Implicitly Conveying Emotion While Teleconferencing

David Giovanni Marino

Department of Electrical & Computer Engineering

McGill University

November 26, 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Electrical Engineering

Abstract

Implicit cues from the body and situational environment contribute to our sense of emotional awareness of our conversation partners while communicating. Yet many of these cues are lost while videoconferencing. With a loss of these cues comes a loss in emotional understanding. This thesis presents two technologies that enhance emotional communication between videoconferencers in an implicit manner. The first technology focuses on 1:N communication, and visualizes the expressive responses of groups of people. It was validated through two studies: a qualitative user study that had participants use the device to give live presentations (N=20), and a quantitative follow-up study to evaluate the emotional effects of the visuals (N=12). The qualitative study revealed that participants reported a feeling of alignment of emotions with the audience, and that they felt the emotions of others affected their own. The quantitative follow-up study showed that participants affectively perceived concomitant animations differently from noise. The second technology focuses on 1:1 communication, and visualizes the acoustic environments in which users are situated. The device was evaluated quantitatively and qualitatively

(N=12), and found evidence for emotional correlation between modalities. Affective computing often assumes a basic theory of emotions—where emotional states can be said to be psychologically primitive. In contrast, this thesis presents two affective interfaces that are designed assuming a constructed theory of emotions. Designing for a constructed theory of emotions is difficult because emotions are not assumed to be composed of atomic psychological states or processes, and there is no deterministic way of sensing/representing emotion. The systems ultimately provide users with a sufficiently ambiguous signal for which emotional meaning can be constructed. It was discovered that the devices could be used to systematically elicit subjective emotional perceptions from users without the use of discrete representations of emotions.

Abrégé

Les signaux implicites envoyés par le corps et l'environnement dans lequel une conversation se tient sont autant de facteurs qui permettent de prendre conscience de l'état émotionnel de notre interlocuteur. Cependant, nombre de ces signaux sont perdus lorsque la communication survient en vidéoconférence. La perte de ces signaux provoque la perte d'une compréhension émotionnelle mutuelle qui pâtit à la qualité de l'échange. Ce mémoire présente deux technologies qui améliorent la communication des émotions entre deux interlocuteurs virtuels de façon implicite. La première se concentre sur la communication 1:N en visualisant l'expressivité des réactions de l'audience. Deux études l'ont validé: une étude utilisateur qualitative dont les participants utilisaient le système pour donner des présentations en direct (N=20), et une étude quantitative de suivi pour évaluer les effets affectifs des éléments visuels eux-mêmes (N=12). L'étude qualitative a révélé que les participants ressentaient un alignement de leurs émotions avec celles de l'audience, et qu'ils étaient affectés par les émotions des autres personnes. L'étude quantitative a montré que les participants ont perçu affectivement les animations concomitantes différemment du

bruit. La seconde se concentre sur la communication 1:1, et restitue l'environnement des utilisateurs à travers diverses modalités. Le système a été évalué quantitativement et qualitativement (N=12), et a permis de démontrer l'existence d'une corrélation émotionnelle entre modalités. L'informatique affective fait souvent l'hypothèse d'une théorie rudimentaire des émotions dans laquelle les états émotionnels sont dits psychologiquement primaires. Par contraste, ce mémoire présente deux interfaces affectives conçues avec l'hypothèse d'une théorie élaborée des émotions. Une telle théorie implique des difficultés de conception liées à la nature des émotions qui ne sont plus supposées composées d'états ou de processus psychologiques atomiques, et donc qui ne peuvent plus être détectées et représentées de façon déterministe. Les systèmes proposés donnent aux utilisateurs un signal suffisamment ambigu pour qu'ils puissent reconstruire une signification émotionnelle. Il a été découvert que ces dispositifs peuvent être utilisés pour provoquer de façon systématique des perceptions émotionnelles subjectives chez les utilisateurs sans faire usage de représentations discrètes des émotions.

Acknowledgements

I would like to thank my supervisor Jeremy Cooperstock for his patience and support through my masters program. I would also like to thank all the mentors (intentional or not!) that helped me grow as a researcher before coming to McGill: Eric Vatikiotis-Bateson taught me much about how science and Language worked, Oliver Schneider and Karon MacLean guided and supported me through my entry to HCI research, Paul Bucci who was a close early collaborator and someone who still I make art and/or bad jokes with. Max Henry and Pascal Fortin were collaborators and friends at the Shared Reality Lab, who contributed to two of the studies included in this thesis. Clara Ducher, Hyejin Lee, and Yaxuan Li were early friends at the lab who supported one another Maurício Fontana de Vargas and Antoine Weill–Duflos were a driving force in haptics language research in the lab. A special thanks to my parents for keeping me curious about the world, and finally, thanks to Vesta Sahatçiu for pushing me to follow my passions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Teleconferencing has become an integral part of how we communicate. However, major teleconferencing platforms offer a degraded conversational experience compared to in-person communication. This can in part be due to the technical limitations of the platform. But it also stems from a history of design decisions that make erroneous assumptions of how human conversation works. This is best exemplified by the following misguided quote from the 1977 book *Evaluating New Telecommunications Services*:

> "Computer-based teleconferencing is a highly cognitive medium that in addition to providing technological advantages, promotes rationality by providing essential discipline and by filtering out affective components of communications. That is, computer-based teleconferencing acts as a filter, filtering out irrelevant and irrational interpersonal 'noise' and enhances the communication of

> highly-informed 'pure reasoning'—a quest of philosophers since ancient times"
>
> – Johansen et al. [1]

Since then García et al. has criticized this view, pointing out that the so-called "noise" of the signal actually "constitute[s] an integral and irreplaceable part of communication" [2].

Indeed, "noise" such as the "affective components" are essential to *grounding conversation*. The term "grounding" when applied to communication has been defined as the process where interlocutors coordinate to establish a shared understanding of their conversation [3]. This coordination can be done through many modalities—such as verbally or visually through posture or facial expression. Grounding in communication has been extended to encompass the notion of *affective grounding*, where interlocutors—artificial or otherwise—must coordinate to establish common emotional meaning [4].

When modeling the meaning of an utterance, a popular method is to formulate it in terms of its truth conditional semantics independent of context. Such an approach to modeling meaning is popular among philosophers, linguists, lawyers, and engineers alike. For example, Wolfram Alpha's Natural Language Understanding (NLU) System utilizes a formal class ontology to determine natural language semantics.[1] Of course, meaning is more than its truth conditions—the field of pragmatics demonstrates this by analyzing meaning in context [5]. The norm of conversation is implicature, where meaning doesn't obey a formal logic. For example, if Alice said "I ate at least one cookie", it is implied that there are a few cookies

[1] https://www.wolfram.com/natural-language-understanding/

left. However, if there were instead zero cookies left then this would be logically consistent with her utterance, but pragmatically unsound [6]. Affective components can also change the meaning of an utterance. If, independent of history, someone angrily said "David closed the door", that may mean that David did something wrong and it should be rectified, whereas if they said it joyfully, it may mean something entirely different. It is thus paramount to convey affective information for pragmatically meaningful conversation. Affective information can be conveyed in many paralinguistic or nonverbal aspects of communication, such as vocal prosody, body posture, and facial expression. However, it is often the case that these are suppressed while using major videoconferencing platforms. For example, eye contact is lost, and perceived head motion, and body posture is greatly degraded. It is such issues that still make in-person communication more desirable than videoconferencing when it comes to aspects of communication beyond semantics, or the "pure information" of words [7].

This thesis presents two technologies pertaining to augmenting remote conversation to enhance affective understanding. The first technology, called *CoHere*, aggregates and visualizes audience facial expressions for 1:N remote conversation. The second technology visualizes the auditory scene of interlocutors based off its acoustic and semantic properties, with an intended use case of 1:1 remote conversation.

1.1 Emotion, Affective Computing, and Affective Interaction

In order to enhance emotional communication while teleconferencing, affective components must be sensed from the environment and represented in a human understandable manner. It is a longstanding goal of affective computing to be able to understand, represent, and respond to human emotion [8]. But human emotion is poorly understood. This section will first outline a number of key theories behind how emotions emerge. It will then relate said theories to how emotion is currently computed, and theoretically motivate this thesis's approach to sensing and representing emotion.

1.1.1 Psychological Theories of Emotion

Early modern psychologists William James and Carl Lange have been associated with the James-Lange theory of emotion, where emotion is posited to be a physiological response which is perceived, and then experienced [9]. Many textbook accounts of this theory interpret James to be saying that the physiological response *is* the emotion, however James also crucially includes interpretation to be a necessary step in experiencing an emotion [10].

The Cannon-Bard theory of emotion emerged as a response to James's theory, and posited that after an event, a physiological response and emotion happen *in parallel*, and that the emotion and physiological response are distinct [11]. The Cannon-Bard theory elevates

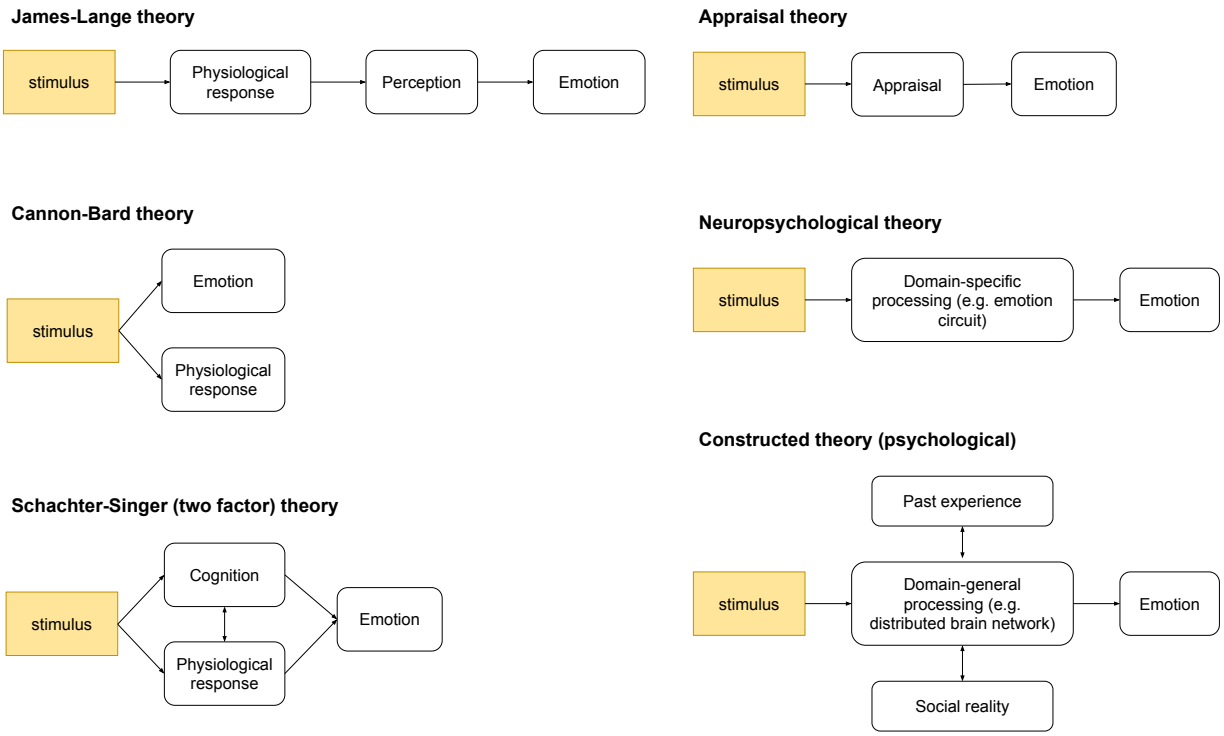


Figure 1.1: The different emotion theories discussed in section 1.1

the role of the central nervous system (CNS) in emotion by highlighting that severing the connection between viscera and the CNS did not alter emotional responses in cats and dogs. Through ablation studies, Cannon offered the thalamus and hypothalamus, as a coordinating center for emotion [12].

The Schachter-Singer theory of emotion, also called the "cognition-arousal" or "two factor theory", was the first to consider cognition as an explanatory factor in emotion [13]. In this theory, an event elicits a physiological response which is then appraised cognitively. A

cognitive label is applied to the response, which constitutes the emotion. This approach addresses how a single physiological response could be attributed to many emotional states (so-called "misattribution of arousal").

Building off these previous approaches, appraisal theories of emotion conceptualize emotions as evaluations of events. These theories stress that it is the interpretation of events, not the events themselves, that cause emotion to arise [14]. Such a formulation, like the Schachter-Singer theory of emotion, explains how two people could experience the same stimulus yet have very different emotional responses. The distinction between an appraisal theory of emotion and Schachter-Singer theory is that appraisal theories hold cognition to be *necessary and sufficient* for emotion, which implies that physiological responses are not explanatory to the emergence of emotion [15].

1.1.2 Biological and Neuropsychological Theories of Emotion

Neuropsychological theories of emotion offer a an explanation for emotion rooted in our neuroanatomy. The limbic system is generally associated with emotion, and has been shown to be a correlate of various emotional responses [16] [17]. The anatomy of the limbic system is complex, with no universal agreement of its constituent structures [18].

Emotional behaviour has been evoked via stimulating or disrupting components of the limbic system, with many early neuropsychological emotion studies focusing on affective correlates of subcortical structures such as the hypothalamus, septal nuclei, and

amygdala [19]. For example, stimulating the amygdala can induce rage [20] [21]. In gelastic epilepsy, a benign tumor near the hypothalamus can cause laughter during seizures [22]. A description of such laughter disassociates the behaviour from the emotion: "the laugh was less happy, being grimly determined and mechanical, like a machine gun" [23]. Today it would be radical and inaccurate to say that emotions *are* brain regions—or for that matter "primitive" activation patterns. Modern neuropsychological research paints a complex picture of emotion as the interplay between subcortical and cortical systems intermixed with learned behaviour patterns [19]. That said, some proponents of neuropsychological emotion theories *do* assert the existence of biologically based "emotion modules" [24] [19]. These are often conceptualized as anatomically distributed networks that explain certain affective constructs such as valence and arousal [19]. There is also considerable research on discovering discrete emotional circuits such as a "fear circuit" or "anger circuit" in the brain [25] [26] [27] [28].

On the level of behaviour, Ekman has hypothesized that there are universal facial expressions shared between all cultures, and that these facial expressions each correspond to a discrete basic emotion [29]. This notion lays the foundation of Basic Emotion Theory, where there is hypothesized to be a number of "elementary" emotions (six according to Ekman, two according to Frijda [30]) that constitute our emotional experience [31]. Emotions are considered to be discrete short behavioral and physiological responses that covary with subjective experience [32]. For emotions to be considered "basic", it is argued

that they must be distinct, hard wired, and serve a evolutionary function [31]. In essence, this theory argues for a biological basis of emotional behavior.

1.1.3 Constructed Theories of Emotion

There are a number of issues with a purely biological explanation of emotion. A large amount of evidence for the claim that an emotion can be associated with some neurological structure is derived from electrical stimulation studies. However electrical stimulation of the same location can produce different mental states depending on the individual and environmental context [33]. Perhaps even more troubling is that electrical stimulation rarely produces a discrete subjective experience consistent with how we label genuine emotion—rather a general experience of pleasure or arousal [33]. Another method used in biological explanations of emotion is in lesion studies—where a region of the brain may be deactivated either organically such as by a tumor or artificially such as with TMS. Barrett points out that it is common to have inconsistent findings with the same anatomical brain lesions, and that people with brain lesions rarely have deficits in single emotions [33]. She also calls into doubt the emotional specificity of certain brain regions, saying: "meta-analytic summaries of functional imaging results show clearly that amygdala activity is not specific to fear, insula activity is not specific to disgust, and orbitofrontal cortex activation is not specific to anger" [33].

Other biological explanations of emotions will attempt to associate emotions with

certain behaviors (e.g. freezing, grooming, laughing), or physiological responses (e.g. heart rate). However, there is no one-to-one mapping between a behaviour and emotion category [33]. Additionally, there are no physiological studies that have found consistent and specific signatures of a given emotion [33]. While many studies have shown that certain emotions are *associated* with a specific change in the face, body, or brain, these changes are not consistent [34]. As such, they cannot be wholly predictive of an emotion. Additionally, further research on Ekman style faces and emotions found that language influences our ability to perceive emotion from faces [35]. Indeed, society and culture generally influence one's perception of the world—and emotions are no exception to this [36]. For example, a single face or voice can elicit different affective perceptions between cultures [37].

