Considerate Systems

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

Rahul Rajan

B.S., Electrical and Computer Engineering, Georgia Institute of Technology
M.S., Electrical and Computer Engineering, Georgia Institute of Technology

Carnegie Mellon University
Pittsburgh, PA

August 2016

Acknowledgements

Thanks!

Abstract

This is an abstract.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

People are naturally aware of social situations. As a species, we draw upon a collective consciousness to interpret and respond to these situations in socially appropriate ways. The fields of comparative psychology and sociobiology inform us that non-human species also use social behaviors to support their interactions From ants and bees to wolves and baboons, many species have developed strategies for social behaviors that support cooperation and survival [16]. Humans in particular have developed languages, and are excellent at interpreting each others signals. People have an innate ability and desire to communicate intent verbally and non-verbally, and have a shared behavioral "language" (or common sense) on how to react to emotional and social input.

Communication is fundamental to the social process. Technologies that enable people to communicate better are more often than not embraced by society. Paper and the printing press set off an early social communication revolution. Email drove adoption of the personal computer. Communication through mobile phones are changing the economic landscape – even in developing countries, and social media is touted as one of the big successes of the new millennium. Each of these technologies might be seen as enhancing how we communicate with each other, disrupting practices and procedures already set in place. When people begin interacting through a new technology, be it smoke signals, writing on paper, email, phones or social network technology, they develop new shared expectations, and derive social meaning from when and how it is used [113]. Eventually, a disruptive technology succeeds in becoming an established communication channel when people develop a shared body of knowledge that can be seen as an evolutionarily stable strategy for communicating using that technology. For instance, with paper, writing techniques and etiquettes were developed, and a new shared understanding arose on what to communicate and expect from a job offer letter, or in the prose of a mystery novel. In the case of interactive computing systems, people are still grappling with a medium that is varied in communication pace and dimensions of interactivity.

Computing systems that encourage communication that builds on the letter-writing paradigm that we are already comfortable with, have grown phenomenally as attested by the success of text messaging and email. With systems that support communications or interactions at a more natural conversational pace, success is hard-pressed as these technologies are more prone to “breakdown” [61]. This might be because the common sense knowledge of this space-time is so deeply embedded in people that any deviation from the expected social protocol is immediately picked up on. It causes a shift in focus from the task at hand to the technology at hand. For example, delayed responses on a telephone call (technology) can detract focus from the topic of conversation (task). The same phenomenon occurs in interactive games, which is why successful game designers for products like Kinect and Wii try to mask the inherent latency of these systems. Pace and feedback are just one of the dynamics in social engagement. Status and societal structure also drive social interactions. For example, smart TV products tout their integration with Facebook and Twitter, but rather than simply displaying the same information on a bigger screen, a more desirable feature would be to display information based on the situated context, like the people in the living room and the relationships between them.

To come to terms with the disruption that newer technologies present, we need to peel back the curtain of time and peer into the future. There are three trends in particular that become apparent. First, *natural user interface* (“NUI”; like pointing, speech, & gaze) technologies are getting better, and might be the preferred mode of interaction in post-desktop, public or complex scenarios, like driving. This leads us into our second observation that as such interfaces get better, they will enable technologies to seep deeper into the fabric of society, making them more *situated*. Third, attention is a limited human resource. In an always connected post-desktop world, the *strain on attention* is only going to become more tenuous. To be successful in such situated, attention-starved environments, it will be imperative for future technologies to be aware of the shared practices and social protocol that humans engage in, and respond in ways that allows it to be perceived as unobtrusively as possible. Evolutionary psychologists argue that engaging in these behaviors establish trust, foster goodwill, and enable cooperation [73]. To not engage in the same manner creates breakages and hinders communication. Systems need to embody this common sense knowledge and act upon it in ways that are respectful of the collective consciousness [41]. Collective consciousness might be thought of as the shared awareness of a situation between actors in an interaction. If there is a death in the family, for instance, it is common sense, and the agreed upon protocol, to not crack a joke. We believe that embodying this common sense knowledge of communication in human-computer interactions can enhance trust, cooperation and reciprocity.

A way forward towards this goal is to move from a user-centric design approach to a socio-centric one, central to which are the notions of embodied interaction and situated knowledge. These refer to the idea that people's understanding of the world, themselves, and their actions are strongly influenced or perhaps even constructed by varying physical and social situations [68]. People are embedded in a network of social, cultural and organizational meaning, fluid and negotiated between other people around them. Meaning is derived and constructed from the ground up, unfolding in real time through our interactions with each other and the system. Along with meaning co-construction, people are also implicitly, negotiating the rules of engagement. They temper their responses and adapt their behavior to achieve the most desirable outcome. Computing systems need to engage with people in the same way, being considerate of the web of social actions they are embedded in. Fundamentally, a socio-centric design approach emphasizes that any action is necessarily both utilitarian and affective, objective and subjective, and it would be amiss to divorce one from the other.

Affective and social computing have made great progress in allowing computers to *recognize* emotional states and social signals [62, 154]. Considerate systems will build on this awareness and anticipate how to *respond* to users in socially appropriate ways. Humans are considerate when they are thoughtful of the other, and this is displayed in their behavior. This paper will present a functional model of what this entails for human-computer interactions. Considerateness is a virtue, in that its goal is to create goodwill while successfully traversing relationships. For considerate systems, this virtue is a functional awareness of the user's cognitive and affective state, and the collective consciousness of a group. To succeed at its communication goals, the system displays this awareness by behaving in socially appropriate ways, i.e., by supporting and not disrupting the interactions that it is actively or passively part of.

We define Considerate Systems as,

> *A system that displays an awareness of the state of the user/s, and their situated social context, by behaving in ways that support their interactions.*

Considerate is shorthand for a shared knowledge that is implicitly negotiated through social actions and behaviors to improve a communication outcome. A considerate system must model such aspects of social feedback, in order to act in socially acceptable ways. Our hypothesis is that by being considerate, situated systems are more effective in achieving their goals. To prove this we start by developing a framework of what it means for a system to be considerate. We note that the framework is descriptive rather than prescriptive, and that its relevance and application are context-specific. To evaluate the framework, we wanted to hold it to a standard higher than simply improving human-computer interaction, which guided the choice of the application scenarios explored in this thesis.

For the first scenario, we chose the conference call setting as it presents a communication constrained domain where we might even improve human-human communication. We ground the framework by applying it to the design of a mediating agent for this popular collaboration tool, and perform a number of experiments to validate our hypothesis. For the second scenario, we wanted to broaden the application beyond the audio modality, and chose a domain where the consequences of being inconsiderate are more drastic. The distracted driving setting presents such a domain where the margin of error is very small, and where any inconsiderateness on the part of the system gets magnified. In this way, the first scenario allows us to investigate various facets of system actions, whereas the critical nature of the second scenario brings the focus more squarely on the timing of its actions.

The main tenet of this thesis is that while today's systems are incapable of adequately representing the tacit unit of context [30], and hence cannot fully capture the richness and complexity of situated interactions, there are techniques we can employ *now* based on the considerate model to improve system effectiveness and overall user-experience.

Chapter 2

Related Work

A body of work on social appropriateness leading to Considerate Systems falls under the context awareness and interruption & disruption problem space. As technology got better, our interactions with computers become more situated in post-desktop scenarios. The social aspects of these interaction, or the shortcomings therein, started becoming more apparent. We review work from both of these areas in more detail below.

2.1 Interruption & Disruption

An early paper on considerate systems [57] focused on *when* to communicate, an important but small part of the commonsense of being considerate. It describes projects like AuraOrb [2], CarCOACH [7], and Notification Platform [82], where the focus was on endowing computing systems with an understanding of the user's focus of attention, workload and interruptibility [80]. The idea is that part of being considerate is associating a cost with an interruption and comparing it against the user's state of interruptibility which the system assesses. These projects were often successful when the knowledge of when to communicate could be easily represented, and when reasonable inferences could be made as to the importance of an interruption and the cognitive load of the user. The use of commonsense knowledge to compare communication streams can reduce the impact of interruptions [6]. Depending on user responses, the system could regulate its interactions with the user using a blackboard architecture [7], or Bayesian statistics in order to mitigate social feedback mistakes [82].

Another area of work that is focused on *how* one might more considerately interrupt or notify a user is peripheral interaction [10]. The strategy is to minimize the attention required to interact with devices, so as to make them less obtrusive. Humans are naturally able to perform peripheral activities, like tying shoelaces while engaged in conversation. The goal of peripheral computing is to design interactions

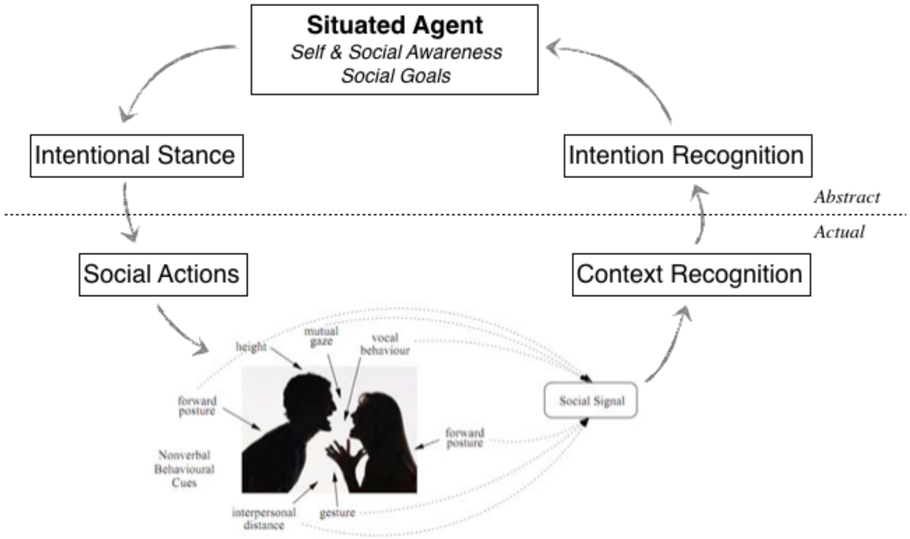


Figure 2.1: The figure depicts the major blocks needed to build a situated agent that demonstrates social intelligence in its interactions with users.

that take advantage of these inherent human capabilities. The approach of employing the periphery of attention in human-computer interaction has been explored under various terms such as calm technology [159], ambient information systems [127], and peripheral displays [109]. Such approaches have also been explored in a number of application domains including the home, office, classroom and car, using a number of interaction styles, such as tangible interaction, gestures, and wearable devices.

Indeed, systems that capitalize on commonsense knowledge of when and how to communicate can improve interaction, and can be beneficial across domains. Previously, their widespread adoption might have been hindered by the fact that most users put up with any (inconsiderate) interaction that would allow them to solve their problem. Computing systems are now, however, becoming an indispensable part of our social lives at the individual (smart-phones), small group (smart-televisions) and community (social networking) levels. The effects of such interruptions or social faux pas are going to be felt more strongly, especially in social settings (see Figure 2.1). This explains the increased interest in research towards conferring these systems with basic emotional and social intelligence. These research efforts, broadly-speaking, fall into two categories — 1) affective and social computing, i.e. being able to *sense and contextualize* the social environment the system is embedded in; and 2) intelligence, i.e. being able to *plan and react* to the sensed context.

2.2 Social Sensing & System Behavior

The issue of sensing the social setting can be addressed through three frameworks as was highlighted by Vinciarelli et al. [154]. The first framework comes from *cognitive psychology*, and is based on emotion and affect. The second framework comes from *linguistics*, and uses vocal prosody and gesture which are treated as annotations of the basic linguistic information. The third more recent framework is called *Social Signal Processing* where the amplitude and frequency of prosodic and gestural activities is used to understand speaker attitude or intention. The difference between this and the linguistic framework is that it uses common non-linguistic signals about the social situation, and is different from the affect framework in that social relation is communicated rather than speaker emotion.

Advanced response behaviors have received more attention in the field of embodied conversation agents. The focus here has been on endowing these systems with facilities to communicate and respond through the production of language and associated non-verbal behavior (gaze, facial expression, gesture, body posture) [18]. A commonly used planning approach is the *Do the Right Thing architecture* [104] that provides the ability to transition smoothly from deliberative, planned behavior to opportunistic, reactive behavior in order to maximize task efficiency while maintaining trust. Similar models have been successfully used in multiparty dialog systems to regulate turn-taking, and other conversational dynamics [21]. The authors further identify numerous improvements that can be made to the behavior models that would better support interaction, which informs some of the work in this thesis.

There has also been an effort towards long-term behavior adaptation through the use of emotion and memory. Francis et al. [50] describes a reflective architecture for agents where detection of emotional stress or frustration can trigger re-evaluation of past behavior, which sets new strategies and goals. They showed that such adaption can extend the range and increase the behavioral sophistication of the agent without the need for authoring additional hand-crafted behaviors. A vast majority of ubiquitous systems, however, would be even more distracting with animated embodied agents, but their social response remains as important. Perhaps work that is closest in spirit to the material described in this thesis, posits the cognitive user interfaces that respond appropriately under uncertainty [165]. However, the focus of that work is not on the social aspects of interaction.

2.3 Beyond Behavioral Response

In the following chapter, we describe a framework by which systems can be designed to be more considerate of the user and the social setting, thus improving trust and perceived intelligibility of the system. We identify a number of ways of incorporating commonsense approaches that can be employed without the need for complex models and techniques. In particular, we focus on the communication channel that is enabling the co-construction of meaning in which the actors are involved. This will motivate our design choices for a Considerate Mediator, which serves as an embodiment of the proposed framework.

We will review more related work with respect to the specific scenarios considered for this thesis in their corresponding chapters. As mentioned in the introduction, these distinct and pertinent scenarios are the conference call setting, and the distracted driving scenario. The Considerate Mediator will be grounded and evaluated in each of these scenarios as will be described in chapters 4 & 5, respectively. In this way, we can highlight how the application of the framework reveals areas for interaction breakdown in prior approaches, and the potential avenues to address them.

Chapter 3

Considerate Systems: How Systems *Respond* to Context

The Computers as Social Actors (CASA) [118] is a well researched and evidence-based paradigm that suggests that people respond to technologies as social actors, applying the same social rules used during human-human interactions. People display a tendency to anthropomorphize non-human agent figures, attributing human characteristics and even personality to these systems [117]. This tendency creates a dichotomy between perception and reality. People are naturally inclined towards believing that systems will respond when it is appropriate, but they haven't and won't until we explicitly model considerate behavior and appropriate feedback; this leads to unmet expectations. People are well known for being terrible at accommodating others that don't respond appropriately in social circumstances. Still somehow people are expected to accommodate computers with these shortcomings. They try to ignore the lone phone ringing in a lecture, or the app notification in the middle of a conversation. It is when they are cognitively or socially loaded — as when following GPS directions while driving, or performing in front of an audience — that these shortcomings can lead to communication breakdown or worse. In such situations the consequences of inconsiderateness can critically impact performance, and even safety. The indiscretion shown by the system when it pops an email notification, while the laptop is connected to the projector, can be a cause for frustration and embarrassment, and detract time and attention from the meeting. It is this interactional friction caused by partially-formed social actors that we want to obviate.

Miscommunication caused by poor modelling of social feedback is a problem that gets magnified when technologies encroach into the natural interaction space-time, as argued in the introduction. On the one hand, the potential for technologies that operate in this space-time is tremendous. The intransigence of human behavior has emerged as the root of most of the world's biggest challenges, be it in the health, social or economic spheres. Aggressive driving, drug addiction, excessive water consumption are just a few examples. There are multitudes of people that are motivated to tackle such problems in their lives. The multi-billion dollar weight-loss industry is a testament to this. Behavioral psychologists have uncovered numerous cognitive biases that hinder progress. Building on these insights, the behavior research community has proposed multiple models to effect behavior change, which they have been successfully demonstrated [49, 63, 101].

A key component in providing the right kinds of triggers or nudges in-situ lies in the choice of how the feedback is framed. This is where context-aware technology that takes into account place, time and situation through sensors and virtual sensors (model based interpretation of sensors) can play a big role. For example, there is evidence showing that imagery techniques at the time of need have been successful in reducing food cravings for weight management [83]. If people perceive such technologies as annoying, their potential will likely not be realized. The role that such technologies have to play will be that of bona fide social actors — "An action is social insofar as its subjective meaning takes account of the behaviors of others and is thereby oriented in its course [157]." Not only will systems have to be context-aware, they will also have to modify their action when a potential reaction is not desirable, i.e. they will have to be consequence-aware.

We use the term *Consequence-Aware* to represent the risk-reward strategy implicitly used by humans to achieve their goals, while simultaneously minimizing undesired consequences. It encompasses the various facets of intelligence we demonstrate including emotional/social intelligence, intuition, and the ability to learn through interaction. It requires endowing machines with a notion of *self*, or reflexivity, and a notion of *other*, or Theory of Mind. Reflexivity refers to decision-making on the basis of continuous reflection on the conditions of one's action [59]. Theory of Mind is the ability to infer another agent's full range of mental states (beliefs, desires, intentions, imagination, emotions, etc.) [128]. Building consequence-aware systems then is a goal, which needs to be broken down and tackled incrementally. We can start making progress on these fronts by having a model of the user (*other*), the system (*self*), and their shared task goals. Our approach here is not to be prescriptive but descriptive. These models serve as a framework that can inform the sensing, reasoning, and action capabilities a consequence-aware agent should be provided with. As machine learning, knowledge representation and automated reasoning get better, these models might be fleshed out to allow for semantic formulations of appropriate social action. In the mean time, models can still serve as a scaffolding to guide the design of commonsense considerate response at the surface or syntactic level, which we will cover in more detail in section 3.4.

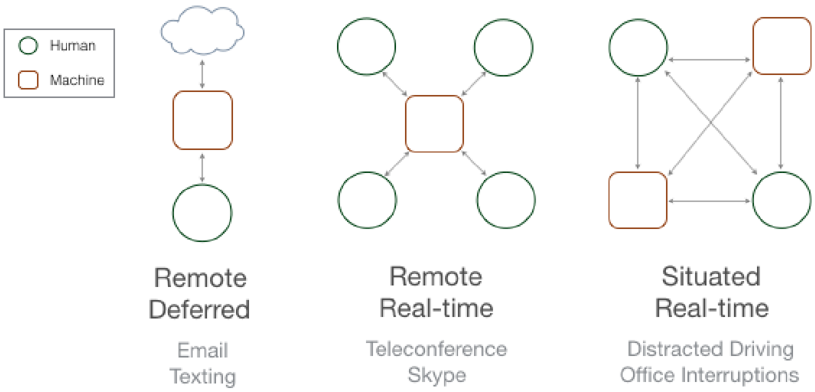


Figure 3.1: Social Group Models

3.1 User Model: User-centric to Socio-centric

When we consider an interaction with a user, we must acknowledge the different levels at which this interaction occurs. Humans are stimuli perceiving, emotionally charged, cognitive processing entities. Their experience is colored by the social situation they are embedded in. By modeling this phenomena in a user, we can support interactions at these different levels:

- *Perceptual* (dimming a bright display),
- *Pre-cognitive* (identifying and reducing stressful external stimuli the user is not consciously aware of like heat, noise, etc),
- *Cognitive* (users are easily distracted from the task at hand, and have definite limits on memory and learning),
- *Behavioral* (accommodating for user mood/comfort by adjusting conversational agents tone, etc).

To understand the social level, we start with basic models to represent social groups. We might start with two simple characteristics in particular that might help define the nature of their interactions — whether the participants are collocated or not, and the rate of interaction. When participants are collocated their interactions are situated in the physical world, for example at a golf club where people gather and socialize. When they aren't collocated, they interact in virtual domains, such as on Facebook. In both the physically situated and virtual domains, when interaction is happening at a conversational pace, i.e. the response times are around a second or so, we say the interaction is happening in real-time. When response times are slight delayed, like during a chat session, the interaction is happening in near real-time. When responses are delayed longer than this, with emails for example, we can say that the interaction is being deferred.

Based on these characteristics we present three types of social group models. The first model shown in Figure 3.1 (a) represents the interaction between a single user and a system that connects participants who are virtually situated. Interaction here is not necessarily real-time or bidirectional, and includes internet forums, social networks, etc. The second model in Figure 3.1 (b) represents the interaction between users and a mediating system such as Skype. Such interactions are necessarily real-time and can be physically or virtually situated. This category of interactions includes presentations, conference calls, etc. The third model in Figure 3.1 (c) represents interactions in which the system and the user are equal participants and co-collaborators engaged in a task. These are physically situated and real-time, and can include interactions between the user and computing systems in an automobile, aircraft, or gaming consoles like Wii and Kinect. In this work, we are mostly concerned with the (b) and (c) models of social groups.

It is also in (c) models, where interactions are situated and occurring in real-time, we are arguing that the social layer of a user is invoked, even if it is just a single user interaction. Interactions in this space-time are inherently social. Human expectations at this space-time are so ingrained that any deviation from what is considered normal behavior might agitate or aggravate users. The uncanny valley is a hypothesis in the field of aesthetics which holds that when features look and move almost, but not exactly, like natural beings, it causes a response of revulsion among some observers. We can make the argument that the uncanny valley hypothesis extends to not just aesthetic features, but also to communication behaviors.

3.2 System Model: A Communication Paradigm

Earlier we argued that in a social setting, every action is necessarily both utilitarian and affective. It would be clearer, perhaps, to think of action in a social setting as a communication. Even the simple act of closing the door can be used to communicate aggression if the door is closed hard, or politeness if it is closed softly. By taking action (or not), a participant is communicating their mental and affective states. The communication, for both the sender and the receiver, is happening at the conscious and subconscious levels, using verbal and non-verbal signals. Social signal processing tries to make sense of these signals. The modus operandi for considerate systems is to provide the expected feedback to support these communications — to confirm it, to enhance it, and to reduce its overhead.

For a system to be considerate in a social setting, it needs to be supportive of the communication in the social group. To support communication, the system should be aware of the different types of communication (verbal and nonverbal), and the mediums through which people communicate (audio, visual, touch, etc.). Visual nonverbal communication includes facial expressions, body movements and posture, gesture and touch, eye contact, and interpersonal space. In audio, nonverbal communication happens using the intensity, timing and pace of speech, and through the tone, pitch, volume, inflection, and rhythm in voice. Humans tune these parameters of communication based on the nature of the interaction, for example speaking in hushed tones when talking about something very personal and private, or in a library. Systems need to reciprocate in a similar fashion for successful integration.

Communication can be further be analyzed at the different structural and organizational layers at which it happens. Here again each level shapes the nature of the interaction and the commonsense rules entailed. These levels are listed below in order of increasing scope:

- *Turn-taking* (Dominant vs. Dormant)
- *Roles* (Sibling vs. Friend)
- *Status & Power* (Employer vs. Employee)
- *Type of meeting* (Instructional vs. Collaborative)
- *Social Setting* (Formal vs. Informal)

Moving on to the actual use of language to communicate, we can differentiate between some of the strategies humans use to maneuver around and lubricate social interactions. We list some of these below, and illustrate them with examples on how to embed considerateness into communications in order to better support interactions:

- *Superficial:* Social niceties like “thank you”, and “please” acknowledge and celebrate interdependence, reduce friction in an interaction, and can make requests, confirmation, and successes more explicit.
- *Embellishing:* Using intonation and spacing between words in a manner that can improve the users retention of information.
- *Ancillary:* A preface or a follow-up to a part of the conversation. Often we find people using qualifiers to make their communication more easily accepted.
- *Multi-modal:* Communication can occur across modalities as well, for example, visually displaying information when the audio modality is being used.
- *Orthogonal:* Euphemisms are indirect expressions which replace words and phrases considered harsh and impolite, e.g. kick the bucket, downsizing.
- *Functional:* Highlighting the key concepts in the current conversation to provide a quick summary for a new entrant, or repetition to accommodate for lapses of attention.

Table 3.1: Task Model: A taxonomy for pursuing system goals and minimizing breakdown. The table lists the two approaches for each.

Pursue System Goals	Minimize Breakdown
Feedback	*Appropriateness*
Social	User
Conversational	Social
Direct Manipulation	Communication
Reinforced	
Feedforward	*Differential*
Suggestions	Preset Rules
Influence	Adaptive
Motivation	Preferential
Gamification	Priorities

3.3 Task Model: Purse System Goals & Minimize Breakdown

Successful use of a tool requires that not only were the goals met, but they were done better or easier than they would have been otherwise, with breakdowns or interactional friction kept to a minimum. For example, the goal while driving is to reach a destination. This task can be completed even when some kinds of breakdown is encountered (getting stressed by strange controls or indicators, having someone honk at you, being cutoff at an intersection, almost hitting something or someone, getting a police citation, etc.). We can say, however, that the task is completed more successfully when one minimizes the risk of such encounters. Therefore, an optimal action by the system is one that maximises desired consequences (goals), while minimizing undesired ones (breakdown).

The task model aids in this action selection problem by presenting a taxonomy of approaches for pursuing scenario goals, and minimizing breakdown. Scenario goals refers to situated task goals (such as ensuring participants get equal stage time in a collaborative meeting), as well as interaction goals (such as transparency and efficiency). These goals can be pursued using the *feedback* and *feedforward* family of approaches. Feedback is how a system orients itself to the user, giving them some indication of their past actions. On the other hand, feedforward is when the system orients the user towards a goal, giving them indications of advisable future actions. Methods to minimize breakdowns fall under the *appropriateness* and *differential* family of approaches. We discuss each of the four approaches as part of the taxonomy below:

3.3.1 Feedback

Feedback is interesting in that it falls on a spectrum that ranges from the subtle to the overbearing. A simple greeting can feel considerate, and is a social feedback of acknowledgement. Feedback of noticing that you are hoarse or sick is a stronger statement of considerateness. Without social feedback, people depend on a preset audience model to communicate, such as while lecturing to a distance learning class, and often don't know if they are being understood. This paper will use the already constrained human/human communication channel of a conference call to explore computer mediated considerate feedback. For instance, we show how such social feedback can be used to bring about awareness of the presence of participants who are still on the line but haven't contributed in a while. Apart from social feedback, other forms of feedback available to considerate systems include conversational, direct manipulation and reinforcement feedback.

