Spatialized Audio and Hybrid Video Conferencing: Where Should Voices be Positioned for People in the Room and Remote Headset Users?

Jeremy Hyrkas*†
University of California San Diego
La Jolla, CA, USA
jhyrkas@ucsd.edu

Andrew D. Wilson*
Microsoft Research
Redmond, WA, USA
awilson@microsoft.com

John Tang
Microsoft Research
Redmond, WA, USA
johntang@microsoft.com

Hannes Gamper
Microsoft Research
Redmond, WA, USA
hannes.gamper@microsoft.com

Hong Sodoma
Microsoft
Redmond, WA, USA
hongsodoma@microsoft.com

Lev Tankelevitch
Microsoft Research Cambridge
Cambridge, United Kingdom
t-levt@microsoft.com

Kori Inkpen
Microsoft Research
Redmond, WA, USA
kori@microsoft.com

Shreya Chappidi†
University of Virginia
Charlottesville, VA, USA
shreyarchappidi@gmail.com

Brennan Jones†
Reality Labs Research, Meta
Redmond, WA, USA
brennanj@meta.com

ABSTRACT

Hybrid video calls include attendees in a conference room with loudspeakers and remote attendees using headsets, each with different options for rendering sound spatially. Two studies explored the listener experience with spatial audio in video calls. One study examined the in-room experience using loudspeakers, comparing among spatialization algorithms spreading voices out horizontally. A second study compared varying degrees of horizontal separation of binaurally rendered voices for a remote participant using a headset. In-room participants preferred the widest spatialization over monophonic, stereo, and stereo-binary audio in metrics related to intelligibility and helpfulness. Remote participants preferred different widths of the audio stage depending on the number of voices. In both studies, rendering sound spatially increased performance in speech stream identification. Results indicate spatial audio benefits for in-room and remote attendees in video calls, although the in-room attendees accepted a wider audio stage than remote users.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; *Sound-based input / output*.

KEYWORDS

spatial audio, teleconferencing, hybrid meetings

ACM Reference Format:
Jeremy Hyrkas, Andrew D. Wilson, John Tang, Hannes Gamper, Hong Sodoma, Lev Tankelevitch, Kori Inkpen, Shreya Chappidi, and Brennan Jones. 2023. Spatialized Audio and Hybrid Video Conferencing: Where Should Voices be Positioned for People in the Room and Remote Headset Users?. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3544548.3581085

1 INTRODUCTION

Teleconferencing services that connect users via video and audio internet calls have become an integral part of corporate and educational environments. Since the onset of the COVID-19 pandemic, video conferencing calls have become an increasingly important tool in people's personal and professional lives. The growth in usage necessitates measures to improve the user experience in video calls to help conversations feel more natural and avoid so-called "Zoom fatigue" [23].

One potential improvement is the use of spatial audio in teleconferencing software, with the aim to distribute sound sources spatially in the user's environment. Spatialized audio is commonly used in movie theaters, video games, and virtual reality, and can be achieved over loudspeakers or headphones. For loudspeaker playback, common spatial audio techniques range from simple stereo rendering to multichannel methods including vector-base amplitude panning [22], Ambisonics [15], and wave field synthesis [4], as well as proprietary formats including Dolby Atmos. For headphone playback, spatial sound synthesis typically involves applying head-related transfer functions (HRTFs) that capture the interaural phase and level differences observed at each ear entrance of the listener for a sound source position relative to the listener's head [14]. As video call systems add spatial audio functionality, such as Apple's FaceTime, it is important to understand the impact

*Both authors contributed equally to this research.
†Authors contributed while interning at Microsoft Research.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581085

on the user experience of call participants. Hybrid meetings add further complexity, as participants in a meeting room typically experience sound through loudspeakers while remote participants often use headsets. We present two studies that look at spatialized sound delivered through loudspeakers and headsets to inform the audio experience design for hybrid meetings.

The benefits of spatial audio in teleconferencing have been explored for decades, largely in the context of audio-only scenarios. In these contexts, spatial audio (either over loudspeakers or headphones) has been shown to increase a listener's ability to identify who is speaking [2], to lower their concentration effort [26], and to reduce the cognitive load for following a conversation [12]. While the inclusion of video in teleconferencing software may dominate the aforementioned user experience aspects [17], spatial audio in video conferences can have a positive effect on fully remote or hybrid meetings [13, 25]. However, there are a number of open research questions around how to deliver the benefits of spatial audio in teleconferencing while accounting for the implicit relationship between video streams and their corresponding audio streams.

In mixed reality and entertainment scenarios that deploy spatial audio, audio streams may be co-located with a visible object that acts as the sound emitter; in other scenarios, the sound location may be displaced relative to the corresponding visual object for dramatic effect or to exaggerate sonic cues. Video call scenarios comprise a very different set of goals and interactions, and therefore may require a unique correspondence between the audio stream and corresponding visual stream. A straightforward solution for video calls may be to co-locate video and audio sources spatially, but this approach may be challenging in practice. All methods for spatial audio over loudspeakers incur some degradation as the listener moves outside of the "sweet spot," a location in the room (usually centered among all speakers) in which the perception of spatialized sound works best [1]. In a typical conference room, many attendees will be seated outside of the sweet spot of any spatialization method, so the choice of both the method and parameters of spatialization should consider the experience for all seating locations.

Conversely, remote users using headsets will always be correctly centered in the three-dimensional sonic space created using binaural audio but face a separate challenge based on screen size. Assuming a laptop user is seated facing the screen, placing audio streams in roughly the same location geometrically as the corresponding videos may result in a separation so narrow as to lose the benefits of spatially separated speech streams, including spatial release from masking [20]. Conversely, exaggerating the size of the audio stage relative to the video stage for purposes of good spatial audio separation may disorient users due to a noticeable audio/video mismatch. Finally, the ideal audio stage size may increase with the number of interlocutors to maintain a minimum spatial separation between competing speech sources, despite the increased mismatch in auditory and visual co-location.

Here we present the results of two studies on the spatial placement of voices in video calls, with a focus on two scenarios: in-person attendees in a meeting room with a loudspeaker system, and remote attendees using headphones (see Figure 1). In the first study, in-person attendees watched pre-recorded video calls between four remote participants positioned horizontally across the screen. Different spatialization techniques were compared, and some participants sat in the sweet spot of the room while others sat off to one side. Questionnaires provided insight into users' experiences with each method, as well as their ability to identify audio events in particular audio streams. The second study focused on remote attendees using a laptop and headphones. Participants watched pre-recorded video calls between two or four callers; the voices were rendered spatially at varying degrees of horizontal separation. Participants were asked to rate the spatial alignment between audio and video and to identify streams containing certain audio events. Taken together, we see the benefits of spatial audio for both in-room and remote participants in video calling. We also see differences in the two settings, particularly how in-room participants accepted a wider audio stage than remote users. By looking at these two studies together, we can compare and contrast between in-room and remote user reactions to spatialized audio in video calls and guide the design of hybrid meeting experiences which include participants in both locations.

2 RELATED WORK

2.1 Spatial audio and speech intelligibility

Spatially separated audio sources play a key role in speech intelligibility, as illustrated in the so-called "cocktail party problem" [10]. Humans can identify and discern separate voices in a crowded room, which can be generalized to listening for a distinct audio stream among a series of auditory stimuli. Early work on following a target speech signal among distractor (or *masking*) signals found that placing target and maskers in different ears improves the ability to follow the target [10], later identified as the *spatial release from masking* (SRM) [18]. While earlier research focused on large degrees of separation between target and masker signals due to the better ear effect (also called the head shadow effect), some research suggests that the largest benefits are realized in the first 45 degrees of horizontal separation [20], especially when the target and maskers are speech signals.

2.2 Spatial audio in audio calls

SRM explains intelligiblity benefits in situations where crosstalk, or multiple speakers talking over each other, is present. Crosstalk is common in conversations but does not necessarily represent the bulk of typical conversations, particularly in the workplace. To that end, there is a large body of research demonstrating additional benefits of spatial audio in audio teleconferences.