Barrett considers this an "emotion paradox": we have vivid experiences of discrete emotional events, yet there are no consistent underlying biological mechanisms [38]. The constructed theory of emotion addresses many of these issues [34]. In this theory, emotions are concepts which are actively constructed by integrating past experience and social reality with the entire brain network [39]. Constructionist theories of emotion can be organized on a spectrum with *psychological construction* theories on one end, *social construction* theories on the other [40]. Psychological construction theories of emotion treat emotion as arising from perception and cognition. It sees emotions not as unique special states, or caused by special mechanisms such as an "emotion circuit". Instead, emotions arise from a distributed brain network. Psychological construction is consistent

with emotions being "embodied appraisals"[2] of the world, scripts, or schemas [33]. Social construction theories, on the other hand, differ when it comes to the brain-level explanation of emotions. These theories do not posit specific neural circuitry or networks, and consider the influence of social context being universal to all humans. Additionally they posit that neither emotions or affect, is shared between non-human animals. In contrast, with psychological construction, affect, but not emotion, is considered to be shared between humans and non-human animals. [40].

1.1.4 Affective Computing

As previously stated, it has been a longstanding goal of affective computing to be able to sense and represent emotion [8]. This poses two large challenges: (1) how can emotion be consistently sensed, given the inconsistent signals from the human body? (2) how can emotion be reliably represented, given that there is no consistent way to communicate an emotion? The various emotion theories discussed are summarized in Fig. 1.1.

Sensing emotion

When it comes to sensing an emotion, we must decide which components(s)—represented as nodes in Figure 1.1—that we want to associate with an emotion. There is no one presented theory where it is possible to readily sense all the hypothesized components in a non-invasive

[2]Though constructionist theories are not consistent with all appraisal theories of emotion, appraisal theories that consider emotions to be unique mental states/responses caused by special mechanisms differ from this conceptualization.

manner. “Cognition” and “appraisal” lack a clear physical operationalization and cannot be readily sensed. Specific emotion circuits or brain networks can be sensed, but the current technology is invasive and extremely high cost, and certainly not something that could be readily applicable to a videoconferencing platform. Components like “past experience” and “social reality” are phenomenological entities and as such are difficult to reliably compute. The most public, accessible, and well studied component is the physiological response.

Various devices have been made to categorize emotion from physiological responses with accuracy levels well above chance, e.g. stress prediction [41] [42], and basic emotion classification [43]. Notably, Picard’s team created a wearable device capable of categorizing eight basic emotions, achieving 80% accuracy for a single participant [44]. However, Picard has reflected on her past work mentioning that it would be incorrect to characterize the device as being “80% accurate in classifying emotion” as the emotion categorization task it accomplishes is very low resolution (it only identifies eight emotions instead of the myriad of emotions humans can identify) [45]. She also notes that physiological variation from day to day for the same emotion is larger than variation among different emotions on the same day [45]. This supports Barrett’s position that while it is possible to associate a physiological response with a discrete emotion, it is not possible to do this consistently (i.e., to create a single statistical model that can be deployed for general use).

Representing emotion

The question of how emotions could be reliably represented is complex. For one, there is much debate as to whether emotions can be reliably signaled from the body at all. The notion of "emotion signaling" is often analyzed in terms of isolated bodily channels—for example, focusing on the face or voice [34]. However people perform poorly when identifying facial expressions in isolation. When asked to provide free emotion labels to facial expressions, participant accuracy at identifying emotions from faces was demonstrated to be only 57.7% [34, 46] When asked instead to match congruent faces (e.g., matching two "upset" faces) their accuracy dropped even further to 42% [34]. In the domain of affective vocalizations, participants were only able to correctly identify the correct emotion in speech 60% of the time when selecting from five basic emotions [47]. It is clear that we do not use a single signaling channel, such as the face, to understand emotion. Instead we integrate a multimodal context. A well known example is that a face can be judged to be in pain in isolation, but then judged to be in pleasure if the body attached to it was posed to be celebrating [34]. The social environment, sensory domains, language, and culture, all play pivotal roles in our ability to "effortlessly" perceive emotions in others [34]. Although there is good evidence that emotion cannot be reliably represented, that of course does not mean that it is impossible to communicate emotion. For example, animators are experts at representing emotions—and the animated body itself need not be terribly complicated: a seasoned animator could make you feel the emotions of a circle. Simple 1-DOF robots are capable of communicating

emotion—most easily along the affective dimension of arousal [48]. When it comes to affective computing specifically, Picard treats emotional expression as a difficult problem, but one that could be solved through artistry. What she believes most difficult is *appropriately* signaling emotion [45]. For example, if a computer were to "smile", it wouldn't matter if the smile was a good representation of positive valence if that smile was delivered at an inappropriate time as it would be poorly received by human interactors. To this she offers no apparent solution.

1.2 Implicit Affective Representation

When communicating in-person, affective signals are for the most part conveyed implicitly (i.e., without explicit symbolic representation, such as when using language). It would thus be desirable to convey affect implicitly while videoconferencing as well. Conveying affect implicitly has a few advantages: for one, people are already familiar with contextual affective information as it is how we understand emotion while co-located. The alternative, of explicitly signaling affective information, would bring emotional symbols to the foreground, and may be distracting. Another potential hazard of explicit signaling is that it may require categorical emotion classification, which would be necessarily low resolution and prone to error. The problem of sensing emotion from any signal is best summed up, somewhat dismally for affective computing, by Barrett:

> "[I]t is not possible to literally verify whether or not a person (or non-human animal) is angry, sad, or afraid (or is in any other emotional state) using methods that do not rely on a human perceiver" [33].

Does this mean that it is technologically impossible to sense emotion? This is a difficult question to answer. As mentioned, technologies have been developed that can sense emotions at a level above chance, but have not yet reached human competence. It is conceivable with advances in deep learning, multimodal affective data can be interpolated to a set of emotion words at higher and higher rates of accuracy. Yet the quotation offers some insight in the strength of having a human perceiver, who is uniquely posed to integrate a lifetime of lived experience in their emotional understanding. The designs included in this thesis utilize the natural pattern-generating capacities of the human brain by placing the burden of emotion classification on the human, not the computer. This somewhat reconceptualizes the role of the computer in affective computing. Traditionally, the computer would play a role in sensing, classifying, and representing emotional information. In this thesis, the role of the computer is to transform sensory information into a meaningfully ambiguous signal that the human can come to their own conclusions about. This approach is similar to what is sometimes called "interactional" affective computing [49]. Such an approach has been used outside of the domain of teleconferencing. In the realm of assistive technology, systems have been co-designed to generate music from physiological signals (biomusic) to enhance the affective awareness of members of the autism community [50]. Such an approach has also

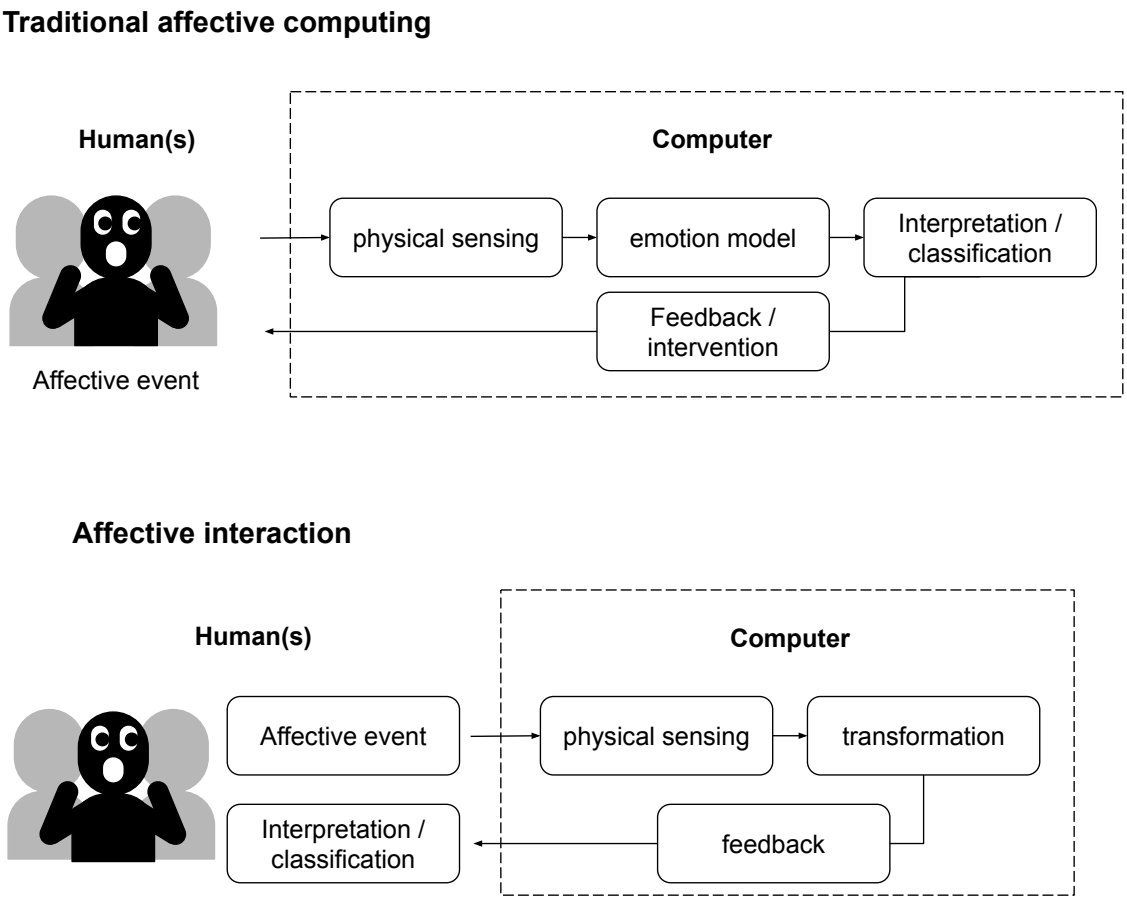


Figure 1.2: Different approaches to the human computer affective loop

been used to reveal a sense of personhood of those with profound mulitple disabilites [51]. In the realm of the workplace, Affector was a device that invited co-located workers to interpret emotional meaning of each other through a shared screen that responded to color and motion [52].

To compare approaches, consider the traditional human-computer affective loop, summarized in Figure 1.2 [53, 54]. A user demonstrates an affective event, which is then

sensed using some physical metric— be it a physiological response such as heart rate, or objective behaviour such as a smile. The system would have a formalized statistical emotion model, which is used to classify a discrete emotional state. This information is then used to symbolically deliver feedback, or provide some sort of intervention to the user. In a non-representational, interactionist setting, the system still utilizes physical sensing but instead of performing classification it transforms the data into a mode that is human understandable. It then delivers transformed physical data as feedback in a non-symbolic manner. The human performs the interpretation of this output. Using such a methodology, many problems in affective computing are circumvented. Sensing problems are bypassed since there is no emotion sensing present in the first place. The crucial problem of *when* to appropriately deliver affective feedback is also solved, since affective responses are grounded in the physical responses of the users.

Since an interactionist approach to affective computing places the burden of classification on the human instead of computer, it is compatible with a constructed theory of emotion. This is because humans are very good at integrating phenomenological events, such as past experience, and social reality, whereas computers are not[3] [55]. This thesis assumes a constructed theory of emotion. As such, the presented systems follow an interactionist approach to affect.

[3]This is of course a hot philosophical debate, but for the purposes of a Masters thesis in engineering is something we can safely assume.

Chapter 2

CoHere: Implicitly Conveying Group Affect While Teleconferencing

Preface

This chapter presents a manuscript that is currently in submission to a peer-reviewed conference. This work presents CoHere, a videoconferencing module designed to enhance the affective awareness of the audience during 1:N calls. CoHere generates a particle visualization that is animated by the facial landmarks of users. Through a qualitative in-situ study, we found that CoHere enabled affective alignment between active participants of the video call. Participants felt that the device enhanced their ability to express themselves emotionally and relieved social pressure. A quantitative follow up experiment

found that CoHere's visualizations significantly affected the user's affect judgements.

Author's Contribution

This work is a collaboration between David Marino, Max Henry, Pascal Fortin, and Jeremy Cooperstock. I architected this system, ran participants, and conducted primary qualitative and quantitative data analysis and wrote the bulk of the submitted paper. Max Henry and Pascal Fortin contributed intellectually to the design of the study, as well as experimental methodology, overall system look and feel, and paper editing. Max Henry also contributed auxiliary scripts for data analysis. Jeremy Cooperstock supervised the research and edited the manuscript.

Abstract

Participants in one-to-many videoconferencing calls often experience a significant loss of affective feedback due to the limitations of the medium. To address this problem, we present CoHere, a videoconferencing module that conveys group affect in an implicit, continuous manner. CoHere consolidates and visualizes nonverbal behaviour of fellow videoconferencers, providing a sufficient medium for others to meaningfully perceive emotion. In a qualitative user study (N=20), users reported feeling a sense of alignment with the emotions of the other participants using the system, and audience members confirmed that the system offered a low-attention alternative to gauging the sentiment of the crowd, though broadcasters had additional attentional requirements. CoHere further created a supportive environment that encouraged users to emote more, and relieved the social pressure commonly experienced in group calls. A quantitative follow up study showed that CoHere's visualizations significantly affected user's judgments of arousal and valence while watching slideshows.

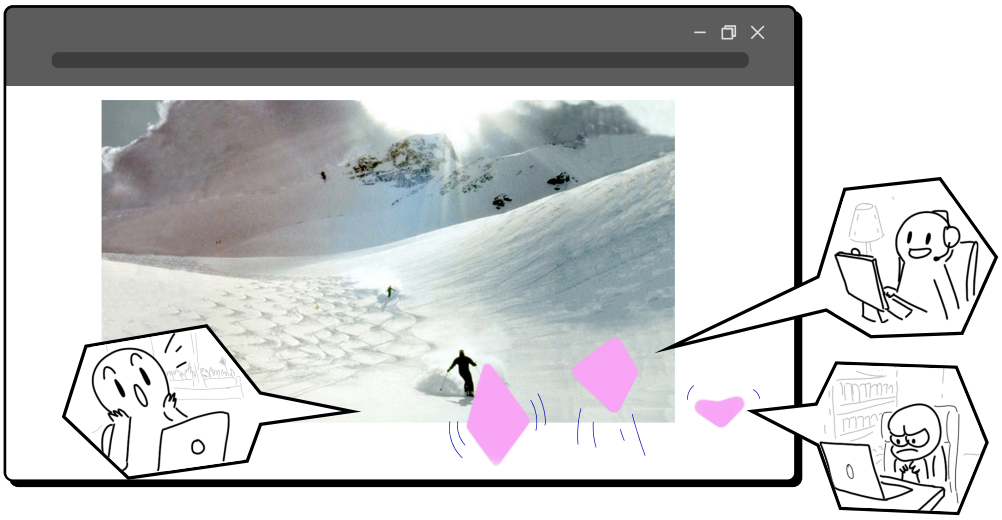


Figure 2.1: CoHere displays the expressions of teleconferencing participants as a particle visualization to implicitly convey group affect during one-to-many calls.

2.1 Introduction

A conversation is more than an exchange of words: it is accompanied by a wealth of paralinguistic, nonverbal, and contextual cues that give a rich interpretive medium for which layers of meaning can be derived [56]. Essential nonverbal cues, such as head motion and gesture, are often lost or degraded while using commercial videoconferencing platforms. This becomes particularly problematic when participating in a one-to-many videocall, such as a presentation or livestream, where there could be a total loss of audience feedback due to disabled cameras, a presenter taking up the majority of the viewport, or extremely low resolution of the viewers. In such cases, there can be a large loss of affective awareness as nonverbal behaviour such as facial expression are key to understanding the

emotional state of interlocutors. If this information is mostly discarded, it may feel to the presenter as though they are talking to themselves.

In the case of a one-to-many presentation, such as a conference talk, comedy show, or musical performance, it is highly desirable to have an understanding of the emotions of the people in the room. For example, a comedian may want affective feedback from the crowd to know if a joke they told landed, or a teacher may want to know if their students are bored or engaged. The audience, likewise, may want affective awareness of the crowd: in a concert, paralinguistic audience behaviour such as body movement, cheering, and entrainment with the performers, provides meaningful cues to enable an emotionally engaging experience [57].

We contribute CoHere, a videoconferencing module that provides a particle visualization of audience facial expression for 1:N video calls. The facial expressions of audience members are mapped to individual particles ("particle avatars") on the screen. Through a qualitative in-situ study, we demonstrate that CoHere provides an emotionally encouraging environment, inviting users to express themselves more freely and to a greater degree. CoHere also gives participants awareness of each other's emotions in an implicit manner. This enabled psychosocial phenomena such as emotional contagion to occur. A quantitative follow up study showed that experienced participants were also able to discern emotions from the particles that were systematically different from noise. A link to a publicly accessible video demonstration and high level overview is provided in this

footnote.[1]

2.2 Background

Hasib et al. [58] outlined a design space for audience sensing and feedback systems, including the dimensions of sender/receiver cardinality (1:1, 1:N, N:N), audience feedback style (explicit, implicit), audience location (collocated, distributed), and synchrony of feedback (synchronous, asynchronous). Here, we broadly cover audience sensing mechanisms across these dimensions but pay special attention to distributed synchronous communication, which typifies most videoconferencing systems.