Conversational feedback in speech recognition systems can improve user engagement. Feedback such as "uh huh" and "come again?" delivered at a conversational pace can be reassuring and motivating, giving users a sense of whether they are being understood or not. Such dynamic feedback that speakers give each other in a conversation allow them to make fluid adjustments and keep a dialog on-track. Considerate systems should look for opportunities to give such lightweight feedback to support and enhance interactions. Conversational feedback in speech recognition systems might improve user engagement.

Direct manipulation feedback is about equipping users with tools to provide feedback to other users of their own accord, and is similar to back-channeling. For example, the ability to indicate to the speaker to speak louder without having to interrupt them mid-speech on a conference call, is analogous to adjusting the microphone in the physical world. By enabling such forms of feedback, considerate systems can better facilitate interactions among users and achieve task goals more efficiently.

Reinforcement feedback is different from the interactional feedback discussed above, as it is more related to the performance of a user. It is an important feature of educational or motivational systems that can monitor the actions of the user, and provide positive or negative reinforcement. A simple example is that of speed radars that flash back the speed of the driver when they are above the speed limit. Such feedback reinforces positive behavior.

3.3.2 Feedforward

In behavioral and cognitive science, feedforward is a method of teaching and learning that illustrates or suggests a behavior or path to a desired goal. Feedforward is different from feedback in that, instead of describing performance in the past and providing room for reflection, it actually illustrates better action choices that the user should make in the future. For example, in a collaborative session, feedback would be in the form of telling a participant that they talked for 80% of the time, whereas feedforward could take the form of encouraging the user to take turns sharing the floor with others.

Feedforward can also nudge users towards more efficient or productive behavior through persuasive methods, which include influence, motivation, gamification, etc. Influence can be won by trying to match the system's response to the user's affective state in order to establish a level of trust and comfort. Motivational approaches try to change the attitude and actions of people through the use of descriptive social norms. For example, having a card in hotel rooms that state that most guests reuse their towels motivates the guests to do the same [60]. Reward mechanisms, such as gamification, such as the energy flow diagram in a prius have been successfully used to get motorists to drive in a more eco-friendly manner [12]. Recently, such technologies have been focused on behavior changes in health domains through digital health coaching, which seeks to augment personal care delivered to patients [18].

While feedback and feedforward approaches can be used by the system to guide the social group towards its goals, when and how a system communicates this can drastically effect user performance. It was found that delaying reinforcement feedback half a second or more relative to turning performance in a car maneuver could improve learning [7]. Like we argued above, success in situated and real-time interactions might depend on considerate response, which is the delicate coupling between goal-oriented action and its precise parametrization. Approaches to parametrize action so as to minimize breakdown are discussed next.

3.3.3 Appropriateness

Appropriateness can range from the less obvious commonsense knowledge to more salient societal norms (such as not spitting on floors inside buildings). Designing considerate responses requires establishing appropriate statements, actions and reactions to begin with. Thinking about it from the various levels of the user model and system communication model described above allows us to identify opportunities that can be used to design systems that begin to do this.

In the user model, we saw examples of supporting interaction at the different layers including perceptual, pre-cognitive, cognitive, behavioral and social. A lot of early work on ergonomics stressed human performance and efficiency, which brought into focus the perceptual and pre-cognitive layers. For example, the eyelid might react on the order of milliseconds, while some responses to warmth, smell etc. can be dramatically slower. With the increasing use of technology for work, the emphasis shifted to cognitive performance and efficiency.

At the behavioral level, one aspect of appropriateness comes from the *tone* the system takes with the user. Systems can come across as condescending, self-centered and arrogant. One manifestation of this in how the onus is on the user to understand the operation of the system [52]. The successful system might have a dynamic model of the experience level of the user and accordingly match its language, tone and mannerisms. Systems can make straight forward modifications in spoken dialog systems using volume, pitch, pitch range and speaking rate to change how the system is perceived by the user [116].

At the social level, we can recognize that information might be sensitive and not privy to everyone in the group. For example, in a distributed meeting scenario, the system can privately remind the speaker of how much time is left on the call, to get them to reflect and react appropriately, without calling attention to it in front of the other participants.

Similarly, appropriateness as an approach can be thought of from the different layers at which system communication is happening. Most critical, perhaps, is the system's appropriateness at the *turn-taking* level, where inappropriate seizing and ceding of the floor greatly impacts the flow of communication. Turn-taking is negotiated in a number of ways, including through the use of speech intonation and pauses, gestures, body orientation and eye-gaze. This dynamism makes turn-taking difficult to interpret and predict. However, as shown in [20] a receptionist system can act appropriately without having to intimately understand the social context by analyzing surface characteristics. A mediating system should take advantage of similar surface interactions and decide when to interject.

Roles and social status of the participants in a group is another area where appropriateness is important. A user might not want the system to communicate the same information it would to a group of adults, versus, when a child is in the group. Similarly, while dominance during meetings is not desirable, if the interaction is between an officer and his subordinate, it might be appropriate. Other opportunities to be appropriate include ensuring that all participants have the same level of grounding or pre-understanding, i.e. use of shared vocabulary, knowledge of facts, etc. This might be tackled by highlighting keywords & phrases that the system has spotted in the past. Designers should be sensitive to the social and cultural norms as determined by the setting and the type of interaction that is occurring.

3.3.4 Differential Response

Differential (or adaptive) response is an appreciation of the fact that user preferences and circumstances change. What might work for a user one time, might not the next time; settings can be too rigid. One way to be differential in response is to use commonsense knowledge to adapt to the simple differences that it can reliably classify. Successful systems might accommodate for this by modeling the situation. For example, even though the phone is in silent mode during a meeting, the user might want a call from their boss to come through.

Another approach to being adaptive is to learn from past experiences. A system monitoring the effect of its actions on the environment, can learn to optimize its action choice so as to achieve desired results. This is an important feature for considerate systems as user preferences can vary significantly in ways that can't be encompassed with commonsense knowledge. Intra-user preferences themselves change over time and can vary based on the situation. For example, as drivers get more familiar with an area, they might prefer to have less explicit GPS instructions, as compared to when they are new to an area.

In social groups, through differential response a system can be preferential towards certain users depending on their need and the system goals. For instance, a system could be capable of detecting that one user needs more support than the others in a group activity, and provide the necessary assistance. On a conference call this might translate to detecting that a participant is continuously being cut off by another participant. The system could respond by increasing their volume, and decreasing the volume of the other participant.

The nature of information being communicated is also highly differential, and can vary from the frivolous to the critical. In an age where more people are getting fatigued with information overload, a considerate system should have the ability to prioritize communication based on the situation, and bandwidth available at the different communication levels. For instance, at the cognitive level, performance while dual tasking can be compromised if the demands exceed the bounded human capacity. This bound varies with age and experience. Inexperienced drivers, for example, require more concentration while driving. In such a situation, the system could use the user's own psychophysiological indicators to determine if they are cognitively loaded before engaging with them [131].

3.4 Towards Considerate Response: An Analytic Approach

Communication relies on a number of signals occurring at different levels from perceptual to social. These can be supported in a variety of ways depending on the scenario goals. Every action towards these goals has consequences, both functional and affective. A consequence-aware system would act towards achieving intended consequences (goals), all the while minimising undesirable ones (breakdown). We define considerate response as achieving the scenario/task goal without breakdown, which is an undesirable shift of focus from the task at hand to the technology at hand:

$$\text{Considerate Response} \Rightarrow \text{Pursue System Goals} + \text{Minimize Breakdown}$$

The point of such a formulation is to reinforce the idea that in social interactions, neither is it desirable for the system to achieve its scenario goals without considering the social consequences of doing so; nor does it make sense for the system to abandon its functional goal for fear of upsetting propriety and relationship. A trade-off must be made. Optimizing such trade-offs requires metrics of social and functional communication value. An analytical approach is needed, which includes a repeatable process that can be used to break a problem down to the elements necessary to solve it. For Considerate Systems these elements are formulated as the following questions:

- How to pursue goals?
- What is the potential for breakdown?
- How to minimize breakdown?

Each of these questions need to be examined separately, systemically, and sufficiently, with proper consideration given to all alternatives. This is precisely where the user, system and task models developed earlier in the chapter serve as guidelines in addressing these questions. For instance, in a driving situation, the scenario goal might be to deliver a text message to the user. However, there is potential for breakdown at the perceptual level if the text message is visually displayed, or at the cognitive level if the user is in the middle of a mentally demanding episode. Breakdown might be minimized by using the audio medium or by delaying the message till the driver is not cognitively loaded. To demonstrate the utility of the analytical approach towards creating a considerate response, we apply it in two exemplar application domains, which are discussed in section 3.6.

3.5 Architecture: Actuation & Gateway Modules

We start with a basic sense-reason-act control loop (marked in black, Figure 3.2), which can be used to represent any human-computer interaction in a situated environment. The term situated environment refers to all of the real-world context that is embedded in activities like driving or cooking. The agents in this situated environment, both human & AI, are stacked together for ease of representation. They independently sense the environment, and act on it. For convenience, we also include a summing junction, which takes as input the actions of the agents, and outputs their sum to the environment.

To this loop we add the Considerate Mediator (CM), which as explained earlier serves as an embodiment of the Considerate Systems framework. In essence, the notion of considerate social response, adds to the traditional sense-reason-act loop, and modifies it into a sense-reason-consider-act loop. Like the other agents, the CM senses and acts on the environment, but what differentiates it from the other agents is that

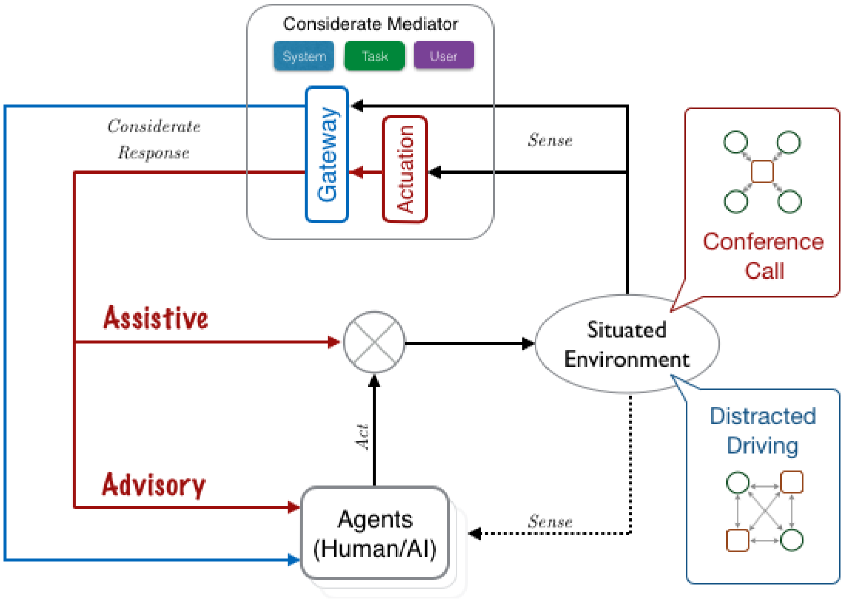


Figure 3.2: Architecture for Considerate Systems

it is allowed to a) communicate with the other agents, and b) control the summing junction. Conceptually, the reason that its allowed these privileges is because its the only agent equipped with the user, system and task models. We will explore this with an example below.

In the driving situation we saw that how the CM parametrizes the action (audio or visual) is as important as the timing of the action (based on the cognitive load of the driver). For this reason, we separate out the timing aspect from action parametrization and represent the CM as consisting of two modules: actuation and gateway. The actuation module is concerned with the action selection problem, and the parametrization of the action. When the action is to communicate with the other agents, its called an *advisory* action, because the agent can't control the other agents; it can only advise them (for example when it shows help information on a screen). When it acts on the environment, or effects it by controlling the summing junction, its called *assistive*. The gateway module is concerned with the timing of the action, and in gating any information in the environment from reaching an agent (for example, delaying transmission of video feeds during live broadcasts). This is why a dotted line has been drawn between the situated environment and the agents.

The purpose of an architecture here is to serve as a tool in the design of considerate response. It allows us to visually think through a problem and consider alternative strategies for mediating a problem. Let us do a quick example to highlight this. Imagine a brain-storming session in a conference call setting, where the participants are represented as agents. Such sessions are more productive when everyone is contributing equally. When one participant becomes dominant, the lack of visual feedback from other participants can prevent the dominant participant from realizing the consequences of their behavior. Adding a CM to this environment, applies the analytic process to break the problem down, and inspect it at the different levels of the user, system, and task models.

Once the CM's considerate response has been determined, the architecture provides the agent with a few options to achieve the desired outcome: 1) the CM could *advise* the participant to share the floor, 2) it could use an *assistive* action to simply subtract out the dominant participant via the summing junction (i.e. reduce the volume or mute them), or 3) the CM could for example use the gateway module to slightly delay the conversation to the dominant participant, so that they feel less empowered to take the floor. Based on the constraints of the problem, the designer or the system can pick the right recourse. The point of this example is to demonstrate that the architecture can be used to shed light on alternative mediation strategies that we might not have considered otherwise.

In the following section, we describe how we plan to evaluate each module.

3.6 Exemplar Applications

Once computers understand speech, vision, emotion and social constructs, it will revolutionize their role in society. The grand technological promise is that they will become indispensable agents in peoples daily lives, adding convenience while making people more effective at work. They will enable individuals and groups to tackle some of the great behavioral challenges facing humankind including health, sustainable living, and economic progress. We have argued that consequence-aware intelligence is needed to fully realize this potential. We motivated the need for considerate systems by articulating the problem, and recognizing that there are parts of it that we can tackle now, given the tools at our disposal. This relies on endowing systems with considerate commonsense behaviors, which can be learned or designed into the system *a priori*. Towards this end, this chapter introduced considerate systems, a conceptual framework that includes a design process & guidelines (user, system, task models), and an architecture.

Our thesis statement is that when systems display considerate behavior, they are more effective. Any evaluation of this thesis statement requires grounding the framework in specific situations that can be technically realized today. As discussed earlier, the application of the considerate systems framework is very context and situation specific. It follows that what it means to be considerate and effective is also situation specific. The rest of the thesis is split into two parts, each of which investigates a specific scenario that is prevalent today, and has been well studied in the literature. These include the conference call, and distracted driving scenarios. We apply the considerate systems framework in each of these scenarios, develop systems that embody considerate behavior, and demonstrate their effectiveness through controlled experiments. Indeed the demonstration of considerate interfaces improving interactions could be done in many domains. A convincing demonstration requires us to show not just the system's successful pursuit of scenario goals, but more deeply that the response succeeds in a natural and social setting with minimal breakdown. The most serious demonstration would come if automated considerate response itself, abstracted from any domain of discourse could be made to help users; thus, improving not just human-computer interaction but human-human communication as well.

The audio conference call provides an ideal starting scenario in which to interject an artificial agent, and evaluate its considerateness. Being a constrained medium, any inconsiderateness on the part of the agent will be magnified, negatively affecting the participants on the channel. This undesirable effect would be reflected in quantifiable performance metrics that can be analyzed. For the second scenario, we chose to broaden the application of considerate systems to a multi-modal scenario, where the consequences of being inconsiderate can be critical. In the driving scenario, a difference of a few milliseconds in reaction times can have far-reaching effects. With regards to the architecture, the first scenario (conference calls) affords itself to the evaluation of the actuation module (marked in red; see Figure 3.2), where the different aspects (parameters) of the actions are explored. On the other hand, because of its time-critical nature, the second scenario (distracted driving) focuses specifically on the timing of the actions, which is represented by the gateway module (marked in blue).

3.6.1 Conference Calls

As the work force becomes more global, we are increasingly embracing technologies that allow us to collaborate in distributed settings. Because of its prevalence, such distributed meetings have been well studied, and a number of problems have being identified. Various solution exist to address these issues but they mostly depend on visual feedback. This means that the participants have to visually focus on a small part of a screen to become aware of, and interpret, this feedback. This becomes inconvenient especially when the visual focus is primarily being used for computer tasks such as editing spread sheets or documents. This is further aggravated by limited screen estate and window clutter.

Given all of these constraints, the most obvious solution then would be to provide audio feedback instead of visual feedback. The user doesn't have to shift their focus from the task at hand which creates breakdown. Audio feedback is however more challenging to get right. It is a constrained medium that is already being shared among the participants in the call. Despite these challenges, we apply the framework towards the design of a mediating agent that tackles five problems prevalent in distributed meetings, and demonstrates measurable success (section 4.5). These include:

- Dominant and dormant participants
- Noisy background environments
- Inability to differentiate between similar sounding participants
- Uncertainty about the presence of participants who have not contributed recently
- Distracting notifications when participants enter or leave the call

We further explore the application of differential response to build a system that can adapt to various types of users (section 4.6.2). We employ reinforcement learning techniques, and simulate multiple users based on their:

- Responsiveness to feedforward and feedback
- Ability to self-moderate without any agent action
- Irritability from consecutive agent actions
- Short-term change in a single user's behavior

3.6.2 Distracted Driving

Another scenario that has become very prevalent are distractions that happen while driving. These include interactions with other people that are mediated by technology (e.g. text messages), and interactions with the technology itself (e.g. notifications). As technology gets better the implications of distracted driving can be dire for road safety. Research has shown that co-located passengers are better than remote partners at modulating their communications, resulting in safer driving performance. Technology needs to be able to mimic passenger behavior in gauging the load of the driver at a fine-grained level, in order to determine if its safe to interrupt or engage with them.

The first study tries to determine if notifications are indeed distracting (section 5.3). We are also interested in exploring the effect that the modality of the notifications has on performance. This takes into account appropriateness at the perceptual level of the user. We compare audio notifications with visual ones, displayed through a heads-up display. While it can be quicker to read a sentence than listen to it being read, sharp visual detail, i.e. foveal vision, is of primary importance for both driving and reading tasks. This can lead to conflicting demands on the visual perceptual resource, which can effect performance. We investigate whether this detriment in performance can be thwarted if this conflict happens only when the user is under a low cognitive load. For comparative purposes, we perform the same study for audio notifications.

Next, we train our sights on autonomously mediating notifications. This takes into account appropriateness at the cognitive level of the user. One approach to determine this is based on psychophysiological metrics. The advantage of this approach is its potential to be applicable in other domains, apart from driving, like handling heavy machinery or performing surgery, without extensive instrumentation of the environment. Generally, it would be of valuable importance for any system that interacts with people in a situated and real-time manner, e.g. spoken dialog systems. For instance, while listening, dialog systems need to have a certain level of comfort with disfluencies when the speaker is momentarily cognitive loaded with another task. It should not prompt the driver to repeat themselves, when they have stopped mid-sentence because the driving task has suddenly occupied their attention. Similarly, while speaking it should be able to match its pace and timing to the varying attention capacity of the driver. In this part of the thesis, we focus on developing a fine-grained measure of cognitive load (section 5.4), and demonstrate its utility in a multitasking driving scenario (section 5.5).

Chapter 4

Scenario 1: Conference Calls

This chapter presents the situated environment of audio conference calls as the first setting in which the Considerate Mediator is grounded and evaluated. We chose increasing the load on an already constrained natural communication channel to highlight the considerate aspect of the system response. Such situations can easily be made worse without careful consideration of the consequences. It is possible that use of visual feedback could work, but to get a baseline of people doing the best they can and improving it, we chose the simplest change to the simplest remote natural collaboration medium: audio.

Conference calls happen over a very constrained audio medium which presents an opportunity to more easily delineate the effects of any inconsiderate behavior (breakdown) on the part of an artificial agent. Conversely, it presents a challenging environment in which an agent might interject itself to achieve its goals. This work explores how technology can aid audio communication by better accommodating these social signals, and by creating new ones. The focus is on group communications in a distributed audio-only conference call where the participants are collaborators in a problem-solving/decision-making meeting.

To ground and evaluate an agent in this setting we focus on five commonly occurring problems on conference calls. We make use of the design guidelines (3.4), and the advisory & assistive actions of the actuation module (3.5) to formulate responses, and demonstrate their effectiveness in addressing these problems.

4.1 Introduction

Communication is a type of social action [69]. It can be verbal and non-verbal in nature. From a suggestive glance to an admonishing tone, people rely on all sorts of cues to assess the situation and regulate their behavior. Particularly while collaborating, people orient themselves and coordinate in creating a shared reality. They engage in this process to seek understanding, and to be understood. Feedback is pivotal to this process, and it propels and directs further communications. It helps in creating a shared awareness and mutual understanding.

When the communication is mediated by technology there is a reduction in these social cues or feedback. This creates a sense of disengagement and psychological distance. It is interesting to note that both video and audio-only conferencing suffer from the attenuation of these cues, albeit in different measures [142]. Hence, we find that the popularity of their use lies on a spectrum depending on the situation, the participants, the nature of the task, and the social setting. For instance, audio conference calls are widely used in business meetings [67], whereas desktop videoconferences are more popular in personal settings [22]. Even so, the reasons for users choice and preference are nuanced and complex, involving multiple tradeoffs related to intrusion, amplification of inattention, and mobility.

Globalization and technology have brought radical changes to business practices, and meetings in particular. With increasing frequency, teams are being composed of members from geographically different locations so that they can bring to bear their expertise on pressing problems, without the travel and associated costs. These distributed teams collaborate by holding meetings on conference calls and other networking solutions. By the very nature of the distributed setting, a host of technical, organizational and social challenges are introduced into these meetings that have been well documented and studied [164, 70]. A number of these challenges are associated with the missing or attenuated channels of non-verbal communication, which affects basic interaction constructs, such as, turn-taking, speaker selection, interruptions, overlaps and backchannels [66]. While there has been significant work and progress to preserve social cues in video communications [119], audio-mediated technologies have not received the same share of attention, and depend largely on visual cues like participant lists to buttress communications [164].

4.1.1 Why Audio?

Audio conferencing is a popular tool and ranks only behind telephone, fax and email in terms of most often used collaboration technologies [121]. In this work, we explore how social cues can be restored on the audio channel, while addressing some of its most often cited drawbacks [148, 142, 164]. These are predominantly social in nature, and focus on the process of interaction. These include dominant or dormant participants, extraneous noise, the ability to accurately identify speakers, the notion of personal space for remote participants, and the issues of awareness about the presence of other collaborators. We design different types of audio cues, and experiment with advisory and assistive actions to better understand how these might support human communication. Our goal is to build a considerate agent that would know when and how to apply these techniques appropriately. Audio interfaces have a low bandwidth for natural synchronous communications between multiple people. The difficulty in improving these communications with an agent further loading this constrained channel should then be maximally hard, making it an ideal place to demonstrate the utility of an agent being considerate and appropriate [141].

The reason we chose to provide feedback in audio comes from considering the potential for breakdown at the perceptual and cognitive levels of the user model. Many discussions and meetings involve documents and physical artifacts that occupy users visual attention. This makes display space expensive, and switching between display views task-intensive. Besides, the visual channel is not the best medium to convey awareness information because human visual field is limited to the frontal hemisphere, and the foveal region in particular. This creates inherent limitations in the use of visual displays, wherein the user must see and attend to the display. Noticing visual changes also gets harder as tasks get more demanding, or if the display is cluttered.

Using audio minimizes the potential for breakdown since people can perceive multiple audio channels simultaneously, and do so with considerable ease. In particular, we have the perceptual ability to hone in on a particular channel while filtering out the rest, commonly referred to as the "cocktail party effect" [5]. Thus, the audio channel can be used as an effective mode of transmitting background information (e.g., [54, 29]). In addition, audio can be used for conveying temporal information like whether an activity is occurring right now, and when actions start and stop. It can also be used to indicate spatial and structural information, like where the actions are happening, the type of activity and the qualities of the action (e.g., [134, 28, 64]).

4.1.2 Advisory & Assistive Actions

In order to explore the idea of pursuing scenario goals by proactively injecting considerate responses on an audio channel, we built CAMEO (Considerate Audio MEeting Oracle), a multiparty conference call facilitator. In the following sections, we survey related work and discuss the design goals and features of CAMEO. We then describe the architecture and implementation details of a testbed system that facilitates audio conference calls between two or more people. We demonstrate how this considerate mediator uses advisory actions to resolve conversational dominance, and disruptive extraneous noise [129]. We then show that through the use of assistive audio cues, users were more assured, and less distracted about the presence of other collaborators. In particular, we focus on resolving issues related to speaker identification, participant presence, and entry & exit announcements [130].

4.2 Related Work

4.2.1 Feedback in Collaborative Settings

In a collaborative setting, teams with constructive interaction styles achieve better performance (e.g., solution quality, solution acceptance, cohesion) than teams with passive/defensive interaction styles [126]. Team interaction styles are a reflection of the aggregate communication traits of the individual members. Higher variations in extroversion between team members lead to less constructive and more passive/defensive interaction styles within teams [11]. Shared leadership is a critical factor that can improve team performance [26]. This leads to the question: is it possible to influence the group dynamics by encouraging extroverted people to share the floor when their dominance is pronounced, without disrupting meeting flow.

Erickson and Kellogg formulated the concept of social translucence to improve online group interactions [46]. They advocate that the three properties a socially translucent system must possess are visibility, awareness, and accountability. Visibility and awareness brings about a collective awareness creating an environment where individuals feel accountable and responsible for resolving problems.

These ideas were employed by Yankelovich et al., in the design of their Meeting Central system to address the problems with audio conferencing which were documented in a series of studies [164]. They grouped the problems into three categories: *audio, behavior* and *technical*. In their assessment, the top problems that affected meeting effectiveness were surface communication properties that included too much extraneous noise (*audio*), and difficulty in identifying who was speaking (*technical*). Among the other reported problems, participants had difficulty knowing who had joined or left the meeting (*technical*), and speakers not being close enough to their microphones (*behavior*). More interestingly, the authors note that *"most audio problems are, in fact, behavioral. They are compounded by the difficulty remote participants have, both technically and socially, in interrupting to indicate that the problem exist."*

The idea of using feedback to influence group dynamics and behavior in distributed meetings was further explored by Kim et al., where they focused primarily on the effects of dominance [95]. Their Meeting Mediator system computes group interactivity and speaker participation levels, and uses a visualization to feed this information back to the participants on their mobile phones. Dominant people had a negative effect on brainstorming as fewer ideas were generated during these sessions. They also found that dominant people caused more speech overlaps in distributed meetings. Since spoken communications fill the channel and are so dynamic, the question arises as to how facilitation can be achieved at the turn-taking level to manage these interrupts or overlaps. A solution is to use virtual meeting facilitators in the form of embodied agents.