Baldis demonstrated that spatial audio over loudspeakers has a positive effect on memory, listening comprehension, and focal assurance in audio calls [2]. In these experiments, memory can be seen as closely related to speaker identification and focal assurance as related to attention or listening effort. Baldis used static pictures of the speakers in some experiments, but not video; audio from speakers was presented as either monophonic, co-located with speaker pictures (a horizontal spread from -20 to 20 degrees off-center) or a wide spread with an audio stages as wide as +/- 60 degrees off-center. Spatial audio was shown to be beneficial over monophonic audio in all cases, and in some experiments the benefits of a wider stage were found to be statistically insignifant over the shallower stage. This finding is congruent with Litovsky's

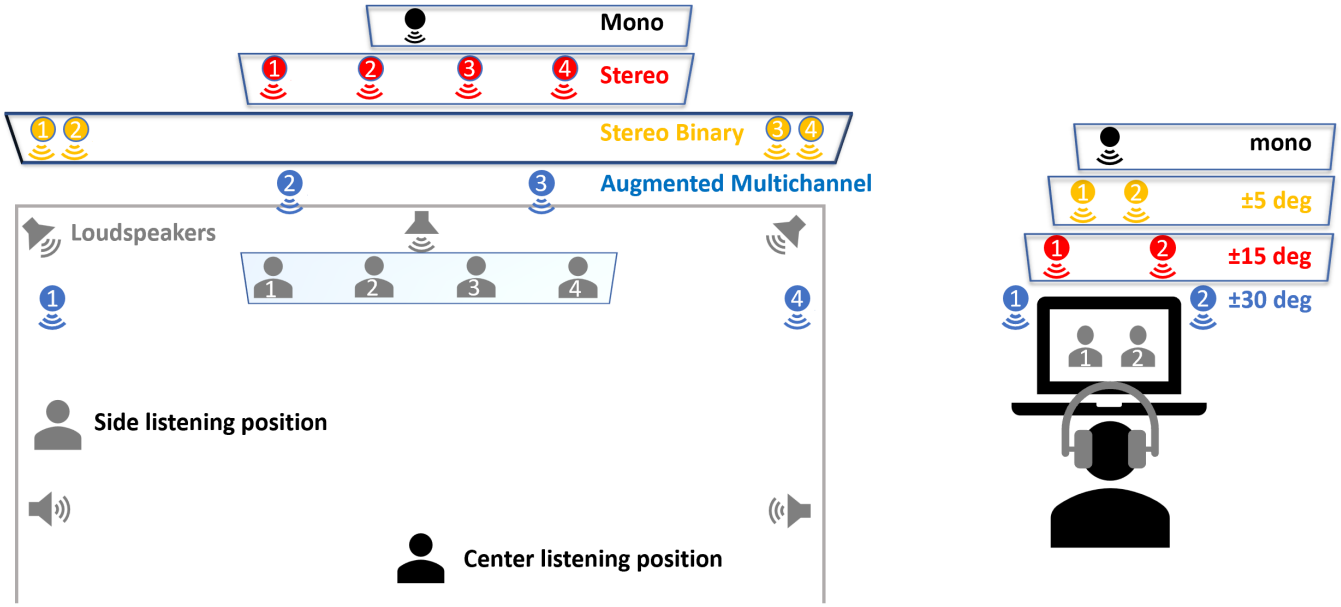


Figure 1: Spatial audio rendering conditions for a) the in-room participants, from the center seat perspective showing Augmented Multichannel in blue, Stereo Binary in orange, Stereo in red, and Mono in black and b) the remote participants for the 2-person videos showing mono and spacing at +/- 5, 15, and 30 degrees.

finding that SRM benefits occur most in the early stages of azimuth separation [20].

Spatial audio over headphones has been found to be broadly beneficial in audio teleconferences when head tracking is enabled [26]. Binaural audio was associated with higher speaker recongition and lower cognitive load, confirming results from Baldis [2]. In addition, binaural audio was associated with higher perception of overall quality, confidence, confidence in speaker identification, perceived connection quality, speech intelligibility and a lower cognitive load and listening effort as reported by listeners. While the results were largely positive, the case of spatial audio without head tracking was not tested. Binaural audio delivered with head tracking requires special hardware, and is therefore not widely available in most audio or video teleconferencing services. Thus, we wanted to examine the benefits of spatial audio without head tracking using the headsets commonly used in video teleconferencing scenarios.

While cognitive load is often self-reported, a recent study demonstrated a reduction in cognitive load when listening to spatialized conversations through direct measurement [12]. Participants were asked to listen to a conversation between two people while performing a secondary task, and answer a questionnaire about the conversation to assess their level of understanding. While there was no significant difference in understanding between groups who listened to monophonic audio and those who listened to spatially separated audio, the latter group successfully completed more trials of the secondary task, indicating a higher bandwidth for additional tasks while listening with spatial audio.

The Telecommunication Standardization Sector of the International Telecommunication Union (ITU) focuses on qualitative evaluation of audio and video teleconferencing systems. The group's recommendation P.1310 outlines practices for conducting evaluations of call systems that utilize spatial audio [21]. The ITU notes that spatial audio benefits are realized more strongly when separating voices with similar frequency content, which corresponds with gender identity. Additionally, they recommend test groups with a variety of experience with spatial audio, as limited exposure can lead to increased emphasis on benefits (the "wow factor") or a uniformly negative reaction due to unfamiliarity with a spatial listening experience. Thus, our studies allow participants to compare among a range of different spatial audio layouts.

2.3 Spatial audio in video calls

The addition of video in teleconference calls adds complexity to the inclusion and implementation of spatial audio. De Bruijn conducted experiments connecting two meeting rooms acoustically and visually using 2-D video projection and loudspeaker-based spatial sound reproduction [11]. The experimental results revealed the effects of participant position in the room, the sound rendering setup, and interactions between audio and video perception such as the "ventriloquist effect" [5] on various aspects of user experience, including speaker identification, speech intelligibility, and perceived discrepancy between audio and video positioning. For a thorough review of quality of experience aspects in teleconferences, including the role of spatial audio in mixed-reality communication scenarios, see the work by Skowronek et al. [27]. Inkpen et al. showed that the inclusion of video in a spatial audio teleconferencing system substantially improved user experience, perhaps overshadowing benefits previously observed for spatial audio in audio-only scenarios [17]. This finding reflects previous research on speech communication where conversational tasks are known

to be positively affected by the ability to see other speakers [7]. The fact that visual perception dominates auditory perception in certain scenarios in terms of the perceived quality or location of sources [3, 5] may be a contributing factor in the relatively slow adoption of spatial audio in commercial video call solutions.

More recent research reevaluating the role of spatial audio in videoconferencing found user benefits. Remote participants in hybrid meetings showed increased conversation comprehension and confidence when audio was spatialized to reflect the position of in-person participants [25]. The effects on comprehension and confidence were stronger when in-person participants wore medical face masks and when simulated head tracking was used.

Fleming et al. found selective attention to a target speaker among masking signals to be most successful when the target video and audio were rendered in the same hemifield, i.e., spatially aligned [13]. Participants were asked to follow one speech signal while another speech signal was present; in some cases video of each speaker was presented, and audio for each speaker was either spatially aligned or misaligned with the corresponding speaker video. Correctly aligned spatial audio and video was better than all audio in mono, and correctly aligned spatial audio with no video was slightly more beneficial than cases where video was present with misaligned spatial audio. The authors did not compare spatial audio and monophonic audio in video examples, however, so it is difficult to assess the benefit of taking a video system with monophonic audio and adding spatialization.

Commercial and open-source video call systems have recently begun incorporating spatial audio. Wong and Duraiswami outline a prototype video conferencing service that is built from the ground up with spatial audio [28]. Similarly, there are recent efforts to support spatial audio conferencing based on the open-source communication software Jitsi Meet [16]. Since these explorations have not yet included user evaluations, we wanted to systematically investigate the effects of different spatial audio layouts using loudspeakers and headsets.

Previous evaluations do not cover listening environments that are most common in current video conferencing scenarios. Baldis [2] and Inkpen et al. [17] investigated using one loudspeaker for each individual in a call, which is not scalable in practical teleconferencing. De Bruijn used wave field synthesis, which requires a large number of speakers to accurately reproduce spatial sound, which are unlikely to be used in most scenarios. Therefore, we explored user experiences when voices are spatialized using the most commonly available method of stereo loudspeakers, as well as more scalable forward-looking methods such as Dolby Atmos which can adapt spatial reproduction to a variety of loudspeaker arrangements. Similarly, the benefits of spatial audio for headphone users has typically focused on users with head tracking headphones [25, 26], which most current headphones do not support. While horizontally aligning a speaker's audio and video has been shown to improve a listener's selective attention [13], our study investigates how users experience the angular positioning of voices binaurally in the most common scenario where head tracking is not available. Our study explores how to layout spatialized audio in hybrid meetings, including whether there should be differences in the layout for the in-room attendees and those joining remotely.

3 EXPERIMENT DESIGN

3.1 Study 1: Spatial placement of audio streams for in-room attendees

3.1.1 Design and procedure. To explore the spatial placement of voices for meeting participants in the room, we set up a lab to serve as a conference room that included a projection screen at the front of the room and an Atmos-driven loudspeaker audio system (see Figure 1a). We wanted to explore the listening experience according to position in the room, so one chair was located centrally in the room, whereas a second chair was located along the left side of the room. The study explored four spatial audio conditions: mono, stereo, stereo binary, and augmented multichannel. As stimuli for the study, we recorded the video and isolated audio of four different scripted conversations of four people that lasted about 11-13 minutes. The conversations were highly interactive, including substantial overlapping talk, on topics such as: Where would you like to travel post-COVID? and What one personal item would you take on a trip to Mars? In all four conditions, the video recordings were displayed in the same positions on the front screen, but since we had isolated audio recordings, we could configure them in any of the four spatial audio conditions. The four conversations were always played in the same order, but the audio conditions were rotated through a different starting point in the mono, stereo, stereo binary, and augmented sequence, so that in the aggregate, the same number of people experienced each audio condition in each of the four experimental condition orders.

