I See What You're Hearing: Facilitating The Effect of Environment on Perceived Emotion While Teleconferencing

DAVID MARINO, McGill University, Canada
MAX HENRY, McGill University, Canada
PASCAL E. FORTIN, McGill University, Canada
RACHIT BHAYANA, Indraprastha Institute of Information Technology, India
JEREMY COOPERSTOCK, McGill University, Canada

Our perception of emotion is highly contextual. Changes in the environment can affect our narrative framing, and thus augment our emotional perception of interlocutors. User environments are typically heavily suppressed due to the technical limitations of commercial videoconferencing platforms. As a result, there is often a lack of contextual awareness while participating in a video call, and this affects how we perceive the emotions of conversants. We present a videoconferencing module that visualizes the user's aural environment to enhance awareness between interlocutors. The system visualizes environmental sound based on its semantic and acoustic properties. We found that our visualization system was about 50% effective at eliciting emotional perceptions in users that was similar to the response elicited by environmental sound it replaced.The contributed system provides a unique approach to facilitate ambient awareness on an implicit emotional level in situations where multimodal environmental context is suppressed.

CCS Concepts: • **Human-centered computing** → *Collaborative and social computing systems and tools*; *Human computer interaction (HCI)*; *Visualization.*

Additional Key Words and Phrases: teleconferencing; visualization; context; multimodal

ACM Reference Format:
David Marino, Max Henry, Pascal E. Fortin, Rachit Bhayana, and Jeremy Cooperstock. 2023. I See What You're Hearing: Facilitating The Effect of Environment on Perceived Emotion While Teleconferencing. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 62 (April 2023), 15 pages. https://doi.org/10.1145/3579495

1 INTRODUCTION

Teleconferencing has become an essential part of our everyday lives. Despite the success of commercial teleconferencing platforms in connecting people remotely, the conversational experience afforded by such platforms is degraded. A fundamental aspect of in-person communication is knowledge about the environment(s) in which interlocutors are situated, yet much of this contextual information is lost in a video call. There are a number of contributing reasons for this: webcams have a restricted field of view, masking most of the visual environment; video quality must be compressed, reducing the fidelity of the signal; and spatial audio cues are lost, as audio is

Authors' addresses: David Marino, dmarino@cim.mcgill.ca, McGill University, Center for Intelligent Machines, Montreal, Quebec, Canada, H3A 0G4; Max Henry, max.henry@mail.mcgill.ca, McGill University, Dept. of Electrical and Computer Engineering, Montreal, Quebec, Canada, H3A 0G4; Pascal E. Fortin, pe.fortin@mail.mcgill.ca, McGill University, Center for Intelligent Machines, Montreal, Quebec, Canada, H3A 0G4; Rachit Bhayana, rachit18301@iiitd.ac.in, Indraprastha Institute of Information Technology , Dept. of Human Centered Design, New Delhi, Delhi, India, 110020; Jeremy Cooperstock, jer@cim.mcgill.ca, McGill University, Center for Intelligent Machines, Montreal, Quebec, Canada, H3A 0G4.

2573-0142/2023/4-ART62 $15.00
https://doi.org/10.1145/3579495

typically captured from a single monaural microphone. Environmental context is further degraded when the speaker in question is using an avatar or using background replacement/augmentation (degradation in the visual modality), or employing aggressive noise cancellation (degradation in the auditory modality). A loss of contextual awareness can result in a diminished sense of co-presence, and impede our ability to relate and fluently communicate with each other remotely. Environmental context also affects how emotions are perceived [3, 5, 18]. For example, a speaker may be perceived as anxious if they uttered a sentence while a dog was loudly barking behind them. But placing the same utterance in the context of a tranquil forest, the speaker may be perceived as more relaxed.

There are times when a loss of contextual information may be desirable—for example, someone who is using a messy bedroom as a home office may opt to hide their environment by using background replacement. Or, members of an expert panel during a remote conference may opt to use noise cancellation so that only speech is transmitted to listeners. These are cases of *information-centered communication*, where conversation is centered around facts and completing exchanges. Yet there are times when it is desirable to convey as much environmental context as possible, such as when bonding with a friend, catching up with family, or engaging in remote group learning activities where situational awareness of other learners is preferable. These are cases of *experiential communication*, where empathy and emotional understanding are at the forefront of conversation. In these cases, it is important to convey the subjective experience of interlocutors, and contextual information is crucial for appropriately understanding it. We designed our prototype for this use case.

This paper presents a system that enhances contextual awareness by visualizing the user's aural environment. While a webcam only captures a narrow segment of the user environment, a microphone offers a much broader view of the space, making the audio modality rich in environmental information. Additionally, by building on the standard videoconferencing audio stream, we eliminate the need for possible high-cost specialized capture devices enabling consumer-level users to use this system with commodity hardware. The visualization system has applications when having experiential conversation remotely. A major design goal of our system is to translate background audio into a visualization that elicits an emotional responses in a way that is similar to its source audio. Our primary research question is: How can we visually represent the acoustic environment in a manner that is emotionally consistent? To explore the answer to this question, we contribute:

- A working prototype of a system that enhances emotional awareness while teleconferencing.
- An evaluation of the system, including the emotional effects of cross modal representations of the ambient environment.
- A preliminary analysis of which audio events during teleconferencing are most relevant to users.

2 BACKGROUND

2.1 Audio visualization

Visualizing audio events has a theoretical precedence in the bouba-kiki effect—where a non-arbitrary relationship is demonstrated between speech sounds and shapes [30][24]. In it, nonsense words like "bouba" are associated with round puffy shapes, while "kiki" is associated with spiky shapes. Aspects of this effect have been shown to be consistent across cultures [7]. The basis of this non-arbitrary relationship has been hypothesized to be based off tacit knowledge of the articulatory processes, such vowel height and lip rounding, that link sound to visuals [12]. However, relationships between sound and visuals is almost certainly not limited to the verbal domain, and they can have arbitary

relationships as well. For example, those with synesthesia have been known to have arbitrary audio-visual associations [17].

