought process to make that connection.

This could suggest that the findings from Baldis, where they found a preference for spatial audio conditions scaled further from the visual source, do not correlate well within video conferencing. This brings up a major point of compromise in deciding to what degree participants' voices ought to be localized. Too far and there may be a confusing and unnatural relationship between a conference member's video and audio. But too close and there may not be enough perceptible distance between the voices to yield a benefit. Indeed, this is an area where further research will be needed.

Many additional comments remarked on the system's usefulness and improvement over standard "Zoom-llke" conferencing applications. Regarding double-talk, one subject said:

> There were few times people spoke over each other. But when they did it was far easier to understand than in a normal [Z]oom meeting. Once two people were speaking to me from opposite sides of the call[. . .] being able to hear them in each separate headphone was beyond helpful.

The prominence of the audio spatialization made users aware of the benefits of localized streams during simultaneous conversation, especially when overlapping voices in mono would be difficult to disentangle and discern. Other comments highlighted the reduced fatigue in following the discussion and more fluid discussion with the spatial condition. Indeed, most participants were excited to take part in the study and were enthusiastic in the adoption of this technology in teleconferencing.

6. Limitations and Future Developments

6.1. Summary

Spatial audio seems to be a serious contender in the roadmap to achieve telepresence in video conferencing systems. Both the academic literature regarding spatial audio, audio, and video telecommunications systems suggests a strong emphasis on audio as the chief medium in task-based communication with video providing needed markers of expression and tone. Specifically, spatial audio is found to offer benefits to cognitive load, intelligibility, and social presence which have been further bolstered by a user study in this thesis. A comprehensive review of the academic literature shows how these systems have run parallel to the technological progress in the last few decades. The development of video conferencing platforms can also be viewed from the market at large. These products are the platforms used on a daily basis and set the standards for interface design and experience that telecommunications applications are known as. Still, there can be insights found in experimental conferencing platforms invoking spatial layouts. The challenges faced by audio telecommunication and spatial audio's integration are clearly highlighted in these.

The proposed implementation is designed to bridge the gap between standard video platforms and those that have made attempts at incorporating spatial audio within group conversations. The intersection of psychoacoustics, telecommunications, and accessibility, highlighted through the body of work discussed, inform the design decisions in integrating spatial audio within a robust conferencing platform. Hopefully, the attempt made within this thesis offers a vision as to how to integrate such an audio system and why it may provide a more realistic communication experience. A condensed form of this thesis can be viewed on the Music, Communication, and Technology program's student blog[34].

6.2. Technical limitations

Although the proposed system met the established features and performance that was decided at the outset of its design there are still areas that warrant further attention. Given the limited time to implement the system, the maximum number of user's tested was five. Even on modern hardware, scaling over five concurrent users in a room, with four spatialized voices, may lend itself to interruptions in the audio thread and potential distortions from the audio track of each user. Real-time spatialization with HRTFs is an expensive operation, even as the audio source never enters into motion in virtual space. There are performance bottlenecks dependent on the web browser and host operating system. Their implementation of various digital signal processing effects (DSP) is dependent on how the audio thread is handled within both the software of the browser and operating system. It is not clear what limits might exist in concurrent spatialization of sound sources with modern consumer hardware and general purpose operating systems. Further studies should be conducted to reveal the capacity of browsers to synthesize in virtual audio spaces and suggest strategies for optimizing the

[34] https://mct-master.github.io/meet-in-space

associated computations. Doubtless, in time the technical possibilities will expand for complex, dynamic, sonic environments for media immersion and communication.

However, for the foreseeable future, high-fidelity spatial audio on commercial hardware will likely only be possible with headphones. Acoustic loudspeaker installations currently exist but are cumbersome, require tuning, and are not affordable for the end-user. But this does not exclude acoustic spatialization entirely. In the case that the user does not have headphones, it might still be possible to perform simple power-law panning of participants between conventional speaker arrangements. A stereo speaker setup is generally available on most laptops and desktop arrangements. Looking forward, these different methods of spatialization could be implemented for a greater degree of accessibility for users who do not have headphones available. It would also be useful from an experimental point of view, to explore user perspectives on stereo lateralization in video conferences and whether it offers any similar benefits in small, distributed conversations.