CoHere approaches affective communication from a constructivist viewpoint. It differs from prior work by using this perspective to provide an ambiguous signal from which users can construct their own meaning. By doing so, it is a system that both *senses* and *conveys* emotion in an implicit manner for 1:N videoconferencing. A final distinguishing feature of CoHere is that it uses non-invasive commodity hardware and displays affective feedback without sacrificing major screen real estate.

2.2.1 Affective audience sensing and teleconferencing

A multitude of approaches have been used to sense audience affect. Audience emotion can be sensed both explicitly and implicitly.

[1] `https://drive.google.com/file/d/1tH-ksp5jd7X1VffYqjnkwqJap5oxz447/view?usp=sharing`

On the explicit side, Live Interest Meter was an app to convey audience engagement by explicitly polling audience members using their smartphones and presenting visualizations of the results [59].

On the implicit side, Biosignal sensing has a long history of use for understanding audience affect. Galvanic Skin Response (GSR) data were used to study the affective response of an audience to performing arts shows and live presentations [60] [61], as well as student engagement in distributed learning environments [62]. EngageMeter was a system that used electroencephalography (EEG) to sense audience engagement in conference settings and displayed engagement levels using a graph or scalar gauge visualization [63]. On the level of body analysis, posutural synchrony was utilized to infer affective feedback from audience members seated in ambient sensing chairs [64].

The human face offers a rich medium for affective feedback that has been often applied to teleconferencing. Simply compositing a feed of a remote conversation partner's face in the center of a user's gaze point, instead of the side of the screen, is sufficient to enhance feelings of emotional interdependence [65]. De Silva et al. [66] used a facial emotion classifier to animate exaggerated 3D avatars in a shared virtual space. Affective Spotlight is a Microsoft Teams extension that operates as a realtime video feed switcher by "spotlighting" user feeds that are algorithmically determined to be emotionally relevant while videoconferencing [67]. Using a similar video switching paradigm, motion detection and speech were used to cut between video feeds of a colocated meeting with the aim of enhancing engagement [68].

Multimodal face and speech data have been used in videoconferencing to classify user affect and display relevant emotion words over their video feeds [69].

During a traditional 1:N videocall, it can be attentionally demanding to to signal your emotions—one could manually select an emoji, or switch focus from the screen to the text chat. It is also has the potential to be attentionally demanding to other viewers as well—a deluge of side expressions in the chat window, or another user interrupting the speaker with emotive vocalizations can break the flow of conversation. It is thus desirable to have implicit input, similar to the prior implicit sensing devices reported in this section. This closely resembles what happens in real life, where there is no distinction between "input" and "output" of your emotions: you simply just smile to indicate some mental state without worrying about inputting the emotion to the system of your conversation. Prior work that utilized specialized devices such as EEG or GSR are invasive for everyday use, or require specialized hardware not commonly available to users. In constrast, CoHere uses everyday sensing devices (i.e., the user webcam) to implicitly translate affective signals. This approach to emotion sensing is not unique in of itself, but the combination of how CoHere both senses and represents emotion, and how that is applied to teleconferencing, is what distinguishes it most from prior work.

2.2.2 Representation of affect

All affective teleconferencing systems presuppose a theory of emotion, but emotions are ontologically difficult to define. Many systems that use emotion classifiers operate on the assumption of basic emotions—defined as a set of discrete, psychologically primitive affective states. Some basic emotion theories posit that there are universal atomic emotions shared between cultures [70]. Many emotion classifiers aim to map a biophysical signal to a primitive emotion, which is then typically presented to the viewer [71]. A popular basic emotion theory posits that there are seven universal facial expressions, known today as "Ekman faces" [29]. Despite its widespread theoretical adoption in AI and HCI research, the notion of universal facial expressions has little empirical support as facial configurations have been shown to map to multiple emotions, and vary across cultures [72]. Additionally, an agent's emotion cannot be be wholly determined by a single signal, such as a smile, or vocal quality; the signal is always situated in a complex context which affects its interpretation [73]. Such an interpretation of emotion is consistent with a constructed theory of emotion. In this theory, emotions are not construed to be atomic universals but instead concepts that arise by utilizing the brain's inherent pattern-generating capabilities, integrating past experience and realtime ambiguous stimulation [39, 74]. We adopt this theoretical position when designing our system. Instead of classifying discrete emotional states, or showing users a representation of atomic emotions, we aim to show the user a sufficiently ambiguous signal for which they can ascribe emotional meaning to themselves. By doing so, we shift the burden of emotion

classification from the computational system to the user, and utilize the brain's ability to form patterns and concepts from ambiguous data [75].

We are of the position that it is important to represent emotion implicitly during a video call, because that is how emotion is represented in real life. Explicitly signaling emotion can break the flow of conversation, and has the potential to divert attention from the topic at hand. It also runs the risk of misrepresenting what users feel because it suffers from low resolution. For example, consider if a user smiled during a conversation, and the system subsequently signaled "HAPPY". Perhaps they were only slightly happy, perhaps they smiled to make others feel more at ease, perhaps they smiled because they were uncomfortable. Regardless of the original intent, viewers would have a unified yet inaccurate picture of the source emotion. There are also times when basic emotions are of an inadequate granularity. For example, one study prototyped a 1:1 videocalling app for individuals with Autism Spectrum Disorder where facial expressions were translated to explicit symbols such as emojis or emotion words based off six basic emotions [76]. Participants found basic emotions unhelpful because they struggled more with understanding complex emotions such as frustration, sarcasm, or confusion [76].

CoHere thus strives to convey emotion to users without using explicit affective symbols. In terms of a design philosophy, CoHere is most aligned with that of affective interaction, where emotion is considered inherent to interaction—it is "dynamic, culturally mediated, and socially constructed and experienceed" [49].

2.2.3 Grounded conversation

Conversation is a cooperative task. As such, the notion of grounding in communication is extremely important for successful dialogue [3]. Grounding conversation entails that interlocutors must continuously coordinate to establish shared common knowledge and beliefs [3]. A closely related concept, that of interactive alignment, extends the notion of grounding—it claims that when successfully communicating, interlocutors align representations among all levels of language: phonetically, syntactically, semantically, and situationally [77]. This is evidenced both behaviorally by shared linguistic constructions between conversants, and also neurobiologically in studies that show a coupling between perception and action between conversants completing joint conversational tasks [78–80]. In the field of human-robot interaction, grounding has been extended beyond conversation to also encompass affect. This has been called affective grounding, where affective ground is established when interactors coordinate on how behavior is to be emotionally understood [4]. Just as grounding in communication conceptualizes conversation as collaborative, affective grounding conceptualizes emotion as collaborative.

Videoconferencing poses challenges to grounding: interlocutors do not share the same physical environment, and sometimes they are not visible [3]. In 1:N communication in particular, the challenge to grounding is even greater as obtaining a shared situational sense between all conversants may be near impossible. A crucial mechanism of coordination that enables grounding is backchannel communication—the presence of verbal

and nonverbal cues like "mhmm" and head nods during conversation [4, 81]. However, in a video call, backchannel communication is often heavily suppressed. A high level design goal of our system is to facilitate the establishment of affective ground between interlocutors. We primarily tackle this problem by visualizing the non-verbal backchannel of head motion and facial configuration of participants.

2.3 System design

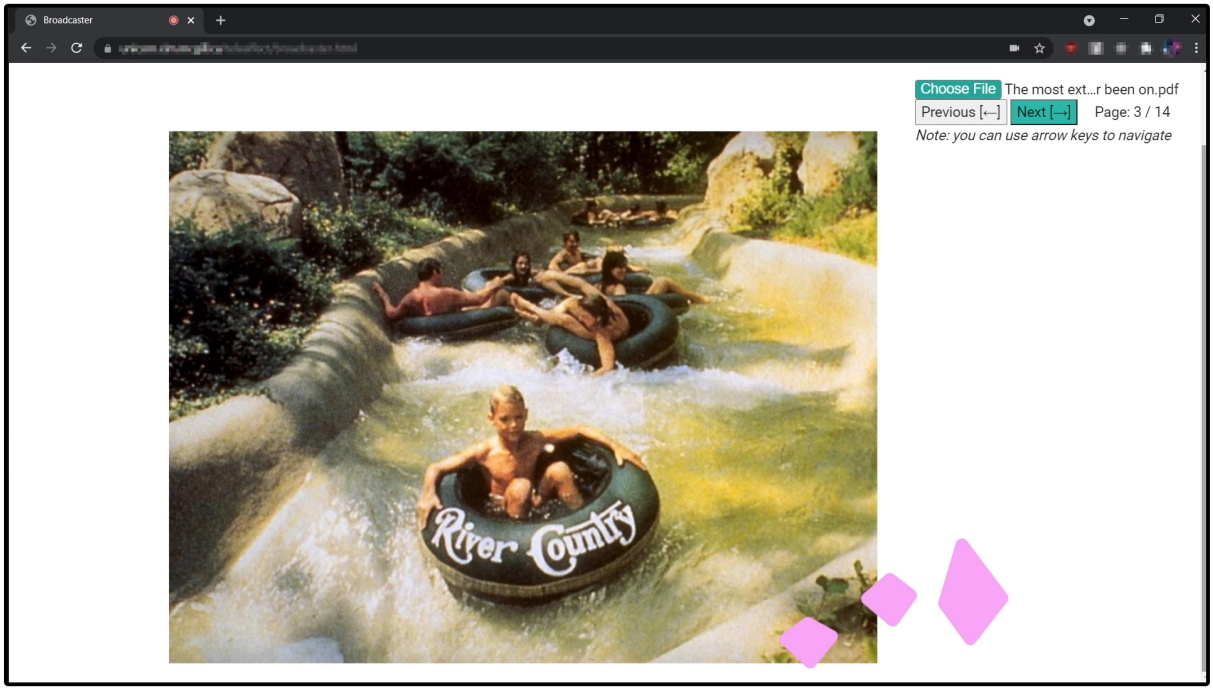


Figure 2.2: A live screenshot of the GUI with 3 participants. The pictured stream is from a slideshow with a high valence, high arousal affective target.

CoHere is a browser- based videoconferencing system that communicates the affective

states of participants to one another. The proposed system is currently designed around the task of giving 1:N presentations, in the form of slideshows. This first application was chosen since it is representative of numerous 1:N remote activities, e.g., teaching and academic presentations. Furthermore, since a slideshow typically takes up the majority of the viewport, it imposes itself as a scenario where the loss of non-verbal cues from the audience can be particularly severe. Facial landmarks were extracted from user video feeds, and utilized to animate the particle avatars. A single particle avatar is animated as follows: absolute positioning of the head affects the avatar's origin point; head roll, pitch, and yaw move the avatar in corresponding directions; eyebrow motion adjust the top size of the avatar; mouth y-distance affects the bottom size of the avatar; mouth x-distance adds sinusoidal motion and expands the sides of the avatar. A sketch of the algorithm showing the relationship between facial landmarks and particle avatar parameters is outlined in Fig. 2.3. While there were a multitude of animation possibilities, the current methods were chosen to convey a natural mapping between live head motion and particle animation. An author who is also an experienced motion graphics designer manually tuned particle avatar animation parameters through trial and error to create animation they found to be compelling. Landmark data was captured at 14 fps and filtered using an autoregressive moving average function to smooth the signal.

New users must first calibrate their facial parameters prior to using the system. This is done automatically by linearly interpolating participant faces to min/max ranges that

are ideal for compelling avatar animations; however users are also given slider control to manually adjust animation settings. This was because the "basic face" used to automatically normalize participant face parameters was that of one of the authors. Since no two people share the same resting face [82], the additional manual controls were provided to ensure that the particle avatar was sufficiently responsive to each user.

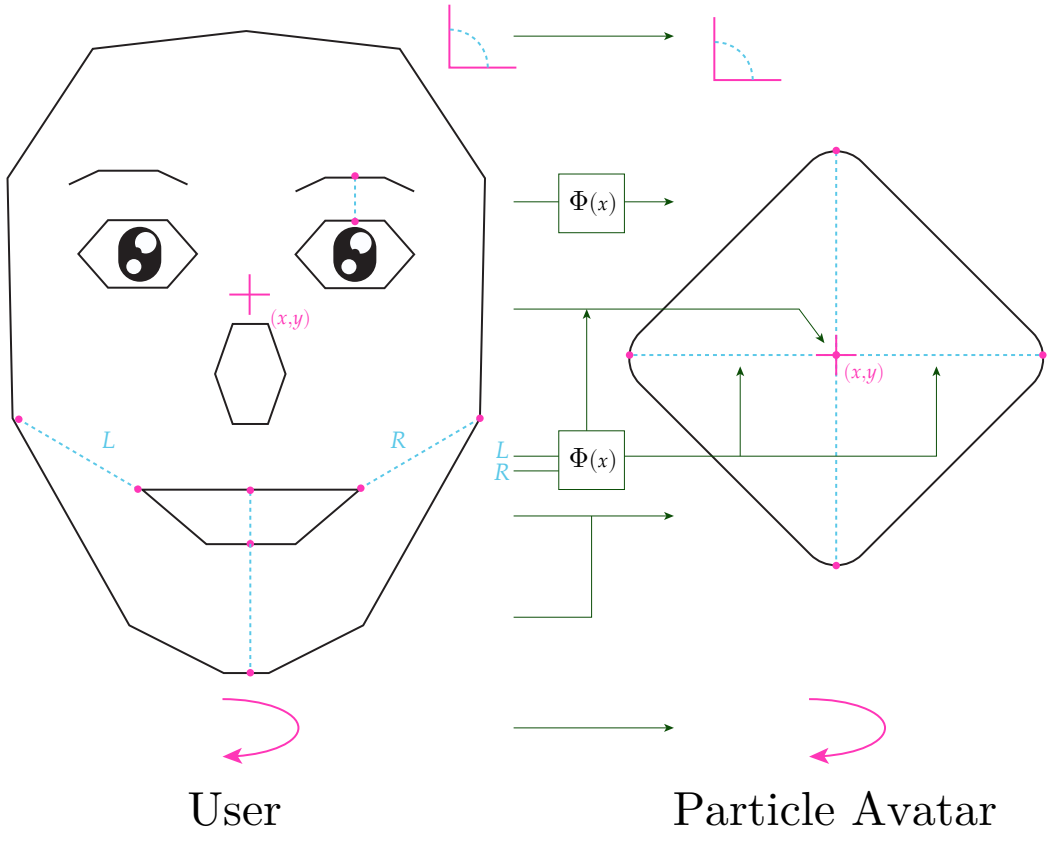


Figure 2.3: Animation parameter mapping between a user and their particle avatar. Lines with $\phi(x)$ utilize nonlinear mappings, where plain lines utilize linear mappings.

When a presentation starts, a number of particle avatars populate the bottom right of

the screen. Every participant of the video call, including the presenter, has a corresponding particle avatar shown on screen. No user video data is shown to other participants, stored, or transmitted. CoHere is designed to augment a traditional videoconferencing system by overlaying particles to a video feed, i.e., of a presentation or screen share. All avatars are of uniform color and size for a sense of anonymity. A high level overview of the system is shown in Fig. 2.4.

2.4 Qualitative User Study

A user study was conducted to investigate the experience and feasibility of using such a system to convey affective feedback while videoconferencing. The system was evaluated qualitatively, focusing on a live presentation task with groups of 3-5 concurrent users.