The videos of speaker participants were displayed on the front wall of the room which was about 4.6 m wide. Videos were arranged horizontally, equally spaced and centered so that they occupied about one half of the available width of the wall. The center chair was placed approximately at the center of the room, facing forward and about 3m away from the front wall. The side chair was placed with its back along the left wall, about 2m from the front of the room. The room was equipped with a Dolby Atmos 7.1 loudspeaker system with a center, two front, two mid, two rear loudspeakers and a subwoofer. The left and right front loudspeakers were placed in the left and right front corners of the room. The left and right mid loudspeakers were placed at about the same distance from the front wall as the center seat. It is worth noting that in this configuration the side chair was to the left of the left mid speaker. Front and mid loudspeakers were placed at approximately the same height as the videos and the listener's head, while the center channel was placed just under the projection area, about 0.6m above the floor. The single subwoofer was placed along the front wall. The rear loudspeakers were placed in the rear corners of the room but were not used in any audio condition.

Audio conditions were chosen to reflect loudspeaker configurations commonly used in conference rooms, as well as explore the possiblity of a wider sound stage afforded by multichannel (Atmos) setups. Figure 1a shows the four different audio conditions for spatializing the voices (color coded) as well as the location of the physical loudspeakers used to render the sound (shown in gray):

- Mono: All audio comes from the center front loudspeaker.
- Stereo: Each speaker's voice is rendered using only left and right front loudspeakers using a constant-power panning

calculation, so that when heard from the center listening position, speaker voices seem to come from their onscreen video: −21°, −7°, +7° and +21° azimuth from the center of the display.

- Stereo Binary: The two speakers on the left side of the display are rendered to the left front loudspeaker only, while the two speakers on the right side of the display are rendered to the right front loudspeaker only. From the center listening position, speaker voices come −37° and +37° azimuth from center. With only two loudspeakers this approach clearly loses much of the precision of stereo panning, but may reduce or eliminate the benefits of sitting in the "sweet spot" and thus lead to a more consistent perception of spatial rendering between the center and side listening positions.
- Augmented Multichannel: Using the Atmos receiver and its *object-* or *position-based* audio rendering codec, speaker voices are placed along an ellipse in the horizontal plane of the listener with major axis 3m and minor axis 2m. From the center listening position, speaker voices are rendered at the mid, front and center speakers, to realize virtual sound sources coming −55°, −18°, +18°, and +55° azimuth from center.

In each conversation, two different distractor sounds, such as a baby crying, dog barking, car alarm, etc., were included into the audio stream of two different people. These distractor sounds were edited into the audio streams after the conversations were recorded, so there was no reaction to the sounds registered in the video recordings. These distractor sounds were included as a test to see how easily and accurately study participants could identify who was associated with the distractor sound during the study.

Study participants were recruited through an email distribution list within our company. Participants were offered to join the study together with someone they knew (half of the participants), or were paired with another volunteer for the study (the other half). Participants came to the study lab site and were assigned which seat to take in the room (center or left side). After signing the consent form, the study began by playing the first recorded conversation. Study sessions lasted about one hour, and participants were given a $50 gift card gratuity for participating in the study. This study design was approved by our institution's Ethics Review Board.

3.1.2 Data collection and evaluation. Participants completed questionnaires after each condition. As an objective measure of audio stream identification, participants were asked to identify which of the four audio streams they thought contained a distractor sound, and if they identified a stream, to rate how confident they were in their assessment. Each scenario included two such distractor questions. Details for each question are provided with the results.

To additionally evaluate the mental demand required in each condition, participants were asked to rate their overall understanding of the conversation, and the ease of determining which individual was speaking and what was being said during overlapping speech. To evaluate participants' preferences among the four audio conditions, they rated the helpfulness of audio placement in each condition, and ranked the four conditions in order of preference. As an additional evaluation of participants' perception of the audio conditions, participants were also asked to sketch the locations on a visual layout of the room where they thought each of the four audio streams were coming from during each condition. Finally, participants were also asked to provide open-text feedback about what they liked about the audio placement in each condition, and what could be improved.

Questionnaires also asked about participants' age, gender, whether they have seeing or hearing difficulties, occupation, educational background, frequency of using video chat for work and for family or friends, frequency of using audio chat for work and for family or friends, and whether they knew any of the people in the conversations. Details of the analysis of specific questions are explained with the results. Overall, distraction identification accuracy and confidence and Likert scale ratings were tested for statistical significance using non-parametric Brunner-Langer analyses [8] (using the ANOVA-type statistic [ATS]), with audio condition as a within-subjects factor, and seated position as a between-subjects factor. Pairwise comparisons for significant main effects or interactions were conducted using Wilcoxon signed-rank tests. All pairwise comparisons were adjusted for multiple comparisons using the Bonferroni correction. Where there are no main effects or interactions with a variable (e.g., seated position), results are visualised collapsed across that variable.

3.2 Study 2: Spatial placement of audio streams for remote attendees

3.2.1 Design and procedure. The second study focused on remote attendees. We asked participants to watch pre-recorded video calls using headphones on their own laptop computers and in an environment of their choice, e.g., at home or in an office. The video calls contained either two or four remote speakers with their videos arranged horizontally across the screen. For the two-person calls, eight conversations between coworkers (one man and one woman) of between 2 and 3 minutes duration were remotely recorded using semi-improvised conversation guides adapted from previous telecommunication evaluation guides [24, 25]. The videos of both speakers were arranged side-by-side and displayed to study participants on their screen. In each pre-recorded call, a short distraction sound (between 2 and 4 seconds) was inserted into one of the two streams at a level roughly 20 dBA quieter than the louder speech stream. All sounds were downloaded from freesound.org with the Creative Commons License CC0. Four spatial audio conditions were compared, with each condition applied randomly to two of the eight conversations. Participants were randomly assigned to one of two groups differing only in the order of spatial treatments experienced. For the monophonic condition, all audio was downmixed to a single channel played back to both headphones. For the spatial conditions, the two voices were displaced symmetrically off-center at either ±5, ±15, or ±30 degrees azimuth (see Figure 1b). The monophonic condition represents the default case for most current video conferencing systems. The rendering with ±5 degrees azimuth off-center was chosen to be close both to the perceptual limits for horizontal localization [6] as well as the visual separation of the two speakers when watched on a typical laptop screen. Conditions rendered with ±15 and ±30 degrees azimuth are expected to be perceived as clearly spatially separate by most listeners. However, they may exceed

the visual separation of speaker videos and cause an audio-visual localization mismatch that may be annoying for certain users [19].

To simulate calls with four remote attendees, the same recordings were used as for the in-room study (cf. subsection 3.1). To reduce the duration of the experiment, the pre-recorded calls were truncated to be between 3 and 5 minutes and contained only one distraction sound rather than two. All speaker videos were arranged horizontally, spanning the width of the screen. As with the two-speaker scenario, each call was rendered either with monophonic audio or spatially with three horizontal spreads, placing speakers at [-15, -5, 5, 15] degrees, [-30, -15, 15, 30] degrees, or [-50, -25, 25, 50] degrees azimuth, respectively. As with the two-speaker scenario, monophonic rendering represents the status quo for video call software; the spatial conditions allowed testing user experience and performance as a function of the horizontal spread of voices in a video call.

Unlike Study 1, participants in Study 2 were recruited remotely and completed the survey with their own equipment in an environment of their choice, mimicking the typical environment of a remote video call attendee. Participants were internally-recruited employees and received all instructions, videos and questionnaires through an online form. Participants received a $25 gift card gratuity for their participation. Participants watched all two-speaker videos followed by all four-speaker videos. Each audio treatment was experienced twice in the two-speaker set, and once in the four-speaker set. This study was also reviewed by our institution's Ethics Review Board.

3.2.2 Data collection and evaluation. As in Study 1, each participant was asked to identify which participant in each call contained the background distraction noise as well as their confidence in their answer. They were also asked to rate the spatial ordering of voices and the relationship between the location of the video and audio in questions modified from the ITU P.1310 guide for evaluating telecommunication systems [21]. A brief pre-study questionnaire collected basic demographic questions and a post-study questionnaire asked for additional open-ended feedback regarding the task, the width of the audio stage, and experience or preference differences between two- and four-person conversations.

4 RESULTS

4.1 Study 1: In-room attendees

4.1.1 Demographics. The data from a total of 40 participants were analyzed for this study. Of these participants, 23 self-identified as female, 16 male, and one did not state their gender. Participants ranged from 18-26 years, with the median age of 21. They were generally familiar with using video chat: 26 reported using it for work every day, 14 use it at least every week, and 14 reported using video chat for family or friends every day, 20 every week, and 6 at least once per month.