Finally, there is still work to be done in reaching cross-browser compatibility. Presently, the application was tested on Chromium and it appears to be compatible with Firefox and Safari as well though this has not been extensively tested. There are still problems that exist in providing unifying JavaScript methods for handling media streams at a cross-browser level. Firefox users of Jitsi Meet have struggled with video conferencing in the past and there were still unrelated issues experienced with the Firefox browsers in testing[35]. These will need to be addressed by the development team at 8x8 and potential Mozilla. As no Apple devices were tested in the implementation's development, it is unknown whether issues exist for the Safari web browser, though compatibility measures have been taken into consideration in code.

6.3. Future work

Societies will continue to move forward with the digitalization of communication for its clear benefits in accessibility, environmental footprint, and convenience. This project hopes to direct that movement into the future with the tenets of telepresence in mind. Within video conferencing software there are still unexplored arrangements and transformations of media that might provide major benefits to these systems. This thesis offers possibilities to bring realism to the regular conference experience but there are surely more. The proposed implementation outlines how it is possible to extend existing frameworks in new directions without a rewrite of a teleconferencing system. In addition, experimental designs such as the one described may be deployed easily for future studies in an entirely digital study. Larger studies might be hosted in this format with greater outreach so as to increase the confidence in the findings reported by this study.

Because the current system is built upon a free, open-source, established software, Jitsi Meet, one can easily suggest fixes and improvements to the project's source or integrate the spatial audio implementation within their own institution's Jitsi Meet deployment. This project is also able to receive updates, features, and fixes from the Jitsi Meet project itself after resolving any conflicts in code. Documentation is provided on the repository page, along with this thesis, to encourage exploration for even the casual programmer or layperson. It might also be a possibility to convert this project into a pull request for the official Jitsi Meet development team at 8x8, a request to integrate new code. For this to be a possibility, more work will need to

[35] https://github.com/jitsi/jitsi-meet/issues/4758

be done to conform the current code to the programmatic syntax and grammar followed by 8x8. If accepted, it might be considered as an optional feature for small groups in the official distribution.

The scope of this thesis has examined the intersections of psychoacoustics, telecommunications, and web technologies. It is a serious challenge to approach topics from a multidisciplinary perspective but these projects can yield richer results. Telecommunications is a new field and in infancy relative to the role it will continue to play within society. From this perspective, and in light of the recent pandemic, there is good reason to advocate for new strategies in telecommunications. This project is just one of many attempts to chip away at the rough block of telecommunications so that we may, one day, have alternatives to in-person interactions that we can genuinely enjoy.

References

Arndt, S., Antons, J., & Möller, S. (2014). Is low quality media affecting the level of fatigue? *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, 47–48. https://doi.org/10.1109/QoMEX.2014.6982286

Arndt, Sebastian, Schleicher, R., & Antons, J.-N. (2013). Does Low Quality Audiovisual Content Increase Fatigue of Viewers? *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*, 69–72. https://doi.org/10.21437/PQS.2013-14

Asturiano, V. (2020, April 11). *Remote work, regional lockdowns and migration of Internet usage*. The Cloudflare Blog. https://blog.cloudflare.com/remote-work-regional-lockdowns-and-migration-of-internet-usage/

Baldis, J. J. (2001). Effects of spatial audio on memory, comprehension, and preference during desktop conferences. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 166–173. https://doi.org/10.1145/365024.365092

Blauert, J. (1983). *Spatial hearing: The psychophysics of human sound source localization*. The MIT Press.

Brown, C. P., & Duda, R. O. (1998). A structural model for binaural sound synthesis. *IEEE Transactions on Speech and Audio Processing*, *6*(5), 476–488. https://doi.org/10.1109/89.709673

Buxton, W. (1992). *Telepresence: Integrating Shared Task and Person Spaces*. 7.

