Hear We Are: Spatial Audio Benefits Perceptions of Turn-Taking and Social Presence in Video Meetings

Kate Nowak
kate.nowak@microsoft.com
Microsoft
Redmond, WA, USA

Lev Tankelevitch
lev.tankelevitch@microsoft.com
Microsoft Research
Cambridge, UK

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

Sean Rintel
serintel@microsoft.com
Microsoft Research
Cambridge, UK

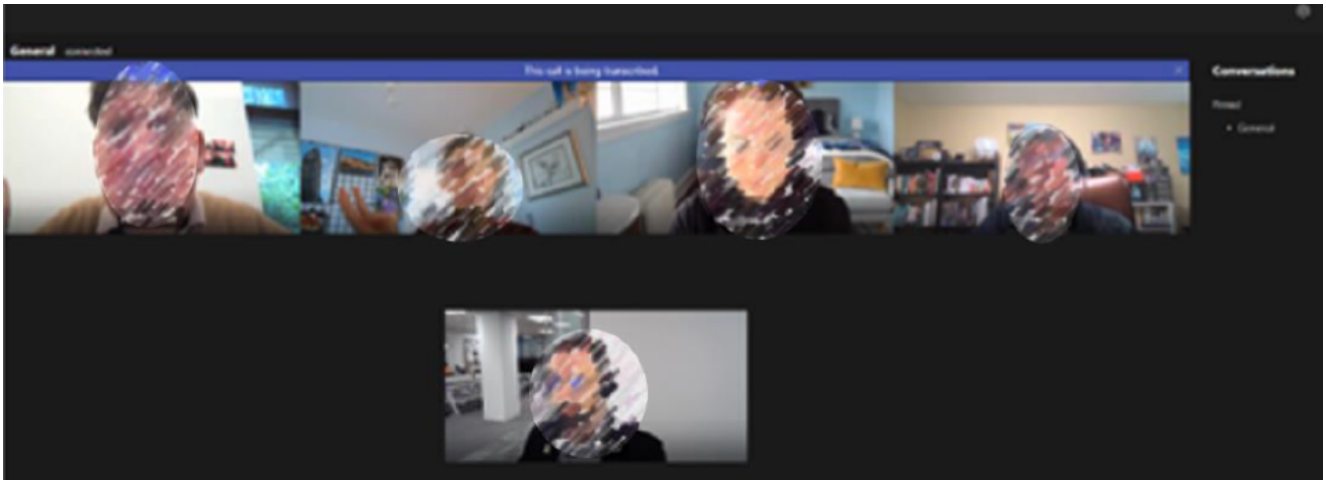

Figure 1: Video meeting layout used for spatial audio study. Self-view shows underneath a line of all other participants. When spatial audio was enabled, the left to right visual position of participants matched the left to right placement on the audio stage.

ABSTRACT

Relative to in-person meetings, conversations in video meetings have long been reported as stilted. Spatial audio in video meetings can simulate the way we hear the world by separating audio streams based on speakers' virtual locations. We report on a within-subject experiment in which 75 employees of a global technology company completed two group survival tasks with spatial audio enabled or disabled. Spatial audio increased perceptions of interactivity, shared space, and ease of understanding. Women experienced effects for social presence while men experienced effects for turn-taking. We discuss implications for inclusion, task performance, fatigue, and future research.

CHIWORK 2023, June 13–16, 2023, Oldenburg, Germany

ACM ISBN 979-8-4007-0807-7/23/06…$15.00
https://doi.org/10.1145/3596671.3598578

CCS CONCEPTS

• **Human-centered computing Empirical studies in HCI**; • **Human-centered computing Collaborative interaction**; • **Applied computing Sound and music computing**;

KEYWORDS

spatial audio, video meetings, turn-taking, social presence, gender, fatigue, task outcomes

ACM Reference Format:
Kate Nowak, Lev Tankelevitch, John Tang, and Sean Rintel. 2023. Hear We Are: Spatial Audio Benefits Perceptions of Turn-Taking and Social Presence in Video Meetings. In *Annual Symposium on Human-Computer Interaction for Work 2023 (CHIWORK 2023), June 13–16, 2023, Oldenburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3596671.3598578

1 INTRODUCTION

Video meetings played an important role in preserving productivity during the COVID-19 pandemic for many people who were working remotely [18]. However, increased usage of video came with costs, especially video meeting fatigue [44] stemming from both overuse and long-standing technical constraints that make

natural conversation difficult. In the post-pandemic hybrid work era in which some people work remotely all or some days per week, improving video meetings continues to be important to support distributed collaboration. Of these improvements, enabling easier conversational turn-taking with better audio is fundamental, as in most cases meetings can continue if video is disrupted but less so if audio is disrupted [45, 62].

The human auditory system is designed to localize the sources of individual sound streams and this capability helps people identify speakers, direct attention, and make decisions regarding when to speak [15]. However, the use of mono or stereo audio in most video meeting systems is vastly different from our in-person experience. Spatial audio consists of variable-volume audio streams spatially separated based on user-controlled location in a virtual room. This simulates the way sound is perceived by humans in a natural environment. Commercial attention has turned to spatial audio in video calls (e.g. Apple FaceTime). While there is research into the value of spatial audio in video games (e.g. [54]), virtual environments (e.g. [64]), and films (e.g. ([63]), there is sparse modern research to validate what aspects of conversation and feelings of social presence spatial audio may improve in video meetings. The few studies that exist (e.g. [2, 27, 46] tend to use simulated meetings to control the sound, and also tend to investigate comprehension and memory more than presence and turn-taking.

As such, this paper reports a controlled field experiment of spatial audio in video meetings using a within-subject design. Employees from a globally-recognized technology corporation were assigned two group survival tasks - one with spatial audio enabled, and another with spatial audio disabled. The results indicate that those who experienced and engaged with spatial audio reported an increased sense of interactivity, shared space, and ease of understanding. Furthermore, some gender-specific effects were observed, with women noticing greater effects for social presence and men experiencing benefits in relation to turn-taking. Our study presents implications related to virtual meeting fatigue, inclusion, and task performance, highlighting the need for further research.

2 PRIOR WORK

Taking turns is fundamental to conversation, and the main properties of the turn-taking system are that turn size and order vary, one person speaks at a time, speaker transitions have no or slight gaps, speaker overlaps are common but brief, and the system works without visual contact [48]. Turn-taking cues include the verbal, prosodic, breathing, gaze, gestural, and body posture [31]. Verbal cues provide the strongest basis for predicting turn-transition, complemented by prosodic voice cues (intonation, loudness, and rate) that help deal with ambiguities [56]. While smooth conversation is understood to be 'no-gap-no-overlap' between speaker turns, gaps and pauses can provide important time for thinking during cognitively effortful conversations, and, on the other hand both competitive and collaborative overlaps have important uses in conversations [50]. Competitive overlaps are disruptive attempts to take over the conversational floor while cooperative overlaps tend to be encouraging of the current speaker.