The interfaces discussed so far depend on GUIs and visualizations adding separate and new information on a separate communication channel. Graphical interfaces to computers were developed over the last few decades as a high bandwidth parallel communication channel. A computer interface can change the look of any part of a screen at any place in a fraction of a second, and the eye can notice millions of stimuli simultaneously. A keyboard allows directly coded symbols to control the computer, and a mouse or touch screen allow a person to react to concrete interface items directly. An audio interface has none of the afore-stated advantages. All information is layered on a low-bandwidth interface and without a keypad there is no coded input. On top of this, if a computer introduces audio into a conference call it is competing with the participants. Introducing an agent into an audio environment implies that they must successfully cohabitate this impoverished interface with the participants. Could a considerate system actually add and not detract from the audio in a conference calling system?

Given its limitations, we would expect audio feedback to distract or block other communication. However, work done by Rienks et al. reveal that participants found voice and visual feedback can be equally efficient [133]. While voice messages were more intrusive than text messages, participants of the meeting appeared to be much more aware of their own behavior when the system provided vocal feedback. This might be because the feedback is overlaid on the primary communication channel as opposed to a peripheral device. They also reported that as the users got used to the audio interface they found it less disruptive.

4.2.2 Auditory Interface Design

There is a rich body of work on the use of audio for user interfaces, which provides the foundation for our work of supporting social cues in conversation. We briefly review how audio interface design has evolved, and the sounds and techniques others have used to provide audio feedback and guide user interactions. We then cover how audio has been used in distributed settings to allow people to coordinate better, and to increase shared awareness of remote events and activities.

Audio interfaces largely use two types of non-speech cues, namely, earcons and auditory icons. Earcons are synthetic tones whose timbre, pitch, and intensity are manipulated, to build up a family of sounds whose attributes reflect the structure of a hierarchy of information. Since earcons are abstract, they require training and need to be learned to be effective. Auditory icons are a more focused class of audio cues, which are carefully designed to support a semantic link with the object they represent, which might make them easier to associate. Furthermore, sounds can be perceptually mapped to the events they indicate using symbolic, metaphorical and iconic methods.

Soundtrack [42] was one of the first auditory interfaces to use earcons and synthetic speech. More recently, Rigas et al. [134] demonstrated the use of earcons to communicate information about the layout of a building. Four different timbres (piano, organ, horn, and clarinet) were used to communicate the sections of the building. Floors were communicated by musical notes rising in pitch. A single note was rhythmically repeated to indicate room number, and combination of timbres was used to indicate hallways. Users successfully located the rooms but were not able to interpret the different hallways, suggesting that combination of two timbres created confusion. Early in our work, we experienced how an overloaded audio dimension could easily be created by assigning multiple tracks of an orchestra to each participant. We focus on methods that prevent such overloading in a single audio dimension.

SonicFinder [53] was the first interface to incorporate the use of auditory icons. A variety of actions made sounds in the SonicFinder, including the manipulation of files, folders, and windows. SonicFinder also made use of dynamic parameterized sounds to indicate temporal and structural activity, like file transfers producing a continuous filling up sounds, and different files producing different pitched sounds. In our work, we seek to explore when we can use the intuitive semantic mappings of auditory icons, over the arbitrary symbolic mapping of earcons.

A number of other works show how audio interfaces can improve interactions. gpsTunes [145] focussed on using adaptive audio feedback to guide a user to their desired location. As the user gets closer to the target, the music gets louder followed by a pulsing track to indicate their arrival. Schlienger et al. [140] evaluated the effects of animation and auditory icons on awareness. They showed that the auditory icons were commonly used to notify a change, and to focus attention on the right object just before it changed. AudioFeeds [39] explored how audio can be used to monitor social network activity, and PULSE [112] evaluated how audio cues can be used to communicate the local social vibes as a user walks around. This paper shows that such indicators might work well and not interfere with a conference call.

Activity Coordination in a Distributed Setting

SoundShark [55] was an auditory interface extension of SharedARK, a multiprocess system that allowed people to manipulate objects and collaborate virtually. It used auditory icons to indicate user interactions and ongoing processes, to help with navigation, and to provide information about other users. Users could hear each other even if they couldn't see each other, and this seemed to aid in coordination. This work motivated the development of ARKola, a simulation of a soft-drink bottling factory [56]. Temporally complex sounds occupied different parts of the audible frequency spectrum, and the sounds were designed to be semantically related to the events they represented. Also, instead of playing sounds continuously, a repetitive stream of sounds were used to allow other sounds to be heard between repetitions.

Gaver et al. observed that the sounds allowed the people to keep track of many ongoing processes, and facilitated collaboration between partners. Users were able to concentrate on their own tasks while coordinating with their partners about theirs, when sound was providing the background information. We seek to employ similar techniques to show how we might improve the process of audio communication itself.

The CSCW community has also paid attention to the use of audio in distributed workspaces. In a shared drawing environment, Ramloll and Mariani [107] played different sounds for different participants, and spatialized the sound in the 2D environment to help with location awareness. Participants complained that the spatial audio was distracting, but it provided them with information about others intentions which helped them with turn-taking. McGookin and Brewster [111] looked into audio and haptic locating tools as well, while extending their single user GraphBuilder to a multiuser interface. They found that shared audio helped in mediating communication, and served as shared reference points, allowing users to refer to events they couldn't see. Our work seeks to extend this to situations where the fact of a person's presence is crucial to the outcome.

Shared Awareness in a Distributed Setting

The Environmental Audio Reminders (EAR) system [54] transmits short auditory cues to people's office to inform them of a variety of events around their building . For example, the sounds of opening and closing doors are used to indicate that someone else has connected or disconnected from a user's video feed. They use stereotypical and unobtrusive sounds to make people aware of events in the workspace without interrupting normal workspace activities. We follow this approach attempting to discriminate in the more delicate domain of presence. ShareMon [28] used auditory icons to notify users about background file sharing events. To indicate the various actions involved, Cohen experimented with three types of sound mappings. For example, to indicate user login he used knock-knock-knock (iconic), "Kirk to enterprise" (metaphoric), and Ding-Dong (symbolic). To some degree, all three methods were intuitive and effective at communicating information, and users found them less disruptive than other modalities, like graphics and text-to-speech. In our work, we try to understand how these mappings affect users when they are used to interject ongoing communication.

For the first several months of mixing in audio to a conversation, our experiments made the conversation more challenging. The OutToLunch system [29] attempted to recreate an atmosphere of "group awareness". It gave isolated or dispersed group members the feeling that their coworkers were nearby, and also a sense of how busy they were, by taking advantage of the human ability to process background information using sound. Each user had a theme that was mixed in with a seamless loop of solo guitar

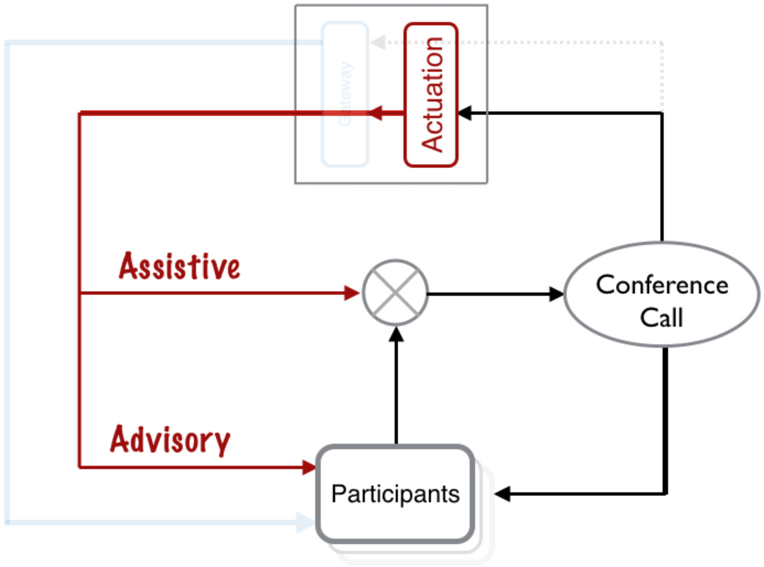


Figure 4.1: The Considerate Systems architecture grounded in the conference call scenario.

music, and would only play when the user was typing on their keyboard. With only six people in the group, the paper reports that users had no trouble associating a theme with the person it represented. We attempted to use a similar approach, but when convolved with conversation, the multitrack instrument sounds overload the channel and can be annoying. Similarly, there has been work in groupware systems to address the issue of awareness through audio. GroupDesign, a real-time multi-user drawing tool, used audio echo to represent user action on another users' interface [15]. In Thunderwire [71], an audio-only shared media space, the audible click of a microphone being switched on or off served to let participants know when people were joining or leaving the discussion. In Chalk Sounds [64], Gutwin et al. used the granular synthesis method to create chalk sounds that were parameterized by the speed and pressure of an input stylus. Our goal is to extend such awareness without the need for users to break the flow of conversation by having to ask about who has joined or left the conference call, for instance.

4.3 Considerate Audio MEdiation Oracle (CAMEO)

The term "considerate" refers to the unstated norms and mechanisms of social behavior that people engage in while communicating with each other [141]. It is analogous to a control system used in any dynamical system — to obtain a desired result, the output needs to be interpreted and the inputs need to be regulated accordingly. Similarly, CAMEO interprets the meeting (sensing) and tries to regulate it through its advisory and assistive actions (Figure 4.1). As a participant, CAMEO becomes a social actor and should strive to respond in socially appropriate ways.

4.3.1 CAMEO's Response

Advisory & Assistive Actions

In order to successfully orient itself with the other participants and positively influence the meeting, we distinguish between CAMEO's *assistive* and *advisory* behavior (Figure 4.1). Advisory actions advises or instructs the user, in the hope of getting them to monitor themselves and change their behavior. For instance, people might not realize how loud they sound to others on the phone. But if CAMEO could hint at this before someone on the line does so, it might save the embarrassment and any disfluencies in communication. Here the influence is indirect. Assistive, on the other hand, is when CAMEO directly influences the meeting through its control of the summing junction. For example, when two participants are speaking over each other, CAMEO can delay or pitch shift one so that the others can make sense of what is being said.

Adaptive Interaction

We also explore how CAMEO might adapt its actions based on the differential approach in order to minimize breakdown. While hand-crafted interaction policies can be used, it is not possible to design a policy for every situation that might arise. Also, not all users will respond to feedback the same way. To circumvent these issues, we model the agent's interaction with the user as a Markov decision process and investigate if an agent can adapt its behavior to different users using reinforcement learning techniques.

4.3.2 CAMEO's Features

Here we describe how CAMEO uses its advisory and assistive actions to resolve five commonly reported problems in teleconferencing [164]. Central to the design of these actions is the notion of considerate response that trades off pursuing system goals with minimizing breakdown.

Dominance Detection

A Dominator is a type of self-oriented behavioral role that participants can occasionally slip into during a meeting [74]. Groups dominated by individuals performing these roles are likely to be ineffective [162]. On the other hand, Dominators also drive discussions and can generate consensus [154]. Thus, it seems that while too much dominance might stifle contributions from the other participants, too little can reduce consensus and decision making because of a lack of a clear social order [58]. This identifies the potential for breakdown at the social level of the user model, where simply telling them or showing them that they are being dominant [154, 37] might not be as effective as encouraging them to use their status more positively at the appropriate times.

In our system, once CAMEO has detected that someone is dominating the meeting, it uses the advisory approach to subtly say "turn taking?" on the dominant person's channel alone. If it is close to the end of the meeting, it privately mentions how much time is left to encourage them to leave space for other participants to contribute. With this considerate response, the goal is to make the user reflect on their behavior and self-correct without publicly interrupting or demeaning them (breakdown). Similarly if someone is being dormant, CAMEO will say "any thoughts?" to encourage their participation. In both cases the advisory action has its purpose embedded within it, and aspires to be suggestive and encouraging.

Extraneous Noise Detection

Sometimes a participant might be in a noisy cafe, or in a moving vehicle. While they might be able to tune out the extraneous noise sources, it can be a larger distraction for the participants on the other end of the line. The nature of the meeting or the roles of the different participants can sometimes make it inconvenient for the others to point this disturbance out. More crucially, it is hard for the participants to identify exactly whose channel is responsible for introducing the noise. To preempt this CAMEO tries to detect high levels of extraneous noise uses non-speech audio buffers. It provides advisory feedback to the offending participant by stating "noisy". It does this in private in order to not disrupt the communication already taking place, thus minimizing breakdown by being appropriate at the social level. As with the dominance feature, the reason that the feedback is so terse is to address the need to economize the agent's footprint on a very constrained medium, thus minimizing breakdown at the perceptual level of the user model.

We chose to use assistive actions for the rest of the features, as these are problems endemic to the audio conference call technology. We used Apple's GarageBand[1] software to work with two pre-recorded ten-minute teleconference calls between five people. These served as the base tracks, which was then overlaid with earcons and auditory icons. This included instrument sounds[2], and nature sounds[3]. In the following paragraphs, we summarize the lessons learnt from the broad explorations in the designs of each of these three features.

[1] http://www.apple.com/ilife/garageband/

[2] http://free-loops.com/

[3] http://www.naturesoundsfor.me/

Speaker Identification

Conference calls suffer from issues that sometimes involves participants joining and leaving a conversation, or unwittingly speaking over each other. Most of these issues are caused by missing in-person cues, which can be disorienting on a multiparty conference call. We tried to obviate this by adding unique background sounds for each participant in order to create a soundscape around which participants can orient themselves. For instance, a distant orchestra could play a unique instrument that corresponds to the participant that has the floor, and ceases playing the instrument when that person cedes the floor. While these background tracks might be distracting (breakdown), the underlying objective is to augment the voice signatures of unfamiliar participants and make it easier for the group members to identify each other.

One of the major reasons people might find video conferencing attractive is because it elevates identifying and distinguishing between speakers to a separate channel with less crosstalk [142, 164]. The question of identity and presence can get even more muddled when people on the line are not familiar with each other, or their accents. So we experimented with different audio cues to support speaker identification in such difficult situations where lack of familiarity and accents could reduce understanding .

We focused on designing **earcons** that were easy to perceive, remember and discriminate. We initially experimented with assigning background *instrumental tracks* to each participant. For example, the bass track might be assigned to participant one, and the rhythm guitars to participant two. When the participants spoke, the track assigned to them would start to play in the background. This, however, proved to be distracting as the instrumental tracks introduced crosstalk on the channel, which caused breakdown at the perceptual level of the user model.

To reduce crosstalk, we experimented with *simple tones* instead of tracks, that pulse while the participant speaks. We were able to achieve good discriminability by using the following *timbres*: tambourine, bongos, and vibraphone for the 2nd, 3rd and 4th participants, respectively. For the 1st and 5th participants we used muted electric bass with tones that were an octave apart.

Once these qualitative design evaluations were done, the next part of the audio design was to determine the temporal nature of the cues, i.e. when they should play and for how long. The cues were designed to play at the beginning of every utterance a participant makes while holding the floor. We found that this worked best when the cues began playing a second into the utterance, as opposed to right at the beginning of the utterance. This duration was long enough to ensure that a participant was contributing more than just back-channel feedback, like uh-huh. In this way, breakdown at the turn-taking level of system communication is minimized.

An interesting side effect from designing the audio cues in this manner was that they also seemed to emphasize a participants hold on the floor, reinforcing personal audio space. These can be thought of as analogous to people's use of hand gestures while speaking. Thus, when a speaker is speaking loudly and at a rapid pace, the cues pulse rapidly too. If the speaker is speaking softly and at a slower rate, the cues pulse at a slower rate.

The timbre from Garageband are high quality, and occupy a large portion of the soundscape when mixed-in with the conference call. To push the auditory cues to the background we experimented with a number of filters and reverb effects. We found that using a high-pass filter, and the "small room" reverb effect worked most effectively in reducing the footprint of the soundscape, thus minimizing the potential for breakdown.

Another technique to aid with the identifying speakers is to **spatialize** them in a 2D environment. We used stereo panning to place the 5 speakers at the -32, -16, 0, +16, +32 positions on the Garageband's Pan Dial. Like in experiments done by Ramloll [107], using stronger panning was actually distracting, and made the listener want to cock their head to the side where the sound was coming from, like when someone taps your shoulder.

Next, we describe the process of designing audio cues for the fourth CAMEO feature, i.e. to indicate presence of other collaborators on the line to a user.

Participant Presence

Feedback is very important for communication. Even the absence of feedback about whether the others on the line can hear you or not, can be distracting. For instance, without the addition of sidetone users have a tendency to attend to their displays to know if their call has been dropped or not. This can be disruptive to the conversation. Sidetone is a form of feedback that is picked up from the mouthpiece and instantly introduced into the earpiece of the same handset. It gives users the assurance that their signal is being registered by the phone system, and is therefore now incorporated into most phone devices. Similarly, the awareness that the other participants are on the line, and are listening is important confirmation which reduces uncertainty about the channel continuing to be functional. For example, if Ron is playing music in the background, or Joe is driving, it becomes immediately obvious when either of them go offline. We are subconsciously aware of their presence on the line even when they are not speaking. Similarly, to indicate presence and annotate participants, we have experimented with CAMEO using an assistive approach to add background tracks like music, tones, or ambient sound during a conversation. There are various methods to present such information visually [164], but the visual channel is a separate interface and might even increases the potential for breakdown at the perceptual and cognitive level of the user model due to its high bandwidth and attention cost as described above.

Initial ideas involved the use of background sounds to create a **soundscape** around which users could orient themselves. For example, if Ron is playing music in the background, or Joe is driving, it becomes immediately obvious when either of them go offline, even if they aren't talking. Their absence becomes conspicuous because of the sudden change in the soundscape. To test this idea in our system, we assigned different *instrumental tracks* to the participants, similar to the OutToLunch system [29]. The tracks would play while the user was online, but they overloaded the conversational channel and were, therefore, highly distracting.

An alternative idea was to employ a roll call, i.e. to periodically announce the presence of the participants either by name or **auditory icons**. Announcing the names of participants would be the easiest to understand, but people may find it annoying to hear their names being called out periodically. Instead, auditory icons were created by sampling *backchannel* like participants laugh, and other characteristic sounds. These were inserted in the channel at periodic intervals when a participant had been silent for a while. However, they were too subtle and weren't noticed. We then experimented with recorded *ambient environmental sounds* of someone typing on a keyboard, clicking a mouse, opening and closing a drawer, and thumbing through papers. These auditory icons were found to be distinct, perceptible, and natural in a work environment.

The last significant example of a technical communication breakdown described in Yanankovitch's work was multiple entry & exit announcements for CAMEO's Entry/Exit feature, which we consider below.

Entry & Exit

It can be hard to tell when participants get dropped from or reenter a conference call. Existing conference call systems can commit considerable time to loudly announcing entry and exit, which can be annoying. This introduces breakdown at multiple levels, but the most critical is at the perceptual level of the user model. We used CAMEO's assistive approach to explore the use of earcons [19], abbrevicons [72], and different intonations for entry and exit. The goal is to convey the most information in as small an audio footprint as possible.

To avail of **iconic** mapping, we attempted to use the sounds of a *door opening and closing* to indicate entry and exit. We initially tried to superimpose the name of the participant with the sounds of the door opening and closing, but it was hard to discriminate between the sound of the door opening and the door closing. We obviated this by appending the sounds to the name either at the beginning of the sound or the end. We usually hear the door open and then see the person enter, so an entrance is announced by the sound of a door opening followed by the name. On the other hand, when leaving we see a person head to the door and then hear the door close. So an exit is announced by the name followed by sound of the door closing.

To make the audio footprint even shorter, we wanted to compare this to more **metaphorical** mapping approaches, like using *fade* and *intonations*. We modified the TTS in Audacity[4] by using the amplitude fade-in and fade-out effects for entrance and exit, respectively. But this reduced the understandability of the name. For intonations, we chose to map entrance to a normal intonation, and exit to an upward intonation. This was partly due to convenience as Apple's TTS engine automatically intonates a word when it is punctuated with a question mark. For instance, "Armstrong?" is automatically intoned upwards and indicates that a participant named Armstrong has been disconnected from the conference call.

4.4 System Design & Implementation

CAMEO needs to prioritize its actions in order to appropriately communicate with the user. For this consider phase, conflict resolution is performed by a blackboard system to prioritize and schedule how CAMEO responds in a meeting. It consists of different knowledge sources (KSs) that update themselves to the Channelizer and Globalizer blackboards. The Channelizer represents a participant and hence is local in its scope (P1, P2, ... Pn in Figure 4.2). The issues of an individual participant are resolved here. The Globalizer represents the meeting. It interprets the different dynamics of the meeting and organizes CAMEO actions.

4.4.1 Channelizer

CAMEO creates a channelizer object for each participant, which consists of four KSs:

The first KS classifies microphone input audio as speech or non-speech. The audio is sampled at 16kHz each, with 64 frames generated per period and 2 periods in a buffer. These 8 ms buffers are pushed into a sliding window that is 30 buffers long. The window step size of one buffer, i.e. there is no overlap. The KS calculates the average energy level of each window and classifies it as speech and non-speech using a threshold value.

The second KS determines the 'Participant State'. A participant can be in one of four states:

- Not Talking: The participant is silent and does not have the floor

[4] http://audacity.sourceforge.net/

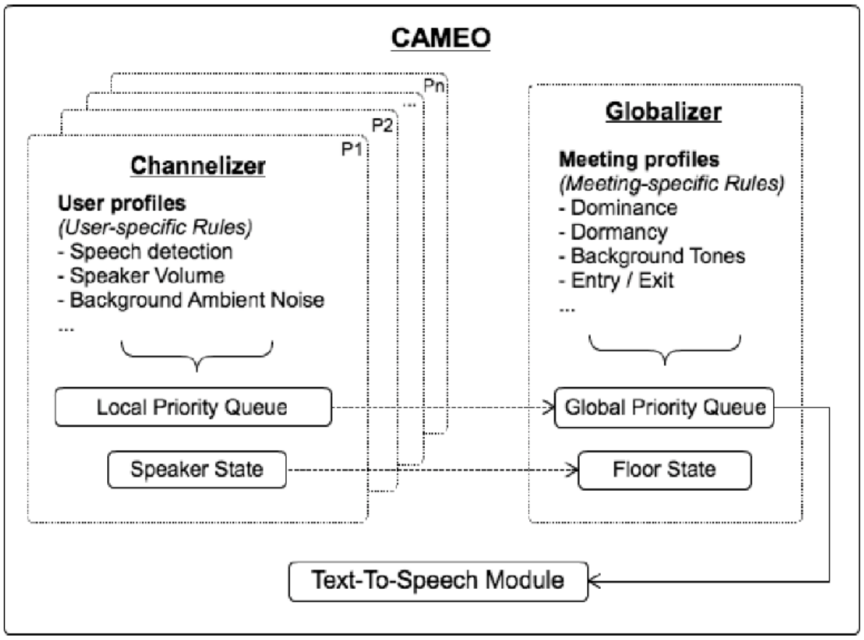


Figure 4.2: Flow Diagram Representation of CAMEO's Blackboard System.

- Start Talking: The participant begins to talk and wants the floor
- Still Talking: The participant is still talking, and if he does not have the floor, still wants it
- Stop Talking: The participant is no longer talking, and if he has the floor, relinquishes it

The third KS calculates signal-to-noise ratios which are used to detect extraneous noise, and to determine if a participant is speaking too loudly or softly. Speech buffers are used to determine signal values, while non-speech buffers are used to determine noise floor values. This KS establishes a noise floor, and uses empirically determined thresholds to set the flags for the Extraneous Noise Detector and Volume Meter features.

The last KS generates speech prompts for flags that have been set by other KSs, which are buffered in an internal priority queue. These are pushed out to the Globalizer's global priority queue (Figure 4.2) based on a reinforcement scheduler that knows how many times the participant has been prompted in the past and how long it has been since the last prompt.

Since each Channelizer object is participant specific, different user profiles (e.g. CEO, guest speaker, student, etc.) can be used that contain KSs with different parameters.

4.4.2 Globalizer

The Channelizer can be seen as a knowledge source for the Globalizer, which is where CAMEO behaviors are negotiated. The Globalizer currently includes six other knowledge sources.

The first KS determines each speaker's dominance by calculating how active each person is relative to the activity level of the other participants. The Globalizer calculates each participant's dominance as their contribution to the sum of all participants:

$$Dominance_{Px} = TSL_{Px} / \sum_{i=1}^{n} TSL_{Pi}$$

TSL is the Total Speaking Length of a particular participant. This dominance measure is useful in resolving conversational conflicts such as interruptions, as well as monitoring how effective CAMEO is in fostering collaboration across all participants (Figure 4.3).

The second KS manages the meeting floor, which the participants take turns in holding. It translates a 'Participant's State' into a 'Floor Action', to determine which participant is currently holding the floor. This establishes a model of floor control. The floor can only be taken by another participant when the floor owner releases it. The four floor actions are:

- No Floor: The participant is not speaking
- Take Floor: The participant starts to speak
- Hold Floor: The participant is still speaking
- Release Floor: The participant is done speaking

This mapping from Participant State to Floor Action gives a detailed picture of turn-taking in the meeting. It allows the Globalizer to detect and measure non-verbal cues easily in order to identify what actions CAMEO should take to socially enhance the conversation. For example, when a participant has the floor, their background track is activated to annotate their speech for the Speaker Identification and Presence feature.

The third KS aggregates several audio cues that are non-verbal, and have proven to be effective in distinguishing a speaker's social activity during a meeting [91]:

Total Speaking Length (TSL) The amount of time a person speaks over the course of the entire conversation.

Total Speaking Turns (TST) The number of distinct times a person speaks over the course of the entire conversation.

Total Speaking Turns without Short Utterances (TSTwSU) The number of distinct times a person speaks, not including any short utterances or affirmations made.