Through conceptual metaphor, people are able to conceptualize abstract phenomena cross-modally [25]. Conceptual metaphor sees understanding a concept in a *target domain* in terms of a *source domain*. For example: in LOUDNESS IS BRIGHT, brightness (the target domain) is being understood in terms of loudness (the source domain). Conceptual metaphor has specifically been applied to non-verbal cross-modal mappings between auditory and visual source and target domains [15]. The notion of conceptual metaphor claims that metaphor forms the foundation of much of human thought [16]. Using this theoretical framework, we can understand audio visualization as a multimodal conceptual metaphor.

2.2 Enhancing contextual awareness

Early efforts to enhance contextual awareness and presence in teleconferencing began with devices such as the small-form-factor Hydra units, consisting of a camera, microphone, video monitor, and speaker, which served as spatially distributed avatars for each of the participants they represented [32]. Physical approaches such as MeBot embodied remote users in a robot that conveyed nonverbal cues [1]. Virtual reality (VR) teleconferencing using head mounted displays (HMDs) or CAVE Automatic Virtual Environments (CAVE) environments provide immersive near-360 degree views of user environments, or offer a "common context" through shared virtual environments [19] [8]. An early effort to provide environmental awareness was exemplified in Portholes, where a ubiquitous network of cameras was deployed to support "whole office" ambient awareness of distributed work groups [11].

The idea of visualizing the audio modality to enhance awareness and feelings of presence in conversation is not new. The Visiphone was a spherical display that animated the audio of remote callers, enabling them to come to conclusions about their conversation, such as volume levels, and conversational rhythm, they may not have otherwise realized [10]. Audio visualizations have also been used in video conferencing to calibrate vocalization levels between interlocutors [23]. In non-remote settings, the Conversation Clock visualized audio patterns of interlocutors in the same physical space, providing them with a shared *social mirror* visualization [22] designed to provide "insight into the participants' culture and status" [4]. Visualizations of the aural environment have also been used with populations living with auditory impairments and autism. Realtime audio visualizations have been used to encourage spontaneous speech-like vocalizations with users with autism spectrum disorder [20]. Non-speech sound visualizations have also been used to convey aspects of the physical environment in which Deaf users are situated [27].

In terms of providing ambient emotional awareness, SmartHeliosity provides an emotional feedback loop with users by producing coloured light based off their facial expressions [33]. A described ideal use case for this device is as an ambient workplace display for emotion regulation. Similarly, BioCrystal is a physical ambient device that uses physiological feedback to produce coloured light [31]. The reactive colours were found to be useful in self monitoring of emotions and interpersonal communication.

The aforementioned technologies are effective in their respective domains; however many require high-cost specialized hardware, and none are designed to convey cross-modal ambient auditory environmental context while videoconferencing with a general population. The system described in this paper focuses specifically on conveying user environmental context as opposed to other contextual aspects lost in videoconferencing such as eye contact or posture. The system has particular use in situations where aspects of the user environment is occluded and limited environmental audio is transmitted, such as when using noise canceling, speaker prioritization, or muted microphones.

3 SYSTEM DESIGN

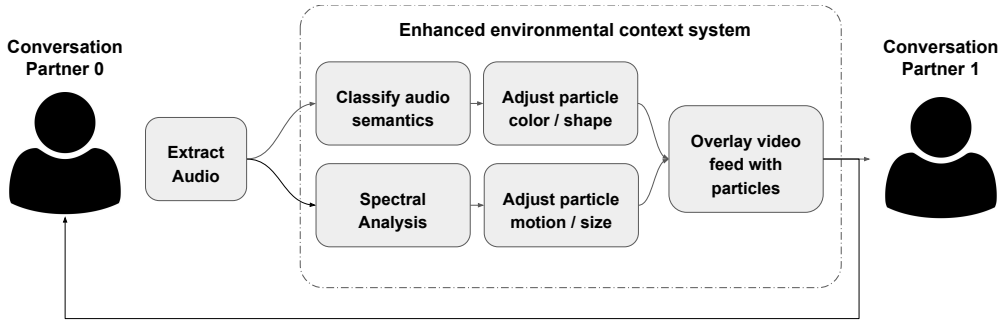


Fig. 1. High level system architecture.

The system displays a particle animation driven by both low-level acoustic, and high-level semantic, features of the user's auditory environment. A convolutional neural network (CNN) classifies the ambient audio into six non-mutually-exclusive semantic categories: *artificial*, *natural*, *foreground*, *background*, *interior*, and *exterior*. This process is described in detail in Section 3.1. The particle animation is also affected by low-level acoustic features of the environmental sound, which is further described in Section 3.2. Semantic information adjusts particle color and shape, while acoustic information affects particle size, animation speed and trajectory. The video feed is overlaid with the particle animation, which is then broadcast to all users of the videoconferencing app. The same process is repeated for all conversation partners. Our initial design decisions are evaluated in a user study. We conclude the user study with suggestions for a future version of the system.

3.1 Semantic features

We classify the soundscape in realtime based off of high level semantic features. A soundscape may be considered as having three key features: *geophony*, *biophony*, and *anthrophony* [29]. Geophony are geophysical sounds, biophony are organic sounds, and anthrophony are human sounds. This schema informed the basis of our semantic classification system. We utilized Google's YAMNet,[1] a pre-trained neural network that classifies audio events based on a taxonomy built on YouTube audio [21]. YAMNet can predict 521 classes of sounds.

For a given audio excerpt, the network predicts a probability that the sound belongs to any of the 512 classes. Class predictions are not mutually exclusive, and one sound can belong to many classes. A recording of piano, for example, might generate a high probability in `'piano'`, `'classical music'`, and `'soundtrack music'`. Such predictions provide an ideal basis for pooling similar classes into broader *semantic features*. To calculate the semantic feature score, one can simply sum the probabilities from each of its constituent classes.