4.1.2 Distractor identification and confidence. Participants were asked to answer the following questions:

(1) Which participant had [*video-dependent distraction sound*] in their background? (*names of the four conversation participants*, "I did not hear it", "I could hear it, but could not tell where it came from", "I could hear it, but could not remember where it came from")
(2) *[For those providing a guess]* How confident are you in this assessment? (Not at all confident, Somewhat not confident, Somewhat confident, Very confident)

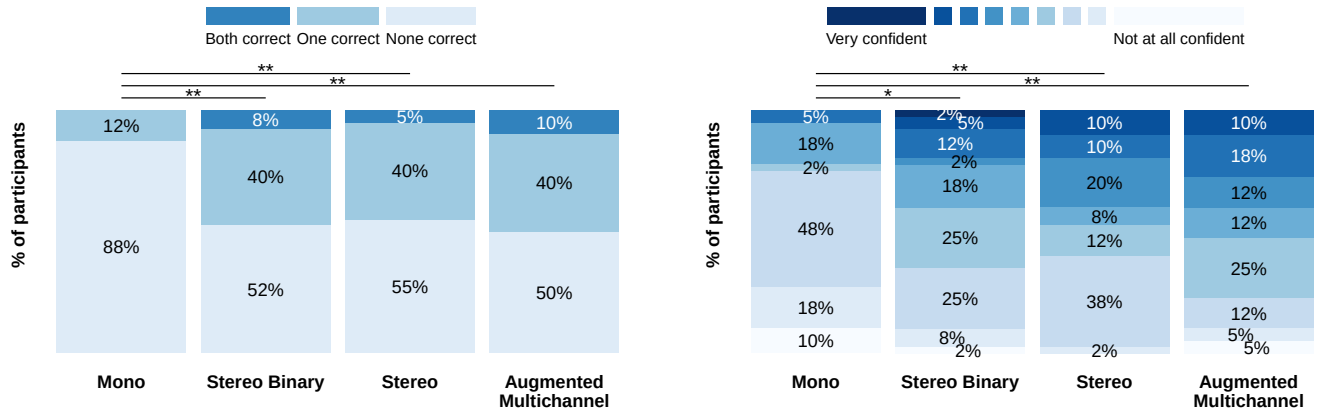
Results are summarized in Figure 2. The data were analyzed for how audio condition affected participants' accuracy in identifying the speaker whose background included a distractor sound. Accuracy for each distractor identification question was coded as 1 if participants identified the correct audio stream, and 0 if they identified the incorrect audio stream or did not provide a guess. These data were averaged across the two distractor questions for each participant. As the data values could only include 0, 0.5, or 1, (i.e., it was more similar to ordinal than continuous data), a non-parametric Brunner-Langer analysis [8] was conducted (using the ANOVA-type statistic [ATS]), with audio condition and seated position as within- and between- subjects factors, respectively. Participants' confidence ratings were coded as 0.25 ("Not at all confident"), 0.5 ("Somewhat not confident"), 0.75 ("Somewhat confident"), and 1 ("Very confident"). As confidence ratings were only asked of participants who provided a guess, those who responded "I did not hear it" were coded as a 0 and "I could hear it, but could not remember where it came from" and "I could hear it, but could not tell where it came from" as a 0.25. As above, confidence data were averaged across the two distractor questions and analysed using Brunner-Langer analyses. Pairwise comparisons were conducted using Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Bonferroni correction.

There was a significant effect of audio condition on distractor identification accuracy (ATS (2.6) = 6.4, $p = 0.001$); there was no main effect or interaction with seated position (both $p > 0.2$). Compared to the Mono condition, participants were much more likely to correctly choose the stream with the distractor in the Augmented Multichannel ($p = 0.005$), Stereo Binary ($p = 0.003$), and Stereo ($p = 0.007$) conditions (Figure 2a; note that, as there was no main effect or interaction with seated position, results are visualised collapsed across this variable).

The analysis of confidence in audio stream identification showed a similar pattern, with a significant effect of audio condition (ATS (2.82) = 10.93, $p < 0.001$), and no main effect or interaction with seated position (both $p > 0.5$). Again, compared to the Mono condition, participants' confidence was much higher in the Augmented Multichannel ($p = 0.001$), Stereo Binary ($p = 0.025$), and Stereo ($p = 0.001$) conditions (Figure 2b).

4.1.3 Mental Demand. Participants were asked to answer the following questions:

(3) During the conversation, determining which individual was speaking was... (1 – Very difficult, 5 – Very easy)
(4) During the conversation, determining what was being said when multiple people talked at the same time was... (1 – Very difficult, 5 – Very easy)
(5) My overall understanding/comprehension of the conversation was... (1 – Very poor, 7 – Very good)



(a) Distractor identification accuracy

(b) Distractor identification confidence

Figure 2: Distractor identification (a) accuracy and (b) confidence. Pair-wise conditional comparisons computed using Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Bonferroni correction. *** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.1.

Results are summarized in Figure 3. As above, rating scale data were treated as ordinal and analysed using non-parametric Brunner-Langer analyses with Wilcoxon signed-rank test pairwise comparisons, adjusted using the Bonferroni correction. There was a significant effect of audio condition on participants' perceived ease of determining which individual was speaking (ATS (2.86) = 3.28, p = 0.02); there was no main effect or interaction with seated position (both p > 0.38). Pairwise comparisons showed that participants rated the Augmented Multichannel condition as easier than the Mono condition (p = 0.012; Figure 3a). Likewise, participants' perceived ease of determining what was being said during overlapping speech showed a main effect of audio condition (ATS (2.8) = 8.47, p < 0.001), and no main effect or interaction with seated position (both p > 0.7). Pairwise comparisons showed the Augmented Multichannel condition as being rated higher than the Mono condition (p < 0.001), the Stereo condition (p = 0.043), and the Stereo Binary condition (p = 0.056; Figure 3b). There were no significant differences in participants' ratings of overall understanding of conversations as a function of audio condition (p = 0.08), location (p = 0.58), or their interaction (p = 0.56).

4.1.4 Helpfulness of audio placement and preferences. Participants were asked to answer the following questions:

(6) The placement of remote individuals' voices was... (1 – Not at all helpful, 4 – Very helpful)
(7) Please rank your preference for the audio conditions of the conversations you have listened to... (1 - Most preferred, 4 - Least preferred)

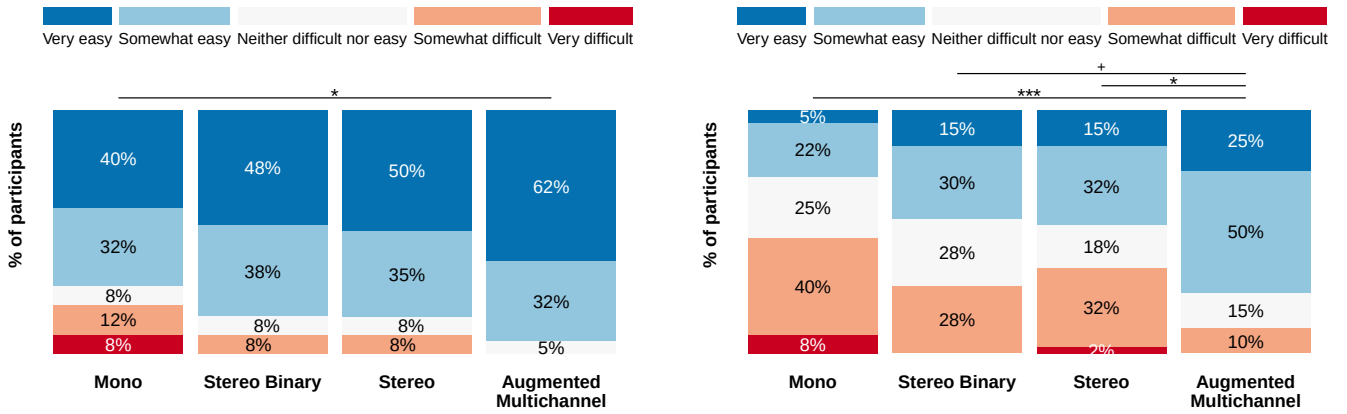
Results are summarized in Figure 4. As expected, participants' ratings of how helpful audio placement was in each audio condition significantly differed (ATS (2.89) = 8.86, p < 0.001); there was no main effect or interaction with seated position (both p > 0.19). Pairwise comparisons showed that participants found the audio placement in the Augmented Multichannel condition as more helpful than in the Mono condition (p < 0.001; Figure 4a). Similarly, placement in the Stereo and Stereo Binary conditions was also rated as more helpful than Mono (respectively, p = 0.04 and p = 0.01). In line with this, more participants ranked the Augmented Multichannel condition as their top preference (48%), compared to 25% for Stereo, 18% for Stereo Binary, and 10% for Mono, with this distribution of top preferences differing significantly from an equal preference across conditions (χ^2 Goodness of Fit test, $\chi^2(3) = 12.6$, p = 0.006; Figure 4b).

Open-ended comments from the Augmented Multichannel condition helped explain users' preference for that condition: "It felt like I was in the middle of a live conversation taking place. ...it felt almost like I was in a live theater or improv show" "[Augmented Multichannel] was by far the best placement of speakers so far, the separation was very clear and made it very easy to tell speakers a part [sic] from one another."

As an additional measure of participants' perception of the audio streams in the four conditions (i.e., the mental model they maintained of the speakers' audio streams), we coded and analysed participants' sketches of the audio stream locations in terms of accuracy. Figure 5 shows the coded accuracy of the sketches according to condition and seated position. Sketches were considered accurate in each condition if they were in the correct left-to-right order and:

- Mono: Voices were clustered around the center
- Stereo: Voices were aligned near where the video images appeared
- Stereo Binary: Left and right voices were clustered around speakers located just to the left and right edges of all the video images, respectively
- Augmented Multichannel: Voices were spread across most of the width of the room, beyond where the video images were located.