Carpentier, T. (2015, January). Binaural synthesis with the Web Audio API. *1st Web Audio Conference (WAC)*. https://hal.archives-ouvertes.fr/hal-01247528

Chen, T., Peng, L., Jing, B., Wu, C., Yang, J., & Cong, G. (2020). The Impact of the COVID-19 Pandemic on User Experience with Online Education Platforms in China. *Sustainability*, *12*(18), 7329. https://doi.org/10.3390/su12187329

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Crawford, J., Butler-Henderson, K., Rudolph, J., Malkawi, B., Glowatz, M., Burton, R., Magni, P., & Lam, S. (2020). COVID-19: 20 countries' higher education intra-period digital pedagogy responses. *Journal of Applied Learning & Teaching*, *3*(1), 1–20. https://doi.org/10.37074/jalt.2020.3.1.7

Daly-Jones, O., Monk, A., & Watts, L. (1998). Some advantages of video conferencing over high-quality audio conferencing: Fluency and awareness of attentional focus.

International Journal of Human-Computer Studies, *49*(1), 21–58. https://doi.org/10.1006/ijhc.1998.0195

Divenyi, P. L., & Oliver, S. K. (1989). Resolution of steady-state sounds in simulated auditory space. *The Journal of the Acoustical Society of America*, *85*(5), 2042–2052. https://doi.org/10.1121/1.397856

Estes, A. C. (2020, March 25). *Why the internet (probably) won't break during the coronavirus pandemic*. Vox. https://www.vox.com/recode/2020/3/25/21188391/internet-surge-traffic-coronavirus-pandemic

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S., Aragão, L. E. O. C., Arneth, A., Arora, V., Bates, N. R., … Zaehle, S. (2020). Global Carbon Budget 2020. *Earth System Science Data*, *12*(4), 3269–3340. https://doi.org/10.5194/essd-12-3269-2020

Gebhart, G. (2017, March 28). *Privacy By Practice, Not Just By Policy: A System Administrator Advocating for Student Privacy*. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2017/03/privacy-practice-not-just-policy-system-administrator-advocating-student-privacy

Graham-Cumming, J. (2020, April 23). *Internet performance during the COVID-19 emergency*. The Cloudflare Blog. https://blog.cloudflare.com/recent-trends-in-internet-traffic/

Gunawardena, C. N., & Zittle, F. J. (1997). Social presence as a predictor of satisfaction within a computer-mediated conferencing environment. *American Journal of Distance Education*, *11*(3), 8–26. https://doi.org/10.1080/08923649709526970

Huttunen, T., Vanne, A., Harder, S., Paulsen, R. R., King, S., Perry-Smith, L., & Kärkkäinen, L. (2014, August 26). *Rapid Generation of Personalized HRTFs*. Audio Engineering Society Conference: 55th International Conference: Spatial Audio. https://www.aes.org/e-lib/browse.cfm?elib=17365

Inkpen, K., Hegde, R., Czerwinski, M., & Zhang, Z. (2010). Exploring spatialized audio & video for distributed conversations. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 95–98. https://doi.org/10.1145/1718918.1718936

Jo, D., Kim, K.-H., & Kim, G. J. (2017). *Effects of Avatar and Background Types on Users' Co-presence and Trust for Mixed Reality-Based Teleconference Systems*. 10.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Citeseer.