2.4.1 Methods and Overview

Participants were asked to use the system to give presentations to a live audience. A total of 20 participants were recruited through a combination of social media ads and snowball sampling. We hosted 5 sessions, each containing 3-5 participants. In a single session, participants took turns giving 2-5 minute presentations to one another. At the conclusion of the presentations, a semi-structured interview was conducted to discuss their experience using the system. During interviews, we used the guiding questions: “How was using this system in comparison to your everyday videoconferencing experience? ”, “Were there

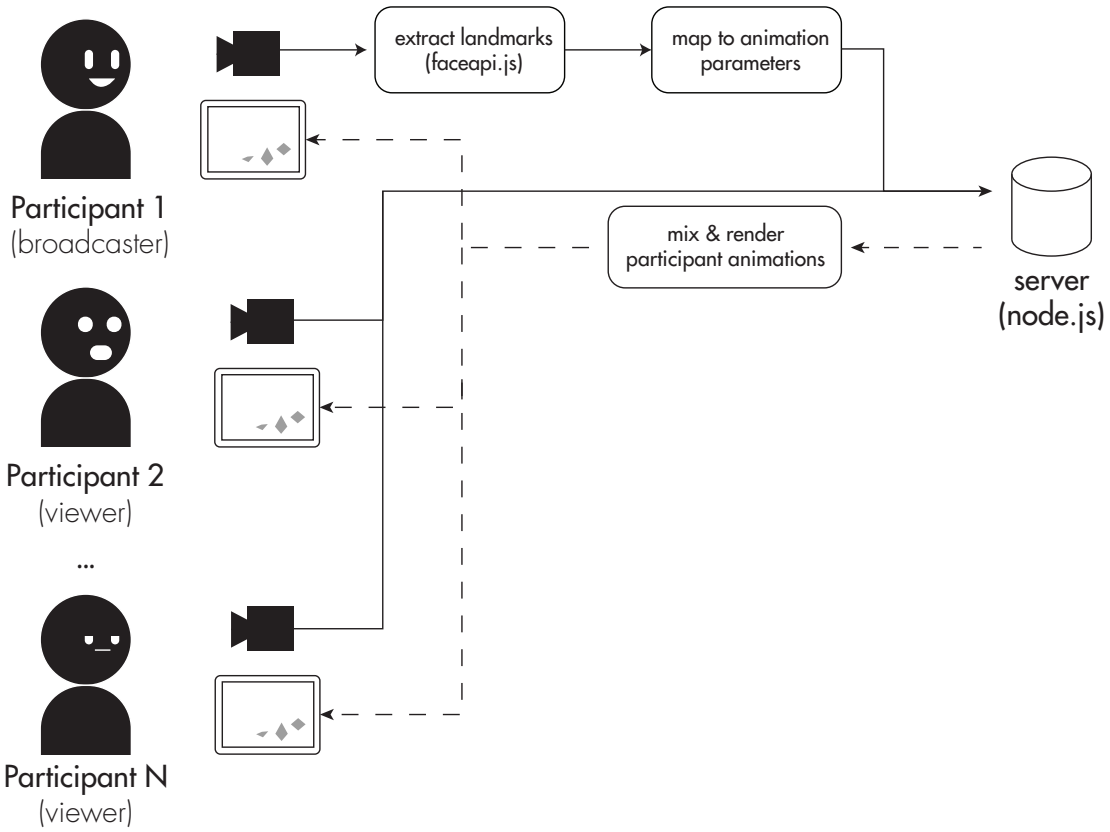


Figure 2.4: High level system architecture. Participant face landmark data is extracted from their webcam feed. Analysis is conducted to map landmark values to animation parameters for their particle avatar. A server broadcasts all participants' animation parameters. Values are then mixed together and animated onto the screen for all users.

moments that stuck out to you in terms of how people were feeling and reacting to the presentations?", "How did the system affect your attention during presentation?", and "What do you consider the ethics of using such a system to be in everyday use?". These questions were designed as starting points to create unstructured conversation for which more precise and circumstantial follow-up questions could be asked.

To fully understand the expressive limitations of the system, we needed to investigate a large affective range. Accordingly, presentations were designed to target quadrants of Russell's 2D circumplex model of affect [83]. To create a presentation, a single researcher selected validated images from the International Affective Picture System (IAPS) that were consistent with the affective target [84]. The goal of selecting validated images was to ensure that the affective target was reached at least once in a presentation. Images were sequenced to lay the foundations of a 2-5 minute slideshow story. Images from the database that were subjectively deemed to be inappropriate for day-to-day teleconferencing (e.g., explicit sexual content or gore) were excluded. Filler images to "complete" the story were sourced from Google Images. Natural, compelling stories and presentations have dynamic and changing emotions. Thus, a single slideshow was not assumed to be emblematic of a single emotion. The slide shows were designed to be ambiguous as to leave room for improvisational user interpretation. As such, they incorporated little to no text and the connections between successive images were not explicitly stated. Participants were given the option of improvising a story to go along with the slide show, creating their own notes beforehand, or, if they were uncomfortable, asking the researcher for a script. For the participants who asked for a script, a single researcher wrote one stream of consciousness to the selected images. Participants were also given the option to modify the slide show in any way to serve their version of the story so long as it was consistent with the affective target (i.e., the slideshow contained validated images that were consistent with the affective

Theme	Categories
A social space between mediums	Experience Promotion of Emotional Expression Attention Alignment
Situational dynamics affect design requirements and considerations of use	Design considerations Identity Ethics

Table 2.1: A high level overview of the categories and themes that emerged from content analysis of interview data

target). One participant opted to include their own photos.

In a single session, slideshows were assigned to participants in a way such that there were no duplicates of affective targets. Over the course of the entire study, 45% of slideshows had positive valence targets, and 50% of them had positive arousal targets.[2]

The interviews were analyzed using Content Analysis, a qualitative research technique in which data is coded and grouped by affinity into categories, which are then inductively grouped into larger themes [85]. In this process, a single researcher transcribed all the interview audio. Then, they coded the interview audio according to literal meaning. Afterwards, the codes were grouped by similar meaning into larger categories. Finally, broad explanatory themes were induced to connect clusters of categories. The emergent themes and categories are outlined in Table 2.1 and reported below.

[2]Despite having 20 participants, these numbers were not perfectly 50/50 due to participants canceling/dropping out and new slideshows having to be scheduled ad hoc.

2.4.2 Theme I: A social space between mediums

Participants felt that CoHere presents a unique social space, somewhere between an phone call and a video call, and between having cameras on and off (Category 1A). The anonymous nature of the space helped to relieve pressure and encouraged participants to emote more freely (Category 1B). For viewers, it provided a lower attention alternative to assessing the feelings of the crowd (Category 1C). CoHere is an empathetic experience, which encouraged participants to align in emotions with one another (Category 1D). The qualitative data that supports these claims is elaborated in the categories below.

Category 1A: Experience

Participants felt that the visualization was "fun to watch" (P7), with P4 feeling that the interface "felt like it's alive, it's not just a disembodied hand [makes thumbs up emoji gesture]". There was the general feeling of the interface feeling "in the middle" between mediums, P11 said " it replicates a little bit more giving a presentation in real life just because when it's over a video call, it's so zoomed in on people's faces, it's not like that they get blended into a crowd". The device was also experienced as being "in the middle" of having one's camera off and on (P4). P8 mentioned that they could "see it being a nice medium between total video off and video on". P14 said that it felt like "a nice in-between...[where] you want people to know that you're present and listening but you're not super comfortable turning on your camera on".

P5 felt that the device added an element of gesture or side channel communication to a video call saying that "it gives that ability to kind of go like pssst" to express emotions. P5 further expressed that "I think it's almost unanimously agreed that speaking in a [video] conference always feels like you're interrupting somebody...there's no room for sharing a moment or sharing an emotion with somebody who's not necessarily speaking. So being able to have this little 'oh we're communicating' like little blobs and it's like smiling...make it feel like you're in the community space a little bit or I guess approximating towards that".

Category 1B: Promotion of Emotional Expression

Many participants felt that the feeling of "in-betweeness" took the pressure off of communication and by doing so encouraged them to emote more. P9 said "it took a lot off a lot of the pressure on my reactions so I could express freely. I feel like there's like this unspoken code where you shouldn't laugh too much or you shouldn't look too pissed off...but if you're anonymous and you can't hear them you actually express as much as you want and I think there's a freedom to it." P11 said "I feel like that takes off the pressure for both parties (the viewers and the broadcasters) so the broadcasters are not frantically looking at every single person's face...and it allows you to be less anxious, and then for [viewers] I feel like you get more genuine reactions because they know that you can't see them and you don't know who's reacting to what". P3 extended that they were "actually emoting a lot more than I usually do...it took out the peer pressure in a good way". P5 said

that the alternate communication modality encouraged them to emote more: “I like having something that I can control on the screen, so I like [that] as soon as I know that I can display emotions and react [it] makes me want to react more and engage more”. P4 had a similar sentiment, and adjusted their emoting style to meet system parameters “I wanted to it to be more emotive you know, sometimes when I was really reacting I get really close to the camera then my thing [avatar] would get really big just because I was like ‘I need them to know like that I’m really happy about hearing this’...I was trying to kind of push it to the limits to kind of show the presenter [how] I felt about it”. P1, who is a comedian, brought up that emoting is often a form of “support”, and that at a comedy show emoting in the audience is a form of supporting the comedian. P9 felt this sentiment while using the system, saying that “as a presenter it was also almost a like a subtle encouragement to see people reacting on the side...[and] it takes off the pressure when everyone’s anonymous, not seeing everyone’s faces made me feel less like the eyes were on me”.

Category 1C: Attention

Participants found that the system was a low attention alternative to traditional video tiles, on the basis that it aided in reducing distractions and mental workload for viewers.

P3 said that “I do think that this is less distracting than have cameras on because...I feel like when I go to big meetings I just I go through people’s photos see what you’re doing like what room are they in like it’s a lot of it’s more information that way so this is to me

less distracting". P15 shared a similar sentiment, and said they were always drawn to what's in the background and particulars of what fellow videoconferencers are doing. P16 said "I prefer [it] over having actual video feeds of myself and of people. I find video of just people's faces really distracting and when I'm a viewer I end up getting really concerned with how I look". P4 felt that it helped them focus on the task at hand: "you're just kind of like focusing on what you're doing instead of trying to read people's faces and like match them to a person".

That said, there was a novelty effect that some participants found distracting, illustrated by P9's statement: "it's still a new experience...my immediate association with it is that talking mirror from Shrek because it looks like that and it moves like it and so it feels cartoony...it would take practice with it to to understand it, and to for not to be so distracting". P3 said that they felt it was "distracting for me at least at the very beginning because [it was] new and I'm trying to see everything I can do with it", though later said that it was altogether less distracting than traditional videoconferencing.

There was also differing levels of distraction between broadcaster and viewer roles. While broadcasting, most participants tended not to look at the visuals, opting to focus more on their notes. P18 said that while broadcasting "I gave less attention to [CoHere] because my focus was more on speaking stuff and getting everything on track". Of note is a distinction between participants who decided to improvise their lines, and participants who pre-wrote scripts. Some participants who improvised their lines had different feelings about using

CoHere as a broadcaster. P6, who improvised her lines, said: “as a presenter I was pretty tuned in and I was like oh no I hope they’re not like dead pan or like tough crowd or whatever”. Contrarily, P10, who also improvised her lines, said: “when I was giving the presentation, part me was kind of like oh it would be interesting to see if it helps gauge people’s reactions, but I found that because I was so focused on being ‘okay wait what slide is going to come up next’ I didn’t find myself looking down and kind of knowing how people react”. P10 is a trained improviser and mentioned that “in person I find that usually it is audio feedback that’s the most prominent when your eyes are distracted by your presentation”. A future version of CoHere could thus utilize alternate modalities to aid in visual-attentional saturation while presenting. P1, a comedian, said that as a broadcaster they didn’t look down at the visualization much because “I’m not necessarily as familiar with all this material. This isn’t material that I thought of. I don’t have like an expected reaction necessarily”, and emphasized that if it they had material they knew intimately it would greatly impact their user experience because they would be looking for specific reactions. P16 said that while they were a broadcaster, they only really noticed the visuals “when I was trying to make a joke...or saying something that had a little more of an impact, like that’s when I would notice if people were moving or not, which I think is nice because again it’s less distracting, because in a real space ...everyone’s just around a table or in a room you don’t have this really close-up view of everyone’s face”.

Nonetheless, viewers found that the anonymous nature of CoHere also took pressure off

and aided in focus, as P16 states: “I wasn’t worrying as much about what my reaction looked like and it was a lot easier to focus on the presentation for me and be more comfortable with ‘oh are my reactions matching everyone else’s’ ”.

Category 1D: Alignment

Many participants had the urge to align their facial expressions and emotions with others in the audience. Facial expression and emotion are correlated and form the basis of the facial feedback hypothesis, where one’s facial expression directly impacts their emotions [86] [87]. Taken together this is evidence of emotional contagion, which sees the emotions of the groups of people converge [88].

P9, on feeling their emotions converge with others, said: “it was kind of like a unison kind of thing, like I felt encouraged to want to join in and on that same emotion when seeing those visual cues”. P7 mentioned “I kind of felt influenced to match or replicate what I thought other people were doing...like a big smile or something like that”. P3 further said, “one thing I found very interesting is that I noticed other people reacting a certain way and it made me feel like I wanted to react that way, I was like everyone’s smiling I should also smile”. P15 brought up an example where others emotions affected hers, during a negative valence presentation with a jump scare: “there was like this creepy creature and someone laughed and it made me laugh”. P5 became aware of this effect in the middle of the session and reflected upon it afterwards: “[when] I smile and [it makes] somebody else smile it does

make it feel like you're in the community space a little bit or like I guess approximating towards that".

There was also individual concern if their emotions matched the crowd. P19 said that while watching presentations "I was seeing, is everybody else laughing as well, or just am I?". P10 shared a similar sentiment, wanting to socially contextualize their emotions, saying: "I was like oh this is humorous to me, and [then] I was like 'oh I wonder how other people are reacting to it' ".

2.4.3 Theme 2: Situational dynamics affect design requirements and considerations of use

CoHere was designed for 1:N video calls However, relevant nonverbal cues to communicate between calls varies widely between situational contexts. CoHere's appropriateness to the current use case and further design considerations are discussed in Category 2A. The nature of anonyminity in CoHere's design is explored in Category 2B. Finally, we elicited any ethics related concerns participants had, and reported them in Category 2C.

Category 2A: Design considerations

There was much discussion among participants about how the context of use affects device requirements.

Participants agreed that 1:N presentations was the best use case of such a system. P2

further stated that it felt best suited for emotion centered presentations: “if you’re giving a serious presentation where emotion is not really part of the picture then I feel like this won’t be as useful vs if you’re doing a twitch stream or something and emotions are a big part of that ”. The notion of “emotional support” or “encouragement” was at the crux of determining CoHere’s situational appropriateness. P1 said “there’s some cases where everybody...is muted for a reason” such as in work meetings where they feel emotional support isn’t as needed. P5 felt that the device was helpful in online theatre or keynote talks “to encourage the speaker”. In professional contexts, the need of intentional control was expressed. P9 told an anecdote of giving a presentation in a work meeting, where a client yawned and “it immediately sunk my confidence, which wasn’t very much to begin with...but if we were to have a platform like this I think I would have preferred not to have seen him yawn. I would have preferred him to have his own little avatar and reacting however he wants and he can convey whatever emotions or questions that he wants behind a mask”.

P12 highlighted the fact that relevant nonverbal feedback differs between presentation types, and thus may require different animation mappings: “[the] kind of data I’d be looking for might be different to if I was in a work meeting, and it was a collaborative project and I was looking for people shaking their head or nodding”.

The use of the visual modality was mentioned to be not as useful in some situations. As previously mentioned in Category 1C: Attention, the high visual mental workload when presenting meant that some broadcasters didn’t have the capacity to attend the visualization,

and thus the audio modality was suggested to represent affective audience feedback, as typically experienced in colocated performances. P16 said "I don't think I would want to use it for a watch party because...if I'm watching a movie in person I'm not usually looking at what other people are doing, I'm watching a movie and then people make comments sometimes it's mostly an audio thing". Participants were split on the specific topic of watch parties saying "for something like a watch party where it's low stakes...it's a nice way to just get a sense of 'yeah I want to feel what the vibe is"' (P2), "I think it could be a better version of Netflix party where it's just the text-based conversation people have" (P18). Regardless of its appropriateness to a watch party, non-visual modalities offer an interesting venue for future work.

The general feeling of "taking the pressure off" sparked debate about its appropriateness in a classroom. P20, who is a high school teacher, saw value in his classroom relating it to his experience teaching online during the pandemic "students just like to be hidden...[maybe] they don't want to show their messy room...[maybe] they just like that anonymity...and I think there's just a lot that has to do with self esteem". P8 extended that there are also circumstances where students wouldn't want to have their cameras on because of family and environmental concerns, so a visualization such as this is ideal. P14, a business and computer science undergraduate student, felt "it's another way for students...to be more interactive in even a lecture based format" but stressed that there were different degrees of interactivity required between their computer science and business classes. However, P6, a

masters student in computer science, said that there are times when she prefers to be fully hidden in class: “sometimes in lecture, I prefer [to have] camera off because I can listen to it like a podcast, and I don’t have to be hyper attentive, and it helps with Zoom fatigue”.

Taken together, system is thus concluded to be best suited for informal emotion centered 1:N casual presentations such as twitch streams, or informal conferences.