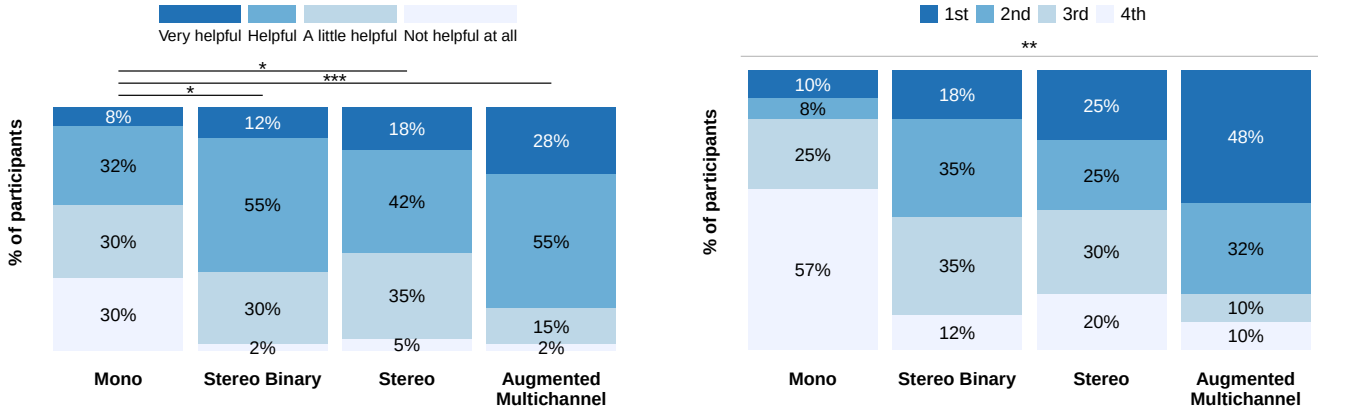
Common errors in the sketches included showing the voices as closer to their own position or closer to the speakers than where the



(a) Understanding individual speakers

(b) Understanding overlapping speech

Figure 3: Perceived ease of understanding (a) individual speakers and (b) overlapping speech. Pair-wise conditional comparisons computed using Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Bonferroni correction. *** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.1.



(a) Helpfulness of audio placement

(b) Preferences for audio conditions

Figure 4: Participants' (a) ratings of helpfulness of audio placement and (b) ranked preferences for audio conditions. For helpfulness ratings, pair-wise conditional comparisons were computed using Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Bonferroni correction. For ranked preferences, a χ^2 goodness of fit test was computed of top preferences against an even preference across conditions. *** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.1.

voices were acoustically rendered, or projecting distinct positions in the Mono condition when the voices were all coming out of the center speaker.

As the sketch accuracy measure was binary, it was analysed using a mixed-model logistic regression (using glmer in R), with audio condition and seated position as fixed effects and participant ID as a random intercept [9]. There was a significant effect of audio condition on sketch accuracy ($\chi^2(3) = 12.9$, $p = 0.005$; Figure 5), and no main effect of seated position or interaction (both $p > 0.1$), although sketch accuracy was directionally higher for participants in the central location in all four audio conditions. Pairwise comparisons showed that participants in the Stereo Binary condition drew more accurate sketches than those in the Stereo condition ($p = 0.002$). It is important to note that while even off-center participants were more accurate in locating the Stereo Binary condition as having voices clustered around the speakers, that layout did not help in distinguishing between the speakers coming from the same loudspeaker. This difficulty was reflected in one of the open-ended comments on this condition: "Sometimes it could be difficult to differentiate between the two ppl on the same side, especially if there were multiple ppl speaking". Thus, Stereo Binary was not ranked higher in terms of helpfulness or preference, despite this higher accuracy.

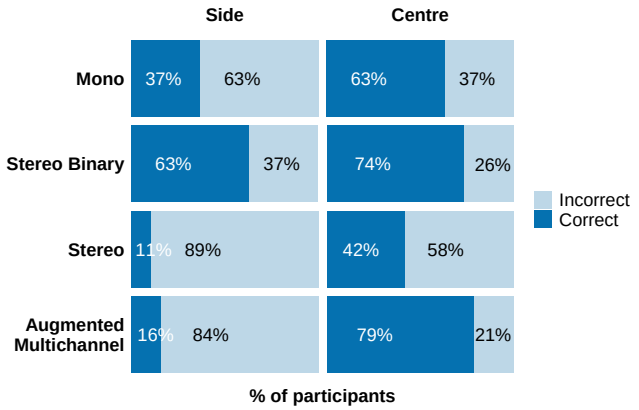


Figure 5: Percent of participants' sketches of perceived audio stream locations coded to be accurate, by audio condition and participants' seated position (side and centre). Pairwise comparisons between audio conditions were computed using the estimated marginal means and adjusted for multiple comparisons using the Bonferroni correction. Although data is shown for each position, there was no significant main effect or interaction with position. *** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.1.

4.2 Study 2: Remote attendees

4.2.1 Demographics. 23 participants were recruited for this study, including 16 men, 6 women, and one non-binary or gender diverse individual. Twelve of the participants were between the ages of 25-34 years old, with five between 18-24 years old, three between 35-44 years old and three between 45-54 years old. 16 participants use video call services daily, and the remaining seven use it 2-3 times per week. Only three self-identified as spatial audio experts, with five very familiar with the technology, ten somewhat familiar, and five not at all familiar. 15 participants used over-ear headphones, with the remainder using earbuds. Laptop screen size responses were varied and more difficult to assess, but most ranged from 13" to 16" diagonally.

4.2.2 Spatial ordering and audio/video position correspondence. Participants were asked to answer the following questions:

(1) How adequate is the spatial ordering of the voices that you hear? (1 – Very inadequate, 5 – Very adequate)
(2) How adequate is the relationship between the location of the speakers on the screen and location of their voices? (1 – Very inadequate, 5 – Very adequate)

Results are summarized in Figure 6 for two-person videos and 7 for four-person videos. As each audio condition was experienced twice in the two-speaker set, responses were averaged across the repetitions prior to analysis. Non-parametric Brunner-Langer analysis [8] on responses revealed a significant effect of audio condition for both the rating of spatial ordering (ATS (2.87) = 19.66, $p < 0.001$) and the rating of audio and video positional correspondence (ATS (2.77) = 23.6, $p < 0.001$) in two-person video scenarios, as well as in four-person video scenarios (Spatial ordering: ATS (2.42) = 11.49, $p < 0.001$; A/V correspondence: ATS (2.37) = 8.48, $p < 0.001$). Pairwise comparisons between conditions for each set of videos were tested using the Wilcoxon signed-rank test. In the videos with two-speakers, there is generally an upward trend in ratings as the width of the audio stage increases, with the largest differences observed between mono and +/- 30 degrees. Increasing from 15 degrees separation off-center to 30 degrees does not significantly change either ranking.

The ratings follow a similar trend in the four-person videos. Using the same analyses as above, there is generally an upward trend as azimuth increases, although every pair-wise step is not necessarily statistically significant. As above, the ratings plateau in the third audio condition, in this case an audio stage ranging from -30 to 30 degrees off-center; increasing the width to 50 degrees off-center in either direction does not show a signfiicant change.

4.2.3 Distractor identification and confidence. Participants were asked to answer the following questions:

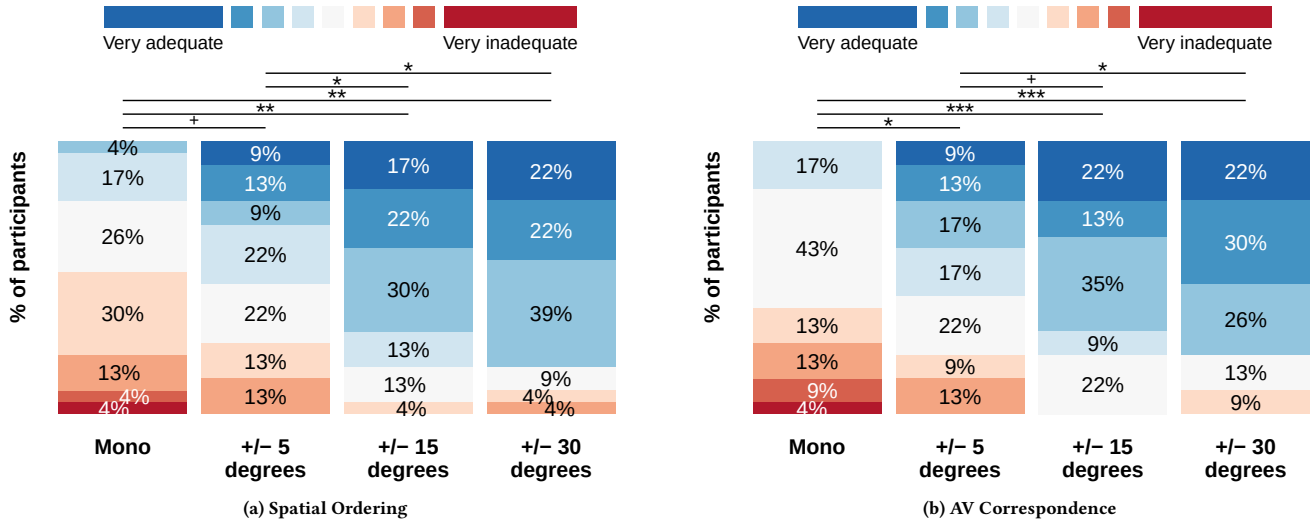
(3) Which participant had *[the video-dependent distraction sound]* in their background? *[List of speakers in order from left-to-right, with pictures]*
(4) How confident are you in this answer? (Not at all confident, Somewhat not confident, Somewhat confident, Very confident)

Results are summarized in Figure 8 for two-person videos and Figure 9 for four person videos. In the case of two-person videos, the random chance level for choosing the correct distractor in a stream is 50%. Brunner-Langer analyses showed a significant effect of condition on identification accuracy (ATS (2.61) = 9.49, $p < 0.001$). Participants were much more likely to correctly choose the stream with the distraction in cases where audio was separated off-center by 15 or 30 degrees; similarly to results found in section 4.2.2, no significant benefit is observed by increasing the width from 15 to 30 degrees. Confidence similarly increased with width (ATS (2.62) = 19.99, $p < 0.001$), again plateauing with no additional benefit observed beyond 15 degrees of horizontal spread.