Total Successful Interruptions (TSI) The number of times a person has successfully interrupted another person. A high TSI is an indication of dominance.

Jayagopi showed that using a combination of these cues to classify conversational dominance yielded an 88% accuracy on a fairly typical meeting corpus [91], which is why we chose the above metrics in our evaluation process.

The fourth KS detects and resolves any conversational collisions, or interruptions. Since we have been experimenting with collaborative problem-solving meetings, CAMEO favors the more dormant participants in case of these events. This triggers the Dominance Detector feature, which then checks the dominance levels from the first KS.

The fifth KS detects entry and exits, which sets flags for the Entry and Exit feature.

Similar to the Channelizer, the last KS generates speech prompts for CAMEO features that have had their flags set by the other KSs. It then buffers them into the Globalizer's priority queue. This priority queue plays the role of the gateway module, and determines when CAMEO communicates the actions from the different knowledge sources, based on its considerate response goals. It combines messages that are repeated. It also delays messages based on their importance, the time since the last alert, and the number of alerts. It can choose to announce messages based on the Floor Action state. For example, for entry and exit events, it waits for the floor to be empty before making an announcement.

This mechanism allows for different meeting profiles (e.g. instructional, planning, etc.) to be used because of the adjustable priorities for the different prompts. For example, in a collaborative scenario, CAMEO will give higher preference to dormant participants, whereas if the system was setup to facilitate an instructional scenario, CAMEO will give higher preference to the instructor. This flexibility in reasoning that the architecture allows us will be useful in adding and testing more considerate features, and meeting scenarios.

4.4.3 Research Platform

The architectural implementation discussed in this section allows it to be easily extended to support further research on audio interactions. CAMEO is built on CLAM (C++ Library for Audio and Music), a framework for audio processing, along with the JACK Audio Connection Kit on Ubuntu Linux. CLAM was chosen for its robust processing feature set and modular API. JACK is used as the audio server for its real-time network streaming capabilities. Together they form a considerate systems test bed, which allows us to build and test features for CAMEO.

Currently CAMEO has a number of actuators. It controls who talks to whom, and can use this to manipulate floor control. It can change the amplitude and frequency of the input channels, and can mute or delay them. It can also overlay background sound or introduce reverb and other effects. It can do this for any combination of the participants. For example, it can introduce a feedback that only participants two and three can hear, or it can delay speech from participant one, to participant three.

4.5 Experiments

To begin evaluating our considerate agent, we tested the five features separately. Many of the initial explorations highlight how easy it is to degrade communication in the audio domain.

For resolving dominance, we initially started off with CAMEO prompting users with the message, "You have been talking for a while. Please give others a turn". While this seemed acceptable the first time, every time after that it became less and less tolerable. It was not just that in a meeting participants have low cognitive bandwidth for a third-party. It was also "nagging", as one of the participants put it. Changing the message to "turn-taking?" (a suggestion), allowed users to consider but almost not notice the agent.

As an example of identifying people with background sound annotation, we tried mapping an instrument track to each person. Besides the music being distracting, it was difficult to remember any mapping between individual and instrument. Generic background sounds of flowing river, an office setting, and traffic proved to be too distracting. Less distracting tones like a marimba at different frequencies (C4, E4, G4), one for each participant proved to be more subtle, and yet, distinctive enough to distinguish.

4.5.1 Study 1: Dominance & Dormancy

Our study attempts to see if CAMEO can demonstrate that a computer agent can decrease the difference between dominant and non-dominant people, i.e. to lower the variance in dominance as the meeting progresses. It also attempts to see if CAMEO can successfully encourage more interactivity, i.e. the turn-taking will be more balanced, and the speech utterances of all participants will be shorter on average.

Methods

The first set of experiments evaluated the Dominance/Dormancy feature of CAMEO. During turn-taking conflicts, the agent uses an advisory approach to remind the dominant participant to share the floor by saying “turn-taking?” on that user’s channel. Similarly if someone is being dormant, the agent will say “any thoughts?” to encourage their participation.

Participants, Procedures, and Task: We conducted a study with twelve groups of three participants each. The groups were formed by drawing from a pool of nineteen volunteers. The participants were students and research scientists (5 females & 14 males) belonging to the same campus, with the youngest being 21, and oldest 43. The collaborative behavior of people with different partners is dramatically different. Group interaction styles reflect aggregation of communication traits of its team members [11]. Even though a participant took part in multiple groups, the 12 groups had unique compositions drawing different dynamics from the 19 participants.

The participants were located in physically different locations with computer terminals that had screen sharing and control enabled. Each group went through two problem-solving sessions, one with CAMEO and one without, for a total of twenty-four sessions. The sessions were held back-to-back and were five-minutes long each. During each session, the participants collaborated on playing Hangman, a multi-step word revealing puzzle. The duration for solving the puzzles was chosen so as to simulate a slice of an actual meeting where everyone is an equal collaborator, and higher group extraversion would be beneficial to the groups performance. The protocol was that the three participants would agree on a letter before entering it; the last person to agree would input the letter into their terminal.

Study Design: We performed a within-subject experiment comparing how the groups behaved with and without CAMEO. On half of the groups we ran the control condition first (CAMEO Off), while on the other half we ran the test condition first (CAMEO On).

Measures: For the purposes of this experiment we measure a participant’s dominance level as a fraction of their Total Speaking Length (TSL) divided by the TSL of all participants which was also shown to be a reasonable measure of dominance in [91]. The dominance percentage threshold we use is 40%. The prompt is only activated when there is a turn-taking conflict after this threshold has been reached. A turn-taking conflict is a speech overlap between the dominant user and another user that is longer than one second. These values were heuristically determined to work well. To compare meetings, we calculate the variance in dominance between the participants, which measures how spread apart the are. The lower the variance, the more collaborative a meeting was.

For interactivity we calculate Turn Taking, which is the ratio of the TSTwSU (Total Speaking Turns without Short Utterances) of the most dominant person to the TSTwSU of the least dominant person. TSTwSU includes only utterances that were longer than simple feedback like “umm” or “yea”. Turn Taking gives us a measure of how well the floor is being shared between the participants. A Turn Taking value of one, implies that the floor was being shared equally.

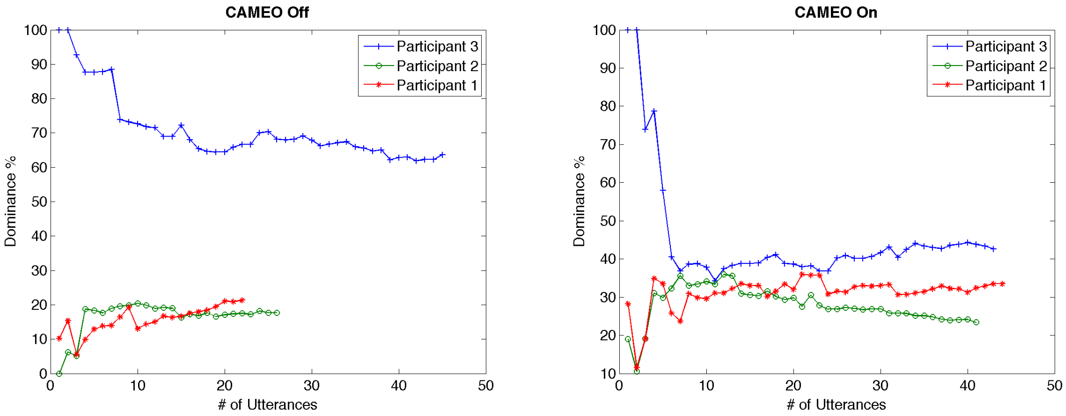


Figure 4.3: Dominance (%) vs. Number of Utterances during the Dominance Resolution evaluation for one of the test groups: With CAMEO On, the Speaker 3 becomes less dominant and Speakers 1 and 2 become less dormant. Also, they contribute more equally, i.e. the number of utterances from each is around the same, with CAMEO On.

	1 min.	End
CAMEO Off	23.07	13.31
CAMEO On	20.87	7.50

Table 4.1: Variance in dominance levels of all participants across all groups one minute into the meeting and at the end of the meeting.

Results

CAMEO had a strong effect on the dominance levels. As the meeting progressed, dominant participants began to give more room for the other participants to contribute. Dormant participants began to contribute more as well. To quantify these results, we calculated the variances of the dominance levels of all participants at the one minute mark and at the end of the meeting, across all groups with CAMEO On and CAMEO Off (Table 4.1). The table shows that the meetings start with similar variance in dominance between the participant. At the end of the meeting, there is a bigger and statistically significant drop in variance with the CAMEO On, than with the CAMEO Off. The standard deviation[5] in dominance among members of a group reduced with statistical significance (N=12, p<0.01, 1-tailed t-test, Figure 4.4).

CAMEO appeared to have a positive effect on interactivity. The most dominant person seemed to be taking the floor less when CAMEO was facilitating the meeting, but there was not enough experimental data to show statistical significance (one-tailed paired T-test: p = 0.05. There was no notable difference in the average speech utterance of the participants (one-tailed paired T-test: p = 0.36; Table 4.2).

[5] In a three-person meeting, the ideal contribution is 33.3%, which is also always the average. The standard deviation gives a measure of how close to ideal participant contributions are in each condition. A standard deviation of 0 implies that all the participants contributed equally.

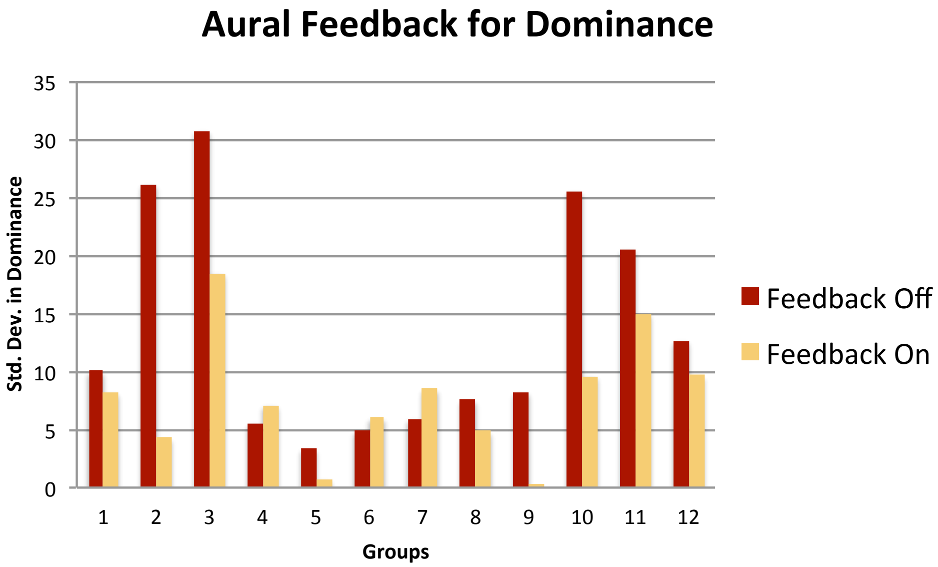


Figure 4.4: The results of the aural feedback experiment across twelve groups. They demonstrate a reduction in standard deviation of dominance among members of a group, when aural feedback was provided.

	Turn Taking	Avg. Speech Utterance
CAMEO Off	2.51 (0.08)	1.68 (0.53)
CAMEO On	1.84 (0.07)	1.63 (0.23)

Table 4.2: Variance in dominance levels of all participants across all groups one minute into the meeting and at the end of the meeting.

4.5.2 Study 2: Extraneous Noise

The second set of experiments attempts to show that CAMEO can reduce the impact of extraneous noise in a conference call by providing feedback about it. The hypothesis is that an agent can allow a meeting to run smoother with participants interrupting each other fewer times.

Methods

Participants, Procedures, Task: We conducted a study of three groups of three participants each. The groups were formed by drawing from a pool of seven volunteers. They were all male, with the youngest being 21, and oldest 28. The participants were located in physically different locations with computer terminals that had screen sharing and control enabled. Each group went through six problem-solving sessions, three with CAMEO and three without, alternatively, for a total of eighteen sessions. The sessions were held back-to-back and were four-minutes long each. During each session, the participants collaborated in discussions around solving chess-move puzzles presented on their screen, of the mate-in one/two/three variety. The game and its duration were chosen so as to simulate a slice of the meeting where the cognitive load on the participants is high, requiring concentration and memory. The protocol included that the three participants would agree on a move before executing it; the last person to agree would move the piece on their terminal. This created a dynamic where a participant would have to guide the other participants through multiple levels of reasoning, before being able to generate a consensus.

The participants were instructed to respond naturally as they would if the extraneous noise on a telephone line was too loud, and that if it was disrupting the meeting they should ask for it to be turned down. At different intervals in the game, a TV program would be played close to one of the terminals to introduce extraneous noise into the meeting. If the participant on that terminal was prompted to reduce the volume either by CAMEO or by one of the participants, they would do so by pressing a button on the provided remote control. After an interval of thirty seconds to a minute, the extraneous noise would be introduced again. CAMEO prompts on a reinforcement schedule, i.e. subsequent prompts would be further and further apart, unless a sufficient amount of time had lapsed since the last prompt.

Study Design: We performed a within-subject experiment comparing how the groups behaved with (experimental condition) and without CAMEO (control condition).

Measures: The experiment measures extraneous noise on the flow of the meeting, as Total Successful Interruptions (TSI) metric, i.e. the number of times a participant successfully interrupts another. An interrupt occurs when a participant is speaking and another participant talks over them.

Results

CAMEO was able to positively impact the flow of the meeting. With CAMEO On, the average number of times any participant interrupted another in a four minute session was reduced by almost half, and was shown to be statistically significant (one-tailed paired T-test: p = 0.007, Table 4.3).

	CAMEO On	CAMEO Off
TSI	7.05	12.44

Table 4.3: Average number of interrupts (TSI) among participants solving chess-puzzles in four-minute sessions.

In the next section, we formally evaluate a final set of audio designs and protocols that came out of the preliminary explorations. We conducted three studies that we discuss below for each of the three features.

4.5.3 Study 3: Speaker Identification

The goal of this study is to understand how the addition of audio cues to a conference call can help people differentiate between different speakers. To highlight difficulties in recognizing people in a conference call we chose five non-native english speakers (males, 22-28 years old), and had them remotely collaborate on a sub-arctic airplane crash scenario commonly used in team building exercises. According to the scenario, the five of them had just survived a plane crash in Northern Canada, and had managed to salvage some items. Their task was to list the items in order of importance, and to come to an agreement on this order as a group. The audio from each participant was recorded in separate files, which was then processed in Audacity. Garageband was used to create three separate versions: a simple downmixed version; one mixed with the speaker identification earcons discussed in the previous section; and another which arranged the speakers spatially in a 2D environment. We compare these three and see how well participants do on each. Our hypothesis was that the participants will do better with the spatialization and earcons aids, than without any audio cues.

Methods

The speaker identity study asked participants to listen to a segment of a pre-recorded conference call while answering questions related to the conversation at hand, and speaker contributions.

Participants, Procedures, and Task: Thirty-two people were recruited for the study (8 female, 24 male). Participants were between 20 and 30 years old, and all reported having no hearing impairments. Participants had the choice to take part in the experiments either remotely or in the lab

This study had two stages, a training stage and a test stage. During the training stage, the participants were first asked to listen to the recorded introductions from the five speakers on the recorded conference call. They were presented with five colored buttons with the number and name of the different speakers. Upon clicking the button, they would hear a recording of the corresponding speaker saying their name and a fun fact. The next page allowed the participants to practice speaker tagging. They could click on the practice button which would cause the program to randomly play a short segment of the recorded conference call. The participant was asked to identify the speaker in the short segment by clicking on the speaker's corresponding button. After every attempt both the right answer and the selected answer were displayed. The participant could choose to go to the next screen whenever they felt confident of successfully differentiating between the different speakers. In the test stage, participants listened to a two minute and thirty second clip of the pre-recorded conference call. As they were listening, questions would appear about the conversation that had to be answered within five seconds.

Apparatus and Sounds: Participants were provided with Logitech headsets. Remote participants were requested to find a quiet place and use headsets. They were provided with the address of the server where the experiment was hosted, and were asked to access it using the Chrome browser.

The **earcon** audio cues were obtained from Garageband. Speakers one to five were assigned muted electric base (low tone), tambourine, bongos, vibraphone, and muted electric base (high tone), respectively. Similarly, for creating a **spatial** 2D environment, speakers one to five were placed at the -32, -16, 0, +16, and +32 units on Garageband's pan knob (2D spatial positioning).

During the training stage when participants were introduced to the five speakers, their corresponding instrument or spatial location was also displayed, both visually and aurally.

Study Design: We used a between group study, where half the participants answered the questions with the aid of musical instrument earcons, while the other half used 2D spatialization. For each group we also included a within-subject condition to compare the test condition (with earcons) to a baseline (with no earcons). To balance out learning effects, half the participants started with the baseline, while the other half started with the test condition. Different two-and-a-half minute segments of the pre-recorded conference call were used in the within-subject study. The first segment had nine questions, while the second segment had eight questions.

To keep the test conditions same across study participants, and to isolate only the participant's perception of the audio cues, we used the same pre-recorded tracks in our evaluations. A limitation of this approach is that the audio cues are evaluated by third-party observers, and not by active participants of a meeting. We tried to account for this by asking questions that were of a "who said <something related to conversation>" nature, which is different from asking who just spoke. The aim was two-fold. First, to keep the participants engaged in the conversation, and to prevent them from simply matching audio cues to the speaker. Second, to cognitively load the user (as they might be while participating in a conference call) so that the distractive effects of audio cues might come to bear on the results.

Results

We present our results below in terms of participants being able to accurately identify the speakers on a conference call, and their response times. Attempt rate is the fraction of questions users answered in each condition. For accuracy, we report two metrics: Overall Accuracy, which includes questions that were not answered, and Attempt Accuracy, which only includes questions that were answered. Together, these metrics should account for distractions that audio cues might introduce causing participants to take longer than five seconds to answer a question.

	Accuracy (Overall)	Attempt Rate	Accuracy (Attempted)	Response Time (ms)
Spatial	0.573	0.841	0.691	2187.4
No cues	0.435	0.788	0.570	2663.0

Table 4.4: Accuracy metrics and average response times for speaker identification with and without 2D spatialization.

	Accuracy (Overall)	Attempt Rate	Accuracy (Attempted)	Response Time (ms)
Earcons	0.538	0.772	0.703	2257.8
No cues	0.386	0.819	0.475	2537.2

Table 4.5: Accuracy metrics and average response times for speaker identification with and without earcons.

Spatialization vs. No audio cues: Participants ability to identify speakers increased significantly, with greater than 20% improvement using spatial audio cues. They were able to do this almost half a second quicker on average when compared to the condition without audio cues ($p<0.05$, 1-tailed t-test, Table 4.4). Overall Accuracy: SEM=(0.024, 0.041); Attempted Accuracy: SEM=(0.046, 0.044); Response Time: SEM=(151.2, 163.7); N=16.

Earcons vs. No audio cues: With earcons, participants were also able to achieve an increase in accuracy of 30% on average over the condition with no audio cues ($p<0.05$). Participants also appeared quick to respond but the difference was not significant ($p<0.1$, 1-tailed t-test, Table 4.5). Overall Accuracy: SEM=(0.052, 0.024); Attempted Accuracy: SEM=(0.054, 0.054); Response Time: SEM=(169.0, 111.5); N=16.

We were able to show that speaker identification improved with the addition of either spatial cues, or earcons. A between group analysis did not reveal any difference between these two conditions. Furthermore, there was no significant difference in the number of questions that were attempted across the three conditions from which we might infer that the addition of audio cues was not notably distracting.

4.5.4 Study 4: Participant Presence

The goal of the audio presence study is to investigate whether the addition of audio cues to a conference call can help reassure people that the other participants are still on the line, and haven't been disconnected. A different segment of the pre-recorded conference call described above was used in this study. Garageband was used to create two separate versions: a simple downmixed version with no audio cues added; and one mixed with the auditory icons for audio presence discussed in the previous section.

Methods

Participants were asked to listen to a segment of a pre-recorded conference call while answering some questions related to the conversation. The participants were also asked to indicate if they thought a participant had been dropped from the call.

Participants, Procedures, and Task: Twenty people were recruited for the study (4 female, 16 male). Participants were between 20 and 30 years old, and all reported having no hearing impairments. Participants had the choice to take part in the experiments either remotely or in the lab

The participant was asked to listen to a five-minute clip of the pre-recorded conference call. As they were listening, questions would appear about the conversation that had to be answered within ten seconds. Participants were also instructed to periodically ensure that everyone was online. They could do so by pressing the "nudge" button which simulated feedback from each participant stating that they were still there (like a ping test).

Apparatus and Sounds: The apparatus used by the participants is identical to the first study. The cues in this study were recorded using an iPhone, and processed in Audacity. As motivated in our exploration experiments above, these include **auditory icons** of ambient environmental sounds like someone typing on a keyboard, clicking a mouse, opening and closing a desk drawer, and shuffling through papers. These cues were then added to the segment of the pre-recorded conference call used for this test.

Study Design: Our working hypothesis was that adding audio cues like keyboard sounds and mouse clicks acted to reinforce the presence of people who had not spoken in a while, but were still online. In other words, we wanted to show that like the sidetone, adding audio cues improves awareness about the presence of other collaborators.

We used a between group study, where half the participants were presented with audio cues (test condition), and the other half was not (baseline condition). Participants had to answer eight multiple-choice questions while listening to the conversation. This was to simulate a real meeting where participants would be paying attention to the conversation, and not actively tracking the presence of other collaborators. Participants were told that because of some collaborators being in weak signal areas, there was a high chance that they might accidentally drop off the call. They were asked to virtually "nudge" the other participants if they suspected that one of them was not present.

Results

We investigate our hypothesis by comparing how often users "nudge" others to check if they are present, with and without the auditory icons discussed above. We found that the number of nudges was reduced

Table 4.6: Average number of nudges, attempt rate, and error rate with and without auditory icons.

	# of Nudges	Attempt Rate	Error Rate
Auditory icons	3.50	0.90	0.33
No audio cues	5.63	0.81	0.35

by 37% in the condition where the auditory icons were used ($p<0.05$, 2-tailed t-test, Table 4.6). There was no significant difference in the attempt rate or error rate. # of Nudges: SEM=(0.72, 0.64); N=10.

4.5.5 Study 5: Entry & Exit

The goal of this study is to understand the effects that different conference call entry & exit announcements have on the participants, and meetings in general. We focus on three kinds of prompts, namely, speech, iconic and metaphoric. Our hypothesis is that the metaphoric prompts using different intonations will have the least impact on participants cognitive capability (i.e., their ability to follow game protocol in this particular study).

Methods

To bring out the effects, we designed and built a memory card game for four people that can be accessed remotely from the browser. Participants are paired off into two teams that take turns in choosing two cards from the sixteen that are shown face down on a GUI screen. If the two cards chosen by a team match, the team wins the turn. The team that matches the most number of pairs, wins the game.

Participants, Procedures, Task: We recruited 21 participants for this study (4 female, 17 male). Participants were between 20 and 30 years old, and collaborated remotely on the game. Six unique groups of four participants each were tested (some participants repeated).

When the participants join the meeting, the administrator would introduce them to the game, and the protocol they were to follow. During a team's turn, both team members are required to select a card. The selected card is revealed only to its selector. Thus, the first team member to click open a card has to communicate its content and position, based on which their partner picks the second card. The protocol specifically requires the team partners to alternate who gets to pick the first card at every turn. The protocol was designed in this way to encourage discussion.

After a practice round, the administrator would notify the participants that the experiments were going to begin. They were told that during the experiments, participants would randomly be dropped from the meeting. If they happened to be dropped from the conference call, they were requested to rejoin as soon

Table 4.7: Entry & Exit prompts using different mappings in each test condition.

	Entry & Exit Prompts
Speech	*<participant_name> has joined the conference* *<participant_name> has left the conference*
Iconic	*sound of door opening + <participant_name>* *<participant_name> + sound of door closing*
Metaphoric	*<participant_name>* (said with normal intonation) *<participant_name>* (said with raising intonation)

as possible. During the course of such an event, a prompt would play to notify the rest of the participants that someone had left the conference call, while another prompt would play to indicate that they had joined back.

Apparatus and Sounds: The apparatus used was identical to the first two studies. Mumble[6] was used to host the conference call. All the participants were requested to download the Mumble client and follow the instructions that were provided.

A mac mini was used to run the python script that generated the prompts. The three entry and exit prompts that were used are speech-based, iconic, and metaphoric (Table 4.7). The prompts are dynamically created using Apple's text-to-speech engine, and pre-recorded audio of a door opening and closing. The Python script was also set up to use the Mumble server's Ice remote procedure call interface to arbitrarily disconnect people every thirty seconds.

Study Design: We used a within-subject study where each group played three rounds of the memory game, one for each of the three conditions. To balance out any learning effects, different sequences of the conditions were used for each group (Table 4.7).

Results

We wanted to investigate the effect that the different prompts would have on the participants ability to observe protocol, i.e. team members switching turns to pick the first card. We only take into account turns where both participants are online. We found that the metaphoric prompts had the lowest error rate at 15% in participants ability to maintain protocol compared to both the iconic and speech prompts ($p<0.05$, 2-tailed t-test, Table 4.8). The iconic prompts affected the participants as badly as the speech prompts did with error rates larger than 25%. Error Rate: SEM=(0.05, 0.05, 0.05); N=6.

[6] http://mumble.sourceforge.net/

Table 4.8: Average error rates in following the protocol and game duration across the three conditions.

	Error Rate	Duration (sec)
Speech	0.29	262.3
Iconic	0.26	258.6
Metaphoric	0.15	222.0

We also wanted to understand how the different prompts affected the game. We hypothesized that the shorter prompts would create less disruptions allowing the participants to finish the game quicker. There wasn't a significant difference in the durations, but the participants do appear to finish the games faster in the condition with the metaphoric prompts. The average durations are shown in Table 4.8. Duration: SEM=(34.4, 22.3, 16.3); N=6.