Selective listening is also part of the human auditory system, and while its actual role in turn-taking is not well understood, the ability to selectively attend and identify speakers is likely to play an important role. First coined by Cherry [9], the "cocktail party problem" was defined as being able to understand "what one person is saying when others are speaking at the same time". Cherry [9] found that listeners who are played different sound streams into each ear can choose which to listen to and cannot describe what was played in the unattended ear. Moray [35] found that though the power of selective attention to reject the content of a sound stream is very strong, a few things may 'break through', such as hearing your own name. This became known as the "cocktail party effect". As Blauert [7] notes, this isn't just about physics - choice is important too: to a certain extent we can choose to listen to a chorale's sound mass or choose to listen to individual voices.

Given the above, disruptions to the mechanics of turn-taking lead to stilted conversations [53]. Even small delays can distort cues needed to create shared meaning and mutual understanding [47]. Poor quality audio is rated worse than video issues [62] and people are more likely to attempt to repair audio than video perturbations in video meetings [45]. Poor audio quality is often cited as the main culprit for turn-taking problems in video meetings. In two seminal studies comparing turn-taking in video-mediated, audio-only, and in-person meeting scenarios [51, 52], video-mediated conversations were found to be more formal and created the feeling of being 'distanced' from others. The liveliest conversations were in-person with naturally spatial sound, featuring more competitive and cooperative overlaps and fewer formal handovers than those with full duplex and imperceptible lag.

The Hydra system used in the Sellen [51, 52] studies used small separate AV devices for three remote participants, placed on a desk in front of a fourth participant, to represent a 4-way round-table meeting and create a natural spatial sound effect. Hydra conversations did not show significantly different turn-taking features than the other conditions, but observations showed participants having side and parallel conversations, and participants reported enjoying selective listening and gaze. More akin to current video meeting systems, Inkpen et al. [25] set up a three-way system with remote participants on the left and right with their sound streamed from left and right speakers respectively. This system also did not show significant differences from mono condition [25]. Visual spatial separation had a greater effect. This may be due to the use of open speakers, which may have weakened the spatial audio effect in comparison to headphones. It may also have been due to participants being acquainted and thus able to identify one anothers' voices, further weakening the value of spatial audio. However, some participants reported that separation helped identify the source and that there was no delay when people were talking simultaneously. The SharedSpace [65] web-based spatial audio video meeting system is reported as receiving "positive" ratings compared to standard monaural systems, but an empirical report is not provided.

Spatial audio has received somewhat more attention in virtual environments, as it is one of several factors that researchers explore in order to enable virtual social presence. Social presence in a technology-mediated context is a psychological state defined as "the degree to which a user feels access to the intelligence, intentions, and sensory impressions of another" [4, 29]. The greater the feeling of presence, the greater the chance human behaviors in a virtual environment will resemble those observed in the real

world [26]. Most social presence studies on immersive technologies have focused on manipulating modality and type of visual stimuli [42], but a meta-analysis shows that sound quality provides a small to medium-sized effect on presence [11]. Dicke et al. [14] show that in an audio-only system, spatial sound is the best audio-based method for creating a strong sense of social presence. Research on immersive gaming and social environments shows that spatial audio improves spatial presence and social richness [22, 26, 55]

Research on the effects of spatial audio on task performance is very limited. Collaboration research has shown the spatialization of a remote participant's voice based on the location of that person's rendered image improved the local participant's sense of engagement in a collaborative experience [67]. Another related study shows that meeting success is correlated with the overall sound of meetings: subjectively effective meetings are short and matter of fact, whereas objectively productive meetings are longer and have a lively speech melody [40]. Much of the existing relevant audio research has centered on auditory signals to aid navigation or help locate objects in an environment [10, 66]. Studies have shown that spatial sound in a 3-D or augmented environment can improve the ability to detect visual targets [34, 67]

As noted above, direct research on the effect of spatial audio in video meetings is quite sparse. Early studies of simulated audio-conferencing scenarios show a benefit of spatial audio for comprehension and memory [2, 27]. Rosset et al. [46] set up a simulated hybrid meeting study, with remote participants watching pre-recorded discussions of co-located actors and found that participants' comprehension was improved with binaural audio, particularly when speakers' mouths were obstructed with a facemask.

The broad picture suggests an accord with the sociological, linguistic, and psychoacoustic findings that the spatiality of sound is important to in-person human conversation. If monaural or simple mixed stereo systems strip away spatial sound cues, then restoring them via spatial audio should, in turn, make mediated encounters more naturalistic and potentially improve meeting outcomes. Thus, our research was designed to explore whether spatial audio compared to mono audio in video meetings improves turn-taking, social presence, and group task performance.

3 METHODOLOGY

This study was a controlled within-subjects experiment. Participants met remotely in assigned groups to complete two survival tasks in video meetings under conditions of mono audio and spatial audio. Participants completed post-task questionnaires after each meeting. Ethics approval was granted for the study. While ecological validity would have been improved with real teams having real meetings, we are not aware of any baseline studies of spatial audio in meetings that do not use simulated data. Further, gathering enough data from real teams to establish validity would be extremely difficult. Thus we felt that the generation of data using naturalistic conversation within known parameters was the most important baseline to be established.

3.1 Procedure

The key factor of interest was spatial audio: participants engaged in two 30-minute group video meetings completing a survival task (see Section 3.4.1) with spatial audio either on (spatial audio) or off (mono, i.e., the standard video meeting experience). The experimental design was within-subjects, meaning all participants experienced both audio conditions and performed both tasks. This was done to minimize variance related to differences in network, device configuration, and audio latency. To minimize task ordering and learning effects, groups were counterbalanced in the order of audio condition and task they experienced. Since group members did not know each other, the first task included a warm-up exercise to stimulate conversation and help participants become comfortable with one another.

Participants were given 15 minutes to independently complete a questionnaire after each meeting. The sessions were not moderated, and calendar invitations were used to help participants stay on schedule. Meeting links were added to calendar invitations the morning before the study began, which determined whether spatial audio was turned on or off during the meeting; nothing in the user interface indicated this to participants. Study instructions were emailed to participants days in advance, detailing how to log on to the meeting experience. Survival task instructions were sent 15-minutes before each meeting and survey links were added to the calendar invitations during the meetings.

3.2 Participants

An initial recruitment pool of 213 information workers in the United States working for a large global technology company were randomly assigned to 40 groups of 4-6 people to complete the two survival tasks. After participant attrition due to failure to complete the survey, data loss, and meetings with significant technical challenges (e.g., loss of video and audio perturbations), the final data for analysis included 75 participants, representing 15 meeting groups and 30 meetings. For the final analysis sample, demographic data sourced from consent forms showed that participants included 25 women and 49 men representing ages ranging from early twenties to seventies and both managers and individual contributors. Participants came from a variety of functions, including engineering, sales, and product management. Participants were all familiar with video meeting tools and computer literate.