The overall trend is similar in responses to the pre-recorded calls with four speakers, with identification accuracy (ATS (2.74) = 23.02, $p < 0.001$) and confidence (ATS (2.08) = 11.81, $p < 0.001$) increasing with azimuth spread and plateauing in the condition where audio is spread between -30 and 30 degrees off-center. Unlike all other results thus far, this analysis shows a degradation when moving to the widest audio condition of -50 to 50 degrees off-center; however, it is unclear whether the widest condition was unfairly penalized by having a distrator sound with the person displayed at -25 degrees, whereas in the other conditions it was displayed at the extreme ±15 or ±30 degrees. We can conclude that increasing the width corresponds with an increase in correct responses and confidence for the stream identification up to a spread of 30 degrees off-center; from these results we cannot confidently comment on the effect of increasing spread from 30 degrees to 50 degrees.

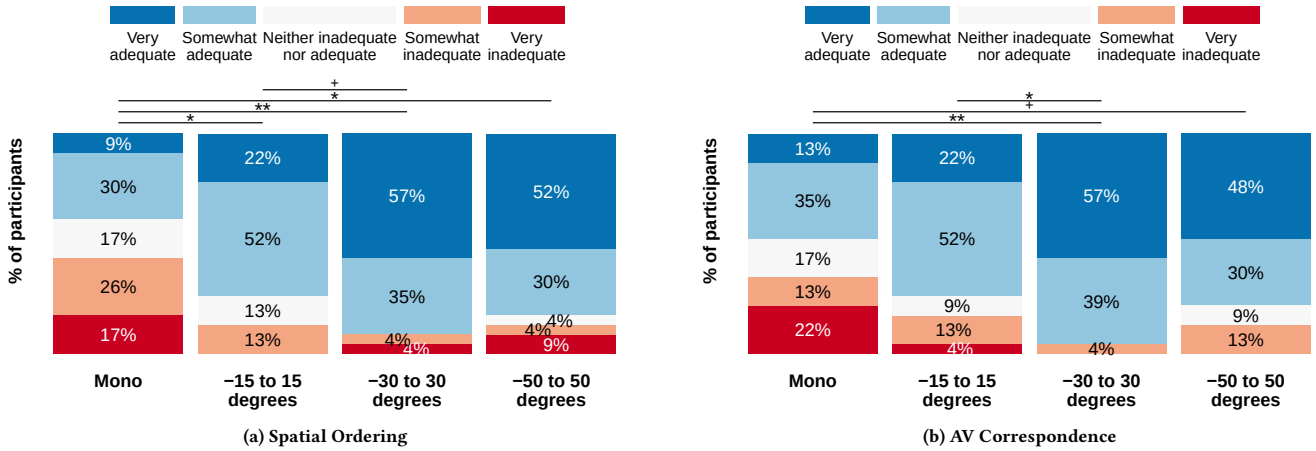
4.2.4 Open-ended feedback. After completing all videos, participants were asked for free-form answers to the following questions:

(1) Describe your overall experience with the audio in this experiment.



(a) Spatial Ordering (b) AV Correspondence

Figure 6: Spatial ordering ratings and audio/video positional correspondence ratings for two-speaker videos. All ratings range from 1-5 and are averaged across two video repetitions. Pair-wise conditional comparisons computed using Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Bonferroni correction. *** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.1.



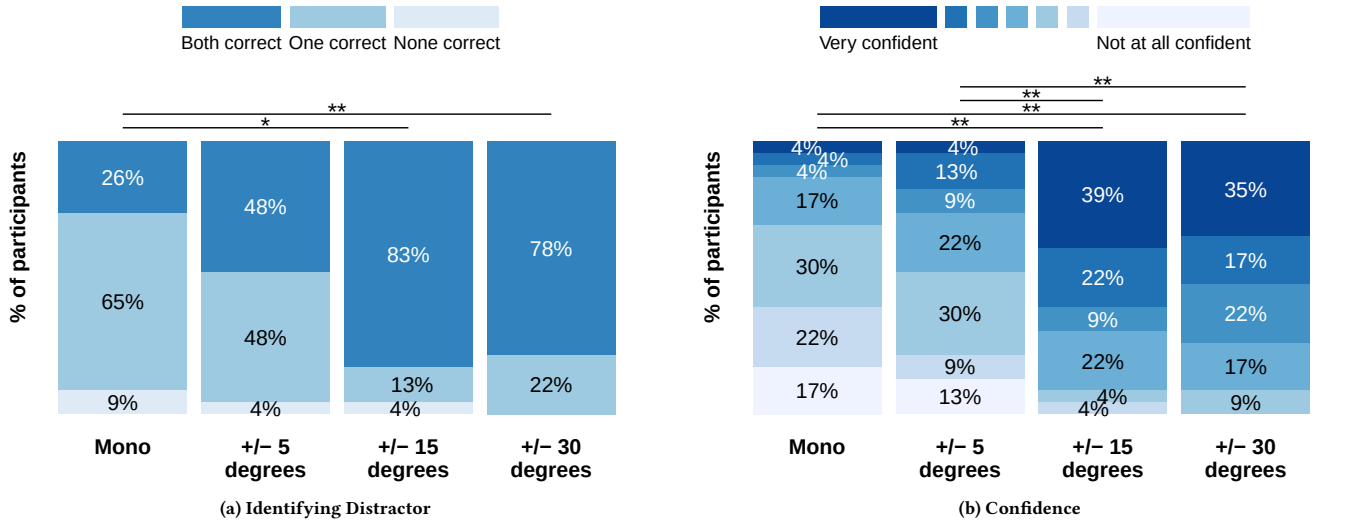
(a) Spatial Ordering (b) AV Correspondence

Figure 7: Spatial ordering ratings and audio/video positional correspondence ratings for four-speaker videos. All ratings range from 1-5. Pair-wise conditional comparisons computed using Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Bonferroni correction. *** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.1.

(2) How did you feel the visual and audio positions corresponded? Did they ever feel matched, mismatched, too narrow, too wide, etc?
(3) How was your experience different between the videos with two speakers and the videos with four speakers?

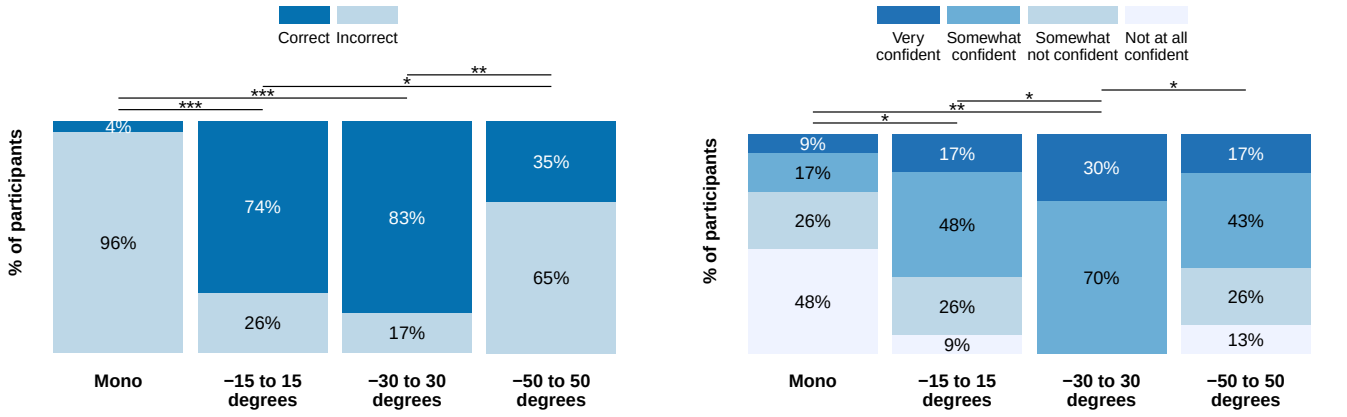
Participants had a range of responses with some common themes emerging. Table 1 collects some repeated sentiments for each question that provides some context for trends found in the post-video questionnaire analyses.