Kilgore, R., Chignell, M., & Smith, P. (2003). *Spatialized audioconferencing: What are the benefits?* 135–144. https://doi.org/10.1145/961322.961345

Le Quéré, C., Jackson, R. B., Jones, M. W., Smith, A. J. P., Abernethy, S., Andrew, R. M., De-Gol, A. J., Willis, D. R., Shan, Y., Canadell, J. G., Friedlingstein, P., Creutzig, F., & Peters, G. P. (2020). Temporary reduction in daily global CO 2 emissions during the COVID-19 forced confinement. *Nature Climate Change*, *10*(7), 647–653. https://doi.org/10.1038/s41558-020-0797-x

Lee, G. W., & Kim, H. (2018). Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear. *Applied Sciences*, *8*, 2180. https://doi.org/10.3390/app8112180

Liu, Z., Ciais, P., Deng, Z., Lei, R., Davis, S. J., Feng, S., Zheng, B., Cui, D., Dou, X., Zhu, B., Guo, R., Ke, P., Sun, T., Lu, C., He, P., Wang, Y., Yue, X., Wang, Y., Lei, Y., … Schellnhuber, H. J. (2020). Near-real-time monitoring of global CO 2 emissions reveals the effects of the COVID-19 pandemic. *Nature Communications*, *11*(1), 5172. https://doi.org/10.1038/s41467-020-18922-7

Minsky, M. (1980). Telepresence. *Omni*, *New York*, 45–51.

Murphy, K. (2020, April 29). Why Zoom Is Terrible. *The New York Times*. https://www.nytimes.com/2020/04/29/sunday-review/zoom-video-conference.html

Nguyen, D., & Canny, J. (2005). MultiView: Spatially faithful group video conferencing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 799–808. https://doi.org/10.1145/1054972.1055084

Nguyen, D. T., & Canny, J. (2007). Multiview: Improving trust in group video conferencing through spatial faithfulness. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 1465–1474. https://doi.org/10.1145/1240624.1240846

Nokia. (2021, March 1). *Network Intelligence Report*. Nokia. https://www.nokia.com/networks/solutions/deepfield/network-intelligence-report/

Nordlund, B., & Lidén, G. (1963). An Artificial Head. *Acta Oto-Laryngologica*, *56*(2–6), 493–499. https://doi.org/10.3109/00016486309127442

Nordlund, Bertil. (1962). Physical Factors in Angular Localization. *Acta Oto-Laryngologica*, *54*(1–6), 75–93. https://doi.org/10.3109/00016486209126924

Oh, C. S., Bailenson, J. N., & Welch, G. F. (2018). A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Frontiers in Robotics and AI*, *5*. https://doi.org/10.3389/frobt.2018.00114

Ortiz, A. L. P., & Orduña-Bustamante, F. (2015). Improving speech intelligibility for binaural voice transmission under disturbing noise and reverberation using virtual speaker lateralization. *Journal of Applied Research and Technology*, *13*(3), 351–358. https://doi.org/10.1016/j.jart.2015.07.001

Pazour, P. D., Janecek, A., & Hlavacs, H. (2018). Virtual Reality Conferencing. *2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 84–91. https://doi.org/10.1109/AIVR.2018.00019

Petrone, J. (2018, May 7). *Google's got our kids*. The Outline. https://theoutline.com/post/4436/google-classroom-education-free-software-children-school-tech

Plomp, R., & Mimpen, A. M. (1981). Effect of the Orientation of the Speaker's Head and the Azimuth of a Noise Source on the Speech-Reception Threshold for Sentences. *Acta Acustica United with Acustica*, *48*(5), 325–328.

Poinsignon, L. (2020, March 17). *On the shoulders of giants: Recent changes in Internet traffic*. The Cloudflare Blog. https://blog.cloudflare.com/on-the-shoulders-of-giants-recent-changes-in-internet-traffic/

Raake, A., Schlegel, C., Hoeldtke, K., Geier, M., & Ahrens, J. (2010, October 8). *Listening and Conversational Quality of Spatial Audio Conferencing*. Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space. https://www.aes.org/e-lib/online/browse.cfm?elib=15567

Ramachandran, V. (2021, February 23). *Four causes for 'Zoom fatigue' and their solutions*. Stanford News. https://news.stanford.edu/2021/02/23/four-causes-zoom-fatigue-solutions/

Rayleigh, Lord. (1907). XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *13*(74), 214–232. https://doi.org/10.1080/14786440709463595

Rendell, J. (2020). Staying in, rocking out: Online live music portal shows during the coronavirus pandemic. *Convergence*, 1354856520976451. https://doi.org/10.1177/1354856520976451