Using Google's AudioSet class ontology as a starting point, three authors thematically organized the classes into 6 high level semantic features, roughly aligning them with high level soundscape features according to their applicability. The authors coded these classes together, and had discussions when conflicts arose. We decided to keep semantic features as coarse grained as possible, as we didn't want to assume what specific sound events were most relevant to teleconferencing users without first running a user study. One relevant aspect of soundscapes with regards to videoconferencing is the distinction between anthrophony in the foreground, such as someone

[1] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

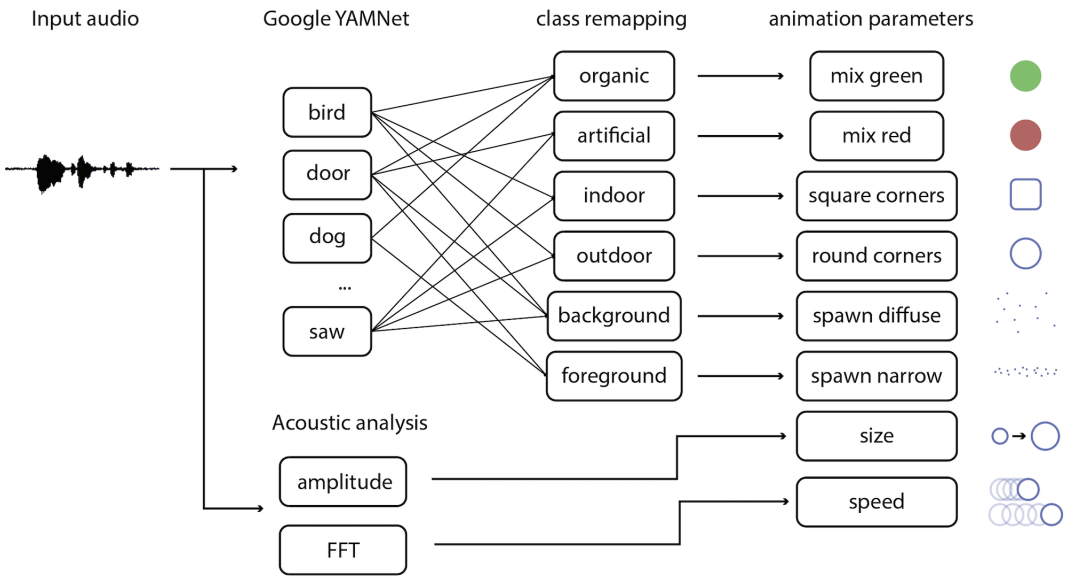


Fig. 2. Input audio is classified using Google YAMNet. We then remap Google YAMNet's classes to our own class ontology. These classes are then mapped to animation parameters. Acoustic analysis is conducted in parallel, which in turn maps to animation parameters.

speaking directly to you, and the background, such as a baby crying. Another relevant aspect further subdividing anthrophony that is relevant for videoconferencing is the distinction between organic sounds such as other humans speaking, and artificial sounds, such as a vacuum cleaner. Finally, soundscape ecology often does not consider transitioning between environments, where this may happen in a video call. We wished for a label to assist in determining if sounds were coming from inside or outside. Incorporating these videoconferencing demands, our semantic features are:

- Artificial — artificial sources such as machines and tools.
- Natural — organic sources such as animal noises, rivers, or wind.
- Interior — normally found inside, such as espresso machines and pen on paper.
- Exterior — normally found outside, such as cars and birds.
- Foreground — sounds that are likely to be from the primary subject in a meeting; mainly, direct human speech.
- Background — sounds that are not from the primary subject. These include sounds that are not human speech, however non-linguistic vocalizations and non-speech human sounds such as burping are also included here.

We evaluate the suitability of these features in a user study, reported in section 4.2.2. In our evaluation we ultimately discovered that a simpler class ontology of "living creatures", "outdoor" and "indoor" suited users needs, and would be ideal basis for a future iteration of the prototype.

As an example, a soundscape might include bird sounds, waterfalls, and a person speaking. In this case, YAMNet could output a probability of 0.8 for `'bird vocalization, bird call, bird song'`, a 0.9 probability for `'waterfall'` and a 0.7 probability for `'female speech, woman speaking'`. As the former two categories fall under "natural" and "exterior" semantic features

categories, our system would tally scores of 1.7 for both; human speech falls into the "foreground" category, and so this frame would count 0.7 in this semantic feature.

We note, however, that semantic features are highly context dependant, and ultimately a design choice had to be made about the intended target in our categorization. For example, a group of people talking could be a "foreground" element in a call with family, but a "background element" in a one-on-one meeting. We used the one-on-one meeting use case as a basis to frame our semantic feature coding.

For each semantic feature, f, and each auditory event, e, picked up by the microphone, the CNN calculates in realtime a confidence score $c \in [0, 1]$ that e is associated with f. These scores are accumulated over a temporal window of length one second for each feature. The features are then translated in proportion to their confidence score to continuous mixable properties of the particles, as follows:

- Colour — Particles begin with a blue base, but are additively blended with red for artificial sounds and green for natural sounds.
- Spatial distribution — spawned particles are spaced closer together for foreground and further apart for background sounds.
- Shape — interior sounds make the particles look more square, while exterior make the particles look more round.

The mappings were designed to be partially motivated signs [9]—that is that the visualizations (signifiers) were inspired by physical aspects that co-occur with semantic categories (signified): background sounds being more spread out than foreground sounds reflect the spatiality of sound sources. The colour green frequently occurs in nature and commonly associated with naturalness in pop culture. Of the primary colours in RGB colour space, the colour red appears less frequently in nature, forming the basis of the artificial mapping. Objects indoors tend to be designed with angular shapes, while objects found outdoors tend to be not designed and curvaceous, which taken together forms the basis of the square:indoor and round:outdoor mapping.

This mapping was intended to be a first exploration of semantic features in the form of particles. Its evaluation, and subsequent suggestions for iteration, are described in our user study.

3.2 Spectral analysis

Spectral characteristics of the source audio were used to modify particle spawn rate, motion, and size. Audio is captured at a sample rate of 44.1kHz. We then calculate a Fourier transform of the signal and take its spectral magnitude. If a bin magnitude is above a threshold value, it generates a particle. The threshold value—initialized at -10 dB—is adjustable to accommodate different user's mic sensitivities. Higher frequency bins generate particles that move faster across the screen. Particle size scales with the bin's spectral magnitude. The acoustic parameters utilize the conceptual metaphors of "loud is large", and "high pitch is fast" [25].