Participant Preferences: During the pilot experiments, participants strongly preferred the speech prompts to the metaphorical ones, which they found to be ambiguous. They were largely ambivalent about the iconic prompts. To help disambiguate the prompts in general, we began playing each of them at the start of their respective test conditions. This practice saw an increase in the number of participants who preferred the metaphoric prompts as they found it to be less distracting. They remained neutral with regards to the iconic prompts, although some of them claimed that it was hard to distinguish between the door opening and closing sounds when the line was noisy. This might explain poor participant performance under the iconic condition.

4.6 Adaptive Feedback

Hand-crafted feedback policies can be designed based on psychological insight as we saw in sections 4.3 & 4.5. These can be brittle — different users might react differently, and even an individual user's response might change over time, and is dependent on the situation. We therefore use the adaptive method from the differential family of techniques to minimize breakdown by adapting the agent's action to the user and situation. A similar problem for cognitive orthotics was addressed using reinforcement learning techniques [135].

While the techniques discussed in this section could be used for any of the social problems described previously, in this thesis we focus on applying these technique towards addressing conversational dominance. Once the agent recognizes the existence of a social problem it attempts to provide feedback based on its interaction policy. The feedback can be parametrized in a number of ways, including its timing, frequency, tone, volume, translucence [46], etc.

4.6.1 Learning Algorithm

We use reinforcement learning to improve social feedback policies. The agent will consider the meeting state based on duration of the meeting, detected social problems, timing and nature of previous feedback, and user's mood. It will also consider feedback actions available to the agent including what type of feedback to give, if any. An agent's action yields some reward $r \in R(s,a)$ and leads to a new state $s' \in S$. In some cases, the desired state in meetings (e.g. non-dominant participants) might occur as a result of several interactions. Such interaction with delayed rewards are well modeled as Markov Decision Processes (MDP).

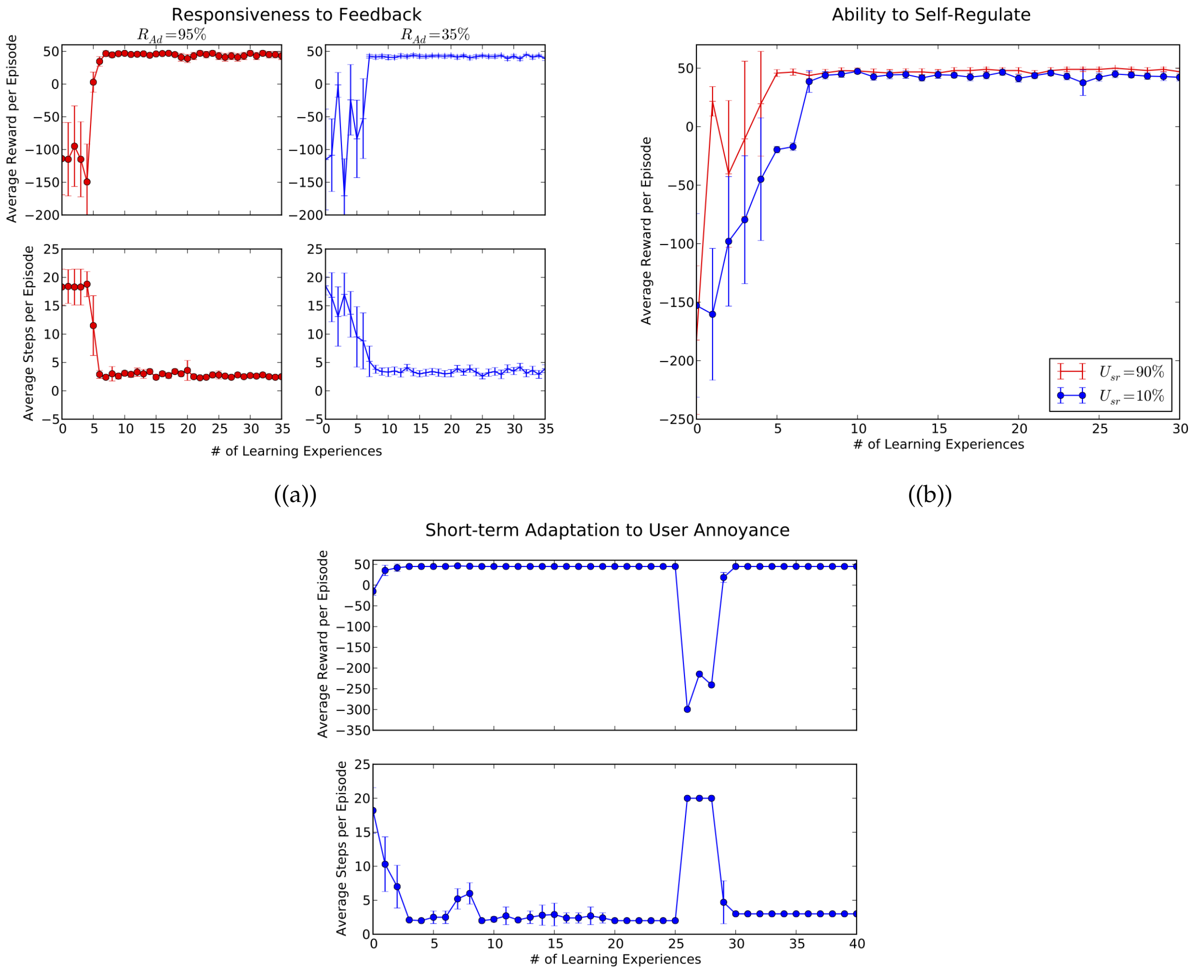
Solving a Markov process, however, requires knowledge of the possible state transition probabilities (interaction model), which is not known in advance. One way to approach the problem is to use a model-free class of algorithms known as *temporal difference* methods. In particular, we use the Q-learning algorithm [156] which is typically easier to implement, where we define $Q^*(s,a)$ as the expected discounted reinforcement for taking action a in state s, then continuing by choosing actions optimally. The Q-learning rule is:

$$Q(s,a) := Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)), \tag{4.1}$$

where α is the learning rate, and γ is the discount factor. $< s,a,r,s' >$ is an experience tuple, as described above. If each action is executed in each state an infinite number of times on an infinite run and α is decayed appropriately, the Q values will converge with probability 1 to Q^* [156]. The optimal policy then becomes $\pi^*(s) = \arg\max_a Q^*(s,a)$.

Payoff Function

We focus on **three binary state features**, which are (i) is the participant dominant?, (ii) have they received feedback?, (iii) are they annoyed?. The agent has a choice of **three actions** to get a dominant user to reduce their dominant behavior: *No Action, Advisory, Assistive*. The agent might provide aural *advisory* feedback to the user that they are being dominant. Alternatively the agent might take an *assistive* action: reducing the volume of a dominant person, or muting them when they interrupt a less dominant participant. For meeting flow, it is preferred that the agent chooses (a) no action unless necessary, and (b) advisory over assistive actions. We therefore give the assistive and advisory actions a cost of -5 and -1, respectively. If the user gets annoyed with consecutive feedback actions, the agent incurs a cost of -10. If the agent is able to get a dominant user to change their behavior, without annoying them, it gets a **reward of +50**.



((a))

((b))

((c))

Figure 4.5: The results of adapting the agent's feedback policy by modeling a user based on their (a) responsiveness to feedback, and (b) ability to self-regulate their behavior. (c) shows the results of the agent's short-term adaptation at the 25th learning experience to knowledge that a user gets annoyed with consecutive feedback.. Each figure shows results with (95% confidence interval) error bars averaged over 10 test episodes, for every learning experience. In all cases, the agent learns an optimal policy in under 10 experiences.

4.6.2 Evaluation: Simulation Study with an Adaptive Policy

/

The Q-learning algorithm was validated for adapting an agent's feedback policy for different users by conducting a set of experiments with a simulated user and environment. In the experiment, an episode starts in any state where the user is dominant. The episode ends when the user is in the goal state, i.e. they are not dominant or annoyed. Thus, there is a trade-off when providing feedback between getting the user to be non-dominant and making sure not to annoy the user. The simulated episodes demonstrate feasibility of the approach; experiments with real users would be a valuable next step for validating and possibly improving the feedback policy.

User Model: A model of potential users focused on their responses to the agent: how did they respond to advisory actions (R_{Ad}), how likely they are to get annoyed (U_{an}), and how well are they able to self-regulate their behavior without feedback (U_{sr}). We would expect that the optimal policy is to do No Action when the user is not dominant or when they are annoyed. This was the case in all the optimal policies that were learnt. Thus, we are left with two states, i.e., (i) $S_{D\bar{F}}$: user is dominant and has not gotten any feedback, and (ii) S_{DF}: user is dominant and has received feedback, where the agent learns different policies.

Results

For the following experiments, an agent is trained using a series of *learning experiences*. A learning experience can consist of one or a batch of episodes. During the learning experience the agent uses an ϵ-greedy explorer to choose an action. An ϵ-greedy explorer chooses a random action with ϵ probability, and choose an action using the learnt policy with 1-ϵ probability. After each learning experience, the new policy is tested over 10 test episodes. During these, the agent always chooses an action based on the learnt policy. The rewards the agent receives over the 10 test episodes is averaged and plotted in Figure 4.5.

Responsiveness to Feedback: Two types of users were simulated with different responsiveness to advisory feedback, i.e. the probability with which a dominant user will become non-dominant when they get advisory feedback: $R_{Ad} = \{95\%, 35\%\}$. The user always responds to assistive feedback with a probability of 95%. If the user has been provided feedback, the likelihood of them responding to advisory feedback drops by 10% in the next attempt. The optimal policy that was learnt was to provide advisory actions in $S_{D\bar{F}}$ and S_{DF} when $R_{Ad} = 95\%$, i.e. the agent learns that the user is likely to respond to advisory feedback. When $R_{Ad} = 35\%$, the agent chose assistive actions in both states, because it learns that the user is unlikely to respond to advisory feedback, and that it has to pursue the less desirable (more costly) assistive actions. Figure 4.5(a) plots the rewards and the number of actions the agent took as it learnt an optimal policy for $R_{Ad} = 95\%$ & $R_{Ad} = 35\%$. These results are averaged over 10 test episodes after every learning experience.

Ability to Self-Regulate: Next, we model a user who is dominant only for short periods of time. In this case, we include a likelihood that the user becomes non-dominant when the agent takes no action (U_{sr}) to the existing ($R_{Ad} = 35\%$ + U_{an}) user model. The agent was trained for two cases: $U_{sr} = \{10\%, 90\%\}$. When $U_{sr} = 10\%$, the agent learns the same policy as the $R_{Ad} = 35\%$ model, since the user does not self-regulate and needs to receive assistive feedback to become non-dominant. When $U_{sr} = 90\%$, the agent chooses to do no action in every state because it learns that the user is likely to self-regulate, and does not need feedback. Figure 4.5(b) plots the rewards earned as the agent learns an optimal policy for $U_{sr} = 90\%$ and $U_{sr} = 10\%$. The higher rewards for $U_{sr} = 90\%$ are indicative of the agent choosing no action (no cost), while the lower rewards for $U_{sr} = 10\%$ indicate the agent choosing assistive actions (-5 cost).

Short-term Adaptation to User Annoyance: In this experiment, we also test the agents short-term adaptation to new information once it has already learnt an optimal policy. We add to an existing user model ($R_{Ad} = 35\%$) the likelihood of them getting annoyed with consecutive feedback (U_{an}). In state S_{DF}, the agent should learn to take no action instead of providing assistive feedback. Figure 4.5(c) plots the results as the agent learns an optimal policy for $R_{Ad} = 35\%$. After the 25th learning experience, the user begins to get annoyed with consecutive feedback (U_{an}). The plot shows how the agent is punished for following the optimal policy when this happens, and how it adapts after 3 to 4 learning experiences to learn a new optimal policy for $\{R_{Ad} = 35\%, U_{an}\}$.

4.7 Discussion

This chapter shows many ways a domain independent assement of a confrerence call can be improved through considerate feedback. We presented five commonly occurring problems in conference calls, and described the design and implementation of CAMEO (an instance of the Considerate Mediator) to address these problems. We examined how the formulation of CAMEO's responses were informed by the design process and guidelines, and the architecture. In particular, we evaluated the advisory and assistive actions of the actuation modules and showed how the agent was successful in pursuing its goals while minimizing breakdown.

The results from the first two experiments for Dominance & Dormancy, and Extraneous Noise demonstrate how CAMEO's advisory approach can positively influence participant behavior in collaborative meetings. Being an audio-only solution, it presents a departure from the more complex methods of using separate modalities, such as, peripheral visual interfaces, to improve conversations. Although we do not present a comparison between the two, we see the advantages of using audio as two-fold based on the perceptual, cognitive and social levels of the user model.

At the perceptual level, since audio is orthogonal to the visual information, the users don't have to shift focus between multiple windows, displays or peripheral devices. They can remain focused on their task. It is also presents an advantage if the user is engaged in an activity like driving, where visual focus is critical.

At the cognitive level, audio feedback is just-in-time and unambiguous in its intent. In the experiments, the participants seemed so engaged in their task that they were not actively conscious of how their behavior was impacting the meeting. The use of an always-on visualization might make them conscious of their behavior, serving as an added cognitive. Instead, by prompting them at just the right time we can reduce this load; but we hold that the use of this more direct approach needs to be socially oriented and considerate to be successful. The CAMEO feedback tested was subtle enough that most users claimed not to pay attention to the prompts while engaged in their collaborative tasks. It is here the system makes its strongest statement, as the results show that the system's actions still did have a significant impact on their behavior. This is an important and novel finding that points to the utility of pro-active feedback approaches in the audio domain that gently nudges users towards desired behaviors.

At the social level, CAMEO's advisory feedback is private. Communications between a participant and CAMEO occurs on their audio channel unbeknownst to others, creating a different dynamic compared to when the information is broadcast to all participants [95, 38]. Participants did not appear to be intentionally dominant, or negligent of extraneous noise. In such cases, a private suggestion is possibly more considerate than a socially-translucent public notification. The system did not negatively affect performance between the test and control case; the number of words or moves attempted, the number of wins and loses, and the number of wrong guesses was not distinguishable between the test and control conditions.

Evaluating the Extraneous Noise feature brought out the most interesting observations. When extraneous noise was introduced into a conference call, a participant would invariably ask for it to be reduced. However, they would be more tolerant to subsequent re-introductions of the extraneous noise, pointing it out only when the noise was increased to a volume much higher than before. Participants sensitivity to extraneous noise was also related to how active the discussion was at the time. The more engaged the participants were in the discussion, the less the extraneous noise distracted their conversation; they would simply start to talk louder to counter it. But as they got louder, they also became more aggressive in trying to get their point across. It has been shown that loud noise elevates a person's state of arousal [92]. What is remarkable is that most participants stated that they were not affected by the noise, when the data clearly shows that they were. CAMEO's prompts about the extraneous noise precluded the noise from effecting the meeting. Without CAMEO, the number of times people interrupted each other was almost double.

The results from the next three experiments show that a constricted audio communication channel can be augmented with assistive social feedback cues, even in highly dynamic environments. In specific, we empirically showed that these cues allowed users to identify speakers more accurately, increased awareness about the presence of other collaborators, and improved participant performance. The successes of automatic considerate feedback with constraints of improving human human communication in the already overloaded auidio channel worked.

To demonstrate the utility of the audio cues, we simulated particularly difficult and stressing situations. It is hard for the average person to distinguish between five people of the same gender with similar accents. Keeping track of multiple things while coordinating with others is difficult when there are a lot of distractions in the environment. For instance, the first time we are introduced to a team that we are collaborating with, is when our understanding of their speech is most important; but it is also when their accents and behaviors are most difficult to interpret.

Similarly, in an increasingly mobile and global workforce, a user might be in a noisy environment and have trouble distinguishing between some of the other collaborators on the conference call. In this case, the user could choose to add cues to some of the other collaborators, which would play only on their own channel. The sounds might also act as aids to users who might choose to associate meta-information (like the person's location or function) with an audio cue. Likewise, when to use speech, iconic or metaphoric prompts to announce events might be dependent on the situation. Developing an understanding of how these cues affect participants, and their applicability in different situations, allows us to build a vocabulary of actuators that a considerate agent like CAMEO would know when and how to use.

Lastly, we showed how we might implement a differential approach to minimize breakdown using reinforcement learning techniques. We simulated different types of users and showed how an agent might adapt its actions to these users, and also how it might adapt to changing preferences of the same user over time.

The successes of automatic considerate response with constraints of improving human-human communication in the already overloaded audio channel worked. Less constrained communication channels could be less challenging; we expect that automated considerate response can be helpful throughout user interface design.

4.7.1 Future Work

The creation of a considerate architecuture was successful at helping communication. Still we started with the simplest models we could for the experiments. In the future, more complex models could be used to more accurately predict when a group member is negatively affecting a meeting, including cues like interrupts and overlaps, meeting type, etc. However, detecting things like the intention of interruptions and overlaps are not straightforward [152]. The participants could be co-constructing a thought as often happens, or could be confirming each other through repetition. Interruptions and overlaps in these cases are not attempts to take the floor, although they might be construed as acts of dominance. Instead, looking for patterns of repeated *acts* of dominance might allow CAMEO to infer a *state* of dominance and respond more appropriately. A state of dominance could potentially be modeled statistically using a Markov process or n-gram model. Future work should compare our simulated users to a corpus from real users for improving the adaptive feedback policy.

For Speaker Identification & Presence, we used the tones of a chord progression. We initially had them sound at the end of an utterance. In a two-person setting, its effect was to subtly moderate turn-taking. Participants paused a little longer for the sound to play. The tone provided some sort of affirming feedback to the speaker that they had been heard. This was not the case when more than two people were on the line. We then moved the tones to the beginning of each speakers turn. Over a prolonged period this creates a musical pattern that reflects the group dynamic, and will be studied as a subliminal audio feedback mechanism.

With every added feature, the timing of each prompt becomes more critical. To experiment with the timing we implemented separate priority queues for each participant, and a global queue for the whole meeting. This allows CAMEO to delay messages to a participant, or prioritize them. However, quantifying and hard-coding ‘too close’ or ‘high priority’ can make the system fragile. Humans are highly sensitive to the situational and temporal context, as in the case where the participants tolerance to extraneous noise increased. To be successful in socially orienting themselves, CAMEO and other proactive agents too will need to display a better awareness of the situational context [68]. For these reasons, in the next chapter we shift our focus towards the timing of a considerate agent’s actions based on situational context, as embodied in the gateway module.

Chapter 5

Scenario 2: Distracted Driving

In this chapter, we address the Distracted Driving scenario. We chose this domain because it allows for a broader multi-modal application in which to ground and evaluate the Considerate Mediator. As another extremely overloaded interface, driving a car can be a mentally consuming activity which has resulted in a large number of fateful accidents, especially when drivers inappropriately multitask with the phone and other infotainment devices. Could a considerate system help people reduce their mistakes when engaged in such dual-task scenarios? For our purposes, we can think of the these technologies that the driver interacts with as agents that need to be mediated by the Considerate Mediator, based on the driver's workload (situational context). To demonstrate considerate behavior then, the timing of the Considerate Mediator's actions becomes critical. Architecturally, this aspect of the Considerate Mediator is embodied in the gateway module. We illustrate how the design process & guidelines grounds and informs our implementation of the gateway module in this specific scenario. We then evaluate its effectiveness by analyzing task performance.

5.1 Introduction

The proliferation of ubiquitous computing platforms like the smartphone is fundamentally changing how we interact with each other, and how we consume information. Messaging apps are making our communications more asynchronous, concise and directed. They are giving rise to new social etiquettes, shaping our expectations of how and when we send messages, and respond to them. Many applications use these same smartphone communication platforms to interact with their users. Polling each application separately to see if there is something that requires action, however, has become cruelly inefficient. Consequently, these applications have come to rely on notifications to request user interaction.

The effects of notifications have been studied in the office-desktop environment when engaged in primary tasks like editing or programming. However, little has been done to understand the nature and effect of mobile notifications in everyday life. Even as we begin to study them, notification strategies are evolving; wearable devices, like smartwatches and head-mounted displays, are being developed that aim to focus our attention on notifications while touting seamless integration. Despite their interruptive nature, notifications might be our only way to keep abreast of time critical requests for our attention. Thus, their effects in mobile situations needs to be understood if we are to develop effective strategies to manage people's attention without subjecting them to undue risk.

The asynchronous nature of notifications afford the user the ability to decide when to take action on a secondary or new activity. While in some cases immediate action is taken by the user, in other cases notifications must be ignored depending on the user's current context. The task of *attending* to these notifications, and making a decision on whether to take an action or not, is currently left up to the user. Prior work typically does not distinguish between noticing, attending and responding to a notification. In this work, we want to understand the cost that is associated with the attending to a notification, and the role played by the modality, i.e. audio or visual. The primary task could be any immersive task involving complex sensorimotor skills, like cooking or even surgery. We present our experiments with ConTRe (Continues Tracking and Reaction) [105], which requires continuous tracking and episodic reactions in a driving-like task.

Automotive cockpits have been gaining a lot of attention because of the real impact that driver distraction can have on road traffic safety. A number of studies have shown how operating mobile devices and other in-vehicle infotainment systems is critically impacting driving performance and is a major factor in automotive accidents. Results of testing these effects in simulator studies has been shown to be replicable in field studies [45]. A broad literature has demonstrated that interacting with telephones and similar secondary activities in the car can adversely affect the primary driving task. This chapter explores how even simply attending to tasks that do not require a response might impact performance. Such tasks might be as simple as attending to notifications, which is the subject of our study.

Distractions from secondary tasks stem from a combination of three sources: manual, visual, and cognitive [147]. Here we are primarily interested in the cognitive sources of distraction, which occur when attention is withdrawn from a complex primary task. Many activities like driving have periods of low and high information processing. Since disabling notifications during such activities is not a blanket solution that we can expect users to realistically employ, notification mediation could be used to present notifications to the user when they aren't overloaded with the primary activity. One of the goals of this work is to study the impact of mediation on performance in both the primary ConTRe task and the secondary notification comprehension task, across both the audio and visual modalities. We report on the results from this study in section 5.3. The other goal of this work is to actually build a system that can autonomously mediate notifications based on task load.

5.1.1 Autonomous Mediation

People have finite mental resources and can only process a limited amount of information without degradation of task performance. Despite this being the case, there is an increasing trend towards computers being proactive and providing information to the user without being prompted. For cognitively challenging activities like driving, divided attention can have dire consequences. Thus, in order to be minimize breakdown in this scenario, we have to be appropriate at the cognitive level of the user model (3.1). There is a need for systems to gauge the load on this mental resource in order to predict or preempt degradation in task performance, while interacting with a user.

While progress has been made towards gauging this load, we are still a long way off from being able to measure it at a real-time fine-grained level. In the future, such capabilities might avert human mistakes in situations of divided attention. For instance, voice interaction might become the most efficient way for a user to interact with a system, when their manual and visual resources are already occupied. By using a rapid and fine-grained cognitive load measure, dialog or proactive agents would be able to track the ebbs and flows of the load being experienced by the user in real-time. This would allow it to preempt disfluencies and other irregularities in speech, as well as to time its responses and other actions, so as to prevent overloading the user. In the driving scenario, it has been shown that passengers adapt their conversation to the driving situation, which leaves the driver with more resources to perform the driving task when it gets difficult [40, 27]. Interactive agents should aim to emulate such considerate behaviors.

Cognitive load can be gauged by directly modeling the driver via psychophysiological measures, or by modeling driving context and its effect on the driver, or by jointly modeling both [94]. Compared to modeling the external driving context, less progress has been made in modeling the driver's internal state in order to identify when to interrupt them. One advantage of the internal state approach is the potential for these models to generalize to other domains. Modeling external context requires specific sensors and techniques to be considered for each domain separately. Furthermore, a physiological-based approach can be tuned for each user individually, as different users might experience external contexts differently. Recent advances in wearable technologies suggest that monitoring at least a few physiological signals in everyday life might become a feasible option.

In section 5.4, we evaluate several signals that might be used as part of a psychophysiological approach to gauging cognitive load. These signals were recorded from users while they participated in the first study (5.3). We found the most success with the pupil dilation measures, which were used to build classification models that can detect which tasks the user is engaged in. We present analysis of how the performance of the model varies with changes in the modality and timing of notifications. We do this for each user, as well as across all users. In section 5.5, we evaluate the feasibility of using such a model built on pupil dilation measures to mediate notifications in real time. We demonstrate its effectiveness by comparing user task performance with and without mediation. In the following sections, we provide background and discuss related work, before describing the two studies and their results in detail.

5.2 Related Work

We start off by describing work that has studied the interruptive nature of notifications, and how compelling they can be. This is followed by work that investigates the effects of multitasking on task performance, and the associated risks. We primarily focus on describing work where driving was the main activity. We then present strategies that others have prescribed to mitigate these risks, and help users better handle task switching, particularly through the use of mediation to manage attention allocation. Finally, we review work that might allow the system to autonomously mediate notifications to the user.

5.2.1 The Interruptive Nature of Notifications

Iqbal and Bailey [86] define a *notification* as a visual cue, auditory signal, or haptic alert generated by an application or service that relays information to a user outside their current focus of attention. A majority of the research on notifications is focused on information workers in a desktop computing environment. Its detrimental effects on primary task performance and efficiency have been highlighted through numerous studies [35, 100]. This effect was shown to be more pronounced when the primary task is cognitively demanding [34]. Interestingly, a study found that while users are aware of the disruptive effects of notifications, they appreciated the awareness provided by them [90].

Notifications play an even more central role in the mobile domain and are becoming the focal point for user interaction. In one study, experience-sampling was used to show that receptivity to an interruption is influenced by content rather than by its time of delivery [47]. Another large-scale study found that users attributed the most value to notifications from communication applications [137]. Regardless of application category, they reported a 50% probability that a user would click on a notification within 30 seconds, which could be indicative of their general capacity to be disruptive. Another study reported a daily average of 63.5 notifications per user, where muting the phone did not increase response times in viewing notifications [125].