Category 2B: Identity

There was much debate among participants as to the value of having anonymous feedback. While many participants agreed that anonymous feedback helped alleviate pressure while videoconferencing, there were some who were nonetheless very concerned with identifying who they were. P4 and P8 both mentioned that they would move their faces in exaggerated manners to try and identify themselves in the crowd. P19 found themselves very preoccupied trying to link particular avatars to people. P14 had a differing opinion, saying: “I feel like there’s more of a novel underlying idea behind keeping it anonymous, because in a way if it’s personalized, it’s more of like a manipulation of what is already happening in the classrooms now, it’s like me coming from a place of like a student. But if you keep it anonymous then it’s easier to find a genuine reaction and it’s easier to be confronted with a genuine reaction of your audience.” Indeed, participants who felt that the system encouraged greater emotional expression attributed it to the anonymity the system afforded, and consequently, a release of social pressure. P13 offered a design solution for those who were preoccupied

with identifying themselves without sacrificing anonymity: privately highlight your personal avatar, while keeping the others anonymous. This will be considered for future iterations, as implementing this could conceivably aid in reducing the novelty effect by reducing the amount of time it takes to find yourself in the crowd, as well as assist users who have a propensity to monitor themselves in video calls.

Category 2C: Ethics

Participants were asked, very generally, of ethical concerns they had of the system's use. Many agreed that the device should follow standard "zoom consent rules" by explicitly asking permission. P11 said: "there's more importance of our consent because it's collecting biometrics data " (i.e. landmark extraction). P11 stressed that there was a fundamental distinction between a raw video feed that you would encounter on Zoom and CoHere because CoHere conducts analysis on the video feed. Some participants had used similar technology before (e.g. a snapchat filter) and had relaxed feelings about using CoHere so long as it was optional and for casual use. P11 said "I would be comfortable using it for entertainment purposes, so like [P12] mentioned like Snapchat, I consent to that and that's totally fine. But I would feel weird about it if it was like required by my school or required by my work". The importance of choice was further illustrated by P5: "I think it's all about choice. If there's an avatar meeting, it's not obligatory to have it, you should be able to choose that". P16 and P18 had concern about facial data being stored, because then it could potentially

be used to conduct further analysis on for which they did not consent to. Currently CoHere does not write facial data to persistent storage, the server keeps landmark data in RAM only for 1/14th of a second after which it is overwritten.

Other ethical concerns were centered around hypothetical oppressive design choices such as if CoHere was modified to detect if someone was looking away from the screen for compliance.

2.4.4 Qualitative Discussion

An unintended yet positive reported experience of using this system was the feeling of "in-betweenness", and its social effects. Indeed, in everyday life, there are many circumstances where one would want something in-between having their camera on and off. This mixes the emotional engagement of a video call with the less demanding nature of a phone call We discovered that the system was not as suitable for presenters as it was to viewers. Presenting a slideshow is a visually demanding task, and CoHere gives mainly visual feedback. This may lead to it being more of a distraction to presenters, or at worst, sensory saturation. During a live, on-stage presentation, the presenter often cannot see the audience because of the lights. They mainly receive affective feedback auditorily. It would be ideal to offload affective feedback to other modalities to presenters. An interesting side effect of the system is that it offered users a greater sense of privacy, enabling them to express themselves more. Of note that this feeling of privacy was predicated on the system being used in a free and

consensual manner. As P11 mentioned, if they were forced to use the system (e.g. for work), their impressions would be different. The qualitative evidence of emotional contagion is an exciting venue for future research. There is a large question as to whether the degree of emotional contagion experienced in with CoHere is comparable to emotional contagion experienced in real life, or in a traditional 1:N video call. In traditional 1:N calls, especially when a slideshow presentation is given, that facial information from other users is either unavailable, or extremely low resolution (not all participants are on screen, and the ones that are have low quality thumbnail feeds). We note the language users had of experiencing others emotions, e.g. "I noticed other people reacting a certain way" (P3), "someone laughed and it made me laugh" (P15). This degree of emotional awareness is not typical of 1:N videocalls—very rarely would someone notice others "reacting a certain way" or laughing. This is positive evidence that the system facilitates greater emotional awareness, but a more in depth study would need to evaluate how this specifically differs from a traditional call and in person. We did not evaluate this particular question in this study because we did not expect emotional contagion to occur when initially designing the study, and intended it to be more exploratory of the phenomena experienced with CoHere.

2.5 Quantitative Experiment

A follow up study was conducted to quantitatively assess the affective perceptions of users familiar with CoHere. All who participated in the prior qualitative study were invited

back to participate. Of the 20 previous participants, 12 were able to successfully complete the quantitative experiment. Two participants opted not to return for the second study, two participants misunderstood the study instructions,[3] and four participants experienced a client browser error that prevented them from sending their data. Approximately a month elapsed between when participants participated in the qualitative study, and the quantitative experiment.

At a high level, we wish to see if the users perceive emotions in natural CoHere animations any differently from noise. However, generating animations based off noise would result in biologically implausible motion, which could be easily identified by any naive user. Therefore, instead of comparing natural animations to raw noise (e.g., white, or Perlin noise), we compare *congruent* and *incongruent* animations—congruent animations co-occured with the slideshow, and incongruent animations were obtained from a different slideshow but are overlaid on the current slideshow.

2.5.1 Methods

Participants from the prior study (N=12) were asked to watch video excerpts from the qualitative study and rate the affect of the audience over time using the dimensions of *arousal* and *valence*. The overlaid CoHere animations that accompanied the presentations

[3]The experiment required participants to watch a video and continuously rate the perceived affect of the audience over time. However, these two participants would watch an entire video and then supply a single affect rating after the fact, instead of giving affect ratings over multiple points in time. Since we were interested in the time-varying affective response of users, and these responses were effectively simple scalar values, their data were discarded.

were manipulated across three conditions, further elaborated upon below.

The experiment interface was a Node.js web app hosted on a private server housed in the *[institution omitted for anonymity]* Participants completed the experiment asynchronously and were compensated 10 dollars.

Before the main experiment, participants were informed about the meaning of the affective dimensions of arousal and valence through textual descriptions with supporting illustrations from the Self-Assessment Manikin—a validated sequence of caricatures of varying arousal/valence states [89]. They were then shown images of the experiment interface, with instructions of how to use it to log arousal and valence over time. The experiment interface consisted of a video player on the lefthand side of the screen, and an affect grid on the righthand side. Users could drag and drop a crosshair on the affect grid to record their affective ratings over time as the video played. We recorded a time series of samples that included: arousal value $\in [-1, 1]$, valence value $\in [-1, 1]$, and the video timestamp at which an affect rating was logged.

Participants were instructed to log how they thought the audience feels, and not how they themselves feel when watching the slideshow. If there were no audience visualizations present, then the participants were told to rate how they imagined the audience would feel. Finally, a video tutorial was supplied, further showing how to use the interface to log arousal/valence data over time while watching a slideshow.

We exposed participants to three animation conditions:

- CONGRUENT: overlaid CoHere animations were derived from real audience data that co-occured with the presentation as it was recorded.
- INCONGRUENT: overlaid CoHere animations were derived from another presentation that differed in target valence and arousal.
- NO VISUALIZATION: only the raw video feed of the presentation without CoHere animations was shown.

A single experiment session used two raw screen recordings of different slideshow presentations, "A" or "B", which were combined with either CONGRUENT, INCOGRUENT, or NO ANIMATION conditions. Each presentation $\times$ animation condition was repeated twice. In total there were 2 presentations $\times$ 3 animation conditions $\times$ 2 repetitions $=$ 12 total trials per participant. Video excerpts were within 20 and 45 seconds in length. Additionally, to ensure that participants did not see slideshows that they encountered in the qualitative study, there were two potential sets of presentations that were shown. Participants with odd IDs were given slideshow group 1, and participants with even IDs were given slideshow group 2. This entailed that the entire experiment ultimately utilized four slideshows, though a single participant only saw two. At the conclusion of the study, participants were asked with a text prompt if there were aspects about the affect grid that made it difficult for them to communicate their perceived emotion.

Data pre-processing

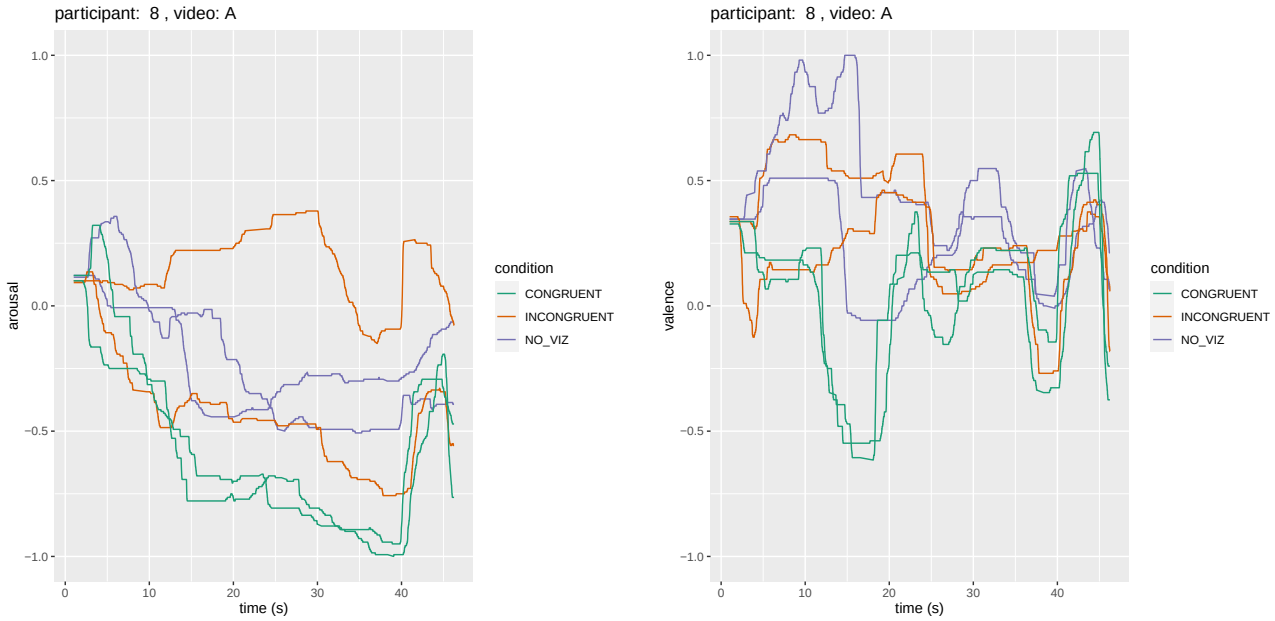
Prior to analysis, each participant's data was processed as follows: first min/max normalization was employed to normalize affective responses between $[-1, 1]$. The min/max values were derived from a participant's entire session, not a single trial. Afterwards, participant responses were linearly interpolated to continuously connect their affect ratings over time. Finally, one second from the head and tail of their responses were cropped to discard noise from moving the mouse to and from the affect grid as the videos start and end. Exemplary samples of participant responses for a single video are given in Figure 2.5.

2.5.2 Analysis

Global affect ratings

Frequency distributions of participant affect ratings for each of the four potential slideshow presentations are shown in Figure 2.6.

Four total two-way repeated measures aligned rank transform ANOVAs were run on each presentation, investigating the effect of ANIMATION CONDITION and AFFECT TYPE on affect ratings. The ANOVA revealed statistically significant $p < 0.00001$ main effects for both ANIMATION CONDITION and AFFECT TYPE on affect ratings across all possible slideshows. Critical values for the ANIMATION CONDITION factor for each of the four slideshows are reported in Appendix A, Table 2.2.



P8 Arousal P8 Valence

Figure 2.5: Affect ratings from P8 for a single slideshow.

Followup paired Wilcoxon signed rank tests were conducted to test if the medians between congruent and incongruent conditions were equal ($\mathbf{H_0} : \eta_c = \eta_i$; $\mathbf{H_A} : \eta_c \neq \eta_i$, where η is the median). After applying Bonferroni corrections, the difference between congruent and incongruent medians was found to be statistically different from 0 for all conditions (all p-values $< 2.2 \cdot 10^{-16}$). This leads us to reject the null hypothesis that median affect ratings are equivalent across all conditions. The critical values for these tests are reported in Appendix A, Table 2.3.

As there is significant variation in affect ratings between congruent and incongruent conditions, we can infer that participants perceive different meanings from the visualizations.

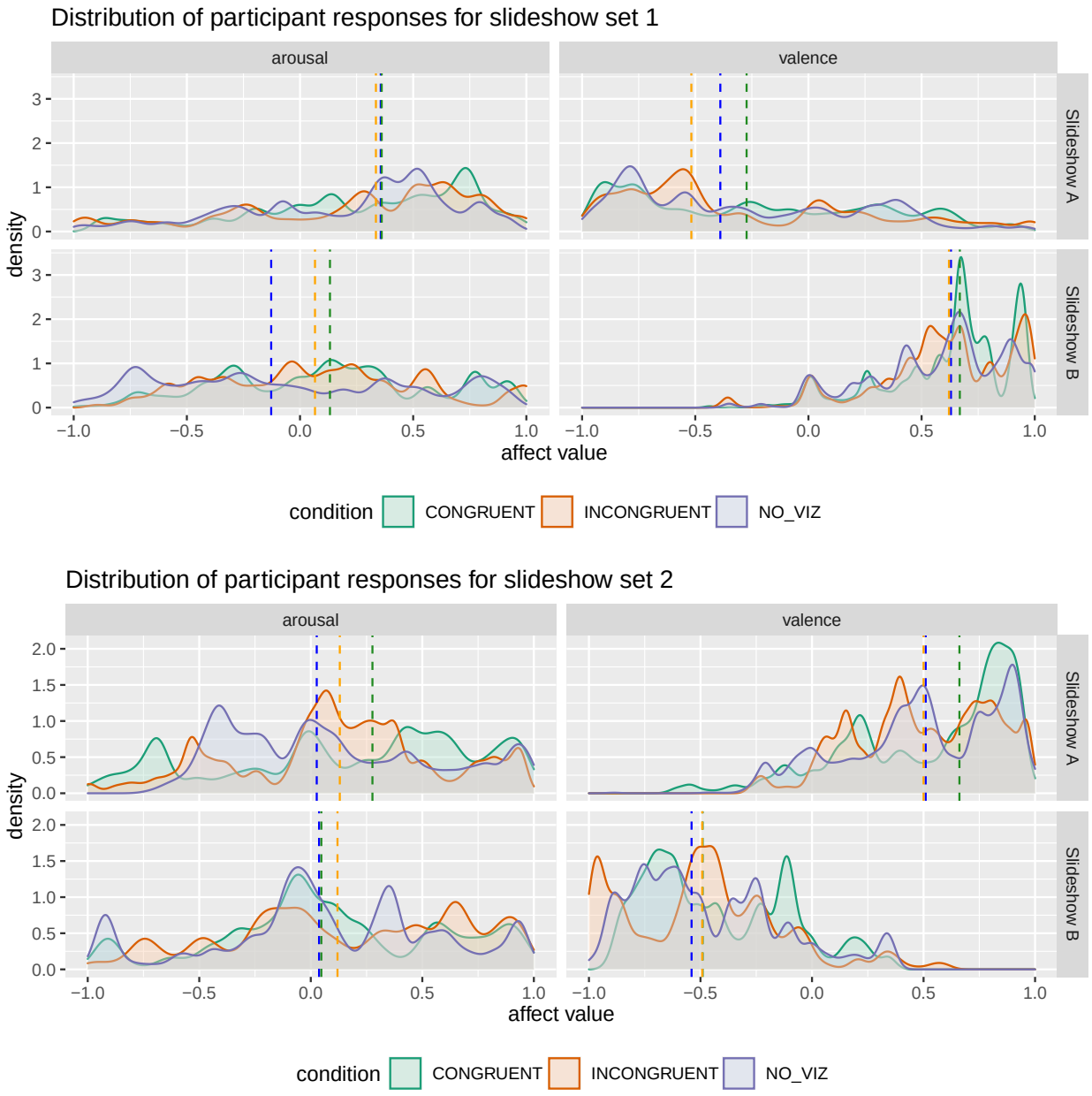


Figure 2.6: Frequency distribution of participant affect ratings. Each row is a unique presentation. The dotted lines demarcate the median.

Post-study question

At the end of the study, participants were prompted to answer "Were there emotional aspects of the videos that you were unable to sufficiently capture with the supplied pleasantness/arousal grid (i.e., did you see emotions that couldn't be expressed through the 2D grid)? If so, can you describe what they were and the situation that occurred? Leave the text field blank if the 2D grid was able to sufficiently capture all the emotions for you." Of the 12 participants who successfully completed the study, four provided feedback to the post-study question. P1 and P9 both had trouble identifying where aspects of "interested"/"disinterested" lay on the affect grid. P9 was uncertain if "disinterested" meant "neutral", and therefore should be at (0,0).

P7 said: "I got quite bored by the end and felt like I would have indicated lower arousal but was at the bottom already". P7 wasn't the only one who became bored—there was a clearly observable fatigue effect from participants. As the experiment progressed, mean trial arousal ratings decreased. A negatively sloped linear model describes the decrease with $\beta = 0.024, t = 2.639, p = 0.00926$. This was not significantly observed for valence ratings, however.