4 USER STUDY

An experiment was conducted to evaluate how participants understood the role of background audio (BGA) visualization. Thirteen participants were recruited from the McGill and Indraprastha Institute of Information Technology Delhi (IIITD) communities. Participants received compensation of CAN $10 for their time. The study was conducted under approval of the omitted for blind review REB, file McGill REB, file #20-08-031. This study utilized a within subjects design, and was broken into two sections: First was a rating task, where participants were asked to assess perceived emotions from pre-recorded videos of two people conversing among different auditory and visual contexts. Lastly, qualitative data was collected via a textbox prompt to investigate what

the visualizations meant to the participants, and what background audio they considered most relevant based off their daily teleconferencing experience.

4.1 Rating task

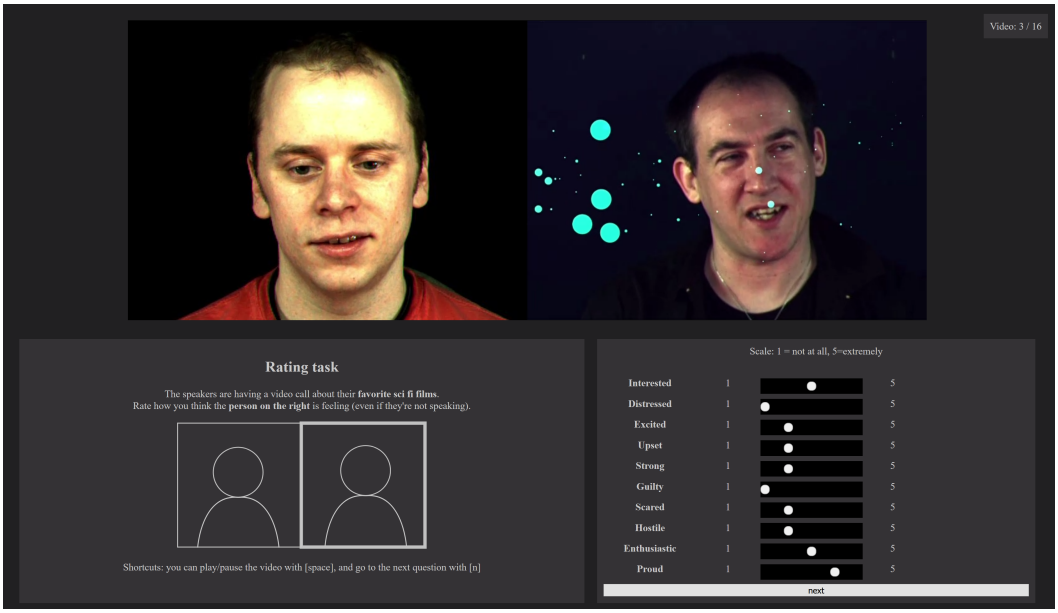


Fig. 3. UI for the rating task

This experiment investigates if we can convey similar emotional contexts between modalities. Participants were shown videos of dyadic conversations and asked to rate the emotional state of a speaker on 10 five-point scales, each corresponding to a validated emotion word from the Positive and Negative Affect Schedule (PANAS) [34]. The Positive and Negative Affect Schedule places emotion words on a single dimension from negative affect (unpleasant emotions like "scared") to positive affect (pleasant emotions like "interested"). The emotion words contained are claimed to be largely uncorrelated, and are derived through factor analysis in a controlled lab environment. We used a decimated form of the PANAS scale consisting of 10 uncorrelated words [26]. The words used were: {Interested, Excited, Enthusiastic, Proud, Strong } (positive affect) and {Distressed, Upset, Guilty, Scared, Hostile} (negative affect). We chose to use emotion words over an approach like an affect grid to encourage participants to engage more with the emotive meaning of the scene. We further used a five point likert scale to allow users to provide continuous ratings over discrete judgements (sometimes a speaker may be a little bit annoyed, vs extremely annoyed). Footage of the speakers was sourced from the Cardiff Conversation Database (CCDb) [2]. All participants were shown a series of excerpts from a single conversation between two interlocutors who were displayed in tile mode (Figure 3). Four short, 10–20s video clips were selected from a single conversation. Participants were randomly shown videos mixed with different conditions: 2 BGA CONDITIONS (on/off) × 2 VISUAL CONDITIONS (on/off) × 4 ENVIRONMENT CONDITIONS (construction, dogs, cafe, and forest) = 16 total videos for a single trial. The order that stimuli were presented was randomized. The independent variables (IV) are: BGA, VISUALIZATION, and ENVIRONMENT TYPE. The dependent variables (DV) are EMOTION RATINGS. The acoustic environments were sourced from YouTube.

The BGA is the background audio that naturally occured with the acoustic environments. The visualizations are the output of our system that are generated from BGA. In conditions where BGA is off and visualizations are on, the BGA is inaudible to the participant, but the visualization system still generates visuals as if BGA were present. EMOTION RATINGS were Likert scales from 1-5. A rating of 1 indicated the emotion was not present, and a rating of 5 indicated an emotion was extremely present.

4.1.1 Main and interaction effects on emotion ratings. Hypothesis 1 (H1) posits that EMOTION RATINGS are affected by BGA, VISUALIZATIONS, ENVIRONMENTS, or combinations thereof ($H1_0$: $\mu_{BGA} = \mu_{viz} = \mu_{env}$; $H1_A : \neg(\mu_{BGA} = \mu_{viz} = \mu_{env})$)

A three-way aligned rank transform (ART) ANOVA was conducted to compare the effect of ENVIRONMENT, BGA and VISUALIZATION on each of the EMOTION RATINGS [35]. An ART ANOVA was selected because emotion ratings were assumed to be ordinal as distance between values in the scale are not presumed to be consistent. Using a Bonferroni corrected $\alpha/10 = 0.005$, the analysis revealed that for a subset of emotions, there were significant effects of ENVIRONMENT, BGA, and VISUALIZATION, as well as interaction effects between ENVIRONMENT & BGA, and ENVIRONMENT & VISUALIZATION. Significant effects with p values are reported in Table 1. Critical F values are reported in Table 2. A boxplot of IVs is shown in figure 4.