Social Pressures of Messaging

Such asynchronous communication channels have become essential for young people in particular [13, 144]. When messaging was restricted as part of a user study, the participants not only showed increased anxiety, but many also did not comply [144]. An increasing number of apps are growing to rely on notifications to draw user attention to new messages or content. Focus groups have uncovered the unspoken rule of immediacy of response, and the pressure felt by people to carry their personal mobile devices at all times [79]. Notifications have become pivotal in propping up such social constructs. Prior work validates our assertion that notifications are too compelling to be ignored, especially in communication settings. To better frame our study on the potential risks of attending to these notifications, we now describe work that has examined the negative effects associated with multitasking.

5.2.2 Effects of Multitasking

Repeated task switching during an activity may lead to completion of the primary task with lower accuracy and longer duration, in addition to increased anxiety and perceived difficulty of the task [9]. Multiple studies have shown how cell phone conversations, texting, or interacting with In-Vehicle Information Systems (IVIS) can be detrimental to driving safety [146, 76, 138, 88, 23]. Drivers engaging in such activities have been shown to have increased brake reaction time [1, 99], failure in scanning for potential hazards in the driving environment [149], and to have accidents with higher likelihood [132].

The distribution of cognitive resources when engaged in such multitasking scenarios is not very well understood. This makes it difficult to assess and predict workload that will be experienced by the user. Theories have been proposed to model how multiple tasks might compete for the same information processing resources [161, 8]. One widely used approach that has been shown to fit data from multitask studies is Wickens' multiple resource theory. This attempts to characterize the potential interference between multiple tasks in terms of dimensions of stages (perceptual and cognitive vs. spatial), sensory modalities (visual vs. auditory), codes (visual vs. spatial), and visual channels (focal vs. ambient) [161]. Performance will deteriorate when demand for one or more tasks along a particular dimension exceeds capacity.

Cognitive Sources of Distraction

Among the three sources of distraction, cognitive is the most difficult to asses. Changes in driving performance associated with cognitive distraction have been shown to be qualitatively different from those associated with visual distraction [3, 45]. As an example, visual distraction has been shown to increase the variability of lane position, whereas cognitive distraction has been shown to decrease the variability of lane position [33]. Similarly, distractions attributed to one source can actually be caused by another. For instance, mandating that cell phone conversations be hands-free while driving has not been found to improve safety compared to hand-held phone conversations [132, 146]. Studies suggest that it is not the motor action of holding the phone, but the cognitive demands of multitasking that degrade task performance [120]. In particular, it was shown that conversations that include information recall challenges have the most detrimental effects on driving [89].

Visual Sources of Distraction

Complex sensorimotor tasks like driving heavily requires visual and spatial working memory to scan the environment, track objects of interests, and judge relative distances to them. A number of studies have shown that operating a device while driving competes for these same resources, to the detriment of the primary task performance [77]. Heads-up displays (HUD) have the potential to share visual resources between the primary and secondary tasks, as information is displayed on transparent surfaces in line with the environment. These have been investigated in the aviation and automotive domains, and while these have shown to benefit users in terms of vehicle control and detection of roadway events, these benefits do not hold under the high workload of unexpected events [75]. Devices like Google Glass (Glass) have spurred on a renewed interest in the impact these devices have on primary tasks. Studies on driving and texting via Glass show that while the device served to moderate the distraction as compared to using a smartphone interface, it did not eliminate it [139, 151].

Driving and Language

Listening and responding to another person while driving a car has been widely studied, and has been shown to effect driving performance, particularly with remote conversants [98]. Passengers sitting next to a driver are able to adapt their conversation to the traffic situation, allowing the driver to focus on driving when it becomes difficult [40, 27]. These findings have motivated research towards building dialog systems that are situation-aware and interrupt themselves when required [97].

In the case of driving and notification comprehension, both tasks compete for resources along the stages dimension. We would expect performance to deteriorate when there is an increased demand for the shared perceptual resources, i.e. when driving is hard and/or when the notification is difficult to comprehend. If the notification is visual, both tasks might also compete along the modality and visual channel dimensions. We would expect performance deterioration to be greater for visual notifications.

5.2.3 Mediating Interruptions & Notifications

Successful dual-task scenarios depend on the availability and requirements of cognitive resources for the secondary task given resource consumption by the primary task [160]. This presents opportunities to increase people's ability to successfully handle interruptions, and prevent expensive errors. McFarlane's seminal work proposed four methods for coordinating interruptions [110], including immediate, negotiated, mediated and scheduled. Mediation has been widely studied in the desktop computing domain [82, 86], but has not been adequately explored in post-desktop, mobile situations.

Bounded Deferral

An example of mediated interruption is to estimate the cost of interrupting a user, and use that in determining when to pass notifications on to them. The bounded deferral technique [81] proposed waiting till a user was not in a busy state, and was shown to be particularly effective at task or perceptual *breakpoints* [87]. Similarly, results from a mobile study suggest that notifications might be considered more favorably when delivered between two physical activities [102]. In the driving domain, informative interruption cues were used to signify the arrival and priority of a new notification, in order to allow the driver to decide whether and when to divert their attention [25].

Shared Context

Another approach is to shape the expectations of the communicating parties by making transparent the availability of the recipient, which is an example of a negotiated interruption. In an office environment, it was shown that using low cost sensors such as a microphone, models could be constructed to infer the interruptibility of an information worker with the same accuracy as humans [48]. In the mobile domain, an interruption library that used user activity, location, time of day, emotions and engagement, was shown to result in increased user satisfaction [124]. In the driving domain, numerous studies have shown that conversations with collocated passengers have less negative impact on driving compared to those with remote callers [76]. This led to work that demonstrated the positive effects of shared context by providing remote callers with the driver's context via video [106] and through auditory messages [88]. In an asynchronous mobile messaging scenario, however, shared context might be perceived as a privacy concern by its users.

The first part of work presented in this chapter focuses on studying the impact of notifications (perceptual and cognitive) on a driving-like task, and the role that mediation has on diminishing the costs (breakdown) associated with it. We will also evaluate the users ability to comprehend a notification with and without mediation. We will conduct this analysis for both audio and visual notifications and compare the results. In the second part, we train our sights on building a system that can autonomously mediate notification.

5.2.4 Autonomous Mediation

In cognitive psychology, there is a general consensus that people have limited and measurable cognitive capacities for performing mental tasks [114]. Furthermore, engaging in one mental task interferes with the ability to engage in other tasks, and can result in reduced performance on some or all of the tasks as a consequence [93]. To characterize the demand on these limited resources, psychologists have employed notions like cognitive load and mental workload, which gains definition through the experimental methods that are used to measure it [96].

Measuring Cognitive Load

Cognitive load can be assessed using data gathered from three empirical methods: subjective data using rating scales, performance data using primary and secondary task techniques, and psychophysiological data from sensors [123]. Self-ratings, being post-hoc and subjective in nature, tend to be inaccurate and impractical to use when automated and immediate assessment is required. Secondary task techniques are based on the assumption that performance on a secondary measure reflects the level of cognitive load imposed by a primary task. A secondary task can be as simple as detecting a visual or auditory signal, and can be measured in terms of reaction time, accuracy, and error rate. However, in contexts where the secondary task interferes with the primary task, physiological proxies that can measure gross reaction to task load are needed to assess cognitive load.

Psychophysiological techniques are based on the assumption that changes in cognitive functioning cause physiological changes. An increase in cortical activity causes a brief, small autonomic nervous response, which is reflected in signals such as heart rate (HR) and heart rate variability (HRV) [51, 115, 163], electroencephalogram (EEG) [136, 163], electrocardiogram (ECG) [136], electrodermal activity (EDA) [84, 143], respiration [115], and heat flux [65], eye movements and blink interval [14, 84, 85, 163] and pupillary dilations. Our dataset includes most of these signals as well as additional signals that have been shown to be sensitive to affect like pulse transit time (PTT), facial electromyography (EMG) and skin temperature [103, 122].

In particular, brain activity as measured through event-related potentials using EEG, or as inferred from pupillary responses have received more attention recently because of their high sensitivity and low latency [4, 108, 96]. There has been very little work that correlates these measures with the other physiological measures, or demonstrates how to effectively align them. Furthermore, to the best of our knowledge this is the only work that has focused on tracking cognitive load that is rapidly and randomly changing, since we are interested in teasing out the dynamic nature of instantaneous cognitive load. Lastly, prior work has typically focused on cognitive load arising in single-task scenarios like document editing [85], and traffic control management [143]. In contrast, there has more recently been interest in studying the effect that complex linguistic processing can have on driving using physiological measures of pupil dilation and skin conductance [36]. We take this further by building models that can estimate realtime cognitive load in this increasingly common multitasking scenario, i.e. distracted driving.

5.3 Study 1: Effects of Mediating Notifications

Our study had multiple goals, all of which are primarily focused on fleshing out a user, system and task model that informs our understanding of the distracted driving scenario. First, we wanted to determine how notifications impacted performance on a driving-like primary task. This relates to the cognitive level of the user model (3.1). Second, we wanted to establish how mediating them relative to task load could improve a user's performance on both the primary and secondary tasks. This relates to understanding the implications of the appropriateness approach in the task model (3.3). Finally, we wanted to understand these effects for both audio and visual notifications, in order to inform communication choices in the system model (3.2). These are formulated in the following research questions:

1. Mediation: How is primary task performance effected when the user is attending to notifications? Can mediation reduce this impact?

2. Modality: How is primary task performance effected by modality? Does mediation have the same effect across both audio and visual modes?

3. How do both of these conditions, i.e. mediation and modality, effect a user's ability to comprehend a notification?

Pilot explorations with a driving simulator, while promising, had a number of limitations. The interaction of full driving experience made it difficult to replicate task loads and added unnecessary dependent variables to data collection. For these reasons we chose instead to use the established ConTRe (Continuous Tracking and Reaction) task [105], which provides a highly controlled yet unpredictable task load for the participant. This allows for consistent and replicable analysis.

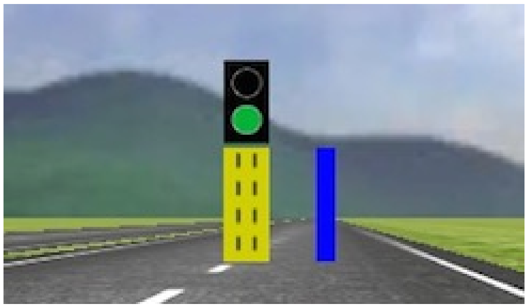


Figure 5.1: Screenshot of the ConTRe (Continuous Tracking and Reaction) Task that displays the yellow reference cylinder with the traffic light on top, and the blue tracking cylinder beside it.

The study was setup so that the primary ConTRe task would randomly switch between low and high workloads. This was done to simulate a typical driving scenario where drivers episodically experience high workload when they are entering/exiting highways, changing lanes, following navigation instructions, etc. For the secondary task, participants attended to notifications that were presented to them, as they performed the primary ConTRe task. Audio notifications were delivered via speakers, while visual notifications appeared through a Heads Up Display (HUD). The audio notifications were created using Apple's text-to-speech engine on OS X Yosemite (Speaking voice: Alex; Speaking rate: Normal). The HUD used was a Google Glass, which projects the screen at a working distance of 3.5 m, approximately 35°elevated from the primary position of the eye.

5.3.1 Experimental Design

The study was designed as a 2 (Audio/Visual modes) X 2 (Mediated/Non-mediated conditions) repeated measures within subjects study. This was done to mitigate individual variance in performance for the primary and secondary tasks. To control for possible effects of order the study was double counterbalanced for mode and condition factors. Additionally, there were two baseline conditions which included performing the ConTRe task in low and high workload settings without notifications.

5.3.2 Participants

20 people participated in our study, recruited through a call sent out to students selected randomly from a graduate engineering school population. There were 10 males and 10 females. The mean age of the participants was 26.4 years, with a standard deviation of 2.7 years. Participants were rewarded with a $40 gift cards for completing the study.

5.3.3 Apparatus

The Robot Operating System (ROS Hydro) was used to synchronize signals from the different components of the experimental setup. This includes data from the simulator, physiological sensors, and the audio-visual feeds, all of which were being sampled at different frequencies, on separate machines. Each component publishes messages via ROS Nodes to the server, which synchronizes the data and writes it to disk. A Logitech camera, a mic, and audio mixer were used to capture audio-visual information. Participants controlled the simulator using a Logitech G27 Racing Wheel.

5.3.4 Tasks

We elaborate below on the design of the primary ConTRe task and the secondary notification task that make up the multitasking scenario.

Primary Task: ConTRe

The ConTRe task comes as an add-on for OpenDS, an open-source driving simulator [105]. It is an abstracted and simplified task that comprises of actions required for normal driving, i.e. operating the brake and acceleration pedals, as well as using the steering wheel. This focuses the user's task and simplifies the recording of tracking behavior. Fine grained measures of performance on the primary task relative to the secondary task requests can be obtained, which is necessary for our investigation.

Here the car moves with a constant speed on a unidirectional straight road consisting of two lanes. The simulator shows two cylinders at a constant distance in front of the car: a yellow reference cylinder, and a blue tracking cylinder. The yellow reference cylinder moves autonomously and unpredictably. The lateral position of the blue tracking cylinder is controlled by the user through the use of the steering wheel. The cylinder moves left or right depending on the direction and angular velocity of the steering wheel, i.e the steering wheel controls the cylinder's lateral acceleration. Their goal is to track the yellow reference cylinder, by overlapping it with the user-controlled blue cylinder, as closely as possible. Effectively, this corresponds to a task where the user has to follow a curvy road. For the low and high task load conditions, the lateral speed of the reference cylinder was set to values that were empirically determined to create low and high workloads for the user, respectively.

Furthermore, there is a traffic light with two colors, placed on top of the yellow reference cylinder. The top light turns on red, whereas the bottom one turns on green. At any time, neither of the lights or only one is turned on. The red light requires that the user respond by depressing the brake pedal, while the green light corresponds to the accelerator pedal. This operates independently of the steering function. As soon as the user reacts to the light by depressing the correct pedal, the light turns off.

Secondary Task: Notifications

The secondary notification task is based on widely used measures of working memory capacity, which include operation span and reading span tasks [31]. Working memory has been purported to be involved in a wide range of complex cognitive behaviors, such as comprehension, reasoning, and problem solving as it is thought to reflect primarily domain-general, executive attention demands of the task [44]. In this work we do not aim to measure working memory, but instead want to measure the effect of processing a notification across the four experimental conditions. Thus, we modify the span tasks for our purposes as described below.

In each condition, drivers were presented with a series of twenty items, which included ten math equations and ten sentences taken from widely used span tasks [31] (see Table 5.1). The math equations and sentences are representative of the symbolic and verbal types of notifications, respectively, that users typically receive. Using standardized stimuli allows for consistency and replicability. Both types of notifications were randomly interspersed, so as to prevent the person from getting into a rhythm of expecting either one. After the subject had read or listened to each item, they verbally indicated if the notification was *true* or *false*. Sentences are true when they are semantically and syntactically correct, while the math equations are true when they are valid.

After each item, the participant was presented with an isolated letter, which represents something they had to remember from the notification. After two, three or four items, the simulator was paused, and they were asked to recall the letters in sequence, which we can liken to *responding* to a text message or some other such notification. Recall tasks are already known to have the most detrimental effects on primary task performance [89]. Pausing the simulator separates the recall effort from the recorded ConTRe task performance (even today's drivers are encouraged to stop their car before interacting with any request from their phone). This focuses the experiment solely on *attending* to notifications and its resulting effect on task performance.

5.3.5 Mediation

Mediation was done relative to task load. In the non-mediated (control) condition, notifications were presented randomly in both the low and high workloads. In the mediated (test) condition, notifications were presented only during low workload. The bounded deferral technique [81] was used, where the

Type	Notification
Math	2/2 + 1 = 1
Sentence	After yelling at the game, I knew I would have a tall voice

Table 5.1: Examples of the two types of notifications

notifications would be delayed while the driver was in a high workload setting. The notification would then be delivered a few seconds into the low workload setting. The mediation was conducted by one of the experimenters who had full view of the simulator and could determine when to deliver the notification. Modality appropriate changes were made if a notification had been delivered, and the workload changed from low to high before the driver responded. For the audio mode, the notification could be paused and continued at the next low workload period, or simply repeated. In the visual mode, the notification could be hidden till the next low workload period, when it would become visible again.

Before each condition, participants were told if they would be receiving audio or visual notifications. However, they did not receive any indication as to whether the notifications would be mediated by task load.

5.3.6 Methodology

Participants arriving at the lab were guided through an informed consent process, followed by an overview of the study. They were aided through the process of having a number of sensors attached to their body for the purposes of recording their physiological responses, including heart rate, electrodermal activity, skin temperature, etc. The participant was then seated in the simulator and shown how notifications would be delivered on the Glass, and through the speakers.

The participant was then taken through a series of practice runs to get them comfortable with the primary ConTRe task. When done with the practice, the low benchmark was recorded using the low workload setting on the simulator. After one minute, they were asked to repeat a series of ten sentences that were read out to them, one-by-one, while they were still performing the ConTRe task. The same routine was performed to record the high benchmark using the high workload setting on the simulator.

This was followed by another set of practice rounds, where the secondary notification task was explained and demonstrated to the participant. After this, a practice trial was run by combining both the ConTRe task (with the randomly alternating workloads) and the notifications task. The notification task included a set of five items, three of which were math equations, with the rest being sentences. This provided the participants with a sense of what to expect during the actual trials. The practice trials could be repeated if necessary.

The participants then moved on to the experimental trials. Each participant participated in a total of four trials, one for each condition. At the end of the four trials, the participant was interviewed about the disruptivity and effectiveness of audio and visual notifications. The entire study lasted approximately 2 hours per user.

5.3.7 Measures

Quantitative performance data on both primary and secondary tasks were collected. From the ConTRe task, we collected the following: steering deviation, i.e. the difference in distance between the reference cylinder and the tracking cylinder; reaction times to respond to the red and green lights, i.e. the amount of time from when the light went off to when the correct pedal was depressed; and the error rate of depressing the wrong pedal. These measures were automatically recorded by the simulator.

In the mediated condition, notifications were presented in the low workload section. In the non-mediated condition, notifications were presented in the low and high workload sections. We would thus expect the ConTRe performance in the low workload sections to be identical for both conditions. Hence, we focus our analysis on the performance data from the high workload sections of the mediated and non-mediated conditions. Steering deviation was being continuously sampled at 570 Hz. To filter out noise and infrequent occurrences of sudden deviations from the trend. The data was filtered using a rolling median. The average steering deviation of each user in each condition was then recorded. For the accelerator and brake reaction tasks, there were an average of 31.5 brake reaction and 31.3 accelerator reaction data points per user for each condition. Like the steering deviation, the reaction times for both the brake and accelerator tasks were low-pass filtered using a rolling median. The mean reaction times were then calculated and recorded for each user in each of the four conditions.

For performance on the secondary notification task, the response times for math and sentences were computed separately. This is the time from when the notification was presented to the driver, to when they respond to indicate true or false. As before, the data is filtered using a rolling median. The mean response times for math and sentences are then recorded for each user in every condition. The errors in the responses were also calculated, as well as the error in recalling the sequence of letters that were presented to the driver after each notification. The sequence could be two, three or four letters long.

At the end of all the trials, participants were interviewed about their preferences regarding the modality of the notification, and the effect of its disruptivity on their primary task performance. They were also asked if they perceived any difference between the two audio or the two visual conditions, i.e between the mediated and non-mediated conditions.

5.3.8 Results

After processing the data as described in the previous section, we arrive at 10 data points per user (one for each measure) for each of the 2 (Modality) X 2 (Mediation) conditions: Audio Mediated (AM), Audio Non-mediated (AN), Visual Mediated (VM), Visual Non-mediated (VN). This totals to 40 data points per user, and a total of 800 data points. Described below are the results from the analysis of these data points, starting with the primary ConTRe task.

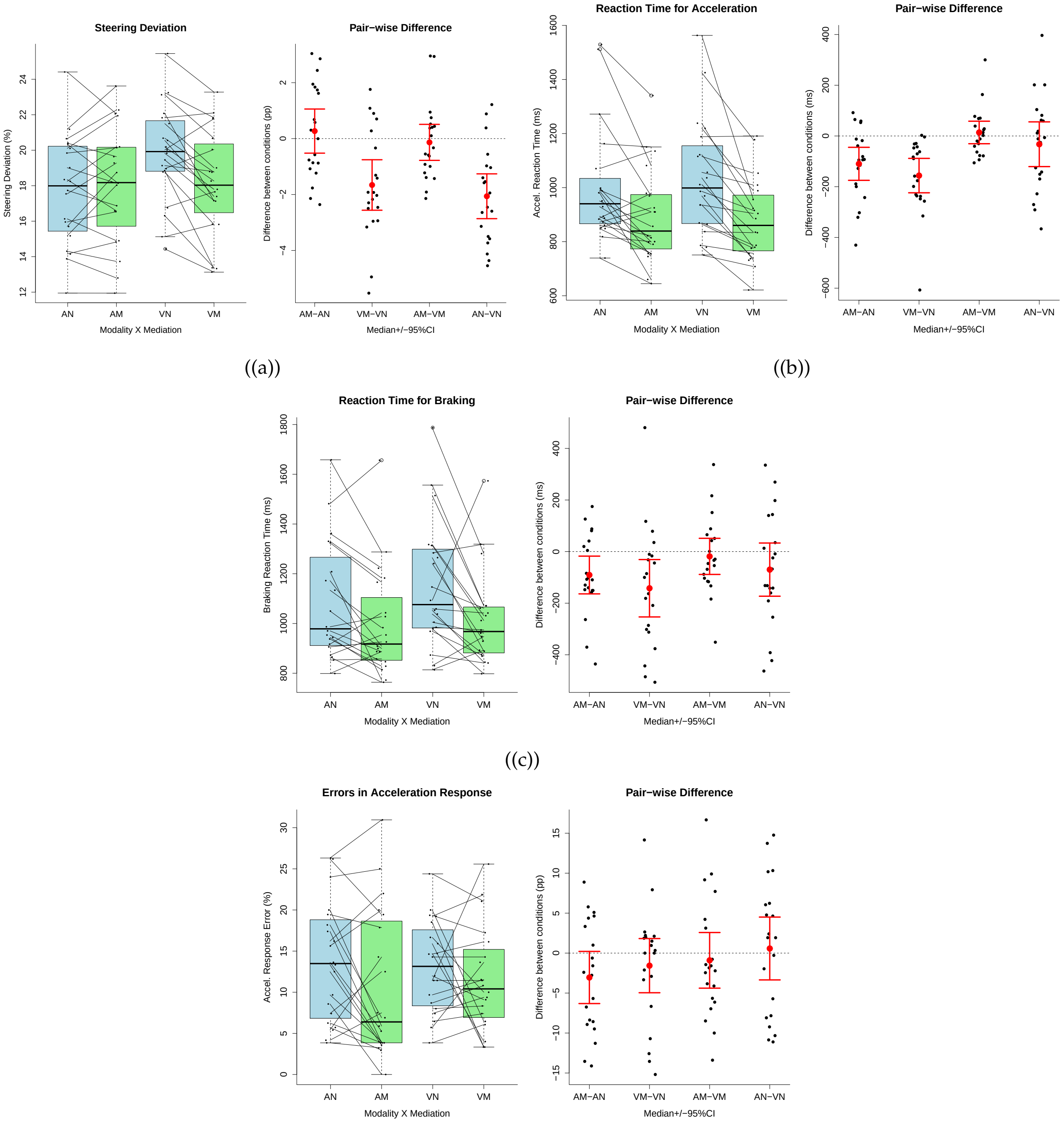
Effects on Primary ConTRe Task

We now review the analysis of the ConTRe performance measures to understand the effects of mediation and modality on the primary task. These measures include Steering Deviation, Reaction Time for Acceleration and Braking, and Errors in Acceleration and Braking. To perform the analysis we use a multivariate ANOVA (MANOVA) using all five driving performance measures as dependent variables. As opposed to running multiple univariate *F* tests for each dependent variable, MANOVA has the advantage of reducing the likelihood of a Type I error, and revealing differences not discovered by ANOVA tests [155].

A two-factor repeated measures MANOVA with within-subject factors (Mediation, Modality) showed a significant effect from Mediation, $F(1,19) = 25.46$, $p < .001$, and no significant effect from Modality $F(1,19) = 1.16$, $p = .29$. There was no significant interaction between the two main effects $F(1,19) = 1.20$, $p = .28$, which validates the main effect analysis. This implies that notifications were distracting and negatively impacted user performance on the ConTRe task. It did not matter if the notifications were audio or visual.

Given the omnibus multivariate *F*-test revealed a significant effect from Mediation, we further analyze the effect on the different metrics separately. Since we are also interested in the effect of Modality, we include its analysis. Thus for each measure, we describe four planned comparisons using paired t-tests: a) Mediated and Non-mediated *Audio* (AM-AN), b) Mediated and Non-mediated *Visual* (VM-VN), c) *Mediated* Audio and Visual (AM-VM), and d) *Non-mediated* Audio and Visual conditions (AN-VN) (see Figure 5.2 & Table 5.2). To control for Type I errors we use the Bonferroni adjusted alpha levels of .0125 per test (.05/4).

The first two planned comparisons (AM-AN & VM-VN) emphasizes the impact of mediation in the audio and visual modalities, separately. The next two planned comparisons (AM-VM & AN-VN) contrasts the audio with the visual modes. In the *Mediated* comparison no notifications were presented in the



((a))

((b))

((c))

((d))

Figure 5.2: Box plots (superimposed with parallel coordinate plots) of the driving performance measures for 2 (Modality) X 2 (Mediation) conditions: Audio Mediated (AM), Audio Non-mediated (AN), Visual Mediated (VM), Visual Non-mediated (VN), along with the pair-wise differences for each of the planned comparisons (AM-AN, VM-VN, AM-VM, AN-VN).