3.3 Spatial Audio Application

An experimental web-based video meeting application was developed to test spatial audio in this study (see Figure 1). No participants had experienced this experimental system, although it was similar to most standard video meeting applications. Its spatial audio feature relied on a combination of Web Audio API built into web browsers and Resonance Audio SDK with Opus high bitrate (300k) audio processing. The typical spatial layout was constructed with users' emitters placed along an angular separation extending 60-degree horizontally, 45 degrees vertically on a sphere (with radius of 1.5 meters). The listener was placed in the centre of the sphere and each audio emitter was 1.5 meters (radius) from the listener so that each emitter had the same volume. The listener's position within the room was slightly off-centre to aid in the stereo imaging of centred sources. This logic was run on each user's web client.

The application spatialized audio by putting users' audio in a virtual room which simulated the acoustics of a physical room.

Users' locations on the screen translated to where they were placed in the virtual room. An ego-centric layout was created for this study to optimize spatialization in a 2-D video meeting environment. Four or five participants were placed in a row at the top of the screen and the self was centred in the bottom row (See Figure 1) This layout was more reflective of sitting around a table and another reason for the meeting group size used in the study.

To maximize the spatial audio effect and to provide some consistency, all participants were required to use wired headphones (those who did not have wired headphones were sent a pair to use). However, in the interests of ecological validity ([38] but cf. [24]), by design this was a self-managed pseudo field-study with participants distributed unevenly, using a range of computing devices, and using variable network connections. Without wanting to over-claim or over-generalize, we believe that the results are more valuable given that spatial audio made significant difference even under non-ideal conditions.

3.4 Tasks

3.4.1 Survival Problems. Study participants collaborated on a classic survival problem in each condition, involving a hypothetical survival scenario either on the moon or in the desert [21, 68]. Survival problems were developed for social psychology research into group effectiveness, with the view to systematically improve the efficacy of team development exercises [29]. Hall and Watson [19] devised the NASA Moon Survival task in the 1960s to explore ways to reduce the confusion, frustration, and time-loss frequently associated with team group activities. Survival problems are commonly used in research on social presence and turn-taking because they induce features of normal conversation by creating natural turn exchanges [8]. A typical survival task provides the narrative of a scenario that sets context and essential clues, and then a list of items from which participants must choose a subset of the most important to group survival in the context. Usually the answers have been ranked by experts from most to least important (as in NASA Survival! curriculum [37]). To fit the within-subjects experimental design for this study, two similar scenarios with different contexts were chosen, in this case desert survival and moon survival [21, 68].

Groups were asked to work together to select the top three items from a list of ten in terms of their importance in allowing the group to survive and be rescued. The list did not include the top three items according to expert ratings, which were instead given to each group in the scenario descriptions to reduce any potential advantage of prior task experience. This design decision and another to remove overlaps from the lists were made to further ensure that the tasks would require critical thinking. Participants recorded their group's selections in the post-task questionnaire. Group answers were scored based on the expert ratings (top-rated item received ten points, the next nine etc.). The total score was calculated by summing the points attributed to each of the top three items.

3.4.2 Post-Task Questionnaires. Participants independently completed a questionnaire after each task, in which they indicated their group's survival task selections, used to evaluate task performance (H3), and completed measures of perceptions of turn-taking (H1) and social presence (H2). These were measured using statements rated by participants on a 7-point Likert scale ranging from 1 = strongly disagree to 7 = strongly agree. Turn-taking was measured using items taken from Sellen [51] which address conversation quality and ease of turn-taking. Statements rated by participants included: "This was a natural conversation" and "The conversation seemed highly interactive" (see Table 1 for details). Social presence was measured using the Networked Minds Social Presence scale and the Temple Presence Inventory [5, 20, 33]. Statements included: "I felt as if I were sharing the same space as the group" and "I paid close attention to others when they were speaking" (see Table 2 for details).

3.5 Variables and Analysis

In addition to spatial audio as an independent variable, gender was used as a moderator given its importance in turn-taking and inclusion in the workplace. The gender inequalities in face-to-face meetings pre-pandemic seem to have been exacerbated in virtual meetings; a study conducted during the pandemic found women have more difficulty than men speaking up in virtual meetings [59]. The overall human turn-taking system is strongly universal, but it can be culture- and context- specific [31]. In this study, culture and setting were not independent variables because culture did not vary significantly within the study population and the meeting context was constant.

The Likert scale questionnaire data were analyzed using Wilcoxon signed-ranked tests comparing the two audio conditions. Survival task score data were continuous and analyzed using paired t-tests, comparing the spatial and mono audio conditions. Due to logistical constraints, the only qualitative data collected was in an open response question at the end of the surveys, and the (fairly limited number of) responses were simply categorised for positive/negative valence. Some of these are quoted below, but, due to the limited data, they do not form a major part of the findings.

4 FINDINGS

4.1 Turn-taking

A descriptive analysis showed the means of all turn-taking variables were consistently higher in the spatial audio condition than the mono audio condition except for S6 regarding perception of few uncomfortable pauses (See Table 1) The higher means served as a preliminary indication that spatial audio improved turn-taking overall. Using the non-parametric Wilcoxon signed-rank test, survey statement S5, "The conversation seemed highly interactive", was statistically significant at the 5% level (z=-2.281, p=0.023). Survey statement S1, "I found it easy to participate in the conversation", was significant at the 10% level (z =-1.782, p=0.075) (See Table 1)

Comparing the survey responses of women to those of men, women (n=25) experienced a larger improvement with spatial audio than men (n=49) in their ability to selectively attend to one person at a time at the 10% level using the Wilcoxon signed-rank test and gender as a moderator (survey statement S7, change in mean of 0.40 vs 0.22; z=-1.76, p=0.084). Compared to women, men experienced a larger improvement with spatial audio in their perceived ease of participation (S1, 0.30 vs. 0.04; z=-2.033, p=0.042), and perception of high interactivity (S5, 0.33 vs 0.28; z=2.021, p=0.043).

Table 1: Differences between audio conditions for turn-taking (n=75, ** indicates statistically significant at the 1% level, * at the 5% level, + at the 10% level)

Survey Statement	Mono: Mean (SD)	Spatial: Mean (SD)	Wilcoxon Signed-rank Test
S1: I found it easy to participate in the conversation.	6.31 (1.11)	6.53 (0.78)	z = -1.782 p = 0.075 +
S2: I was able to take control of the conversation when I wanted to.	6.20 (1.12)	6.29 (0.88)	z = -0.311 p = 0.756
S3: There were few inappropriate interruptions.	4.75 (2.38)	4.90 (2.31)	z = -0.659 p = 0.510
S4: This was a natural conversation.	6.15 (1.02)	6.33 (0.95)	z = 0.009 p = 0.992
S5: The conversation seemed highly interactive.	6.24 (1.08)	6.55 (0.68)	z = -2.281 p = 0.023 *
S6: There were few uncomfortable pauses.	4.24 (2.48)	3.92 (2.51)	z = 0.869 p = 0.385
S7: I could selectively attend to one person at a time.	5.56 (1.60)	5.84 (1.43)	z = -1.453 p = 0.146
S8: It was easy to keep track of the conversation.	6.19 (1.11)	6.35 (0.94)	z = -1.103 p = 0.270

4.2 Social Presence

As above, data for H2 were analyzed using the non-parametric Wilcoxon signed-rank test. Descriptive statistics revealed higher means in the spatial audio condition than the mono audio condition for all social presence items except for S16 regarding others' perceptions of feelings. The differences between the mono and spatial audio conditions for S9, "I felt as if I were sharing the same space as the group", and S13, "It was easy to understand the thoughts of others in the group", were statistically significant (z=-2.687, p=0.007, and z=-2.405, p=0.016 respectively) (See Table 2).