Risoud, M., Hanson, J.-N., Gauvrit, F., Renard, C., Lemesre, P.-E., Bonne, N.-X., & Vincent, C. (2018). Sound source localization. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, *135*(4), 259–264. https://doi.org/10.1016/j.anorl.2018.04.009

Ryan, M. G. (1976). The Influence of Teleconferencing Medium and Status on Participants' Perception of the Aestheticism, Evaluation, Privacy, Potency, and Activity of the Medium. *Human Communication Research*, *2*(3), 255–261. https://doi.org/10.1111/j.1468-2958.1976.tb00484.x

Ryan, M. G. (1981). Telematics, teleconferencing and education. *Telecommunications Policy*, *5*(4), 315–322. https://doi.org/10.1016/0308-5961(81)90039-2

Sellen, A. (1995). Remote Conversations: The Effects of Mediating Talk With Technology. *Human-Computer Interaction*, *10*, 401–444. https://doi.org/10.1207/s15327051hci1004_2

Sellen, A., Buxton, B., & Arnott, J. (1992). Using spatial cues to improve videoconferencing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 651–652. https://doi.org/10.1145/142750.143070

Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. Cocos (Keeling) Islands: McGraw-Hill.

Singer, N. (2017, May 13). How Google Took Over the Classroom. *The New York Times*. https://www.nytimes.com/2017/05/13/technology/google-education-chromebooks-schools.html

Skowronek, J., & Raake, A. (2015). Conceptual model of multiparty conferencing and telemeeting quality. *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, 1–6. https://doi.org/10.1109/QoMEX.2015.7148101

Skowronek, Janto, & Raake, A. (2011). *Investigating the Effect of Number of Interlocutors on the Quality of Experience for Multi-Party Audio Conferencing*. 4.

Skowronek, Janto, & Raake, A. (2015). Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for spatial and non-spatial audio conferencing calls. *Speech Communication*, *66*, 154–175. https://doi.org/10.1016/j.specom.2014.10.003

Spur, M., Guse, D., & Skowronek, J. (2016). *Influence of Packet Loss and Double-Talk on the Perceived Quality of Multi-Party Telephone Conferencing with Binaurally Presented Spatial Audio Reproduction*. 4.

The Video Call Platforms that Dominate the World. (2021, March 22). EmailToolTester.Com. https://www.emailtooltester.com/en/blog/video-conferencing-market-share/

U.S. video call service usage during COVID-19 2020. (n.d.). Statista. Retrieved April 26, 2021, from https://www.statista.com/statistics/1119981/videoconferencing-services-us-coronavirus-pandemic/

Vorländer, M. (2020). Convolution and Binaural Sound Synthesis. In M. Vorländer (Ed.), *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality* (pp. 135–144). Springer International Publishing. https://doi.org/10.1007/978-3-030-51202-6_9

Watson, A., & Sasse, M. A. (1996). Evaluating audio and video quality in low-cost multimedia conferencing systems. *Interacting with Computers*, *8*(3), 255–275. https://doi.org/10.1016/0953-5438(96)01032-6

Wosik, J., Fudim, M., Cameron, B., Gellad, Z. F., Cho, A., Phinney, D., Curtis, S., Roman, M., Poon, E. G., Ferranti, J., Katz, J. N., & Tcheng, J. (2020). Telehealth transformation: COVID-19 and the rise of virtual care. *Journal of the American Medical Informatics Association*, *27*(6), 957–962. https://doi.org/10.1093/jamia/ocaa067

Yankelovich, N., Kaplan, J., Provino, J., Wessler, M., & DiMicco, J. M. (2006). Improving audio conferencing: Are two ears better than one? *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 333–342. https://doi.org/10.1145/1180875.1180926

Yost, W. A. (1997). The cocktail party problem: Forty years later. In *Binaural and spatial hearing in real and virtual environments* (pp. 329–347). Lawrence Erlbaum Associates, Inc.