Primary Task Measures	**AM-AN**	**VM-VN**	**AM-VM**	**AN-VN**
Steering Deviation (pp)	0.27 (.48)	-1.66 (**.001**)	-0.13 (.67)	-2.06 (< **.001**)
Acceleration Reaction Time (ms)	-109.75 (**.002**)	-156.13 (< **.001**)	13.84 (.52)	-32.53 (.45)
Brake Reaction Time (ms)	-90.71 (.017)	-142.05 (.015)	-18.49 (.59)	-69.83 (.17)
Acceleration Response Error (pp)	-3.05 (.06)	-1.57 (.34)	-0.91 (.59)	0.57(.76)
Brake Response Error (pp)	-3.26 (.08)	-0.87 (.65)	-1.92 (.35)	0.47 (.78)

Table 5.2: Average pair-wise difference for each primary task measure, with p-values from paired t-tests in parenthesis.

high workload sections as per protocol. Thus, we would expect ConTRe performance to be identical for the audio and visual modes in the mediated condition as there were no notifications presented to the participants. On the other hand, in the non-mediated condition, the difference in notification modality might be borne out on the ConTRe task performance.

Steering Deviation: Comparing the steering deviation for Mediated (M = 18.02 %, SD = 3.27 pp) and Non-mediated conditions (M = 17.75 %, SD = 3.08 pp) in the *Audio* mode does not reveal any significance, $t(19) = 0.71$, $p = .48$, which matches results from previous work that found cognitive load costs are minimally borne out on steering deviation [76, 24]. In the *Visual* mode, there was a significant difference between Mediated (M = 18.15 %, SD = 2.95 pp) and Non-mediated conditions (M = 19.81 %, SD = 2.75 pp), $t(19) = -3.84$, $p = .001$ (Figure (a)). This indicates that the Visual mode effects the visual requirements of the primary task, i.e. tracking the lateral movement of the system-controlled yellow cylinder. This is consistent with findings in the literature which indicate that tracking will be negatively impacted by glances away from the road [78].

In the *Mediated* condition, comparing the Audio and Visual modes showed no significant differences. As explained before, the *Mediated* condition is equivalent to driving without any notifications. In the *Non-mediated* condition, Audio was significantly less disruptive than the Visual mode $t(19) = -0.43$, $p < 0.001$. Again, this can be attributed to adding a visual source of distraction to a primary task that depends on visual input.

Reaction Time for Acceleration and Braking: In the *Audio* mode, the mean reaction time for acceleration was significantly reduced in the Mediated condition (M = 892.2 ms, SD = 176.4 ms) as compared to the Non-mediated condition (M = 1001.9 ms, SD = 214.6 ms), $t(19) = -3.52$, $p = .002$. The same effect carried on into the *Visual* mode with the Mediated condition (M = 878.4 ms, SD = 152.4 ms) being significantly less than the Non-mediated condition (M = 1034.5 ms, SD = 214.6 ms), $t(19) = -4.81$, $p < 0.001$ (Figure (b)). This is again consistent with findings which indicate that diverting focal attention from the road will result in longer reaction times [76]. The difference between Audio and Visual was not significant in the *Mediated* or *Non-mediated* condition.

Performing the same analysis for the braking reaction times, the *Audio* mode showed a near significant difference between the Mediated (*M* = 995.6 ms, *SD* = 216.8 ms) and Non-Mediated means (*M* = 1086.3 ms, *SD* = 238.8 ms), *t*(19) = -2.59, *p* = .017. In the *Visual* mode, the difference was near significant as well with the Mediated condition (*M* = 1014.2 ms, *SD* = 189.1 ms) being lower than the Non-mediated condition (*M* = 1156.2 ms, *SD* = 254.8 ms), *t*(19) = -2.67, *p* = .015 (Figure (c)). No difference was found when comparing the Audio and Visual modalities.

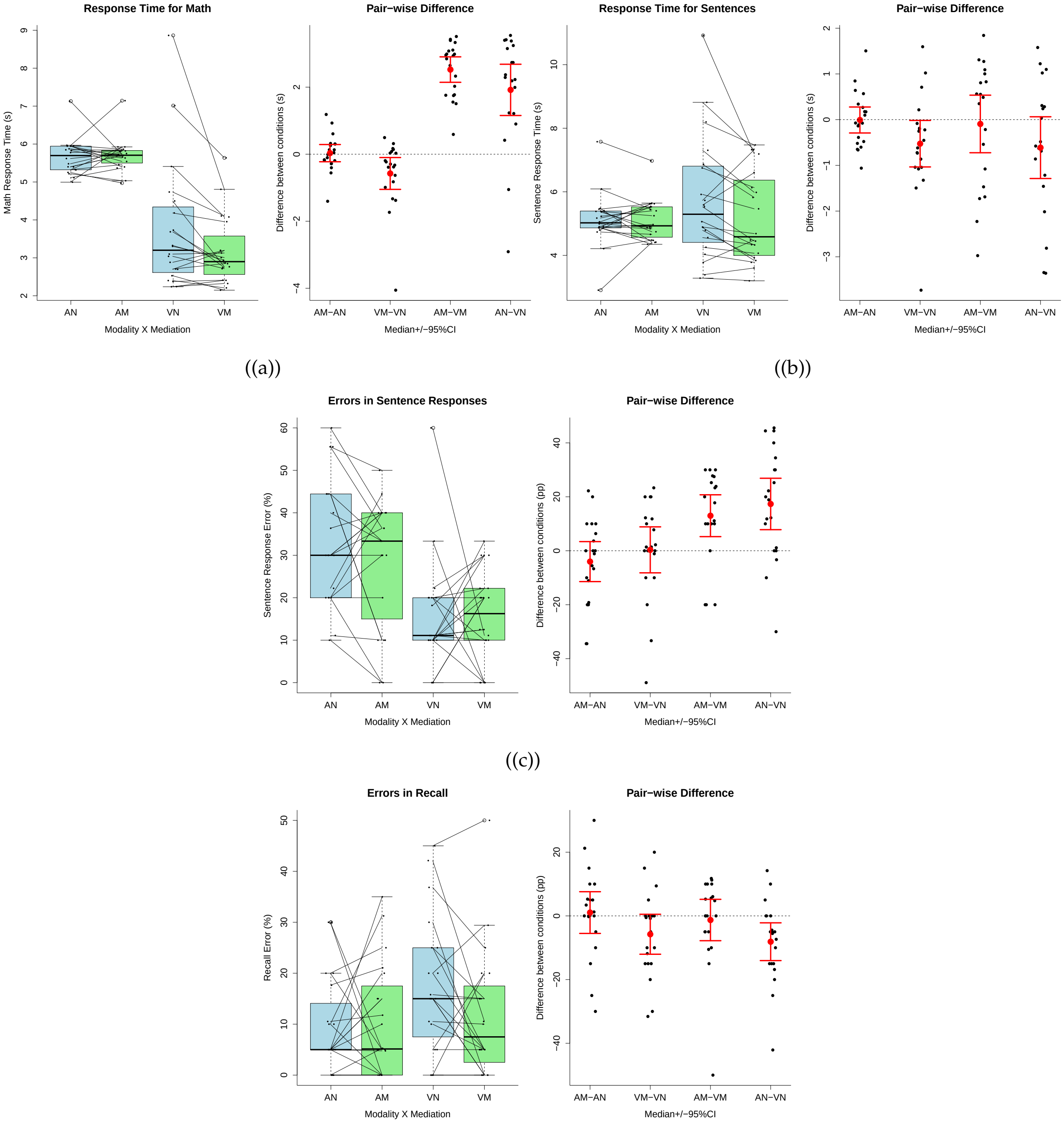
While acceleration and braking test for similar things, the slight increase in reaction times for braking compared to acceleration might be attributed to the extra time it takes the user to move their foot from the accelerator pedal (over which it used to hover by default for most users) to the braking pedal. It is plausible that the act of braking itself introduces a larger manual source of distraction compared to acceleration.

Errors in Acceleration and Braking: As with the reaction time analysis, we begin by analyzing the acceleration results. In the *Audio* mode, there were fewer errors in the Mediated condition (*M* = 10.64 %, *SD* = 8.94 pp) as compared to the Non-mediated condition (*M* = 13.69 %, *SD* = 7.24 pp), *t*(19) = -1.96, *p* = 0.06, but this did not reach significance. In the *Visual* mode there was no difference between the Mediated (*M* = 11.54 %, *SD* = 6.27 pp) and Non-mediated conditions (*M* = 13.11 %, *SD* = 5.56 pp), *t*(19) = -0.97, *p* = 0.34 (Figure (d)). Again fvor both the *Mediated* and *Non-mediated* conditions, no difference was found between the Audio and Visual modalities.

For the errors in braking responses, there was a slight differences between the Mediated (*M* = 10.02 %, *SD* = 7.37 pp) and Non-mediated conditions (*M* = 13.26 %, *SD* = 9.50 pp), *t*(19) = -0.46, *p* = 0.08, in the *Audio* mode, but this did not reach significance. In the *Visual* mode, there was no significant difference between the Mediated (*M* = 11.91, *SD* = 7.10 pp) and Non-mediated conditions (*M* = 12.78, *SD* = 6.59 pp), *t*(19) = -0.46, *p* = 0.65. No difference was found in the *Mediated* and *Non-mediated* conditions across both modalities. Due to the similarity in the acceleration and braking response, only the acceleration response errors were plotted.

Effects on Secondary Notification Task

We now proceed to analyze the main effects of Mediation and Modality on the notification task, whose measures include, Response Times for Math and Sentences, Response Errors for Math and Sentences, and Recall. Similar to the primary driving task analysis, we use a multivariate ANOVA (MANOVA) using all five notification task measures as dependent variables.



((a))

((b))

((c))

((d))

Figure 5.3: Box plots (superimposed with parallel coordinate plots) of the driving performance measures for 2 (Modality) X 2 (Mediation) conditions: Audio Mediated (AM), Audio Non-mediated (AN), Visual Mediated (VM), Visual Non-mediated (VN), along with the pair-wise differences for each of the planned comparisons (AM-AN, VM-VN, AM-VM, AN-VN).

Secondary Task Measures	**AM-AN**	**VM-VN**	**AM-VM**	**AN-VN**
Math Response Time (ms)	0.03 (.80)	-0.57 (.02)	2.52 ($<$ **.001**)	1.92 ($<$ **.001**)
Sentence Response Time (ms)	-0.008 (.95)	-0.52 (.04)	-0.09 (.75)	-0.61 (.07)
Math Response Error (pp)	1.92 (.48)	2.11 (.41)	-2.57 (.29)	-2.38 (.35)
Sentence Response Error (pp)	-4.02 (.27)	0.34 (.93)	12.99 (**.002**)	17.35 (**.001**)
Recall Error (pp)	1.05 (.74)	-5.76 (.07)	-1.29 (.68)	-8.10 (**.01**)

Table 5.3: Average pair-wise difference for each secondary task measure, with p-values from paired t-tests in parenthesis.

A two-factor repeated measures MANOVA using within-subject factors (Mediation, Modality) showed that all effects were significant at the .05 significance level. The main effect of Mediation yielded an F ratio of F(1,19) = 5.49, p = .03. The main effect of Modality yielded an F ratio of F(1,19) = 12.81, p = .002. There was also a significant interaction effect, F(1,19) = 6.90, p = .017. To understand the interaction, we calculate the simple effects for each of the two levels in the independent variables (Mediation, Modality).

We first analyse the simple effects of Mediation by setting the independent Modality variable to *Audio*. A one-way MANOVA with Mediation as the within-subject variable showed no significant effect between the Mediated and Non-mediated conditions, F(1,19) = 0.03, p = .85. Whereas, setting the independent Modality variable to *Visual*, revealed a significant effect, F(1,19) = 7.52, p = .01. This implies that audio notifications are comprehended equally well under low and high workloads. Visual notifications, on the other hand, are comprehended differently under low and high workloads.

Next we analyze the simple effects of Modality by setting the independent Mediation variable to *Mediation*. A one-way MANOVA with Modality as the within-subject variable showed a highly significant effect of mediation between Audio and Visual modes F(1,19) = 28.98, $p < .001$. Setting the independent variable to *Non-mediation* did not show a significant effect F(1,19) = 3.84, p = .06. What we might infer from this analysis is that under low workloads, users comprehend audio and visual notifications differently. Under high workloads, modality of notifications does not effect comprehension ability.

To understand the direction of the differences, and the impact on the different dependent variables, we perform four planned comparisons using paired t-tests, similar to the primary task analysis. For each dependent variable, we describe below comparisons between: a) effect of mediation in the *Audio* mode (AM-AN), b) effect of mediation in the *Visual* mode (VM-VN), c) effect of modality in the *Mediated* conditions (AM-VM), d) and the effect of modality in the *Non-mediated* conditions (AN-VN) (see Figure 5.3 & Table 5.3). To control for Type I errors we use the Bonferroni adjusted alpha levels of .0125 per test (.05/4).

Response Times for Math and Sentences: We first analyze the reaction times for math. In the *Audio* mode, there was no difference in the reaction times between the Mediated (*M* = 5.69 s, *SD* = 0.43 s) and Non-mediated conditions (*M* = 5.66 s, *SD* = 0.47 s), *t*(19) = 0.25, *p* = .8. In the *Visual* mode there was a difference between the Mediated (*M* = 3.17 s, *SD* = 0.9 s) and Non-mediated conditions (*M* = 3.74 s, *SD* = 1.70 s), *t*(19) = 0.25, *p* = .02, but it did not reach significance. There was a highly significant difference in the reaction times between the Audio and Visual modes in both the *Mediated* and *Non-mediated* conditions, $p < .001$ (Figure (a)).

The differences in reaction times for sentences were less dramatic. In the *Audio* mode there was no difference between the Mediated (*M* = 5.10 s, *SD* = 0.63 s) and Non-mediated cases (*M* = 5.11 s, *SD* = 0.84 s), *t*(19) = -0.056, *p* = .95. In the *Visual* mode, there was a slight difference in the Mediated (*M* = 5.20 s, *SD* = 1.41) and Non-mediated conditions (*M* = 5.72 s, *SD* = 1.94 s) , *t*(19) = -2.16, *p* = .04. In the *Mediated* case, there was no difference between the Audio and Visual conditions, *t*(19) = -0.31, *p* = .75. In the *Non-mediated* case, there was a slight difference, but it did not reach significance, *t*(19) = -1.89, *p* = .07 (Figure (b)).

Errors in Math and Sentences: For errors in math responses, there were no differences between all four of the planned comparisons, for which reason they were not plotted. (Audio: *M* = 8.14 %, *SD* = 8.29 pp; Visual: *M* = 6.22 %, *SD* = 8.62 pp; Mediated: *M* = 10.72 %, *SD* = 7.59 pp; Non-mediated: *M* = 8.61 %, *SD* = 10.42 pp).

For the errors in sentence responses, there was no significant difference in the *Audio* and *Visual* modes between the Mediated and Non-mediated conditions. In the *Mediated* condition there was a significant difference in the Audio (*M* = 28.69 %, *SD* = 15.00 pp) and Visual conditions (*M* = 15.69 %, *SD* = 10.78 pp), *t*(19) = 3.51, *p* = .002. Similarly, there was a significant difference in the *Non-mediated* case between the Audio (*M* = 32.71 %, *SD* = 14.59 pp) and Visual conditions (*M* = 15.35 %, *SD* = 13.36 pp), *t*(19) = 3.81, *p* = .001 (Figure (c)). From this analysis, we might infer that modality has a significant effect on users' ability to comprehend sentences accurately, regardless of mediation. They made fewer errors when sentences were presented visually, as opposed to aurally.

Errors in Recall: Analyzing the errors in recall revealed no significant difference in the *Audio* mode (Mediated: *M* = 10.45 %, *SD* = 10.99 pp; Non-mediated: *M* = 9.40 %, *SD* = 9.32 pp), *t*(19) = 0.33, *p* = .74. There was a slight difference in the *Visual* mode between the Mediated (*M* = 11.74 %, *SD* = 12.65 pp) and Non-mediated conditions (*M* = 17.51 %, *SD* = 13.34 pp), *t*(19) = -1.92, *p* = .06, which did not reach significance. Comparing the *Mediated* Audio and Visual conditions did not show any difference, *t*(19) = -0.41, *p* = 0.68. There was a significant difference between the *Non-mediated* Audio and Visual conditions, *t*(19) = -2.86, *p* = .01 (Figure (d)).

Physiological Measures		Performance Measures	
Raw	Derivative	ConTRe Task (T1)	Notification Task (T2)
Electrocardiogram (ECG)	Pulse Transit Time (PTT)	Steering Deviation	Sentence Response Time
Photoplethysmograph (PPG)	Inst. Heart Rate (IHR)	Acceleration Reaction Time	Sentence Accuracy
Impedance Cardiography(ICG)	SKT B − SKT A (SKT)	Acceleration Accuracy	Math Response Time
Respiration		Braking Reaction Time	Math Accuracy
Electrodermal Activity (EDA)		Braking Accuracy	Recall Accuracy
Skin Temp. Nose (SKT A)			
Skin Temp. Cheek (SKT B)			
Electromyography (EMG)			
Pupil Dilation			
Eye Gaze			

Table 5.4: Collection of measures available in the dataset.

Subjective Feedback

After the experiment, the participants were interviewed about their preferences, and for any feedback they might have. While the data showed that visual notifications can be distracting under high workloads, the subjective data indicated that this did not bear on the modality preferences of the participants. Some of the participants who preferred the visual notifications mentioned that they did not find the particular text-to-speech voice appealing. While others who preferred the audio notifications explained that they found it cumbersome to periodically avert their gaze from the simulator screen to the heads-up display.

On being asked if they noticed any differences between the two audio trials or the two visual trials, i.e. between the mediated and non-mediated conditions, none of the participants indicated that they perceived a difference. They did not find any condition to be less difficult or stressful. This is intriguing because the data revealed that non-mediated notifications are distracting and impacted performance measures, but this was not consciously picked up on by the participants.

5.4 Estimating Cognitive Load using Physiological Measures

In this section, we describe the model building process to estimate cognitive load from physiological data that was collected while users participated in Study 1. A modified version of this model will serve as the gateway module (section 3.5), and will be used to autonomously mediate notifications in Study 2.

5.4.1 Physiological Sensors

Physiological signals were captured and recorded using the Biopac's BioNomadix monitoring devices for Electrocardiogram (ECG), Photoplethysmograph (PPG), Electromyogram (EMG), respiration, skin tem-

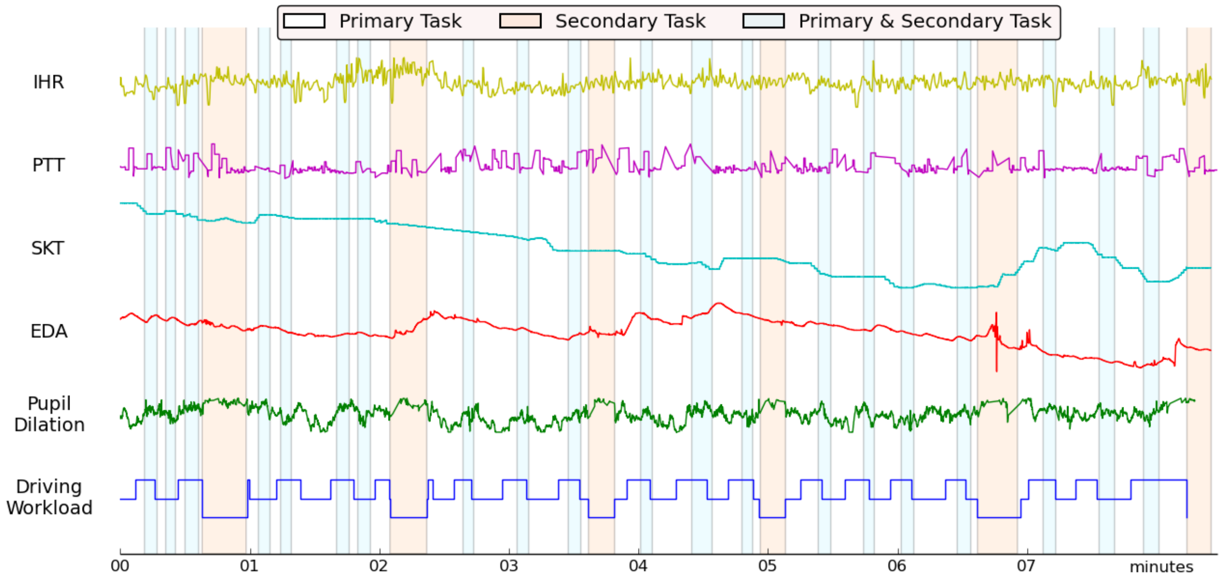


Figure 5.4: Example physiological measures collected during an audio non-mediated experimental condition. Driving workload is represented as a step function (1: High, 0: Low, -1: Pause). Colored regions delineate when the user was engaged in the primary driving task (T1; white regions), secondary notification task (T2; orange regions), or both (T1 & T2; blue regions).

perature, Electrodermal Activity (EDA), and Impedance Cardiography (ICG). Pupil dilation and eye gaze was captured using Pupil Pro hardware[1], which is a head-mounted mobile eye-tracking platform.

Since the hands of the participant were occupied for driving, we placed the PPG and EDA sensors on the participant's left toe & instep, respectively [153]. The facial EMG sensor was placed just above their left eyebrow to measure activation of the corrugator supercilii muscle, which is associated with frowning. Two skin temperature sensors were placed on the tip of the nose and on the left cheek. The ECG, impedance cardiography, and respiration sensors were placed in the default positions, i.e., on the chest and neck.

5.4.2 Dataset

The dataset consists of a number of physiological and performance measures which are tabulated in Table 5.4. We recorded ten psychophysiological signals: EDA, EMG, skin temperatures (nose and cheek), four signals based on cardio-respiratory activity (ECG, PPG, ICG & respiration), and two based on eye activity (gaze and pupil dilation). Apart from the eye-based signals which were sampled at 30 Hz, the rest of the signals were sampled at 2000 Hz.

Three derivative signals were also calculated. Instantaneous heart rate (IHR) was obtained from the ECG signal using the BioSig library[2] which implements Berger's algorithm [17]. Pulse Transit Time (PTT)

[1] http://pupil-labs.com/pupil/

[2] http://biosig.sourceforge.net/

was obtained by calculating the difference in between the ECG R-wave peak time and the PPG peak time, which is the time it takes for the pulse pressure waveform to propagate through a length of the arterial tree. Difference in skin temperature (SKT) was also calculated by subtracting the temperature of the nose from that of the cheek.

The performance measures encompass both the primary driving task and the secondary notification task. Of interest are the reaction times and accuracies to the red and green light stimuli, and the steering deviation in tracking the reference cylinder. Also recorded are the performance measures for the secondary notification task as shown in Table 5.4.

Preprocessing & Labelling

In this exploration, 5 of the 13 psychophysiological signals collected were seen to be the easiest and most fruitful to analyze for dynamic task load modelling. They include IHR, PTT, SKT, EDA and pupil dilation (Figure 5.4). These signals were extracted from the collected data and down-sampled to 40 Hz (except for pupil dilation which remains at its original sampling rate of 30 Hz). Each signal was plotted, and thresholds were determined to filter out unlikely values (from movement artifacts, etc.). Data for each user was standardized (zero mean & unit variance), prior to which outliers that were more than three standard deviations from the average, were filtered out.

Two sets of labels are included in the dataset, a set each for the primary and secondary task. By syncing with the timestamps from both the task logs, we determined the precise primary and secondary task conditions that the participant was under for every physiological sample. The primary task labels denote if the participant is in the low, high, or paused driving workload condition (see Driving Workload in Figure 5.4). The logs from the secondary task allow us to determine the periods during which a participant was attending to a notification, i.e. blue regions in Figure 5.4. The orange regions signify the recall part of the secondary task, when the primary driving task was paused.

Feature Extraction

We derived a number of statistical features on the main signal $(x[n])$, the derivative signal $(x[n+1] - x[n])$, and the percentage change $((x[n+1] - x[n])/x[n] * 100)$. These features include the mean, median, percentiles (10th, 25th, 75th, 90th), ranges (between min and max, 10th and 90th percentiles, and 25th and 75th percentiles), skewness, and standard deviation.

Features were extracted using a sliding window. To capture temporal properties, windows were overlapped, i.e. their step size was smaller than their length. Different window lengths and step sizes were considered. Specifically, the following pairs of window and step sizes (seconds) were analyzed: (7, 1), (5, 1), (3, 1) and (3, 0.25).

5.4.3 Modelling for Multitasking Scenario

Based on the insights from Wilkin's multiple resource theory described in the Related Work section, the loads that the primary and secondary tasks impose on the user are not mutually exclusive. Both tasks compete for resources along the stages dimension, and along the visual channel dimensions if the notifications are visual. Hence it would be more prudent for the classifier to make predictions on the load the user is under for both tasks separately and simultaneously, instead of attempting to make predictions on some notion of composite or absolute load. Thus, we can view this as a multi-label classification problem. In this formulation, each window is assigned two labels, where each label is drawn from the set of labels that corresponds to the primary and secondary tasks. Given the limited data to train the model on, we reduce our task to a multi-label binary classification problem, and ignore the specific states of the ConTRe (low/high workloads) and notification (attending/recall) tasks. Essentially, at this stage, we are simply trying to predict which tasks (T1 and/or T2) the user is engaged in by analyzing the psychophysiological data.

A sliding window is labelled as T1 if for the duration of a window the participant is only engaged in the primary ConTRe task. If the ConTRe task is paused, and the participant is engaged in recall, the window is labelled as T2. If the participant is attending to a notification while performing the ConTRe task, the window is labelled as both T1 and T2. For simplicity, the transitory unlabelled windows were discarded. The features of the remaining windows, and their corresponding multi-label assignments {T1,T2} were fed to a Random Forest classifier, which is an ensemble technique that learns a number of decision tree classifiers and aggregates their results. Models were built across all users, as well as for each user separately to account for individual differences in their psychophysiological response. To evaluate the classifier's performance, we used leave-one-user-out cross-validation for the population models, and 3-fold cross-validation for the individual user models.

The time it takes to comprehend a notification varies by participant. This creates variation in the number of driving and notification task labels generated per participant, which in turn results in a varying baseline accuracy for each user because of the class imbalance problem. Hence, instead of accuracy, we use the Area Under the Receiver Operating Characteristic Curve (ROC AUC) metric to evaluate the classifier, as it is insensitive to class imbalance. ROC curves show the trade-offs between higher sensitivity and higher specificity. Sensitivity refers to the correct detection of a condition or state when it is truly present.