Comparing the survey responses of women to those of men, women experienced a larger improvement with spatial audio in their perceived ability to be understood by others, significant at the 10% level (S14, 0.32 vs. -0.02; z = 1.753, p = 0.080), and to reciprocate when others spoke at the 5% level (S19, 0.36 vs. 0.06; z = -2.162, p = 0.031). While the change in mean in perceived ability to understand the thoughts of others was greater for women compared to men, the positive change men experienced in spatial audio was significant at the 10% level (S13, 0.36 vs. 0.24; z = -1.939, p = 0.053).

4.3 Task Performance

Total scores and number of correct answers chosen in the survival problems were overall higher in the spatial audio compared to the mono audio condition, except for total scores in the moon problem (See Table 3). However, these differences were not statistically significant using the paired samples t-test. For transparency, Table 3 shows paired t-test comparisons separately for the desert and moon tasks. However, given that task scores are calculated at the group level, these analyses are underpowered and should be interpreted with caution. In the moon scenario, total score was lower with spatial audio compared to mono audio, prompting further analysis. Total scores for the moon task were on average higher than those of the desert task. The difference in scores was statistically significant when comparing desert and moon (t=-4.440, p<0.001). This indicated that the desert task was harder than the moon task. It is therefore possible that spatial audio benefits particularly challenging tasks that require more dynamic conversations (i.e., the desert task), and has no effect or a potentially distracting effect in easy tasks (i.e., the moon task).

To test this, the Wilcoxon signed-rank statistical tests performed for H1 and H2 were repeated using survival scenario as a moderator. For the harder desert task, participants experienced statistically significant improvements with spatial audio for perceptions of high interactivity (survey statement S5, z=-2.096, p=0.036), ability to selectively attend to one person at a time (S7, z=-2.037, p=0.042), and ability to understand the thoughts of others (S13, z=-2.017, p=0.044). For the easier moon task, the only statistically significant improvement with spatial audio was for participants' perception of fewer uncomfortable pauses (S6, z=1.994, p=0.046).

4.4 Summary of Results

Spatial audio appears to improve the perception of turn-taking in video meetings compared to similar meetings with mono audio. Specifically, spatial audio induced a statistically significant improvement in participants' perceptions of the conversations as highly interactive, compared to the mono audio condition. This is important given the common complaint that video meetings are less interactive than in-person meetings. Our results also show that spatial audio appears to increase participants' feelings of social presence in video meetings compared to similar meetings with mono audio. Spatial audio induced a statistically significant increase in participants' feeling of sharing the same space and their perceived ability to understand the thoughts of others.

Exploratory analyses also suggested interesting gender differences. In the spatial audio condition, women experienced significant improvements in measures of social presence, particularly in their perceived ability to reciprocate when others spoke, while men experienced significant improvements in turn-taking, specifically in their perceived ease of participation and perception of high interactivity. Notably, there were marginally statistically significant improvements with spatial audio in men's perceived ability to understand the thoughts of others and women's perceived ability to be understood by others. These results suggest that spatial audio might play a role in making video meetings more inclusive for women because while men experience a boost in ease of participation, women feel more present in conversations and potentially understood. These findings need to be explored more deeply, especially considering the potential of video meetings to reinforce gender bias [13].

On average, groups performed better on the tasks in the spatial audio condition. However, these results were not statistically significant. In an exploratory analysis, when focusing on the results of the harder desert tasks, statistically significant improvements were found in the spatial audio condition compared to the mono

Table 2: Differences between audio conditions for social presence (n=75, ** indicates statistically significant at the 1% level, * at the 5% level, + at the 10% level)

Survey Statement	Mono: Mean (SD)	Spatial: Mean (SD)	Wilcoxon Signed-rank Test
S9: I felt as if I were sharing the same space as the group.	5.60 (1.48)	6.01 (1.33)	z = -2.687 p = 0.007 **
S10: My presence was obvious to others in the meeting.	6.25 (1.04)	6.35 (1.08)	z = -0.869 p = 0.385
S11: I paid close attention to others when they were speaking.	6.25 (1.12)	6.45 (0.93)	z = -1.538 p = 0.124
S12: I felt as though people were paying close attention to me when I was speaking.	6.31 (0.99)	6.39 (1.03)	z = -0.819 p = 0.413
S13: It was easy to understand the thoughts of others in the group.	6.23 (1.06)	6.51 (0.67)	z = -2.405 p = 0.016 *
S14: Other group members understood my thoughts.	6.33 (0.92)	6.43 (0.76)	z = -0.675 p = 0.500
S15: I could tell how other group members were feeling.	5.61 (1.37)	5.84 (1.23)	z = -1.114 p = 0.265
S16: Other group members could tell how I was feeling.	5.53 (1.53)	5.51 (1.47)	z = -0.741 p = 0.459
S17: My feelings influenced the mood of the group interaction.	5.61 (1.41)	5.73 (1.24)	z = 0.693 p = 0.488
S18: The feelings of other group members influenced the mood of the group interaction.	5.45 (1.54)	5.62 (1.46)	z = -0.747 p = 0.455
S19: I reciprocated when other group members spoke.	6.13 (1.23)	6.29 (0.87)	z = -0.779 p = 0.436
S20: Other group members reciprocated when I spoke.	6.21 (1.07)	6.31 (0.93)	z = -0.656 p = 0.512

Table 3: Differences between Mono Audio and Spatial Audio Conditions. (N refers to number of scores at the group level, ** indicates statistically significant at the 1% level, * at the 5% level, + at the 10% level)

Survival Problem	Mono: Mean (SD), N	Spatial: Mean (SD), N	Paired Samples T-test
Both tasks combined			
Total Score	20.87 (5.48), 15	21.47 (3.89), 15	t = -0.2902
Number of Correct Ratings	0.33 (0.49), 15	0.47 (0.74), 15	p = 0.776
Desert			
Total Score	17.38 (5.13), 8,	19.14 (3.98), 7	t = -2.838
Number of Correct Ratings	0.38 (0.52), 8	0.43 (0.79), 7	p = 0.025 *
Moon			
Total Score	24.86 (2.12), 7	23.50 (2.56), 8	t = 3.477
Number of Correct Ratings	0.29 (0.49), 7	0.50 (0.76), 8	p = 0.013 *

audio condition for perceptions of high interactivity, ability to selectively attend to one person at a time, and ability to understand the thoughts of others. These results are noteworthy because they indicate differences in the impact of spatial audio on meeting outcomes depending on the difficulty or requirements of the task (e.g., whether it requires critical thinking, socializing, or brainstorming). However, given their exploratory nature, future research is necessary to confirm these findings.