Interpretation

There was significant variation in medians between congruent and incongruent conditions, indicating that animations do affect how audiences are perceived. The findings taken together

indicate that CoHere visualizations are meaningful to participants in some way distinct from noise. The question as to what *specifically* certain visualizations meant to participants is rather complex to answer. There was a large amount of variability in terms of how participants emotionally interpreted slideshows ($\bar{\sigma}_{valence} = 0.586$, $\bar{\sigma}_{arousal} = 0.495$). It could thus be methodologically problematic to treat all participants as a monolithic entity when it comes to assessing the deterministic ways that the visuals affected slide show meaning. A rich answer to the question of "what do the visualizations mean?" is best characterized by an individualistic approach. Nonetheless, the experiment demonstrated that the animations had unique effects on emotion ratings, and that congruent vs. incongruent animations factored into how presentations were perceived.

2.6 Conclusion and Discussion

CoHere shows promise in its ability to facilitate the communication of affect in a continuous, implicit manner during 1:N video calls for. Through in-situ use, we have obtained qualitative evidence that CoHere gave participants awareness of each other's emotions—this was best shown in observed reports of emotional contagion and alignment in expression. This is suggestive of emotions being coordinated between interlocutors, which is a fundamental aspect of establishing affective ground. In this regard, this finding is encouraging that CoHere facilitates the affective grounding of user conversation. As users had shown awareness of others' emotions, CoHere was demonstrated to be able to

non-representationally communicate affective feedback from the audience. Whether the visualizations themselves genuinely represent the real emotions experienced by the audience, and that audience members accurately perceive those emotions, is an unanswered question and subject for future work. This is a shortcoming from the study, as it did not establish "ground truth" of audience emotions while experiencing a live presentation via a questionnaire or interview. As such, we cannot claim anything pertaining to whether a user's emotional perceptions were accurate to the emotions that truly occurred during the presentation. Another shortcoming of the quantitative study is that due to a large participant drop off rate, the sample size is low while we are covering a large affective range. The results should be thus interpreted as encouraging but in need of follow-up studies. CoHere also enabled users to express themselves to a greater degree that they would not otherwise have in a traditional video call. There is strong quantitative evidence that the unique motions of visualizations can influence perceived affect with experienced users. CoHere accomplished this without explicit emotion categorization or signaling but instead providing users with an ambiguous time varying signal motivated by natural facial expression data.

Of note of the emotions experienced in our qualitative analysis is that there was an under-representation of participant discussion around negative valence emotions despite half the presentations having negative valence content. Indeed, inspecting the quantitative data, it can be confirmed that there was a slight positive valence, positive arousal bias in the user

data, with $\bar{x}_{valence} = +0.2$ and $\bar{x}_{arousal} = +0.13$ (Fig. 2.7).

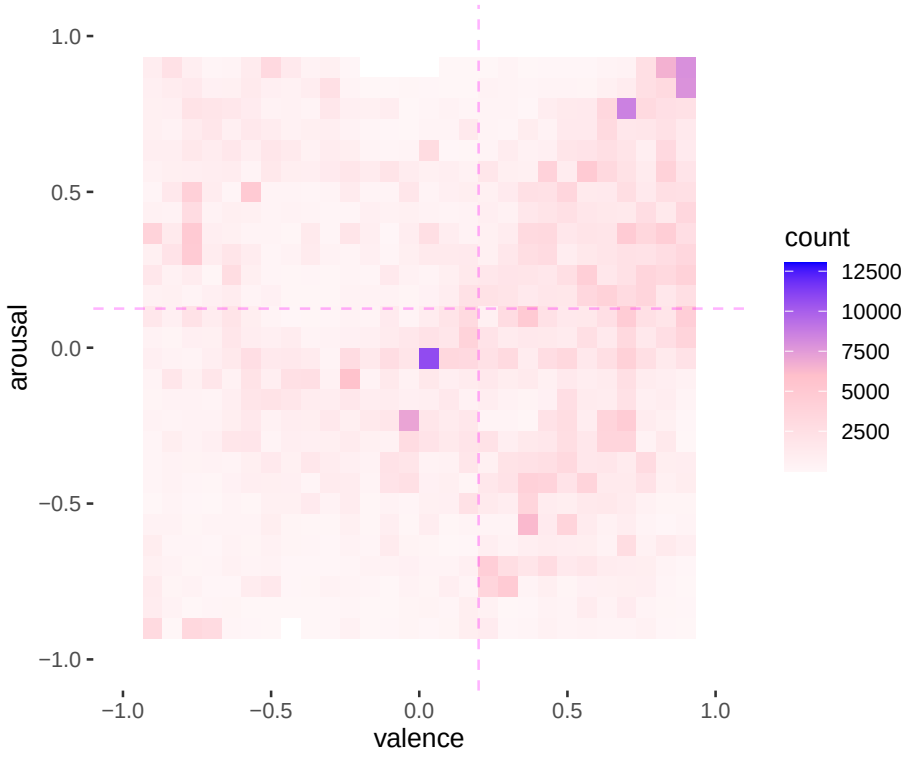


Figure 2.7: A heatmap of all user affect ratings in the quantitative experiment.The vertical bar indicates the global mean valence rating, and the horizontal bar indicates the global mean arousal rating.

This could be because allowing presentations to be improvised meant that some participants were more playful, which could lend itself to more positive affect presentation styles. Many images from the IAPS that were high arousal and low valence were excluded from the slide show presentations due to explicit content that was unsuitable for day-to-day videoconferencing (e.g., body mutilation). This selection criterion, when applied to the IAPS dataset, could have affected the final quality of negative valence

stimuli. Finally, an under-representation of negative valence emotions could also be a shortcoming in the expressive capabilities of the system, where the most detectable emotions skewed positive valence.

It was discovered that CoHere was best suited for audience members. For broadcasters, the animations were largely ignored due to the high visual demand of the slide shows. Future work for such a system could render affective feedback in alternate modalities, such as sound or touch, as to assist in the unique attentional demands between broadcasters and viewers as discovered in this study. The system is also currently only suitable for small crowds (approximately < 10 people); a version of CoHere for massive audiences would require a new rendering and facial analysis techniques, and would be an ideal iteration to accommodate use cases such as remote concerts or large-scale gaming streams. CoHere also need not be limited to analyzing the face—the body as a whole offers a rich basis that future versions of CoHere could use to render affective feedback. The speech signal is also a rich source of affective information and could be further utilized as input.

There is much discussion to be had about the fidelity of the particle avatars used in CoHere. Prior work investigating realism of avatar forms found that visually low-realism avatars similar to the avatar particles employed in this study elicited inferior reports of copresence and emotion identification compared to a video stream [90]. However, this experiment was mainly focused on 1:1 calls, not 1:N video calls, where there are unique constraints on screen space and attention. It is an open question as to whether higher

fidelity avatars will offer greater affective awareness during calls, and what their corresponding attentional burdens may be, especially when the number of active videocallers grows to encompass tens to hundreds of people.

2.7 Appendix: Critical Value Tables

Slideshow	Group	$F()$
A	1	$F(2, 476197) = 564.98$
B	1	$F(2, 646549) = 6687.5$
A	2	$F(2, 651589) = 2784.48$
B	2	$F(2, 354229) = 226.94$

Table 2.2: Results of four ANOVAs ran on unique slideshows. Shown here are results for the ANIMATION CONDITION factor only (levels: CONGRUENT, INCONGRUENT, NO VIZ). All p values are < 0.00001. Each row is an ANOVA run on a unique slideshow. Participants with odd numbered PIDs that were shown slideshows from Group 1, and those with even PIDs were shown slideshows from Group 2. There were four total slideshows used in the study: 1A, 1B, 2A, 2B. Animation condition significantly accounted to variance in affect ratings for each possible slideshow.

Slideshow	Group	affect	V
A	1	arousal	1629250334
A	1	valence	1897284312
B	1	arousal	3363169015
B	1	valence	3224677524
A	2	arousal	3216856834
A	2	valence	3216856834
B	2	arousal	683740432
B	2	valence	1101405917

Table 2.3: Results of multiple paired Wilcoxon tests assessing if the medians between congruent and incongruent conditions are equal. Critical V values for the Wilcoxon tests are reported. All corresponding p-values are $< 2.2 \cdot 10^{-16}$.

Chapter 3

I See What You're Hearing: Enhancing Contextual Awareness in Teleconferencing Through Audio Environment Visualization

Preface

This chapter presents a manuscript that is currently in submission to a peer-reviewed conference. The current work presents a system designed to enhance the contextual awareness of participants in 1:1 calls. The system augments user video feeds with a particle

visualization based off acoustic and semantic features of the environmental soundscape in which remote interlocutors are situated. Through an experiment, we discovered that the system was able to convey similar audio contexts across modalities in terms of how they were affectively perceived.

Author's contribution

This paper was a collaboration between David Marino, Max Henry, Pascal Fortin, Rachit Bhayana, and Jeremy Cooperstock. I programmed the main system, conducted the experiment, performed analysis, and wrote the majority of the paper. Pascal Fortin contributed intellectually at many steps in the process, and contributed to paper writing. Max Henry assisted with software development and paper writing. Rachit Bhayana created the data visualizations for this paper and contributed to paper writing. Jeremy Cooperstock supervised the project and made intellectual contributions.

Abstract

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. 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 elicited emotional perceptions in users that were similar to the response elicited by environmental sound it replaced. We also found that participants were implicitly aware of aspects of the visualized sound, such as whether it was artificial or likely to occur outside. The contributed system provides a unique approach to facilitate ambient awareness on an implicit emotional level in situations where multimodal environmental context is suppressed.

3.1 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 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 the system employs 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 [34, 91, 92]. 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. There are many times when conversing over a video call that the acoustic environment is suppressed—either by active noise cancellation or muted microphones, resulting in a loss of the sense of the environment that the user is situated in. The system translates the aural environment to the visual modality to convey this discarded information. There are a few motivating factors for this approach: First, there are many times when it is desirable to not have background audio present in a video call, but there still may be use in understanding what's occurring in the user's acoustic environment. An example of this is when muting microphones to let others speak without interference. Second, it is also sometimes desirable to discard background audio via noise

cancelling so that interlocutors speech can be better understood. But this comes at the cost of losing environmental information which contributes to a loss of shared understanding of eachother's situational circumstances. 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, instead 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 response in a way that is similar to the source audio. 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.

3.2 Background

3.2.1 Audio visualization

Visualizing audio events has a theoretical precedent in the bouba-kiki effect—where a non-arbitrary relationship is demonstrated between speech sounds and shapes [93] [94]. 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 [95]. The basis of this non-arbitrary relationship has been hypothesized to be based off tacit knowledge of the articulatory processes, such as vowel height and lip rounding, which link sound to visuals [96]. However, relationships between sound and visuals is almost certainly not limited to the verbal domain, and they may also have arbitrary relations. For example, those with synesthesia have been known to have arbitrary audio-visual associations [97].

Through conceptual metaphor, people are able to conceptualize abstract phenomena cross-modally [98]. 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 [99]. The notion of conceptual metaphor claims that metaphor forms the foundation of much of human thought [100]. Using this theoretical

framework, we can understand audio visualization as a multimodal conceptual metaphor.

3.2.2 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 [101]. Physical approaches such as MeBot represented remote users in a robot that conveyed nonverbal cues [102]. 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 [103]. 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 [104, 105].

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 [106]. Audio visualizations have also been used in video conferencing to calibrate vocalization levels between interlocutors [107]. 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 [108] designed to provide "insight into the participants' culture and status" [109]. Visualizations of the aural environment have also been used in clinical populations. Realtime audio visualizations have been used to encourage spontaneous speech-like vocalizations with users with autism spectrum disorder [110]. Non-speech sound visualizations have also been used to convey aspects of the physical environment in which Deaf users are situated [111].

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 are occluded and limited environmental audio is transmitted, such as when using noise canceling, speaker prioritization, or muted microphones.

3.3 System Design

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*,

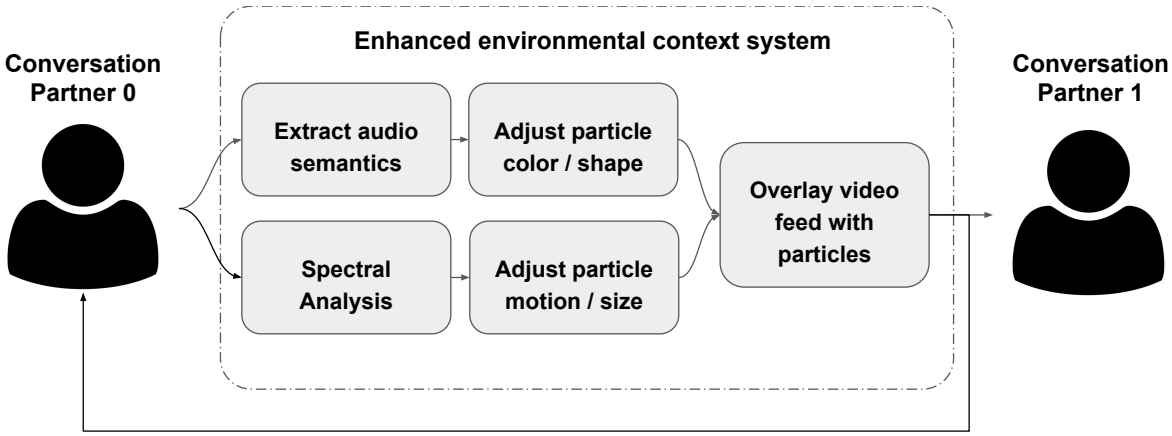


Figure 3.1: High level system architecture.

natural, *foreground*, *background*, *interior*, and *exterior*. This process is described in detail in Section 3.3.1. The particle animation is also affected by low-level acoustic features of the environmental sound, which is further described in Section 3.3.2. 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.

3.3.1 Semantic features

We classify the soundscape in realtime based off high level semantic features. A soundscape may be considered as having three key features: *geophony*, *biophony*, and *anthrophony* [112]. Geophony are geophysical sounds, biophony are organic sounds, and

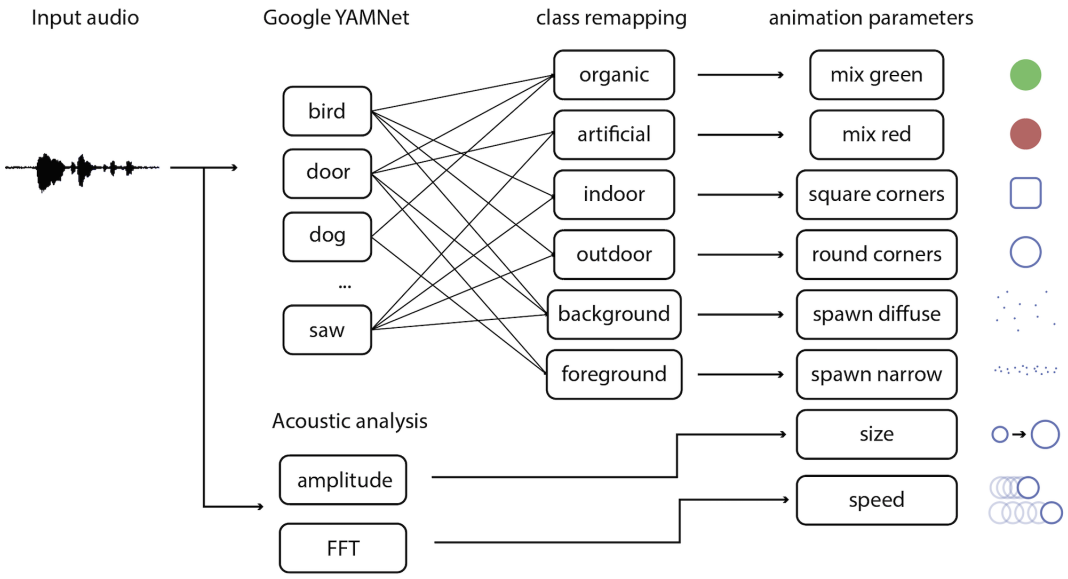


Figure 3.2: Input audio is classified using Google YAMNet. We then remap Google YAMNet's classes to our own class ontology. These classes are then mapped to animation parameters. Acoustic analysis is conducted in parallel, which in turn maps to animation parameters.

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 [113]. YAMNet can predict 521 classes of sounds, which are hierarchically grouped into seven general categories: human sounds, animal sounds, natural sounds, music, sounds of things, source-ambiguous sounds, and noise from recording/playback devices. Using Google's AudioSet class ontology as a starting point, three authors thematically organized the classes into six 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. It was decided to keep semantic features as coarse grained as possible, as we did not 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 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. Additionally, 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

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

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 3.4.2. In our evaluation, we ultimately discovered that a simpler class ontology of "living creatures", "outdoor" and "indoor" suited users needs, and would be an ideal basis for a future iteration of the prototype.