Factor	Interested	Distressed	Proud	Upset	Strong	Guilty	Scared	Hostile
environment	0.0147	0.0005	0.8438	0.0055	0.0067	1.1301e-05	1.0188e-14	3.8809e-12
BGA	0.2151	0.0029	0.0155	0.8363	0.3041	2.5136e-11	0.0001	0.0043
visualization	0.9124	0.2228	0.0641	0.5001	0.2217	2.4186e-11	0.0004	2.7463e-06
env:BGA	0.3179	0.0006	0.4787	0.3889	0.8901	< 2.22e-16	2.3863e-05	1.3157e-09
env:viz	0.2889	0.0436	0.1333	0.0231	0.8404	5.2816e-07	1.5748e-07	0.0005

Table 1. p values for DVs (columns) by factors (rows). Highlighted values are $p < 0.005$. Columns with no significant factors are omitted (enthusiastic, excited). Rows and columns with no significant values are omitted, except for columns that had significant values prior to Bonferroni correction.

Factor	F( )	Interested	Distressed	Proud	Upset	Strong	Guilty	Scared	Hostile
environment	F(4,183)	3.1839	5.3006	0.3501	3.7858	3.6688	7.5815	21.7856	17.4346
BGA	F(1,183)	1.5469	9.1317	5.9697	0.0428	1.0623	50.5589	15.1537	8.3621
visualization	F(1,183)	0.0121	1.4964	3.4697	0.4566	1.5051	50.5589	13.2581	23.4285
env:BGA	F(3,183)	1.1822	5.9975	0.8304	1.0114	0.2091	38.2204	8.5598	16.6490
env:viz	F(3,183)	1.2618	2.7594	1.8869	3.2473	0.2792	11.6146	12.6072	6.2555

Table 2. F values for DVs (columns) by factors (rows). Highlighted values surpass critical F value. Columns with no significant factors omitted are (enthusiastic, excited)

There is evidence to reject null hypothesis $H1_0$ at p< 0.005 for the emotion words: distressed, guilty, scared, and hostile. There is a main effect of ENVIRONMENT between the aforementioned words, and a main effect of VISUALIZATION for guilty, scared, and hostile. There is also an interaction effect between ENVIRONMENT and VISUALS for the same emotions. From these findings, we can infer that visualizations do induce a perceived change of context.

These results are validating to our ground truth assumption that the environment changes emotional perceptions. But do the visualizations elicit changes in emotion ratings the same way BGA does? To answer this question, we calculated Spearman rank correlations on perceived emotion ratings between BGA-only conditions, and visualization-only conditions (Fig. 5).

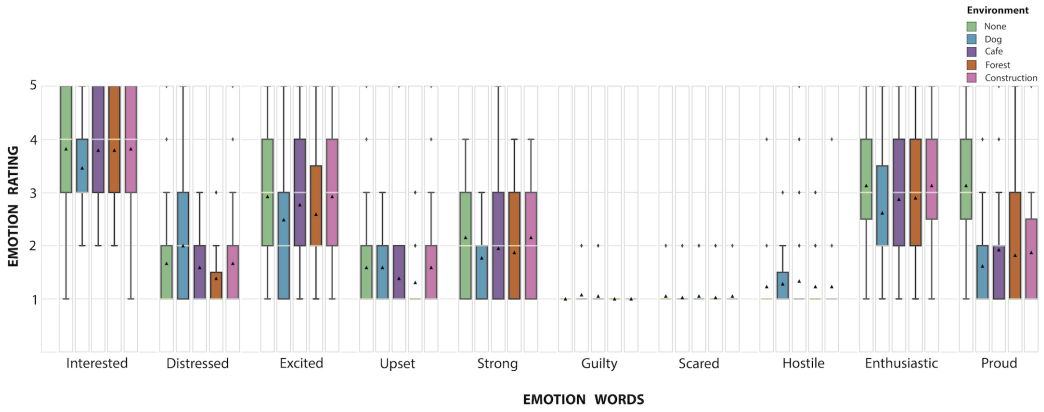
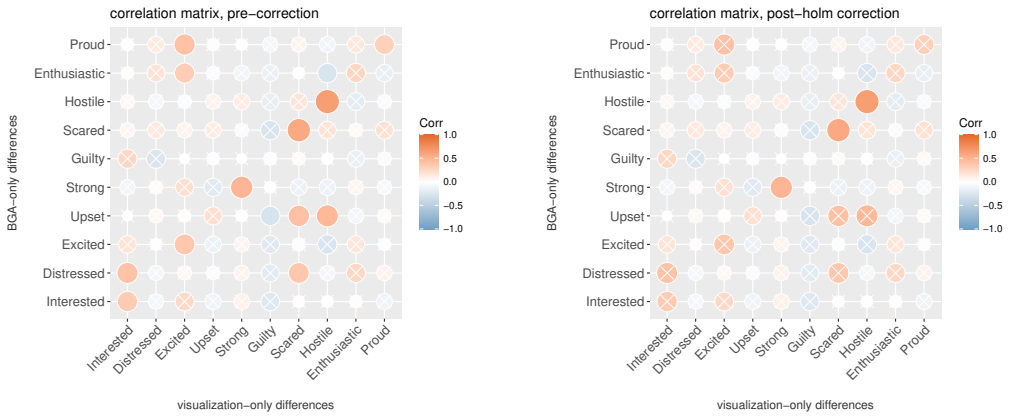


Fig. 4. Box plots depicting emotion rating for each of the emotion word for each of the five environments. Conflating between BGA:on and BGA:off conditions.



(a) pre-correction

(b) post-correction

Fig. 5. Emotion rating between visualizations and BGA. The nodes with an X are not significant. Shows the difference before and after Holm corrections were aplied.

4.1.2 Hypothesis 2. Do the visuals evoke similar emotions as the audio? Hypothesis 2 (H2) posits that the magnitude and direction of emotion ratings between visualization-only conditions and BGA-only are similar. We test by first calculating $\text{corr}(\bar{x}_{base} - \bar{x}_{BGA}, \bar{x}_{base} - \bar{x}_{viz})$, where $\bar{x}_{base}$ are the mean emotion ratings for the base condition (no visuals, no BGA), $\bar{x}_{BGA}$ are the mean emotion ratings of the background-only condition, and $\bar{x}_{viz}$ are the mean emotion ratings for the visualization-only condition. We calculated correlation coefficients for each cross modal combination of emotion words, and constructed a correlation matrix (Fig. 5). If the visuals conveyed similar contexts as the BGA, we would expect to see significantly strong correlations in the diagonal. For example, visualizations associated with "proud" should covary with BGA associated with "proud", and not (necessarily) BGA associated with "scared". There are 10 visualization-only conditions × 10 BGA-only conditions

= 100 possible pairs. Of the 100 pairs, 14 were significantly correlated pre-correction ($p < 0.05$), and 7/10 were significantly correlated in the diagonal (Fig. 5, (a)). Using Holm corrected p-values, there were just three significantly cross modal correlations, all on the diagonal, for the emotions "strong", "scared", and "hostile" (Fig 5, (b)). From our previous test, there were four emotions that were affected by changing environments: "distressed", "guilty", "scared", and "hostile". Of those emotions, 50% of them were significantly correlated cross modally.