Four responses to Question 1 indicated that spatialization aided with intelligibility and following the conversation flow. These responses align with previous research on spatial audio [2, 26]. Four participants stated that they preferred narrower spatialization in general. Two participants expressed discomfort when listening to "hard-panned" sound (i.e., two voices, each isolated to one ear); interestingly, there are no examples present in this study with hard-panned audio and spatial spread peaks at 50 degrees off azimuth center. This response is complementary to participants who preferred narrow spatialization and suggests that listeners may perceive a horizontal positioning as wider than it actually is. It should



(a) Identifying Distractor (b) Confidence

Figure 8: Distractor identification accuracy and confidence in choice for two-speaker videos. Accuracy (left) is coded as 1 (correct) and 0 (incorrect) and averaged across the two video repetitions for analysis. Confidence (right) ranges 1-4 and is similarly averaged across two repetitions. Pair-wise conditional comparisons computed using Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Bonferroni correction. *** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.1.



(a) Identifying Distractor (b) Confidence

Figure 9: Distractor identification accuracy and confidence in choice for four-speaker videos. Accuracy (left) is coded as 1 (correct) and 0 (incorrect). Confidence (right) ranges 1-4. Pair-wise conditional comparisons computed using Wilcoxon signed-rank tests and adjusted for multiple comparisons using the Bonferroni correction. Degradation of results in the widest condition may be due to the distractor in the corresponding video being harder to identify than in other cases (see Section 4.2.3). *** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.1.

be noted, however, that two responses across all three questions indicated a preference for wider horizontal positions.

Question 2 asked participants to comment on if spatial audio felt matched to the video positioning and whether there were any instances where audio felt too shallow or wide. Eight responses indicated that some videos featured audio that felt too narrow and seven that audio felt too wide in some videos, suggesting that many participants notice a mismatch in the width of the audio stage. Three participants reported that the videos with two speakers had instances that felt too wide, while no one reported audio in the four speaker videos felt too wide. This sentiment suggests that the ideal threshold for two speakers is smaller than that for four speakers, which is consistent with the results found in Section 4.2.2.

Responses to Question 3 provided insight into the differences of experiences in calls with two speakers versus four. Four participants felt that spatialization was better in the videos with four

Table 1: Count of common sentiments in the final questionnaire

	Sentiment	# Resp.
Q1	*Describe overall experience*	
	Easier to follow when spatialized	4
	Preferred narrower spatialization	4
	Sometimes felt like hard-panning	3
Q2	*Overall thoughts on positioning*	
	Sometimes too narrow	9
	Sometimes too wide	7
	2-person often too wide, 4-person often just right	3
Q3	*Difference in 2 and 4 speaker videos*	
	Better in four-person conversations	5
	Easier to tell left/right than exact position	4
	More comfortable in 2, more helpful in 4	2

speakers. In the case of picking the stream with the distraction sound, four responses indicated that spatialization helped determine if the sound came from participants from the left or right side of the video call, but that narrowing it down to one stream was still difficult. This sentiment was confirmed when analyzing whether participants chose the correct hemisphere in the four-person video questionnaires. Two participants stated that spatialization was more comfortable in two-person calls, but more helpful in four-person calls, highlighting listening comfort and better conversational cues as separate spatialization benefits that do not necessarily co-occur.

5 ANALYSIS

5.1 Study 1: In-room attendees

Study 1 focused on the experience of spatial audio using loudspeakers, which is how hybrid meeting participants who are joining together in a room would experience audio. The results showed a preference for the Augmented Multichannel audio condition compared to Mono, with some variation of differences with the Stereo and Stereo Binary conditions. Augmented Multichannel, Binary Stereo, and Stereo were all better than Mono in identifying distractor sounds and overall audio placement; Mono represents random choice in the case of identifying distractor sounds. The Augmented Multichannel condition was better than Mono in ease of determining who was speaking. Augmented Multichannel was better than Stereo Binary, Stereo, and Mono in understanding what was said during overlapping talk. Taken together, the data show a clear preference for the Augmented Multichannel audio condition, which has the widest audio stage that goes beyond the visual placement of the video streams. In the event where advanced multichannel audio is not available, some benefits may still be observed for each seating position even in the moderate upgrade from monophonic to stereo audio.

While some measures that showed a difference between the center and side seating position were not statistically significant, there was some evidence that reflected the difference in their listening experience. The analysis of the sketches showed that the side participant was less accurate in locating the positions of the voices, except for the Stereo Binary condition, which did not help them distinguish between the voices coming from the same side. Furthermore, open-ended comments from the side participants indicated some of the challenges they experienced:

- "The audio that came from the speaker closer to me was generally easier to understand and clearer." [Stereo Binary]
- "Some voices were more overpowering than others due to me sitting closer to one individuals voice rather than another" [Stereo Binary]
- "I had to strain a bit more to hear what was coming from the farther speaker" [Stereo Binary]
- "It also felt weighted towards the speaker close to me, which made it harder to line up" [Stereo]
- "The speaker so close to my seat is really overpowering." [Augmented]

It is important to consider that in a hybrid meeting, not everyone will be able to sit in the center position for ideal spatial audio perception. Especially given the strong evidence for user preference for the widest audio stage offered in the Augmented Multichannel condition, it makes sense to maximize the audio spread of the voices for in-room participants as one way to help mitigate the diminished effects of sitting particularly off-center.

5.2 Study 2: Remote attendees

Study 2 focused on the user benefits of speaker placement in a teleconferencing call when using headsets, which is how many remote participants experience hybrid meetings. Spatial audio ratings increased for remote users wearing headsets as azimuth spread increased, with improvements plateauing before the widest observed audio stage in both two- and four-person conversations. Free-form answers from the final questionnaire also indicated a preference for narrower spread and some even perceived voices as hard-panned in stereo, even though audio was never hard-panned in this study (see Table 1). Taken together, it is reasonable to assume that remote participants in a video call will have an improved experience with spatial audio if the audio stage is chosen to be adequately wide but not so wide as to be distracting. These results can be seen as consistent with research suggesting that intelligibility benefits from SRM plateau in the earlier stages of separation [20]. Furthermore, as benefits in the metrics measured here plateaued, it is better to err on the side of narrower spatialization for remote users.

The addition of spatial audio improved participants' ability to identify streams containing a background noise. Their ability to locate distractor sounds in the case of two speakers increased with more azimuth spread; that trend is extant in responses to four-person videos, with the caveat that responses to videos with the widest treatment in this case were negatively penalized by the study design. User comments illuminate that horizontal spread may assist listeners in telling if sounds came from streams in the left or right hemisphere, but additional localization may be difficult. Identifying the location of sound sources is not possible in monophonic scenarios without contextual clues, so the benefit of spatial audio is trivial in this case; nevertheless, while sound location is only one aspect of user experience in video calls, it can be beneficial for conversational understanding. Spatialized sound in headsets without head tracking can lead to problems not present in mono, such as

ear fatigue on one side when a speaker to one side dominates the conversation, and these problems can be addressed orthogonally to the audio width layout.

Participants preferred a narrower audio stage in videos with two speakers than those with four, a trend identifiable in both the post-video survey responses and the final comments. Many participants commented that spatialization was more beneficial with more speakers, with some commenting that spatialization was not necessary in the case of only two speakers. One implication for engineering audio solutions in video calls is that the maximum allowable azimuth separation for audio positioning may need to grow dynamically with the number of speakers. Fewer speakers necessitates less spatialization, which corresponds more closely with the visual geometry for laptop users. One possible explanation for the preference for wider audio stages with more interlocutors is that users may accept a more unrealistic or exaggerated audio stage with regards to visual geometry as increased spatial width introduces more intelligibility benefits.

5.3 Comparing across both studies

Both studies show benefits of any spatialized sound over mono for both in-room and remote participants. While we could not reliably measure cognitive load in such a short experimental study, the evidence for the ease of identifying distractors and understanding speakers suggests that spatial audio can reduce cognitive load in video calling. However, it was interesting to note the *differences* in the experienced and perceived benefits of spatialized audio between using loudspeakers in the room and using headsets remotely.

The Stereo separation condition in Study 1 for in-room participants is somewhere between the +/15 and +/30 degree spans in Study 2 for remote participants with headsets. The strong preference for the Augmented condition in Study 1 indicates that in-person participants accepted a wider audio stage than headset users. Perhaps the ability of in-room users to freely turn their heads and orient to where sounds were rendered in the room enables them to appreciate the wider audio stage without the negative effects of having sounds fixed in extreme positions, which headset users (without head tracking) would experience. Thus, for remote people, the benefits of spatial audio seemed to plateau at less than the widest audio stage. Plus, headset user comments raised concerns about too wide of an audio stage for the widest condition.

Taken together, our studies suggest a different audio spatialization strategy for the in-room and remote participants that are involved in hybrid video conferences. In-room participants appreciated the widest audio stage where the audio streams were located beyond the visual location of the video windows. However, people using headsets preferred a narrower audio stage that provided the benefits of audio separation, but avoided complaints of audio coming from too extreme positions. These results indicate that system designers should be cognizant to these differences and implement distinct strategies for each use case to improve the experience for all users.