There was positive feedback in open-ended survey responses on the spatial audio condition. This included similarity to in-person sound, such as *"it was more natural"*, *"it felt real"*, as well as what in-person sound affords, e.g. *"it was easier for my brain to figure out who was saying something"*, *"we could talk at the same time and still understand each other."* One woman participant summarized the overall effect as both comfortable and inclusive: *"It felt like a natural conversation. It flowed very well. People did not talk over one another, which was nice. It was the closest I've felt to being in the same room as others in a virtual meeting. I walked away happy and felt like I was able to retain information from the meeting. I felt like I had been heard."*

Some feedback in the open-ended responses included suggested improvements to the sound design, e.g. *"improvements to positional alignments [are] needed,"* and *"the sound was too high quality."* However, there was also some negative feedback. Less than half of study participants preferred the spatial audio condition, and almost an equal amount preferred the mono audio condition. The latter group felt that spatial audio was more a gimmick than a necessary feature, e.g. *"[it] was cool, but probably not something I would enable,"*. In addition, many participants were not comfortable with using headphones. This represents a serious obstacle for spatial audio technologies, many of which work best when spatial sound is isolated from ambient noise.

5 DISCUSSION

5.1 Theoretical Implications

This study, though modest in scope and findings, provides some of the first modern direct evidence that spatial audio has a positive effect on people's perceptions of their ability to engage in video meetings. It supports the broader argument that spatial audio could help bridge the gap between in-person and video meeting conversational dynamics. The value of bridging that gap is the restoration of less effortful turn-taking and a greater sense of social presence.

Current theoretical explanations of videoconferencing fatigue (aka Zoom Fatigue) propose that non-verbal factors introduce two poles of unnaturalness: lack of information (body language, eye contact) and too much information (constant self-view, artificial grouping of faces), intensified by repeated exposure without variety [44]. Bailenson [1] does not include mono or mixed stereo audio as one of the input factors to video meeting fatigue. Nadler [36] attributes video meeting fatigue to spatial dynamics that flatten people into a "third skin" comprising person, background, and

technology, and this effect alters how we interact in virtual contexts. Although this study did not ask about fatigue, the results indicate that spatial audio reduces some of the effort of mediated turn-taking, and thus these findings spotlight audio as underemphasized in the theoretical conversation thus far [43, 57]. Spatial audio improvements to mediated turn-taking might help reduce video meeting fatigue.

This study also contributes to research on predictors of effective telepresence and mediated social presence. Effective telepresence relies on an experience that focuses the sense of presence "in the mediated environment, rather than in the immediate physical environment" [60]. Oh, et al. [42] note "this dimension of presence relates strongly to how vividly the user experiences the environmental and spatial properties of the mediated environment". Clearly, then, non-spatialized audio should be – should have always been – explored more deeply as a base-level feature in the effectiveness of video meetings. Similarly, if social presence is "the degree to which a user feels access to the intelligence, intentions, and sensory impressions of another" [4], then the reduced ability to selectively attend to others, participate, and understand mediated conversation, should be – again, should have always been – predictors of reduced social presence. Our point, then, is that auditory stimuli should be regarded as a necessary rather than simply valuable component of systems intended to induce social presence [12, 41]

Finally, this study examines the impact of spatial audio on behaviors and task performance in workplace video meetings. While the results are not statistically significant due to limitations in cohort size (see below), they do point to the potential for improvements in task performance overall, and more-so in harder tasks [46]. Clearly, any task that relies on conversation is fundamentally at the mercy of technological disruption in video meetings, but evidence is limited on how direct the relationship is between conversational dynamism and task effectiveness. One clue comes from Neibuhr et al. [40] who report that meeting productivity correlates with the overall sound of talk in individual meetings. Prosody, the patterns of voice stress and intonation are both the most diverse and most powerful predictors. Meetings characterized by affectively calmer, simpler, and shorter prosody are perceived to be more effective, but meetings characterized by lively, interactive, stimulating prosody generate a higher output of feasible or good ideas. Spatial audio, affording easier participation, greater sense of interactivity, and better understanding of the thoughts of others, should contribute markedly to lively prosody, and hence better task outcomes.

5.2 Implications for Future Research

This study only scratches the surface of understanding the impacts of spatial audio on collaboration in video meetings. We find evidence for isolated aspects of theoretical spatial audio benefits to feelings of social presence, turn-taking, and task performance, but clearly more detailed research is needed on the full spectrum of these issues.

The statistically significant social presence results represented only two dimensions of social presence: co-presence and perceived message understanding. No statistically significant results were found for other dimensions, including attention allocation and behavioral interdependence. The lowest scoring social presence questions were related to perceived affective understanding and emotional interdependence. More research is needed to explore these other dimensions, as well as further theoretical benefits such as greater ability to identify and attend to one person over another, reduced cognitive load and fatigue, and enhanced awareness of virtual space and the activities of people in it, and increased meeting focus and critical engagement.

We would hope that greater workplace inclusion is an expected consequence of increased participation and social presence in meetings. Spatial audio could make meetings more inclusive not only for women struggling with participation, but also neurodiverse communities and blind and low-vision people [61]. Studies of this kind would benefit from triangulation of task outcomes with objective turn-taking measures to test the hypothesis that spatial audio leads to more seamless conversations. Such measures would include better understanding of cooperative and competitive overlaps based on quantifiable time-based measures for overlaps, gaps, and pauses, as well as links to outcome measures. For example, in one study of scientific teams, a high number of turns in 10 minutes involved multiple members sharing ideas and no dominant turn-takers. This was positively correlated with total award dollars submitted and received [32]. It will also be crucial for research to understand how pre-existing group dynamics may affect outcomes and interact with objective measures [30].

For end-user value, the results of such studies could be displayed on post-meeting dashboards of insights to improve attendees' awareness of meeting dynamics and entitativity (the feeling of groupness) [6], with implications for meeting effectiveness and inclusivity. There is a research history of both real-time meeting feedback research (e.g. [28] and post-meeting dashboards [17, 49].

Spatial audio does require substantial investment to fit the many differing conditions of users. However, investments in spatial audio are likely to be faster and cheaper that investments in video, with a large ratio of cost to value, offer people a more immersive video meeting experience, and require the fewest changes to traditional meeting hardware and social/organizational behavior. This is because while video is the unique affordance that differentiates video meetings from other communication systems, for the most part, video of people is less relevant than audio from them. Standaert et al. [58] report that the ability to hear voice and share screens, but not video of participants, is identified as critical to all business meeting objectives. The current push for immersive reality environments ('metaverse/s'), while very likely to enable more naturalistic engagements in the medium long-term bet, nevertheless will be computationally and capitally intensive, as well as requiring wholesale changes in social and organizational behavior.