Conducting Holm corrections on 100 possible pairs greatly increases the chance of type II errors. We are fundamentally interested if most significant correlations are in the diagonal.

Our interpretation of the results are as follows: first, of the emotions "distressed", "guilty", "scared", and "hostile" changed between different environments. The visualization system was able to capture that change in a way that was significantly correlated with the BGA 50% of the time. This suggests that the visualization system was 50% successful at evoking emotional contexts in a way that was somewhat similar to the BGA.Words that were not correlated between modalities could have been because the visualization did not adequately convey environmental meaning between modalities. Another reason is that the emotions may have simply not been present in the scene, as the study used a single conversation in its analysis. If an emotion word was significantly affected by ENVIRONMENT from our prior ANOVA, yet was uncorrelated between modalities, then we interpret this as meaning that the system did not successfully convey the environmental context, as seen with the case of "guilty". If an emotion was not significantly affected by ENVIRONMENT from our prior ANOVA, and it was also uncorrelated between modalities, we interpret this to mean that the emotion was either probably not present in the conversation, or attributable to the auditory modality independent of environmental influence, as seen with the case of "proud".

It should be noted that despite claims that PANAS words are largely uncorrelated, in practice, the emotion words used in this study are not necessarily orthogonal in their meaning. Though in a laboratory environment, factor analysis may reveal that words like "scared" and "upset" may be independent. In practice, when one is situated in an environment that makes them scared, we may also expect them to be upset. Therefore some inter-correlation between emotion words across modalities is to be expected.

We conclude that though our results cannot be a definitive reflection of the general population due to N=13 participants, there is encouraging evidence that the device can render visualizations that are capable of producing affective contexts in a way somewhat similar to the BGA. However, further work is required to refine the device to create stronger cross modal effects.

4.2 Qualitative Analysis

At the conclusion of the study, participants were given a textbox prompt that asked the following questions: "Did the visualizations ever take on any meaning for you? If so, what did some of their qualities represent?", and "aside from someone speaking to you, what are the most important sounds that you encounter while videoconferencing?" We performed inductive qualitative analysis on participant responses, and report findings in the remainder of this section.

4.2.1 Visualization meaning. We utilized the qualitative research method of Content Analysis [14] to investigate participant responses. A single researcher coded the results according to literal meaning, then clustered them by affinity to categories. This process yielded three categories: emotion, lack of meaning, and sound.

[**C1**: *Emotion*] 40% of participants felt that the particles had inherently emotional meaning, and did not identify any explicit relationship between the particles and sound. The properties of the particles had varied emotional significance for the participants. Colour was a common theme—P07 said that "the colours would sometimes affect how [I] viewed the emotions of the speaker." Two

participants felt that blue was more calming, and that red colours were more strong. P06 simply said that "[the particles] conveyed energy and emotion depending on color and how many there were." There were also a number of participants who had singular emotional impressions of the visualizations as a whole, saying that that they represented relaxation, excitement, or happiness. P02 grounded the visualization in real world events, saying the particles were "reflective of lighting in party clubs".

[**C2**: *Lack of meaning*] For 31% of participants, the particle animation did not take on any discernible meaning. Some participants found that the particles got in the way. P13 said "I've paid no attention to the particles, concentrating on the people." P9 said that they were "focusing on the sound" and tried their best to ignore the particles. Some participants mentioned that they found visuals distracting, especially if the they obscured an interlocutor's face. A number of participants in this subset appeared to be fixated on a particular channel of communication, such as the face, or the sound, and attempted to ignore other stimuli.

[**C3**: *Sound*] Three participants identified the relationship between particle size and the loudness of the BGA, though their understanding of this relationship appeared to be somewhat vague—P5 noted that the size of the particles represented "the business of the background." One participant said simply that the particles "represent the type of noise" in the background. No participants explicitly identified how particles reacted frequency or semantic aspects of sound.

4.2.2 Relevant Sounds. We gathered sounds most relevant to participants through a textbox entry. Two researchers independently reviewed participant responses and thematically clustered them. They then met to discuss their clusters and adjusted categories accordingly. The list of relevant sounds is hierarchically organized in Figure 6. Every leaf of the tree is a participant response. We found three overarching categories of sounds: living creatures, outdoor, and indoor. The living creatures category includes human-made sounds such as footsteps, laughter, and typing on the computer; and non-human animal sounds. The outdoor category is comprised of two subcategories: urban sounds of the city, and transportation sounds, such as a single car, or multiple cars in traffic. The distinction between urban sounds and transportation is that the former also include the hustle of the people in the city and non-car city activities such as a hotdog stand. The indoor category includes household features, appliances, and miscellaneous. Household features are the noise-making components and properties of the house, such as the sink, or the door creaking. Appliances include detached items such as a TV, fridge, or vacuum. The miscellaneous subcategory includes technical aspects of telecommunication such as feedback from the microphone and sound events that could not be consistently placed in any other category such as music.

These findings contextualize our semantic features and lay the groundwork for future work in this area. The new overarching categories enable a more parsimonious set of semantic features informed by real user data. For example, a future version of this system could replace our preliminary semantic features with simply: living creatures, outdoor, and indoor. We initially designed our semantic features to be general as we wanted to make as little assumptions as possible with regards to what particular sound events were most relevant to teleconferencing users before collecting user data. The main trade-off of this approach is semantic granularity. These findings can guide the construction of a more nuanced semantic feature ontology. For example, using these qualitative findings as an ontological framework, unique animations can distinguish between human activity and non-human activity, which can be further subdivided into animations for specific BGA events such as a baby crying or a dog barking.