5.4 Limitations and future work

We point out several limitations of our studies that identify opportunities for future work. Because these were two semi-independent studies, they did not always offer direct comparisons between in-room and headset users. There was enough overlap to demonstrate similarities and differences between the two settings, but it was not always a direct comparison. The stimuli used in both studies were simulated conversations, where the study participants were "spectators" to a conversation and did not interactively participate in the conversations. While we believe the standard stimuli allow for exploring the effect of spatializing the speech streams, additional effects may be introduced when people actively participate in conversations.

We only looked at 2- and 4-person meetings, which were enough to see differences in spatial layout, but do not cover larger meetings and a potential ceiling effect of the benefits of spatial audio. More research is neeeded on spatial audio layouts for larger meetings. Furthermore, we note that our participants were from one high tech company, and while we are not aware of ways in which they would be unrepresentative of a broader population, more diversity in the study participants would be valuable. Study 1 was limited to our company because of COVID-19 policies in place when planning the study, since people needed to come in to the same room as a pair for the study. Study 2 was limited to internal recruitment for readily deploying code for the study.

The lack of significant differences between centered and off-center participants in Study 1 is surprising, as the “sweet spot" problem in loudspeaker spatial audio setups is well known [1] and previously observed in experiments with loudspeaker-based spatial audio in teleconferencing [11]. It is possible that the conditions in Study 1 were insufficient to reproduce the problems anticipated for the off-center participant based on previous works. However, video teleconferencing is a unique application among those usually explored in previous works (such as concerts or interactive audio/visual art pieces) and may therefore be more robust to a lack of immersion caused by improper spatialization. Previous research that observed problems with spatial audio for in-person teleconferencing used different sound synthesis techniques than used here, which may also explain the difference in results [11]. Further research could help illuminate the problem of “sweet spots" and teleconferencing.

As future work, we anticipate exploring the use of head tracking techniques to fix the position of sound sources in the environment. While speaker placement was not the focus of inquiry, earlier explorations of teleconferencing and binaural audio tend to favor a wider spread of speakers, closer to those present in the augmented multichannel setup in Study 1 [26], and often used head tracking. Binaural audio without head tracking is currently the most commonly available solution, but solutions with head tracking are likely to increase in prevalence. Head tracking would enable headset users in a teleconference call to turn their head to find a more comfortable listening position, perhaps enabling an even wider sound stage and more closely matching the in-room experience.

6 CONCLUSION

Taken together, both studies look at various spatial arrangements of voices in a video conference meeting for people gathered together in a room using loudspeakers, and people joining remotely using headsets. These two settings are important in the growing

occurrence of hybrid meetings as workers begin to return to the office as we emerge from the pandemic. Both studies found advantages of spatializing the voices in a video conference compared to mono. People in the room using loudspeakers perferred the widest audio stage where the voices extended +/- 55° from center, well beyond the visual positions of the video images. Remote people using headsets appreciated a wider audio stage up to a point, which varied as the number of meeting participants increased. Balancing the benefits of spreading out the voices wider with the qualitative feedback of distraction if the spread was too wide led to a preference for a narrower audio stage for headset users. This difference in preference suggests a different user experience for hybrid meeting participants depending on whether they are joining in the room or remotely. Offering a wider audio stage for those in the room but a narrower stage for remote headset users would reflect the different preferences for those sites without creating confusion in the user experience of the overall meeting. We see an opportunity for designing spatial audio in video conferences in a site-specific way that can improve the user experience in hybrid meetings, and potentially mitigate some of the cognitive effort expended in participating in video conferencing meetings in the long run. Spatial audio rendering should be considered alongside (and particularly in relation to) visual rendering methods to improve intelligibility, conversational flow and reductions in "Zoom fatigue" in teleconferencing, both in current screen-centric calls and emerging technologies such as virtual or mixed-reality conferencing.

ACKNOWLEDGMENTS

We thank the Experiences + Devices organization at Microsoft for providing funding for the internships that contributed to this research. We thank Greg Baribault for partnering as an intern mentor, and Mike Winters for advising in experimental design. We thank Sasa Junuzovic, Justin Kilmarx and Dimitra Emmanouilidou for contributing to the recorded video samples used in these studies.

REFERENCES

[1] Ronald M Aarts. 1993. Enlarging the sweet spot for stereophony by time/intensity trading. In *Audio Engineering Society Convention 94*. Audio Engineering Society.

[2] Jessica J Baldis. 2001. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 166–173.

[3] John G Beerends and Frank E De Caluwe. 1999. The influence of video quality on perceived audio quality and vice versa. *Journal of the Audio Engineering Society* 47, 5 (1999), 355–362.

[4] Augustinus J Berkhout, Diemer de Vries, and Peter Vogel. 1993. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America* 93, 5 (1993), 2764–2778.

[5] Paul Bertelson and Monique Radeau. 1976. Ventriloquism, sensory interaction, and response bias: Remarks on the paper by Choe, Welch, Gilford, and Juola. *Perception & Psychophysics* 19, 6 (1976), 531–535.

[6] Jens Blauert. 1997. *Spatial hearing: the psychophysics of human sound localization*. MIT press.

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

[8] Edgar Brunner, Sebastian Domhof, and Frank Langer. 2002. *Nonparametric analysis of longitudinal data in factorial experiments*. Vol. 373. Wiley-Interscience.

[9] Jerry Brunner. 2019. Repeated measurement analysis of binary responses. http://www.utstat.toronto.edu/~brunner/workshops/mixed/

[10] E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25, 5 (1953), 975–979.

[11] Werner Paulus Josephus De Bruijn. 2004. Application of wave field synthesis in videoconferencing. (2004).

[12] Edina Fintor, Lukas Aspöck, Janina Fels, and Sabine J Schlittmeier. 2022. The role of spatial separation of two talkers' auditory stimuli in the listener's memory of running speech: listening effort in a non-noisy conversational setting. *International Journal of Audiology* 61, 5 (2022), 371–379.

[13] Justin T Fleming, Ross K Maddox, and Barbara G Shinn-Cunningham. 2021. Spatial alignment between faces and voices improves selective attention to audio-visual speech. *The Journal of the Acoustical Society of America* 150, 4 (2021), 3085–3100.

[14] William G Gardner and Keith D Martin. 1995. HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America* 97, 6 (1995), 3907–3908.

[15] Michael A Gerzon. 1985. Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society* 33, 11 (1985), 859–871.

[16] Jackson Montgomery Goode. 2021. *Toward a Telepresence of Sound: Video Conferencing in Spatial Audio*. Master's thesis.

[17] Kori Inkpen, Rajesh Hegde, Mary Czerwinski, and Zhengyou Zhang. 2010. Exploring spatialized audio & video for distributed conversations. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 95–98.

[18] Gary L. Jones and Ruth Y. Litovsky. 2011. A cocktail party model of spatial release from masking by both noise and speech interferers. *J. Acoust. Soc. Am.* 130, 3 (2011), 1463–1474. https://doi.org/10.1121/1.3613928

[19] Setsu Komiyama. 1989. Subjective evaluation of angular displacement between picture and sound directions for HDTV sound systems. *Journal of the Audio Engineering Society* 37, 4 (1989), 210–214.

[20] Ruth Y Litovsky. 2012. Spatial release from masking. *Acoust. Today* 8, 2 (2012), 18–25.

[21] Telecommunication Standardization Sector of ITU. 2017. *Spatial Audio Meetings Quality Evaluation, Document ITU-T Rec. P.1310*. International Telecommunication Union, Geneva, Switzerland.

[22] Ville Pulkki. 1997. Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society* 45, 6 (1997), 456–466.

[23] Alexander Raake, Markus Fiedler, Katrin Schoenenberg, Katrien De Moor, and Nicola Döring. 2022. Technological Factors Influencing Videoconferencing and Zoom Fatigue. https://doi.org/10.48550/ARXIV.2202.01740

[24] Alexander Raake and Claudia Schlegel. 2008. Auditory assessment of conversational speech quality of traditional and spatialized teleconferences. In *ITG conference on voice communication [8. ITG-Fachtagung]*. VDE, 1–4.

[25] Loïc Rosset, Hamed Alavi, Sailin Zhong, and Denis Lalanne. 2021. Already It Was Hard to Tell Who's Speaking Over There, and Now Face Masks! Can Binaural Audio Help Remote Participation in Hybrid Meetings?. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[26] Janto Skowronek and Alexander Raake. 2015. Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls. *Speech Communication* 66 (2015), 154–175.

[27] Janto Skowronek, Alexander Raake, Gunilla Berndtsson, Olli S Rummukainen, Paolino Usai, Simon NB Gunkel, Mathias Johanson, Emanuël AP Habets, Ludovic Malfait, David Lindero, et al. 2022. Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey. *IEEE Access* (2022).

[28] Matthew Wong and Ramani Duraiswami. 2021. Shared-Space: Spatial Audio and Video Layouts for Videoconferencing in a Virtual Room. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 1–6.