These features are not mutually exclusive. A single sound can appear in many different circumstances and contexts—a bird may appear both indoors as a pet and outdoors as a wild animal, and the trickling of water may be from a natural source such as a creek, or artificial

source such as a faucet. The applicability of these sounds to certain semantic features is also context dependent. 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 [114]—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.

3.3.2 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 44.1 kHz. 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 -10dB—is adjustable to accommodate different users' 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" [98].

3.4 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 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. Second, 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.

3.4.1 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 [115]. Footage of the speakers was sourced from the Cardiff Conversation Database (CCDb) [116]. All participants were shown a series of excerpts from a single conversation between two interlocutors who were displayed in tile mode (Figure 3.3). Four

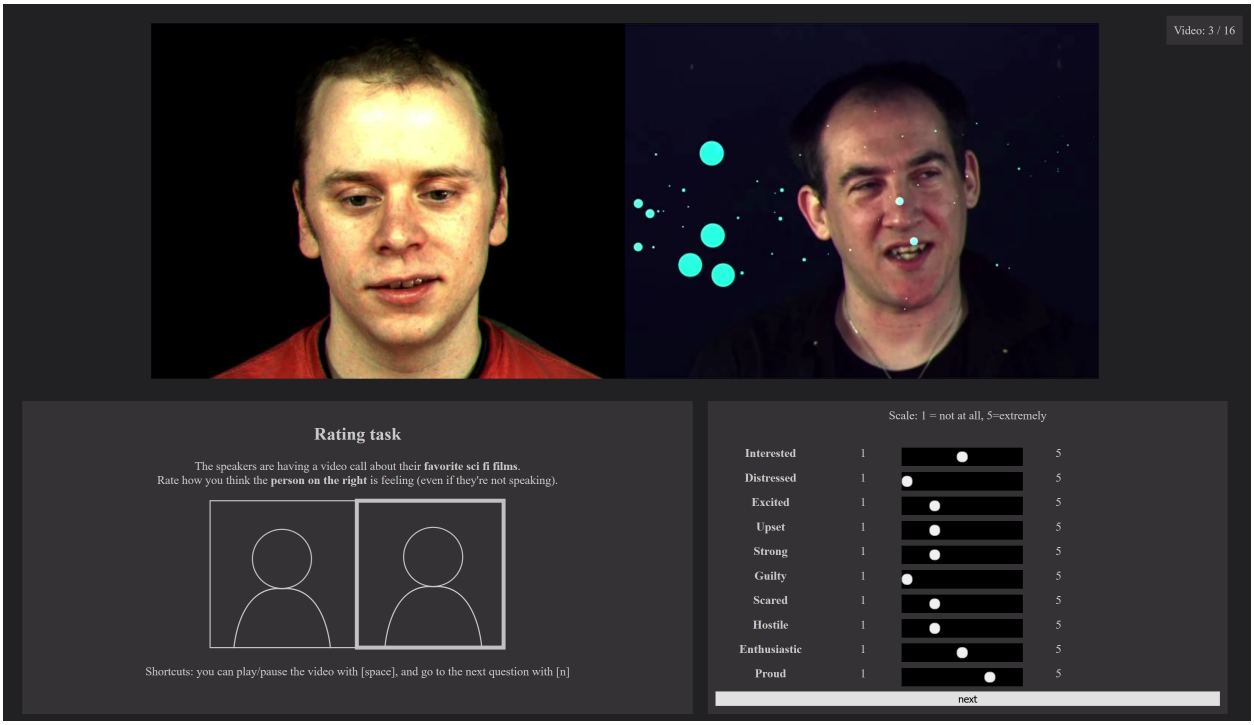


Figure 3.3: UI for the rating task

short, 10–20 s 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 ranging from 1 (emotion was not present) to 5 (emotion was extremely present).

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 [117] was conducted to compare the effect of ENVIRONMENT, BGA and VISUALIZATION on each of the EMOTION RATINGS. An ART ANOVA was selected because emotion ratings were assumed to be ordinal as distance between values in the scale were 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 3.1. Critical F values are reported in Table 3.2. A boxplot of IVs is shown in Figure 3.4.

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.

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 3.1: 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 3.2: F values for DVs (columns) by factors (rows). Highlighted values surpass critical F values after correction (pre-corrected values by row: 2.42, 3.89, 3.89, 2.65, 2.65). Columns with no significant factors omitted are (enthusiastic, excited)

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. 3.5).

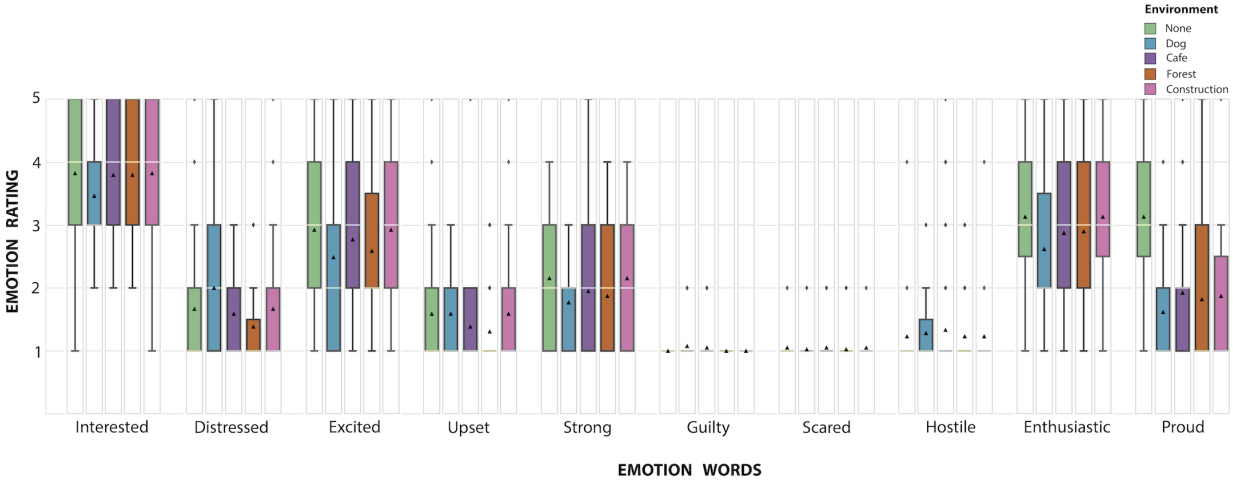
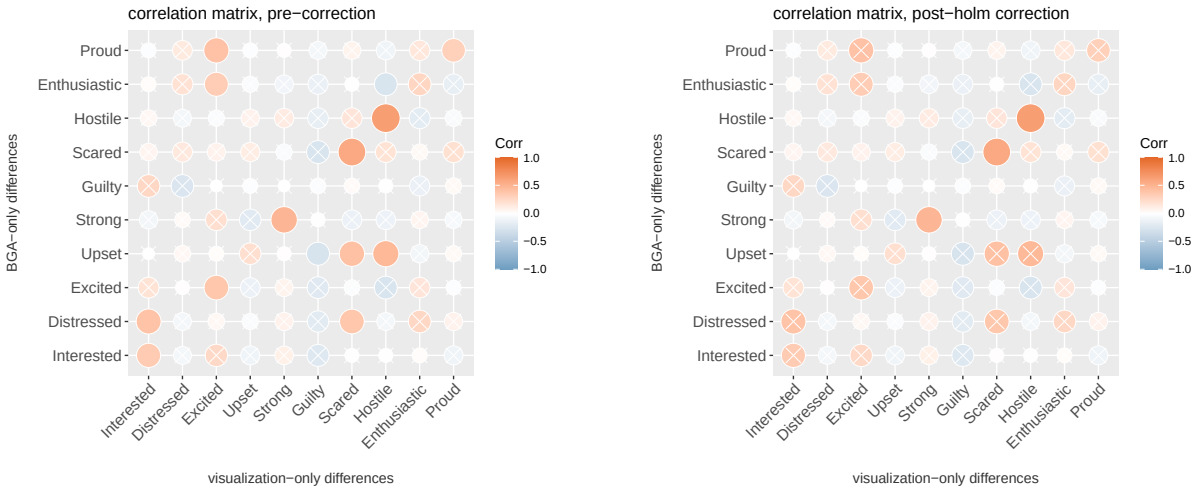


Figure 3.4: Box plots depicting emotion rating for each of the emotion word for each of the five environments, combining BGA:on and BGA:off conditions.

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. 3.5). 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



(a) pre-correction

(b) post-correction

Figure 3.5: Emotion rating between visualizations and BGA. The nodes with an X are not significant. The two figures show the difference before and after Holm corrections were applied.

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. 3.5, (a)). Using Holm corrected p-values, there were just three significant cross modal correlations, all on the diagonal, for the emotions "strong", "scared", and "hostile" (Fig 3.5, (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 in the question of whether most significant correlations are in the diagonal. This is most similar in form to an identity matrix. As such, we also used a Bartlett test that a correlation matrix is an identity matrix. Formally, $\mathrm{H3}_0 : M = I_{10}$, $\mathrm{H3}_0 : M \neq I_{10}$. We obtained $\chi^2(45, 52) = 16.87$, p = 0.99, leading us to retain the null hypothesis that the correlation matrix is an identity matrix.

Our interpretation of the results were as follows: first, 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. That there is no evidence the correlation matrix significantly differs from an identity matrix is an encouraging result (though not a definitive one). This is suggestive that a greater

proportion of emotions may have been cross-modally correlated given less harsh post-hoc corrections; though this is not strictly entailed by this finding. Words that were not correlated between modalities could have been because the visualization did not adequately convey environmental meaning cross-modally 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 the ten emotion words used in this study are not necessarily orthogonal in their meaning, and therefore some inter-correlation between emotion words across modalities is to be expected. For example, scoring high in "excited" may covary with scoring high in "interested"—therefore it stands to reason that visualizations that evoke "excited" are correlated with the auditory contexts that evoke "interested". The same can be said for the pairs of ("hostile", "distressed,") and ("scared", "distressed").

We conclude that the visualizations are capable of producing affective contexts in a way somewhat similar to BGA, but further work is required in refining the device.

3.4.2 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.

Visualization meaning

We utilized the qualitative research method of content analysis [118] 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 colour 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**: *No qualitative effect*] For 31% of participants, the particle animation did not take on any discernible meaning. These participants largely gave no further explanation as to why this was. One mentioned that it was because it was distracting and seemed "additional and not impacting a lot". Other participants mentioned that the particles were distracting or getting 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 not gain any useful information from the visualizations and tried to fixate on a particular channel of communication, such as the face or sound.

[**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 according to frequency or semantic aspects of sound.

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 3.6. 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 few 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, 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.

3.5 Discussion

In this study, we demonstrated the feasibility of employing visualizations to enhance contextual affective awareness in situations where the auditory modality is unavailable or degraded.

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

Living Creatures	Outdoor	Indoor
Human activity	Urban sounds	Household features
⎣ Background speech	⎣ Construction	⎣ Tap water
⎣ Laughter	⎣ Busy street	⎣ Fan
⎣ Others video-conferencing	Transportation	⎣ Air Conditioning
⎣ Roommates	⎣ Car	⎣ Kitchen sounds
⎣ Footsteps	⎣ Traffic	⎣ Doors
⎣ Neighbors		⎣ Doorbell
⎣ Movement		⎣ Door slamming
⎣ Typing		⎣ Door opening
⎣ Babies		⎣ Door closing
⎣ Eating food		Appliances
Non-human		⎣ TV
⎣ Dogs barking		⎣ Vacuum
⎣ Pets		⎣ Telephone ringing
		⎣ Fridge
		Misc
		⎣ Laptop fan
		⎣ Echo
		⎣ Mic feedback
		⎣ Music

Figure 3.6: Most relevant sound events

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.

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.

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 from 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 scenario 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 user 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. 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.

Chapter 4

Conclusion

The technologies described in this thesis are all theoretically aligned with a constructed theory of emotion. As such, they all utilize an *affective interaction* paradigm, as opposed to an *affective computing* one. A challenge of affective interaction is how to represent concepts, such as emotion, without explicitly representing them. In this regard, affective computing has it easy—if you wish to communicate "happy", one could just classify a signal and display some symbol to indicate happiness. For affective computing, the hard part is the classification problem, the easy part is the representation problem. The two presented technologies have demonstrated feasibility in communicating emotional signals during live videoconferencing tasks without explicit representation or classification. CoHere offers a good example of how this could work. It conveyed the emotions of the audience without explicitly sensing emotions. It achieved this by aggregating and visualizing the facial configurations of all

conversation partners. This information was sufficient for participants to "align" themselves to the emotions of the crowd, and even reported experiencing others emotions as influencing their own. The second device, using a combination of semantic and acoustic soundscape visualization, showed that it was also possible to convey affective contexts cross modally for 1:1 remote communication. Though there was some discrete categorization present in terms of its acoustic feature detection (i.e., organic vs. artificial), no *emotion detection* was employed. The algorithm simply classified manifest aspects of the sound instead of highly interpretive aspects. Altogether, this provided a sufficient amount of information to create a cross modal environment with correlated emotion perceptions. Both of these technologies demonstrate different methods of designing for a constructed theory of emotion, where emotion categories are not assumed to be intrinsic to a single signal.

One benefit that a symbolic, affective computing approach has is that it requires little learning in terms of how meaning is inferred (if using familiar symbols). For example, if someone smiles and yellow text appeared that said "happy", then most users would agree that what the system is trying to communicate is that the user is happy. Just as we must learn to understand the affective nuance in a smile, a non-representational, affective interaction approach also requires learning. If a user was to receive haptic feedback based off the muscle activation of another user's face, it would require some learning to understand that the user is smiling, frowning, etc. Although the prior studies did not explicitly discuss it, all of them had clear learning phases. CoHere's initial qualitative in-situ study familiarized users

with the system. As such, the follow up quantitative study was effectively conducted with trained users. Likewise, it could be argued that there was a training effect between phases of the audio context study. When users had to explicitly identify what was happening in the background based off visualizations without the aid of background audio, they had already obtained a familiarity with how the system worked in the previous emotion rating phase. The true expressive capabilities of implicit affect communication would be revealed through longitudinal use.

In terms of an explicit symbolic approach to emotion representation, it could be argued that that there would also be less semantic variation in terms of how users perceive symbols. For example, if 20 users see red text that says "User A is angry", then there will not be much disagreement between users as to what that symbol means. However, if 20 users, even if trained, were to view an ambiguous signal that is correlated with anger, then there would be a lot more disagreement. Although this sounds somewhat as a shortcoming of this approach, it is also close to the reality of how we perceive emotion in the wild. The symbolic approach may cause overconfidence in its unambiguous representation. If end users come to the wrong conclusions about one another based off a false positive, it could have serious detrimental effects to their relationship. A non-representational approach preserves the inherent uncertainty we have when identifying emotion in the wild, and affords the ability to negotiate the complex affective relationship between one another. The technologies presented in this thesis preserve this aspect of affective communication, and offer a new channel for

affectively nuanced remote communications. Future versions of these devices can utilize additional modalities, such as speech, as a basis to generate affective signals. There is also much further work to do in terms of iterating on how to visualize such sensory information in a meaningful way.

Bibliography

[1] R. Johansen, J. Vallee, and K. Collins, "Learning the limits of teleconferencing: Design of a teleconference tutorial," in *Evaluating new telecommunications services*, pp. 385–398, Springer, 1978.

[2] O. García, J. Favela, and R. Machorro, "Emotional awareness in collaborative systems," in *6th International Symposium on String Processing and Information Retrieval. 5th International Workshop on Groupware (Cat. No. PR00268)*, pp. 296–303, IEEE, 1999.

[3] H. H. Clark and S. E. Brennan, "Grounding in communication.," 1991.

[4] M. F. Jung, "Affective grounding in human-robot interaction," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pp. 263–273, IEEE, 2017.

[5] K. Kearns, "Semantics," 2000.

[6] H. P. Grice, "Logic and conversation," in *Speech acts*, pp. 41–58, Brill, 1975.

[7] E. L. Price, E. J. Pérez-Stable, D. Nickleach, M. López, and L. S. Karliner, "Interpreter perspectives of in-person, telephonic, and videoconferencing medical interpretation in clinical encounters," *Patient education and counseling*, vol. 87, no. 2, pp. 226–232, 2012.

[8] R. W. Picard, *Affective computing*. MIT press, 2000.

[9] W. James, "Discussion: The physical basis of emotion.," *Psychological review*, vol. 1, no. 5, p. 516, 1894.

[10] P. C. Ellsworth, "William james and emotion: is a century of fame worth a century of misunderstanding?," *Psychological review*, vol. 101, no. 2, p. 222, 1994.