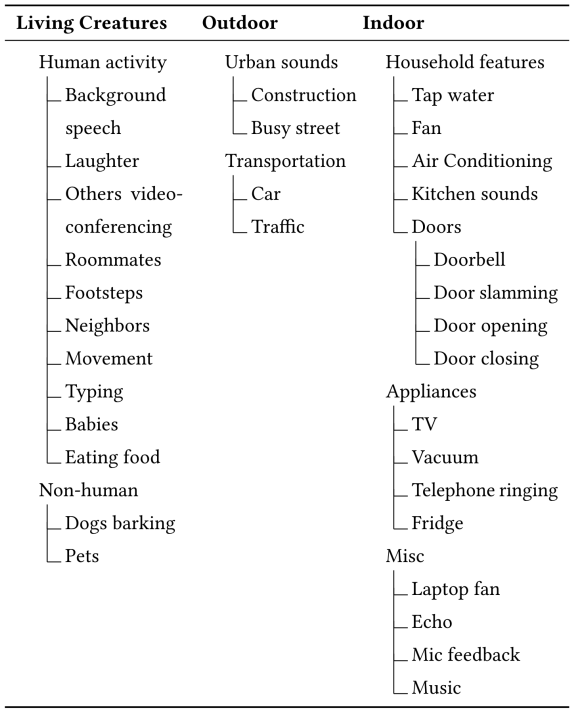


Fig. 6. Most relevant sound events

5 DISCUSSION

In this study, we demonstrated the feasibility of employing visualizations to enhance contexual affective awareness in situations where the auditory modality is unavailable or degraded. We do not claim to have discovered the optimally emotionally consistent way to represent the acoustic environment. Rather, we have demonstrated a working system that can accomplish this, and the implications of our design decisions while constructing it. In relation to our research question, we did find that animations based off the semantics and acoustics did produce similar emotional contexts, however there is much we learned from our evaluation for further areas of improvement and factors to consider. This is further elaborated upon in this section.

As participants were not told how the system worked beforehand, many had their own unique impressions about what the visuals themselves meant. Aside from a general impression that the visuals moved to sound, most participants did not pick up on technical aspects of how the visualizations worked nor explicitly identify the mapping between visualization parameters and semantic features. Most participants instead had a holistic emotional understanding of visualization meaning. This could indicate that there is some emotional meaning of the sound that is being translated cross-modally. But this may also be because participants were primed by performing an emotional rating task early on in the experiment. Participant impressions of the system remained very coarse grained overall, which may be because they had short exposure to the system. There was no "training phase" or orientation, and their exposure was limited to the length of the study—most finished within half an hour. With longitudinal use, participants might be able to determine

more specific reactions between visuals and the acoustic environment, enabling a more nuanced understanding of the particles' meaning.

Beyond the relatively short exposure to the system, it is worth mentioning that since the study was conducted amidst a global pandemic, it is possible that participants' extended use of teleconferencing platforms could have biased their experience with the experiment platform. For example, participants who had spent a significant amount of time on videoconferencing platforms on the day of their participation (e.g., working from home) could be subject to *zoom fatigue*. This in turn could have influenced their emotional and attentional states, as well as their receptivity to the stimuli employed [28]. While we acknowledge this possibility, we argue that the data is still representative of general mid- and post-pandemic videoconferencing platform usage.

Emotional perceptions may also be affected by longitudinal use. Participants may grow more adept at understanding the sound-environment-visualization relationship with repeated exposure, and would have more in-depth insights to the visuals as a result.

In addition to these limitations, it is important to acknowledge that the proposed visualization technique might not be accessible to colorblind and visually impaired users.

There are some ethical considerations when deploying such a system. Though conveying auditory contexts visually can affect users emotions, no explicit emotion classification or emotion representation is being conducted. The semantics and acoustics of the BGA are translated, but as to what that emotionally means is given to the user to decide. The device is far form a perfect reflection of the user's actual aural environment, though it may reveal aspects of the user's environment they did not wish to reveal. There are many circumstances when users may wish to suppress their environments. For example, many twitch streamers will stream in front of a green screen as to not show their room. This device was designed with a specific circumstance in mind: remote 1:1 conversations where both users wish to freely share their situational environment for more emotion centered communication, such as with family members, or close friends.

Our initial set of semantic features was utilized to demonstrate a proof-of-concept of how such a method of conveying context could work, yet the particular features used are not claimed to be optimal. Indeed, the emergent features from our relevant sound findings show a new feature ontology that may be more suitable to users needs. This should be utilized in the next iteration of the device. However, it bears mention that there is no one universal semantic ontology that generalizes to the needs of all participants. A teleconferencer who works in a day care may have a different set of relevant contextual needs compared to a teleconference who works on an industrial shop floor. As such, the ability for users to customize and define their own semantic feature ontology would be a necessary step to align the system with their idiosyncratic needs.

A way to further improve the visualizations would be to preserve the "compositional" or "polyphonus" nature of the soundscape. Sounds can be perceptually decomposed into multiple coherent textures—for example, a listener can decompose the sounds of the city to the noises of the cars, people walking by, or the rain falling. The system currently analyzes the semantics of the sound as a single "audio event". Yet a soundscape may be composed of many different audio events with a layered semantics, such as a bird (organic) and lawnmower (artificial) on a summer's day. This device was evaluated in a laboratory environment, and as such used a tightly controlled set of environmental sounds. In reality, soundscapes are complex and polyphonus. Crowdsourced annotations of soundscapes have shown that more complex sounds tend to have more semantic disagreement between raters, which could prove challenging for the semantic aspects of such a device [6]. An additional technological improvement to the device could see the incorporation of spatalized audio. Auditory direction of arrival is becoming an increasingly feasible sensing task with commodity mono microphones thanks advances in latent variable analysis as described in El Badaway et. al [13].

A future version of the system could better reflect the compositional nature of sound by visualizing parallel, perceptually salient, constituents of the source audio.

The concept of generating visualizations reflective of ambient audio when complete audition of the environment may not be available has relevance to users beyond the average teleconferencer. This system may have applications for people who are hard of hearing, though further iteration with those particular user groups is required to fully understand design requirements.

Visualizing the semantics and acoustics of ambient audio offers a promising way to implicitly convey context in circumstances when the multimodal environment is suppressed while participating in remote conversations centered on experience.