We noted above that, as the best spatial audio experience requires headphones, and usually wired headphones, as the standard Bluetooth protocol adopted for connection does not feature spatial audio, this could be a serious blocker to take-up. It is imperative that wireless audio device connections follow Apple's lead to support spatial audio in video calls (AirPods currently enable spatial audio in FaceTime), although for the greatest good, a common and open or cheaply licensed standard is preferable to Apple's walled garden approach.

Enabling the transmission and reception of spatial audio is, of course, only the beginning of the process. On the technical level,

there is much to be done designing the sound field and position of audio relative to the position of people and doing so in a way that works coherently and consistently on dynamic visual stages and scalable from small to very large meetings. Specifically, we need more research on the sound and position alignments of different layouts, proxemics in video meetings, and the need for personalized HRTFs [3, 16, 39].

While left/right and near/far are reasonably easy to simulate in many video meeting contexts, up and down are significantly harder given that the height of screens is usually less than the width, and as participant numbers increase vertical stacking of participants is necessary. Precise location of sound to visual representation of people will become very hard from the double-digit number of attendees upwards, and compromises will need to be made. These could involve deciding on logical aggregations of people and arranging their sound to come from similar places, prioritizing spatial audio by permanent or situational meeting role, or possibly even decoupling audio from visual placement in some circumstances. Regarding the latter, the assumption is that audio and video representations should be consistent, as they are in person, but, as Hollan and Stornetta [23] argue, we do not need to limit ourselves to recreating 'being there'. Spatial audio needs to achieve the value of naturalistic cues – sound separation and localization – but not necessarily the method. We should not assume that a high level of faithfulness to in-person conversation is necessary without testing other methods. In this regard, future attention should also be paid to designing the broader audio soundscape in which spatial audio of people's voices is to be represented. Some soundscapes may have effects that interact with perceptions of positionality.

Beyond the technical, further research is needed to determine in which collaboration scenarios spatial audio works best to improve outcomes, and help users make decisions on when and how to utilize spatial audio features to maximize impact. Video meetings are highly heterogeneous, and one size does not fit all. Indeed, if the users who reported negative perceptions of spatial audio are considered, time and/or training may be needed to accommodate people getting used to spatial audio, and there may need to be ways in which spatial audio is not used. If we redesign visual stages on the assumption that spatial audio will be available – especially when it may be relied upon for disambiguation – then we may set back those who choose not to use it.

6 LIMITATIONS

Participants were limited to US knowledge workers in a global technology company. As such, these results may apply to Western knowledge workers in organizations that regularly use video meetings, but may differ for other sectors, countries, and countries. Similarly, although survival problems are known for their ability to induce features of normal conversation, the tasks in the study only approximated real-world workplace conversations. All participants were strangers, which accounts for only a subset of meeting contexts; results may differ for participants who are acquainted. Large sample attrition limited the study's statistical power, particularly for the measure of task performance which is scored at the group level and is therefore limited by the number of groups in the analysis.

Given the maturity of the technology and compliance issues, the experimental video meeting application did not perfectly represent real-world spatial audio and some participants complained about spatial position confusion. Due to technical limitations, participants' positions were not fixed, meaning everyone did not see the same person in the same box. All participants were required to wear wired headphones to keep their experience comparable and maximize the spatial audio effect. In addition, while the layout was a better representation of an in-person meeting, the lack of fixed positions likely contributed to inconsistent experiences and audio and location dissonance. The combination of technology limitations could explain why differences between the conditions were not more pronounced.

7 CONCLUSION

This study provides evidence that spatial audio can increase social presence and improve turn-taking and could impact performance on tasks that require critical thinking. It also provides indications, which require more research, that spatial audio can help make meetings more inclusive for women and positively change conversational dynamics related to gender. The study upholds and further validates social presence and turn-taking theories.

There is much more research to be done to understand who wants and needs spatial audio, in which scenarios, and at what level of social presence. The study suggests it is not a one-size-fits-all solution for more collaborative and immersive virtual meeting experiences. The full potential of spatial audio will not be realized without further research and engineering to improve the technology and optimize layouts for spatialization. Metrics to objectively evaluate the impact of spatial audio on speech patterns are also needed to improve our understanding of the conversational dynamics that drive inclusive meetings.

If spatial audio is implemented in video meeting tools in a measured and research-backed way, it can provide an untapped resource for transforming our remote and hybrid meeting experiences. While there is no doubt there is much technical and design work still to be done, based on these results we strongly believe that spatial audio in video meetings could help with the adoption of more immersive environments. And rather than further overloading our visual channel with ever more features, spatial audio could be an important ingredient for more engaging, less fatiguing interactive environments in the future.

ACKNOWLEDGMENTS

The authors thank Chris Krekel, Abigail Sellen, Amos Miller, Ben Cutler, Spencer Fowers, Amber Hoak, David Tittsworth, Whitney Hudson, Steven Dong, Chris Gideon for their contributions and advice. We also thank the research participants for their time, and the reviewers for improvements and clarifications.

REFERENCES

[1] Jeremy N. Bailenson. 2021. Nonverbal Overload: A Theoretical Argument for the Causes of Zoom Fatigue. *Technology, Mind, and Behavior* 2, 1 (Feb. 2021). https://doi.org/10.1037/tmb0000030

[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 (CHI '01)*. Association for Computing Machinery, New York, NY, USA, 166–173. https://doi.org/10.1145/365024.365092

[3] Christopher C. Berger, Mar Gonzalez-Franco, Ana Tajadura-Jiménez, Dinei Florencio, and Zhengyou Zhang. 2018. Generic HRTFs May be Good Enough in Virtual Reality. Improving Source Localization through Cross-Modal Plasticity. *Frontiers in Neuroscience* 12 (2018). https://www.frontiersin.org/articles/10.3389/fnins.2018.00021

[4] Frank Biocca. 1997. The Cyborg's Dilemma: Progressive Embodiment in Virtual Environments [1]. *Journal of Computer-Mediated Communication* 3, 2 (Sept. 1997), JCMC324. https://doi.org/10.1111/j.1083-6101.1997.tb00070.x

[5] Frank Biocca and Chad Harms. 2002. Defining and measuring social presence: Contribution to the networked minds theory and measure. In *Proceedings of the Fifth Annual International Workshop on Presence*.