[11] W. B. Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American journal of psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[12] W. B. Cannon, "Again the james-lange and the thalamic theories of emotion.," *Psychological Review*, vol. 38, no. 4, p. 281, 1931.

[13] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychological review*, vol. 69, no. 5, p. 379, 1962.

[14] I. J. Roseman and C. A. Smith, "Appraisal theory," *Appraisal processes in emotion: Theory, methods, research*, pp. 3–19, 2001.

[15] R. S. Lazarus, "Cognition and motivation in emotion.," *American psychologist*, vol. 46, no. 4, p. 352, 1991.

[16] P. D. MacLean, "Some psychiatric implications of physiological studies on frontotemporal portion of limbic system (visceral brain)," *Electroencephalography and clinical neurophysiology*, vol. 4, no. 4, pp. 407–418, 1952.

[17] M. Catani, F. Dell'Acqua, and M. T. De Schotten, "A revised limbic system model for memory, emotion and behaviour," *Neuroscience & Biobehavioral Reviews*, vol. 37, no. 8, pp. 1724–1737, 2013.

[18] V. Rajmohan and E. Mohandas, "The limbic system," *Indian journal of psychiatry*, vol. 49, no. 2, p. 132, 2007.

[19] G. Gainotti, "Neuropsychological theories of emotion," *The neuropsychology of emotion*, pp. 214–236, 2000.

[20] R. G. Heath, R. R. Monroe, and W. A. Mickle, "Stimulation of the amygdaloid nucleus in a schizophrenic patient," *American Journal of Psychiatry*, vol. 111, no. 11, pp. 862–863, 1955.

[21] E. Fonberg and J. M. Delgado, "Avoidance and alimentary reactions during amygdala stimulation," *Journal of Neurophysiology*, vol. 24, no. 6, pp. 651–664, 1961.

[22] P. Loiseau, F. Cohadon, and S. Cohadon, "Gelastic epilepsy a review and report of five cases," *Epilepsia*, vol. 12, no. 4, pp. 313–323, 1971.

[23] D. Andrewes, *Neuropsychology: From theory to practice*. Psychology Press, 2015.

[24] K. Heilman, "The neurobiology of emotional experience," *The neuropsychiatry of limbic and subcortical disorders*, pp. 133–142, 1997.

[25] R. Marek, C. Strobel, T. W. Bredy, and P. Sah, "The amygdala and medial prefrontal cortex: partners in the fear circuit," *The Journal of physiology*, vol. 591, no. 10, pp. 2381–2391, 2013.

[26] J. E. LeDoux, "Emotion circuits in the brain," *Annual review of neuroscience*, vol. 23, no. 1, pp. 155–184, 2000.

[27] R. J. Nelson and B. C. Trainor, "Neural mechanisms of aggression," *Nature Reviews Neuroscience*, vol. 8, no. 7, pp. 536–546, 2007.

[28] O. M. Klimecki, D. Sander, and P. Vuilleumier, "Distinct brain areas involved in anger versus punishment during social interactions," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[29] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[30] P. Johnson-Laird and K. Oatley, "Are there only two primitive emotions? a reply to frijda," *Cognition and emotion*, vol. 2, no. 2, pp. 89–93, 1988.

[31] R. W. Levenson, "Basic emotion questions," *Emotion review*, vol. 3, no. 4, pp. 379–386, 2011.

[32] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *Journal of nonverbal behavior*, pp. 1–28, 2019.

[33] L. F. Barrett, "Emotions are real.," *Emotion*, vol. 12, no. 3, p. 413, 2012.

[34] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[35] C. M. Doyle and K. A. Lindquist, "Language and emotion: Hypotheses on the constructed nature of emotion perception.," 2017.

[36] R. E. Nisbett and T. Masuda, "Culture and point of view," *Proceedings of the National Academy of Sciences*, vol. 100, no. 19, pp. 11163–11170, 2003.

[37] M. Kawahara, D. A. Sauter, and A. Tanaka, "Culture shapes emotion perception from faces and voices: changes over development," *Cognition and Emotion*, vol. 35, no. 6, pp. 1175–1186, 2021.

[38] L. F. Barrett, "Solving the emotion paradox: Categorization and the experience of emotion," *Personality and social psychology review*, vol. 10, no. 1, pp. 20–46, 2006.

[39] L. F. Barrett, "The theory of constructed emotion: an active inference account of interoception and categorization," *Social cognitive and affective neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.

[40] J. J. Gross and L. Feldman Barrett, "Emotion generation and emotion regulation: One or two depends on your point of view," *Emotion review*, vol. 3, no. 1, pp. 8–16, 2011.

[41] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International journal of human-computer studies*, vol. 65, no. 8, pp. 724–736, 2007.

[42] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on intelligent transportation systems*, vol. 6, no. 2, pp. 156–166, 2005.

[43] E. L. van den Broek, J. H. Janssen, J. H. Westerink, and J. A. Healey, "Prerequisites for affective signal processing (asp).," in *Biosignals*, pp. 426–433, 2009.

[44] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[45] R. W. Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[46] J. A. Russell, "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies.," *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.

[47] K. R. Scherer, "Speech and emotional states," *Speech evaluation in psychiatry*, pp. 189–220, 1981.

[48] P. Bucci, X. L. Cang, A. Valair, D. Marino, L. Tseng, M. Jung, J. Rantala, O. S. Schneider, and K. E. MacLean, "Sketching cuddlebits: coupled prototyping of body and behaviour for an affective robot pet," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 3681–3692, 2017.

[49] K. Boehner, R. DePaula, P. Dourish, and P. Sengers, "How emotion is made and measured," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 275–291, 2007.

[50] F. Grond, R. Motta-Ochoa, N. Miyake, T. Tembeck, M. Park, and S. Blain-Moraes, "Participatory design of affective technology: interfacing biomusic and autism," *IEEE Transactions on Affective Computing*, 2019.

[51] S. Blain-Moraes, S. Chesser, S. Kingsnorth, P. McKeever, and E. Biddiss, "Biomusic: A novel technology for revealing the personhood of people with profound multiple disabilities," *Augmentative and Alternative Communication*, vol. 29, no. 2, pp. 159–173, 2013.

[52] P. Sengers, K. Boehner, M. Mateas, and G. Gay, "The disenchantment of affect," *Personal and Ubiquitous Computing*, vol. 12, no. 5, pp. 347–358, 2008.

[53] M. Vircikova, G. Magyar, and P. Sincak, "The affective loop: A tool for autonomous and adaptive emotional human-robot interaction," in *Robot Intelligence Technology and Applications 3*, pp. 247–254, Springer, 2015.

[54] A. Landowska, "Affective computing and affective learning–methods, tools and prospects," *Stara strona magazynu EduAkcja*, vol. 5, no. 1, 2013.

[55] J. R. Searle, "Minds, brains, and programs," *Behavioral and brain sciences*, vol. 3, no. 3, pp. 417–424, 1980.

[56] S. Duncan, "Some signals and rules for taking speaking turns in conversations.," *Journal of personality and social psychology*, vol. 23, no. 2, p. 283, 1972.

[57] D. Swarbrick, D. Bosnyak, S. R. Livingstone, J. Bansal, S. Marsh-Rollo, M. H. Woolhouse, and L. J. Trainor, "How live music moves us: head movement differences in audiences to live versus recorded music," *Frontiers in psychology*, vol. 9, p. 2682, 2019.

[58] M. Hassib, S. Schneegass, N. Henze, A. Schmidt, and F. Alt, "A design space for audience sensing and feedback systems," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2018.

[59] V. Rivera-Pelayo, J. Munk, V. Zacharias, and S. Braun, "Live interest meter: learning from quantified feedback in mass lectures," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 23–27, 2013.

[60] C. Latulipe, E. A. Carroll, and D. Lottridge, "Love, hate, arousal and engagement: exploring audience responses to performing arts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1845–1854, 2011.

[61] C. Wang, E. N. Geelhoed, P. P. Stenton, and P. Cesar, "Sensing a live audience," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1909–1912, 2014.

[62] C. Wang and P. Cesar, "Physiological measurement on students' engagement in a distributed learning environment.," *PhyCS*, vol. 10, p. 0005229101490156, 2015.

[63] M. Hassib, S. Schneegass, P. Eiglsperger, N. Henze, A. Schmidt, and F. Alt, "Engagemeter: A system for implicit audience engagement sensing using electroencephalography," in *Proceedings of the 2017 Chi conference on human factors in computing systems*, pp. 5114–5119, 2017.

[64] R. Wataya, D. Iwai, and K. Sato, "Ambient sensing chairs for audience emotion recognition by finding synchrony of body sway," in *The 1st IEEE Global Conference on Consumer Electronics 2012*, pp. 29–33, IEEE, 2012.

[65] S. Kim, M. Billinghurst, G. Lee, M. Norman, W. Huang, and J. He, "Sharing emotion by displaying a partner near the gaze point in a telepresence system," in *2019 23rd International Conference in Information Visualization–Part II*, pp. 86–91, IEEE, 2019.

[66] L. C. De Silva, T. Miyasato, and F. Kishino, "Emotion enhanced multimedia meetings using the concept of virtual space teleconferencing," in *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, pp. 28–33, IEEE, 1996.

[67] P. Murali, J. Hernandez, D. McDuff, K. Rowan, J. Suh, and M. Czerwinski, "Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2021.

[68] A. Ranjan, J. Birnholtz, and R. Balakrishnan, "Improving meeting capture by applying television production principles with audio and motion detection," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 227–236, 2008.

[69] I. Duboyskii, A. Shabanova, O. Sivchenko, and E. Usina, "Architecture of cross-platform videoconferencing system with automatic recognition of user emotions," in *IOP Conference Series: Materials Science and Engineering*, vol. 918, p. 012086, IOP Publishing, 2020.

[70] J. L. Tracy and D. Randles, "Four models of basic emotions: a review of ekman and cordaro, izard, levenson, and panksepp and watt," *Emotion review*, vol. 3, no. 4, pp. 397–405, 2011.

[71] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.

[72] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[73] P. Bucci, L. Zhang, X. L. Cang, and K. E. MacLean, "Is it happy? behavioural and narrative frame complexity impact perceptions of a simple furry robot's emotions," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2018.

[74] L. F. Barrett, *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[75] M. Shermer, "Patternicity: Finding meaningful patterns in meaningless noise," *Scientific American*, vol. 299, no. 5, p. 48, 2008.

[76] A. Begel, J. Tang, S. Andrist, M. Barnett, T. Carbary, P. Choudhury, E. Cutrell, A. Fung, S. Junuzovic, D. McDuff, *et al.*, "Lessons learned in designing ai for autistic adults," in *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pp. 1–6, 2020.

[77] S. Garrod and M. J. Pickering, "Why is conversation so easy?," *Trends in cognitive sciences*, vol. 8, no. 1, pp. 8–11, 2004.

[78] S. Garrod and M. J. Pickering, "Joint action, interactive alignment, and dialog," *Topics in Cognitive Science*, vol. 1, no. 2, pp. 292–304, 2009.

[79] K. E. Watkins, A. P. Strafella, and T. Paus, "Seeing and hearing speech excites the motor system involved in speech production," *Neuropsychologia*, vol. 41, no. 8, pp. 989–994, 2003.

[80] L. Menenti, S. C. Garrod, and M. J. Pickering, "Toward a neural basis of interactive alignment in conversation," *Frontiers in human neuroscience*, vol. 6, p. 185, 2012.

[81] M. Pasupathi, L. L. Carstensen, R. W. Levenson, and J. M. Gottman, "Responsive listening in long-married couples: A psycholinguistic perspective," *Journal of Nonverbal behavior*, vol. 23, no. 2, pp. 173–193, 1999.

[82] E. Lee, J. I. Kang, I. H. Park, J.-J. Kim, and S. K. An, "Is a neutral face really evaluated as being emotionally neutral?," *Psychiatry research*, vol. 157, no. 1-3, pp. 77–85, 2008.

[83] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[84] P. J. Lang, M. M. Bradley, B. N. Cuthbert, *et al.*, "International affective picture system (iaps): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, no. 39-58, p. 3, 1997.

[85] B. Downe-Wamboldt, "Content analysis: method, applications, and issues," *Health care for women international*, vol. 13, no. 3, pp. 313–321, 1992.

[86] C. Darwin, "The expression of emotions in man and animals. new york: Philosophical library," *Original work published*, 1872.

[87] F. Strack, L. L. Martin, and S. Stepper, "Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis.," *Journal of personality and social psychology*, vol. 54, no. 5, p. 768, 1988.

[88] E. Hatfield, J. T. Cacioppo, and R. L. Rapson, "Emotional contagion," *Current directions in psychological science*, vol. 2, no. 3, pp. 96–100, 1993.

[89] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[90] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder, "The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction," *Presence: Teleoperators and Virtual Environments*, vol. 15, no. 4, pp. 359–372, 2006.

[91] K. H. Greenaway, E. K. Kalokerinos, and L. A. Williams, "Context is everything (in emotion research)," *Social and Personality Psychology Compass*, vol. 12, no. 6, p. e12393, 2018.

[92] P. H. Bucci, X. L. Cang, H. Mah, L. Rodgers, and K. E. MacLean, "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)*, pp. 1–7, IEEE, 2019.

[93] V. S. Ramachandran and E. M. Hubbard, "Synaesthesia–a window into perception, thought and language," *Journal of consciousness studies*, vol. 8, no. 12, pp. 3–34, 2001.

[94] W. Köhler, *Gestalt psychology: an introduction to new concepts in modern psychology.* Liveright Pub. Corp., 1947.

[95] Y.-C. Chen, P.-C. Huang, A. Woods, and C. Spence, "When "bouba" equals "kiki": Cultural commonalities and cultural differences in sound-shape correspondences," *Scientific reports*, vol. 6, no. 1, pp. 1–9, 2016.

[96] A. D'Onofrio, "Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm," *Language and speech*, vol. 57, no. 3, pp. 367–393, 2014.

[97] A. I. Goller, L. J. Otten, and J. Ward, "Seeing sounds and hearing colors: An event-related potential study of auditory–visual synesthesia," *Journal of Cognitive Neuroscience*, vol. 21, pp. 1869–1881, 10 2009.

[98] G. Lakoff and M. Johnson, *Metaphors we live by*. University of Chicago press, 2008.

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

[100] R. W. Gibbs Jr, "Evaluating conceptual metaphor theory," *Discourse processes*, vol. 48, no. 8, pp. 529–562, 2011.

[101] A. Sellen, B. Buxton, and J. Arnott, "Using spatial cues to improve videoconferencing," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 651–652, 1992.

[102] S. O. Adalgeirsson and C. Breazeal, "MeBot: A robotic platform for socially embodied telepresence," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 15–22, Mar. 2010. ISSN: 2167-2148.

[103] P. Dourish and S. Bly, "Portholes: Supporting awareness in a distributed work group," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 541–547, 1992.

[104] C. Greenhalgh and S. Benford, "Massive: A collaborative virtual environment for teleconferencing," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 2, no. 3, pp. 239–261, 1995.

[105] J. R. Cooperstock, "Multimodal telepresence systems," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 77–86, 2010.

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

[107] A. Kimura, M. Ihara, M. Kobayashi, Y. Manabe, and K. Chihara, "Visual feedback: its effect on teleconferencing," in *International Conference on Human-Computer Interaction*, pp. 591–600, Springer, 2007.

[108] K. Karahalios and T. Bergstrom, "Social mirrors as social signals: Transforming audio into graphics," *IEEE computer graphics and applications*, vol. 29, no. 5, pp. 22–32, 2009.

[109] T. Bergstrom and K. Karahalios, "Conversation clock: Visualizing audio patterns in co-located groups," in *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pp. 78–78, IEEE, 2007.

[110] J. Hailpern, K. Karahalios, and J. Halle, "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*, pp. 453–462, 2009.

[111] T. Matthews, J. Fong, F. W.-L. Ho-Ching, and J. Mankoff, "Evaluating non-speech sound visualizations for the deaf," *Behaviour & Information Technology*, vol. 25, no. 4, pp. 333–351, 2006.

[112] B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti, "Soundscape ecology: the science of sound in the landscape," *BioScience*, vol. 61, no. 3, pp. 203–216, 2011.

[113] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, 2017.

[114] F. De Saussure, *Course in general linguistics*. Columbia University Press, 2011.

[115] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales.," *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[116] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vendeventer, D. W. Cunningham, and C. Wallraven, "Cardiff conversation database (ccdb): A database of natural dyadic conversations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 277–282, 2013.

[117] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 143–146, 2011.

[118] C. Erlingsson and P. Brysiewicz, "A hands-on guide to doing content analysis," *African Journal of Emergency Medicine*, vol. 7, no. 3, pp. 93–99, 2017.