REFERENCES

[1] S. O. Adalgeirsson and C. Breazeal. 2010. MeBot: A robotic platform for socially embodied telepresence. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 15–22. https://doi.org/10.1109/HRI.2010.5453272 ISSN: 2167-2148.

[2] Andrew J Aubrey, David Marshall, Paul L Rosin, Jason Vendeventer, Douglas W Cunningham, and Christian Wallraven. 2013. Cardiff conversation database (ccdb): A database of natural dyadic conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 277–282.

[3] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current Directions in Psychological Science* 20, 5 (2011), 286–290.

[4] Tony Bergstrom and Karrie Karahalios. 2007. Conversation Clock: Visualizing audio patterns in co-located groups. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. IEEE, 78–78.

[5] Paul H Bucci, X Laura Cang, Hailey Mah, Laura Rodgers, and Karon E MacLean. 2019. Real Emotions Don't Stand Still: Toward Ecologically Viable Representation of Affective Interaction. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7.

[6] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P Bello, and Oded Nov. 2017. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–21.

[7] Yi-Chuan Chen, Pi-Chun Huang, Andy Woods, and Charles Spence. 2016. When "Bouba" equals "Kiki": Cultural commonalities and cultural differences in sound-shape correspondences. *Scientific reports* 6, 1 (2016), 1–9.

[8] Jeremy R Cooperstock. 2010. Multimodal telepresence systems. *IEEE Signal Processing Magazine* 28, 1 (2010), 77–86.

[9] Ferdinand De Saussure. 2011. *Course in general linguistics*. Columbia University Press.

[10] J Donath. 2000. Visiphone: Connecting domestic spaces with audio. In *International Conference on Auditory Display, Atlanta, April 2000*.

[11] Paul Dourish and Sara Bly. 1992. Portholes: Supporting awareness in a distributed work group. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 541–547.

[12] Annette D'Onofrio. 2014. Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm. *Language and speech* 57, 3 (2014), 367–393.

[13] Dalia El Badawy, Ivan Dokmanić, and Martin Vetterli. 2017. Acoustic DoA estimation by one unsophisticated sensor. In *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 89–98.

[14] Christen Erlingsson and Petra Brysiewicz. 2017. A hands-on guide to doing content analysis. *African Journal of Emergency Medicine* 7, 3 (2017), 93–99.

[15] Charles Forceville. 2009. Non-verbal and multimodal metaphor in a cognitivist framework: Agendas for research. In *Multimodal metaphor*. De Gruyter Mouton, 19–44.

[16] Raymond W Gibbs Jr. 2011. Evaluating conceptual metaphor theory. *Discourse processes* 48, 8 (2011), 529–562.

[17] Aviva I. Goller, Leun J. Otten, and Jamie Ward. 2009. Seeing sounds and hearing colors: An event-related potential study of auditory–visual synesthesia. *Journal of Cognitive Neuroscience* 21, 10 (10 2009), 1869–1881. https://doi.org/10.1162/jocn.2009.21134 arXiv:https://direct.mit.edu/jocn/article-pdf/21/10/1869/1937500/jocn.2009.21134.pdf

[18] Katharine H Greenaway, Elise K Kalokerinos, and Lisa A Williams. 2018. Context is everything (in emotion research). *Social and Personality Psychology Compass* 12, 6 (2018), e12393.

[19] Chris Greenhalgh and Steven Benford. 1995. MASSIVE: A collaborative virtual environment for teleconferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 2, 3 (1995), 239–261.

[20] Joshua Hailpern, Karrie Karahalios, and James Halle. 2009. Creating a spoken impact: encouraging vocalization through audio visual feedback in children with ASD. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–462.

[21] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 131–135. https://doi.org/10.1109/ICASSP.2017.7952132

[22] Karrie Karahalios and Tony Bergstrom. 2009. Social mirrors as social signals: Transforming audio into graphics. *IEEE computer graphics and applications* 29, 5 (2009), 22–32.

[23] Atsunobu Kimura, Masayuki Ihara, Minoru Kobayashi, Yoshitsugu Manabe, and Kunihiro Chihara. 2007. Visual feedback: its effect on teleconferencing. In *International Conference on Human-Computer Interaction*. Springer, 591–600.

[24] Wolfgang Köhler. 1947. *Gestalt psychology: an introduction to new concepts in modern psychology.* Liveright Pub. Corp.

[25] George Lakoff and Mark Johnson. 2008. *Metaphors we live by.* University of Chicago press.

[26] Jeana L Magyar-Moe. 2009. *Therapist's guide to positive psychological interventions.* Academic press.

[27] Tara Matthews, Janette Fong, F Wai-Ling Ho-Ching, and Jennifer Mankoff. 2006. Evaluating non-speech sound visualizations for the deaf. *Behaviour & Information Technology* 25, 4 (2006), 333–351.

[28] Hadar Nesher Shoshan and Wilken Wehrt. [n. d.]. Understanding "Zoom fatigue": A mixed-method approach. *Applied Psychology* n/a, n/a ([n. d.]). https://doi.org/10.1111/apps.12360 arXiv:https://iaap-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/apps.12360

[29] Bryan C Pijanowski, Luis J Villanueva-Rivera, Sarah L Dumyahn, Almo Farina, Bernie L Krause, Brian M Napoletano, Stuart H Gage, and Nadia Pieretti. 2011. Soundscape ecology: the science of sound in the landscape. *BioScience* 61, 3 (2011), 203–216.

[30] Vilayanur S Ramachandran and Edward M Hubbard. 2001. Synaesthesia–a window into perception, thought and language. *Journal of consciousness studies* 8, 12 (2001), 3–34.

[31] Asta Roseway, Yuliya Lutchyn, Paul Johns, Elizabeth Mynatt, and Mary Czerwinski. 2015. BioCrystal: An Ambient tool for emotion and communication. *International Journal of Mobile Human Computer Interaction (IJMHCI)* 7, 3 (2015), 20–41.

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

[33] Oliver Stefani, Milind Mahale, Achim Pross, and Matthias Bues. 2011. SmartHeliosity: emotional ergonomics through coloured light. In *International Conference on Ergonomics and Health Aspects of Work with Computers*. Springer, 226–235.

[34] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[35] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

Received January 2022; revised July 2022; accepted November 2022