[6] Anita L. Blanchard, Andrew G. McBride, and Joseph A. Allen. 2022. Perceiving meetings as groups: How entitativity links meeting characteristics to meeting success. *Psychology of Leaders and Leadership* 25 (2022), 90–113. https://doi.org/10.1037/mgr0000124

[7] Jens Blauert. 1996. *Spatial Hearing: The Psychophysics of Human Sound Localization*. https://doi.org/10.7551/mitpress/6391.001.0001

[8] Judee K. Burgoon, Joseph A. Bonito, Artemio Ramirez, Jr., Norah E. Dunbar, Karadeen Kam, and Jenna Fischer. 2002. Testing the Interactivity Principle: Effects of Mediation, Propinquity, and Verbal and Nonverbal Modalities in Interpersonal Interaction. *Journal of Communication* 52, 3 (Sept. 2002), 657–677. https://doi.org/10.1111/j.1460-2466.2002.tb02567.x

[9] 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 (Sept. 1953), 975–979. https://doi.org/10.1121/1.1907229

[10] Gregory D. Clemenson, Antonella Maselli, Alexander J. Fiannaca, Amos Miller, and Mar Gonzalez-Franco. 2021. Rethinking GPS navigation: creating cognitive maps through auditory clues. *Scientific Reports* 11, 1 (April 2021), 7764. https://doi.org/10.1038/s41598-021-87148-4

[11] James J. Cummings and Jeremy N. Bailenson. 2016. How Immersive Is Enough? A Meta-Analysis of the Effect of Immersive Technology on User Presence. *Media Psychology* 19, 2 (April 2016), 272–309. https://doi.org/10.1080/15213269.2015.1015740

[12] James J. Cummings and Blake Wertz. 2018. Technological predictors of social presence: a foundation for a meta-analytic review and empirical concept explication. In *Proceedings of the 10th Annual International Workshop on Presence (Prague)*.

[13] Natasha Dhawan, Molly Carnes, Angela Byars-Winston, and Narjust Duma. 2021. Videoconferencing Etiquette: Promoting Gender Equity During Virtual Meetings. *Journal of Women's Health* 30, 4 (April 2021), 460–465. https://doi.org/10.1089/jwh.2020.8881

[14] Christina Dicke, Viljakaisa Aaltonen, Anssi Rämö, and Miikka Vilermo. 2010. Talk to me: The Influence of Audio Quality on the Perception of Social Presence. BCS Learning & Development. https://doi.org/10.14236/ewic/HCI2010.36

[15] Mark A. Ericson, Douglas S. Brungart, and Brian D. Simpson. 2004. Factors That Influence Intelligibility in Multitalker Speech Displays. *The International Journal of Aviation Psychology* 14, 3 (June 2004), 313–334. https://doi.org/10.1207/s15327108ijap1403_6

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

[17] Maria Frank, Ghassem Tofighi, Haisong Gu, and Renate Fruchter. 2016. Engagement Detection in Meetings. https://doi.org/10.48550/arXiv.1608.08711

[18] Michael Gibbs, Friederike Mengel, and Christoph Siemroth. 2021. Work from Home & Productivity: Evidence from Personnel & Analytics Data on IT Professionals. https://doi.org/10.2139/ssrn.3843197

[19] Jay Hall and W. H. Watson. 1970. The Effects of a Normative Intervention on Group Decision-Making Performance. *Human Relations* 23, 4 (Aug. 1970), 299–317. https://doi.org/10.1177/001872677002300404

[20] Chad Harms and Frank Biocca. 2004. Internal Consistency and Reliability of the Networked Minds Social Presence Measure. (2004).

[21] J. Hauber, H. Regenbrecht, A. Hills, A. Cockburn, and Mark Billinghurst. 2005. Social Presence in Two- and Three-Dimensional Videoconferencing. In *Presence 2005: The 8th Annual International Workshop on Presence*. London, UK, 198–198.

[22] Claudia Hendrix and Woodrow Barfield. 1996. The Sense of Presence within Auditory Virtual Environments. *Presence: Teleoperators and Virtual Environments* 5, 3 (Aug. 1996), 290–301. https://doi.org/10.1162/pres.1996.5.3.290

[23] Jim Hollan and Scott Stornetta. 1992. Beyond being there. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*. Association for Computing Machinery, New York, NY, USA, 119–125. https://doi.org/10.1145/142750.142769

[24] Gijs A. Holleman, Ignace T. C. Hooge, Chantal Kemner, and Roy S. Hessels. 2020. The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology* 11 (2020). https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00721

[25] 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 (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 95–98. https://doi.org/10.1145/1718918.1718936

[26] Angelika C. Kern and Wolfgang Ellermeier. 2020. Audio in VR: Effects of a Soundscape and Movement-Triggered Step Sounds on Presence. *Frontiers in Robotics and AI* 7 (2020). https://www.frontiersin.org/articles/10.3389/frobt.2020.00020

[27] Ryan Kilgore, Mark H. Chignell, and Paul W. Smith. 2003. Spatialized audioconferencing: what are the benefits?. In *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative Research, October 6-9, 2003, Toronto, Ontario, Canada*, Darlene A. Stewart (Ed.). IBM, 135–144. https://dl.acm.org/citation.cfm?id=961345

[28] Olga Kulyk, Jimmy Wang, and Jacques Terken. 2006. Real-Time Feedback on Nonverbal Behaviour to Enhance Social Dynamics in Small Group Meetings. In *Machine Learning for Multimodal Interaction (Lecture notes in Computer Science)*, Steve Renals and Samy Bengio (Eds.). Springer, Berlin, Heidelberg, 150–161. https://doi.org/10.1007/11677482_13

[29] Kwan Min Lee. 2004. Presence, Explicated. *Communication Theory* 14, 1 (Feb. 2004), 27–50. https://doi.org/10.1111/j.1468-2885.2004.tb00302.x

[30] Daniel Levi and David A Askay. 2020. *Group dynamics for teams*. Sage Publications.

[31] Stephen C. Levinson. 2016. Turn-taking in Human Communication – Origins and Implications for Language Processing. *Trends in Cognitive Sciences* 20, 1 (Jan. 2016), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

[32] Hannah B. Love, Bailey K. Fosdick, Jennifer E. Cross, Meghan Suter, Dinaida Egan, Elizabeth Tofany, and Ellen R. Fisher. 2022. Towards understanding the characteristics of successful and unsuccessful collaborations: a case-based team science study. *Humanities and Social Sciences Communications* 9, 1 (Oct. 2022), 1–11. https://doi.org/10.1057/s41599-022-01388-x

[33] Matthew Lombard, Theresa B. Ditton, and Lisa Weinstein. 2004. Measuring presence: The Temple Presence Inventory. http://matthewlombard.com/research/p2_ab.html

[34] Radha Nila Meghanathan, Patrick Ruediger-Flore, Felix Hekele, Jan Spilski, Achim Ebert, and Thomas Lachmann. 2021. Spatial Sound in a 3D Virtual Environment: All Bark and No Bite? *Big Data and Cognitive Computing* 5, 4 (Dec. 2021), 79. https://doi.org/10.3390/bdcc5040079

[35] Neville Moray. 1959. Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology* 11, 1 (Feb. 1959), 56–60. https://doi.org/10.1080/17470215908416289

[36] Robby Nadler. 2020. Understanding "Zoom fatigue": Theorizing spatial dynamics as third skins in computer-mediated communication. *Computers and Composition* 58 (Dec. 2020), 102613. https://doi.org/10.1016/j.compcom.2020.102613

[37] NASA. 2009. Exploration: Then and Now - Survival! Lesson. http://www.nasa.gov/stem-ed-resources/jamestown-survival.html

[38] Ulric Neisser. 1976. *Cognition and reality: Principles and implications of cognitive psychology*. W H Freeman/Times Books/ Henry Holt & Co, New York, NY, US.

[39] David T. Nguyen and John Canny. 2007. Multiview: improving trust in group video conferencing through spatial faithfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1465–1474. https://doi.org/10.1145/1240624.1240846

[40] Oliver Niebuhr, Ronald Böck, and Joseph A. Allen. 2021. On the Sound of Successful Meetings: How Speech Prosody Predicts Meeting Performance. In *Companion Publication of the 2021 International Conference on Multimodal Interaction (ICMI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 240–248. https://doi.org/10.1145/3461615.3485412

[41] Rolf Nordahl and Niels Christian Nilsson. 2014. The Sound of Being There: Presence and Interactive Audio in Immersive Virtual Reality. In *The Oxford Handbook of Interactive Audio*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199797226.013.013

[42] Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. 2018. A Systematic Review of Social Presence: Definition, Antecedents, and Implications. *Frontiers in Robotics and AI* 5 (2018). https://www.frontiersin.org/articles/10.3389/frobt.2018.00114

[43] 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

[44] René Riedl. 2022. On the stress potential of videoconferencing: definition and root causes of Zoom fatigue. *Electronic Markets* 32, 1 (March 2022), 153–177. https://doi.org/10.1007/s12525-021-00501-3

[45] E. Sean Rintel. 2010. Conversational management of network trouble perturbations in personal videoconferencing. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction (OZCHI '10)*. Association for Computing Machinery, New York, NY, USA, 304–311. https://doi.org/10.1145/1952222.1952288

[46] 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 (CHI EA '21). Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3451802

[47] Karen Ruhleder and Brigitte Jordan. 2001. Co-Constructing Non-Mutual Realities: Delay-Generated Trouble in Distributed Interaction. *Computer Supported Cooperative Work (CSCW)* 10, 1 (March 2001), 113–138. https://doi.org/10.1023/A:1011243905593

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

[49] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445615

[50] Emanuel A. Schegloff. 2000. Overlapping Talk and the Organization of Turn-Taking for Conversation. *Language in Society* 29, 1 (2000), 1–63.

[51] Abigail J. Sellen. 1992. Speech patterns in video-mediated conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '92)*. Association for Computing Machinery, New York, NY, USA, 49–59. https://doi.org/10.1145/142750.142756

[52] Abigail J. Sellen. 1995. Remote Conversations: The Effects of Mediating Talk With Technology. *Human–Computer Interaction* 10, 4 (1995), 401–444. https://doi.org/10.1207/s15327051hci1004_2

[53] Lucas M. Seuren, Joseph Wherton, Trisha Greenhalgh, and Sara E. Shaw. 2021. Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. *Journal of Pragmatics* 172 (Jan. 2021), 63–78. https://doi.org/10.1016/j.pragma.2020.11.005

[54] Jean-Luc Sinclair. 2020. *Principles of game audio and sound design: sound design and audio implementation for interactive and immersive media.* CRC Press.

[55] Paul Skalski and Robert Whitbred. 2010. Image versus sound: A comparison of formal feature effects on presence and video game enjoyment. *PsychNology Journal* 8 (2010), 67–84.

[56] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (May 2021), 101178. https://doi.org/10.1016/j.csl.2020.101178

[57] Janto Skowronek, Alexander Raake, Gunilla H. Berndtsson, Olli S. Rummukainen, Paolino Usai, Simon N. B. Gunkel, Mathias Johanson, Emanuël A. P. Habets, Ludovic Malfait, David Lindero, and Alexander Toet. 2022. Quality of Experience in Telemeetings and Videoconferencing: A Comprehensive Survey. *IEEE Access* 10 (2022), 63885–63931. https://doi.org/10.1109/ACCESS.2022.3176369

[58] Willem Standaert, Steve Muylle, and Amit Basu. 2021. How shall we meet? Understanding the importance of meeting mode capabilities for different meeting objectives. *Information & Management* 58, 1 (Jan. 2021), 103393. https://doi.org/10.1016/j.im.2020.103393

[59] Willem Standaert and Sophie Thunus. 2022. Virtual Meetings during the Pandemic: Boon or Bane for Gender Inequality.

[60] Jonathan Steuer. 1992. Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication* 42, 4 (1992), 73–93. https://doi.org/10.1111/j.1460-2466.1992.tb00812.x

[61] John Tang. 2021. Understanding the Telework Experience of People with Disabilities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 30:1–30:27. https://doi.org/10.1145/3449104

[62] Anna Watson and M. Angela Sasse. 2000. The good, the bad, and the muffled: the impact of different degradations on Internet speech. In *Proceedings of the eighth ACM international conference on Multimedia (MULTIMEDIA '00)*. Association for Computing Machinery, New York, NY, USA, 269–276. https://doi.org/10.1145/354384.354503

[63] Joseph Williams, Sven Shepstone, and Damian Murphy. 2022. Understanding Immersion in the Context of Films with Spatial Audio. In *Audio Engineering Society Conference: AES 2022 International Audio for Virtual and Augmented Reality Conference*. http://www.aes.org/e-lib/browse.cfm?elib=21878

[64] Julie Williamson, Jie Li, Vinoba Vinayagamoorthy, David A. Shamma, and Pablo Cesar. 2021. Proxemics and Social Interactions in an Instrumented Virtual Reality Workshop. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 253, 13 pages. https://doi.org/10.1145/3411764.3445729

[65] 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)*. 1–6. https://doi.org/10.1109/I3DA48870.2021.9610974

[66] Jing Yang, Yves Frank, and Gábor Sörös. 2019. Hearing Is Believing: Synthesizing Spatial Audio from Everyday Objects to Users. In *Proceedings of the 10th Augmented Human International Conference 2019 (AH2019)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3311823.3311872

[67] Jing Yang, Prasanth Sasikumar, Huidong Bai, Amit Barde, Gábor Sörös, and Mark Billinghurst. 2020. The effects of spatial auditory and visual cues on mixed reality remote collaboration. *Journal on Multimodal User Interfaces* 14, 4 (Dec. 2020), 337–352. https://doi.org/10.1007/s12193-020-00331-1

[68] Mike Z. Yao and Andrew J. Flanagin. 2006. A self-awareness approach to computer-mediated communication. *Computers in Human Behavior* 22, 3 (May 2006), 518–544. https://doi.org/10.1016/j.chb.2004.10.008