Window, Step (s)	Population			Individual		
	T1∨T2	T1	T2	T1∨T2	T1	T2
7, 1	0.85	0.90	0.80	0.84	0.89	0.78
5, 1	0.84	0.89	0.78	0.83	0.88	0.78
3, 1	0.81	0.85	0.76	0.81	0.87	0.75
3, 0.25	0.80	0.86	0.75	0.80	0.86	0.74

Table 5.5: ROC AUC Scores for population and individual models using different window and step sizes

Specificity indicates the correct rejection of a state when it is truly not present. The area under the ROC curve is a measure of adequacy on both. Curves corresponding to random or chance classification of 50% would fall close to the diagonal, and result in an ROC AUC score of 0.5 regardless of class imbalance, while the most successful classifications would have an ROC AUC score close to 1.0.

Being a multi-label classification problem, the classifier outputs two probabilities simultaneously: one for the probability of the sample belonging to the primary task (T1), and another for the probability that the sample belongs to the secondary task (T2). We report the macro-averaged ROC AUC scores for the pair of labels, as a measure of how well the classifier is simultaneously able to predict both labels (T1∨T2). We also report the ROC AUC score for each label, individually, to shed light on how accurately the classifier is able to identify each task.

5.4.4 Results

Of the five physiological signals analyzed, the pupil dilation measures were the only signal to yield results that were much better than random. For this reason, we only list and discuss results using the pupil dilation measures. For the four window and step size combinations considered, mean ROC AUC scores for the population and individual models are shown in Table 5.5. A larger window size tends to provide better results, and this trend holds for both the individual and population models. The population scores are comparable to the average user scores, which tells us that the model based on pupil dilations is generalizable.

Table 5.5 also shows ROC AUC scores for predicting each label individually. The scores indicate that the models are better at identifying when the user is engaged in the primary driving task (T1) as compared to when the user is engaged in the secondary notification task (T2). This might be because of the differences in load induced by equation and sentence notifications, and also from the differences in the notifications being right or wrong. Our model doesn't account for these yet, but each can potentially be treated as a different class under a label in the multi-label framework.

Condition	T1∨T2	T1	T2
Non-mediated			
Video	0.88	0.90	0.86
Audio	0.90	0.92	0.88
Overall	0.88	0.91	0.86
Mediated			
Video	0.82	0.89	0.76
Audio	0.81	0.88	0.74
Overall	0.81	0.89	0.74

Table 5.6: Population-based ROC AUC Scores under different timing and modality conditions.

We also compared how varying the independent variables of timing and modality impacted the ROC AUC scores. The results for these experiments are tabulated in Table 5.6. Only the analysis on the 7 second long windows are presented here as similar trends were observed for the other combinations. It is clear that mediating when notifications were sent had a larger effect than modality on the model's ability to identify the secondary task. To explain this, we must remember that in the mediated condition, the participant is sent notifications when they are in the low driving-workload state. The multiple resource theory predicts that the cognitive load on the user in this state (low driving-workload + notification) is similar to the cognitive load they experience when they are in the high driving-workload state. Thus, in the mediated condition windows where the user is driving with and without notification, i.e. windows labelled {T1} and {T1,T2}, look similar. In the non-mediated condition, this is not the case, as notifications are also delivered in the high driving-workload states. This allows the classifier to better identify the secondary task (T2) in the non-mediated condition.

5.5 Study 2: Autonomous Mediation

In this section, we present the second instantiation of the Considerate Mediator, which autonomously mediates communication with the driver. This function is primarily handled by the gateway module, which "gates" information to the different agents. (see Figure 5.5). A modified version of the model built and evaluated in the previous sections served as the core of the gateway module. The model was modified so as to work with realtime data, and was trained only on data from the non-mediated condition.

The model gets standardized input from the pupil dilation data stream. Moving windows that were 5 seconds long with a step size of 1 were used, as these showed the most promise during pilot experiments. The model outputs a {T1,T2} classification every second. Since at this preliminary stage, we are only detecting which tasks the user is engaged in, we make the assumption that if the user is multitasking, i.e

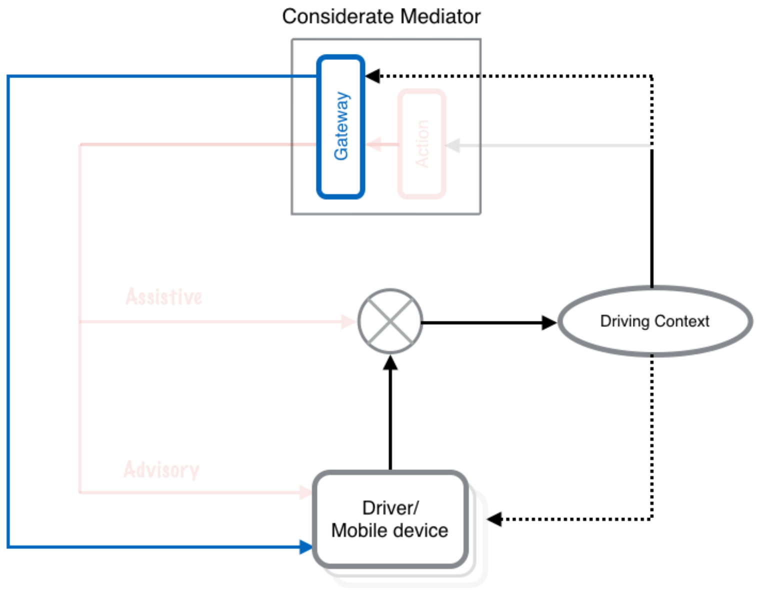


Figure 5.5: The Considerate Systems architecture grounded in the distracted driving scenario.

{T1,T2} = {1,1}, then the user is experiencing high load. If the user is only engaged in driving it outputs {1,0}. Based on the pattern of the classifications streaming out from the model, the mediator decides whether to delay or resume notifications (see Section 5.5.5 & Table 5.8 for more details).

We evaluate this instantiation of the Considerate Mediator, and the gateway module in particular, using an experimental setup similar to the one used in the previous study with some modifications to the design and tasks. These changes are described below.

5.5.1 Design

This study focused on audio notifications only. It was designed as a repeated measures within subject study with only one independent variable, i.e. non-mediated (control) vs. mediated (test) conditions. To control for possible effects of order the study was counterbalanced.

5.5.2 Participants

10 people (all male) participated in our study recruited through a call sent out to students selected randomly from a graduate school population.

5.5.3 Tasks

Since notifications are what the system needs to mediate, they could not be used as the secondary task (T2) that the classifier detects. We therefore increased task load in a different way by using a manual transmission-based gear changing task as the secondary task. The tasks were chosen so as to make it difficult for a user to perform perfectly on the primary and secondary tasks simultaneously.

Primary Task (T1): ConTRe

The primary task remains the same as in the first user study. The participant is engaged in an abstracted driving task, where they track a yellow cylinder with a steering wheel. The participant also has to simultaneously respond to red and green lights on the yellow cylinder by depressing the brake and accelerator pedals, respectively. The ConTRe task was set to alternate between periods of low and high workloads as described in the first study.

Secondary Task (T2): Gear Change

An LCD screen is placed in front of the simulator such that its contents are easily visible below the yellow and blue cylinders presented on the simulator screen. Numbers from 1–6 are presented on the LCD screen, which correspond to the gears on the manual transmission gearbox which is included with the Logitech G27 Racing Wheel. The user was asked to shift to the right gear when the number changed on the screen. To create a high task load for the user, the gear number only changed when the ConTRe task was in its high load setting. The gear number was set to change every 1–3 seconds.

Mediated Task: Notifications

The notification task is a simplified version of the one used in the previous study. To create a continuous task scenario the pause and recall portion of the previous study was eliminated. In this study, notifications only consist of audio math and sentence prompts that the user responds to with a true or false.

5.5.4 Apparatus and Sensors

The apparatus used to conduct, synchronize and record the experiment was the same as before. Only audio notifications were presented to the user. As pupil dilation was the lone physiological measure of interest, the Pupil Pro headset was the only physiological sensor worn by the user.

5.5.5 Methodology

Participants were guided through an informed consent process, followed by an overview of the study. The participant was then seated in the simulator, and was asked to put on the Pupil Pro headset. They were instructed on how to perform the ConTRe task. Once comfortable with the task, the secondary gear

Performance Measures	M	N	*p*
Primary Contre Task			
Steering Deviation (%)	22.0	23.1	.47
Accel Reaction Time (ms)	980	1014	.67
Brake Reaction Time (ms)	1117	1157	.47
Accel Response Error Rate	0.34	0.23	.07
Brake Response Error Rate	0.25	0.32	**.05**
Secondary Gear Task			
Attempts per stimulus	1.15	1.26	**.015**
Response Error Rate	0.22	0.31	**.05**
Mediated Notification Task			
Math Reaction Time (s)	2.02	2.30	.19
Sent. Reaction Time (s)	2.30	2.53	.32
Math Response Error Rate	0.08	0.08	.82
Sent. Response Error Rate	0.22	0.27	.33

Table 5.7: Mean performance measures of the primary, secondary and mediated tasks from both the mediated (M) and non-mediated (N) conditions, along with paired t-test two-tailed p-values.

changing task was introduced. After this the audio math and sentence notifications were demonstrated to the user. Once the user was familiar with all the tasks, a calibration step was performed to determine the parameters needed to standardize the data before classification. This step simply required the user to perform the ConTRe task in its low workload setting for 10 seconds. This was followed by two experimental trials. These included the test condition in which notifications were autonomously mediated based on task load, and the control condition in which notifications were randomly presented to the user regardless of task load.

Notifications were mediated by delaying them if they hadn't started playing. If they had started playing, and then the system detected that the task load on the user was high, the notification would cut off and repeat itself when the load on the user had reduced. A trigger-happy system that cuts off a notification every time a {1,1} is output by the classifier can be annoying to the user. For better user experience, notifications were mediated only when certain patterns of classification outputs were observed. Based on pilot studies, the protocol was set to delay or cut-off notifications anytime a pattern of either [{1,1}, {1,1}] or [{1,1}, {1,0}, {1,1}] classifications was output by the classifier. The system would then wait for a series of five {1,0} classifications before resuming delivery of notifications.

5.5.6 Measures

Quantitative performance data on primary, secondary, and mediated tasks were collected. From the primary ConTRe task, we collected the following: steering deviation, i.e. the difference in distance between the reference cylinder and the tracking cylinder (sampled at 570 Hz); reaction times to respond to the red and green lights, i.e. the amount of time from when the light went off to when the correct pedal was depressed; and the error rate of depressing the wrong pedal. These measures were automatically recorded by the simulator. An average of 23.8 and 13.7 acceleration stimulus points were presented to each user in the mediated and non-mediated conditions, respectively. Similarly, an average of 21.3 and 11.2 brake stimulus points were presented in the mediated and non-mediated conditions, respectively. Since notifications were being delayed in the mediated condition, these trials were longer that the non-mediated ones. For each user in each condition, the mean steering deviation, reaction times, and reaction errors were calculated.

From the secondary gear-changing task, the number of tries the user took to get to the right gear, and the number of times they didn't succeed in reaching the right gear were determined. The mean of these measures for each user in both conditions were then calculated. Per user, an average of 52.2 and 28.3 gear change requests were made in the mediated and non-mediated conditions, respectively.

For performance on the mediated notification task, the response times for the math and sentence prompts were computed. This is the time from when the notification was presented to the driver to when they respond to indicate true or false. The mean response times for notifications are then recorded for each user in every condition. The errors in the responses and the mean per user was also calculated for each condition. An average of 7.5 math and sentence prompts each, were presented to users in both conditions.

The outputs from the classifier, which occur every second, were also recorded for the mediated condition. These will be analyzed to shed light on how the classifier's outputs could inform the system's mediation behavior.

5.5.7 Results

Below we report on results from the experiment. We look at mediation effects on each task by collectively analyzing their corresponding performance measures. Since this presents three sets of comparisons (one for each task), we use the Bonferroni adjusted alpha levels of .017 per test (.05/3) to control for Type I errors. To perform the analysis, we perform a multivariate ANOVA (MANOVA) on the performance measures from each task. As opposed to running multiple univariate F tests on each measure, MANOVA has the advantage of reducing the likelihood of a Type I error, and revealing differences not discovered by

Output Pattern	Sensitivity	Specificity	Accuracy
every *H* and *L*	40	72	61
H	90	19	56
HH or *HLH*	83	42	63
HHH	74	68	71
HHHH	58	82	70

Table 5.8: Evaluation of different output patterns that the Considerate Mediator could be designed to respond to. The first two rows indicate the overly eager and cautious behaviors, respectively. The next three rows represent different patterns of classifier output.

ANOVA tests [155]. We also analyze the classifier output with respect to task load, in order to shed light on how a system might mediate notifications more effectively.

Mediation Effects

The analysis of mediation effect on the primary ConTRe task using a repeated measures MANOVA showed no significant effect, $F(5,5)=1.44$, $p=.35$. The means for each of the five primary task measures in both conditions and the paired t-test two-tailed p-values are listed in Table 5.7.

For the secondary gear-changing task, a repeated measures MANOVA showed a significant effect, $F(2,8)=7.42$, $p=.015$. Further analysis of each of the dependent variables showed a significant difference in the mean number of tries the user took to get to the right gear between the mediated ($M=1.15$, $SD=0.16$) and non-mediated ($M=1.26$, $SD=0.14$) conditions, $t(9)=-3.72$, p=.004. There was also a slightly significant difference in the failure rates between the mediated ($M=0.22$, $SD=0.05$) and non-mediated ($M=0.31$, $SD=0.13$) conditions, $t(9)=-2.26$, p=.05. These are listed in Table 5.7.

A repeated measures MANOVA for the mediated notification task revealed no significant effect, $F(4,6)=0.98$, $p=.48$. The means for each of the four notification task measures in both conditions and the paired t-test two-tailed p-values are also listed in Table 5.7.

Mediation Performance

Since the classifications are done on a sliding window, we can expect a lag from the onset of high task load to when the classifier output indicates so. Another reason for the delay in classifications might be that even though a high load is being imposed on the user, it might take a couple of seconds for them to experience it as such. To find the average delay, the cross-correlation between the alternating load conditions and time-shifted classification outputs was determined for multiple time shifts. Across users the average time-shift at which the cross-correlations were maximum was 4.9 s with a standard deviation of 1.44 s. Thus, to simplify our analysis going forward, we first account for this lag by shifting the classification outputs by 4.9 s, so that the input and output are more directly correlated. We represent a {1,0} classifier output as *L* and a {1,1} classifier output as *H*. The goal of this analysis is to get a sense of how well the classifier was detecting high load situations across users in the study, and how the system's mediation behavior could potentially be improved.

By being overly eager or overly cautious, a system can display two extremes in how it uses the classifier outputs to inform its mediation behavior. The eager system for example reacts immediately to every change in task load *L* and *H* being output by the classifier by playing or pausing a notification. We would expect the system to have high specificity, as it immediately changes its behavior based on classifier output. The cautious system also stops notifications immediately when high task load is sensed *H*, but continues to do so even if an *H* is followed by *L*s for a specified period of time. Thus it displays low specificity. Under the cautious behavior, a single *H* occurring during a high load section is considered as a true positive (correct classification). Conversely, a single *H* during a low load section is a false positive (incorrect classification). The system's sensitivity, specificity and accuracy are calculated by aggregating the true and false positives over the trials from all users. Results for the overly eager and cautious behaviors are shown in Table 5.8, along with a few intermediate behaviors which we describe next.

To trade-off between sensitivity and specificity, the system could be designed to mediate notifications only if it sees a particular pattern of classifier outputs. As described above, if a pattern occurs during a high load section it is marked as a true positive, and if it occurs during a low load section it is marked as a false positive. A few example patterns were evaluated, and their results are listed in Table 5.8. These include patterns such as [*H*,*H*] or [*H*,*L*,*H*] which reduces the sensitivity of the system to the classifier outputs, making it less cautious. This was also the pattern that was actually used by the system in the autonomous mediation study. We can reduce system sensitivity even further by having the system mediate notifications only when it sees [*H*,*H*,*H*] from the classifier. Table 5.8 also lists evaluation results when [*H*,*H*,*H*,*H*] is the pattern that the system responds to. In this way we get a sense of how the system's mediation behavior would have changed if the protocol was set to respond to different patterns of classifier outputs.

5.6 Discussion

This chapter presents the second scenario in which the Considerate Mediator was grounded and evaluated, namely distracted driving. We started by taking the aid of the user, system and task models to formulate research questions for the first study. After identifying potential for breakdown at the cognitive level of the user, we proceeded to implement the gateway module in order to minimize it. At its core, the gateway module uses a model trained on physiological measures to estimate cognitive load. Then based on the model's output, and informed by the appropriateness approach, the mediator decides whether to delay or resume delivery of notifications to the driver.j

The dual-task studies in this chapter consisted of a primary driving-like tracking and reaction task, and a secondary notification-based cognitive task. ConTRe was used as the primary task as it focuses on core driving skills and removes learnable contextual cues. This improves data and repeatability of experiment. Similarly, prompts frequently employed in complex span task experiments serve as the notifications presented to a participant. These represent the symbolic and verbal nature of notifications commonly received by people on their smartphones.

The first study was focused on investigating the effects of attending to symbolic and verbal notifications. The timing and modality of the notifications were treated as independent variables to understand their effects on cognitive load. By mediating the notifications based on the task load, we wanted to understand its effects on both the primary sensorimotor task, and the secondary notification task. The experimental results revealed that notifications are indeed distracting and impacts primary task performance. This effect was significant for both aurally and visually presented notifications. Furthermore, the visual modality did allow users to perform better on the secondary notification task when the notifications were being mediated, i.e. when the notifications were being presented only during low workload.

The analysis of the different measures for the primary ConTRe task performance (Table 5.2) showed that acceleration and braking response times were the most effected by the increased cognitive load of having to process notifications. The braking response times appeared to be longer than the acceleration response times possibly due to the added manual source of distraction in moving the foot to the brake pedal (we noticed that users hovered their foot over the accelerator pedal by default). The added visual source of distraction from visual notifications was borne out on the steering deviation measure during high task workloads. This effect on steering deviation was absent when the notifications were presented in audio. With regards to the acceleration and braking response errors, mediation had a slight but insignificant positive effect in the audio mode, but no effect in the visual mode. It might be the case that even with mediation, there is a competition for visual resources in the visual mode.

From the analysis of the performance measures in the secondary notification task (Table 5.3), it is clear that users respond faster to visual notifications when they involved the symbolic cognitive task of considering an equation. This effect did not hold for comprehending sentences, which might be more representative of communication notifications. Perhaps reading sentences while driving required more glances than reading an equation. If so, the result indicates that people were self-mediating by making trade-offs between performing the ConTRe task and comprehending sentences. With regards to accuracies in responses, the number of math errors were consistent across all the conditions. Users were, however, making more errors in comprehending sentences when they were presented aurally compared to visually. It could be that iconic visual memory is less susceptible to disruption than serial audio memory. The condition with non-mediated visual notifications presented the most difficulty to drivers in which to recall a sequence of letters.

The post experiment debriefing revealed that preferences for modality was not based on its impact on task performance but rather the comfort level of the participants with the modality. Furthermore, none of the participants realized that notifications in two out of the four trials were being mediated. We might infer from these findings that users were not reflexively aware of the extra load being introduced by the secondary task and its impact on their performance, which can have dangerous implications in real life situations. As designers then, we must rely on data and user performance, not preferences in the successful design of cognitively challenging systems.

In this way, these findings informs the socio-centric user model, communication-based system model, and the task model when designing solutions for such scenarios. There is a clear potential for breakdown at the perceptual and cognitive level of the user. Despite the number of studies highlighting the increased risks associated with visual distractions, recent advances in augmented reality (e.g. Navdy[3], Microsoft HoloLens[4]) have further kindled interests in the applications of such display-based technologies. While such advances make these devices more practical and seamless to use, it might not necessarily lead to lower distraction. The eye's fovea only sees the central one to three degrees of the visual field. It is difficult for people to simultaneously focus on two visual stimuli especially if the information contained in both is orthogonal to each other. For these reasons, the rest of the work described in this chapter focused on the potential breakdown at the cognitive level.

A dataset of 13 psychophysiological signals was collected and used to build models that can estimate cognitive load. These signals include ECG, PPG, ICG, Respiration, EDA, nose & cheek skin temperatures and the differences between them, EMG, pupil dilation, eye gaze, PTT and IHR (listed in Table 5.4). These were collected during a dual-task user study that subjected a participant to a series of alternating low and high task loads. The study was designed in this way to mimic the fluctuating loads people experience while driving in the real world. The goal was to capture these fluctuations as reflected in the participants physiological responses. The dual-task scenario can be cast as a multi-label learning problem of the primary and secondary tasks. The approach succeeded at building classification models

[3] https://www.navdy.com/

[4] http://www.microsoft.com/microsoft-hololens/en-us

that distinguish whether the user is engaged in the primary task, the secondary task, or both. The pupil dilation measure gave the most promising results. The model was built using statistical features derived from measurements of pupil dilation, which were fed to a random forest classifier. The model worked for each user and across all the participants, and was used as the core of the gateway module for the second study.

In the second study, we evaluated the gateway module's ability to mediate notifications to users in real-time. The setup from the first study was altered to include a manual gear changing task instead of the notification itself. Pupil dilation data was streamed to the classifier which output a multi-label classification every second. In the test condition, the system would inhibit notifications if it believed that the user was simultaneously engaged in two tasks. In the control condition, notifications were delivered randomly. The effects of mediation were determined by analyzing the performance measures for each task. Mediation allowed users to reach the right gear (their secondary task) with less errors, and fewer number of attempts per gear change request. Notice that the gear-shifting task uses different perceptual and cognitive skills than the verbal notification task, which is what the model was trained on. Our model transferred and performed well on this stressing activity based on mechanical performance.

The system's mediation performance was evaluated using cross-correlation measures between the user task loads and time-shifted classifier outputs. We can interpret the results as there being an average lag of 4.9 s between the onset of high task load for the user, and when the system mediated notifications to them. System mediation behavior was also analyzed based on how it responds to different patterns of outputs from the classifier. The trade-off between the system's sensitivity and specificity was demonstrated for these different patterns.

5.6.1 Future Work

Future work includes building systems that can autonomously mediate notifications relative to not only workload, but also driving skill, expertise, other real-time information along the planned route like traffic, accidents, detours, weather, etc. Progress is being made towards implementing such systems [166, 150, 43]. We might also estimate the cognitive load that a particular notification would induce, based on the sender and content. This takes into account the social level of the user model. These estimates could then feed an overall cost measure that determines if, when, and how, to engage with the user without aggravating the risks associated with distracted driving. Many challenges remain in assessing the role that a mediating system should take, and in the design of its corresponding actions. Furthermore, it is of deep concern that users could not be depended on to choose the condition that reduced performance. We should more deeply explore how to give designers guidance to make systems that keep users from attending to stimuli when it will be detrimental to performance or even dangerous. Conversely, work could be focused on giving users an intuition on situations where their performance is improved or degraded in such situated interactions.

There are a number of more directly relevant directions that future work can take. First, our data analysis did not include moving windows over transitions from low to high workloads and back. We are optimistic that temporal models could be used to detect these transitions, reducing the lag in load detection. Second, improved measures of the load experienced by the users (ground truth) can be obtained by using a composite measure of the different task performance metrics. Reaction times can serve as more reliable proxies for cognitive load than externally imposed task load settings. Third, with more data we can make fine-grained estimations about user load within each task (for example, T1 = 0, 1, 2, 3, etc., based on difficulty of the primary driving task). Fourth, we should explore which physiological signals are more indicative of stress, and which are better suited for estimating cognitive load. Stress is likely to arise when failure at a task is coupled with feelings of lack of control, in situations where participants are evaluated by others [32]. We might hypothesize that stress is an affect. It ebbs and flows at a slower pace than cognitive load, which being reflective of the stages of mental processing, fluctuates more rapidly.

To show that pupil dilation measures can be robust we used an inexpensive off-the-shelf measuring technique. Prior work reports use of expensive eye-trackers with higher sampling rates for pupilometric measurements. Our study succeeded with a consumer webcam (Microsoft LifeCam HD-6000) that has a sampling rate of only 30 Hz. Even when tested outdoors in a car during the day, with no special attempt to control for ambient luminescence (apart from the initial calibration step), the system showed promising results in estimating task load through pupil dilation measures. More work could be done to refine the setup and understand the trade-offs between the fidelity of the equipment, environmental setup, and the robustness of results.

Chapter 6

Discussion & Conclusions

This work responds to calls for a paradigm shift [68] in the way human-system interactions are designed, brought about by the dovetailing trends of ubiquitous computing, and context- and socially-aware computing. With the increasing pervasiveness of such systems, it has become apparent that while highly utilitarian, their lack of social grace often makes them a source of frustration and embarrassment. As these systems have grown integral to everyday life, becoming more usable with fewer steps per action, their inability to change their behavior to reflect the social situation has prevented ubiquitous systems from achieving Weiser's vision of a vanishing technology [158].

A common sense knowledge of how to react in social situations is key to creating socially intelligent systems. This thesis introduces Considerate Systems as a framework of approaches that structure the application of this common sense knowledge through a design process & guidelines (user, system & task models), and an architecture. It motivates the need to design future systems with such knowledge so that they can respond in socially appropriate ways, building trust and improving collaborations. It shows how social appropriateness lies in the trade-off between pursuing system goals and minimizing interaction breakdown. This trade-off is motivated by a taxonomy of approaches, embodied in the task model, which better inform the design of such systems. These approaches include feedback, feedforward, appropriateness, and differential. The application of these approaches are in turn guided by the system and user models.

The system model guides how a considerate actor can augment the communication with superficial, embellishing, ancillary, multi-modal, orthogonal or functional considerate utterances that support interactions, without distracting from the content exchange. The considerate communication can be targeted at the perceptual, precognitive, cognitive, or behavioral levels in the user model of the other actor. In a social setting, it can be targeted at the different structural layers at which the communication is occurring including turn taking, conversant roles, participants' social power and status, the type of interaction, and the social setting.

Two exemplar systems were built to evaluate the Considerate Systems framework, and to demonstrate its application. The framework, as embodied by the Considerate Mediator, was grounded in the conference call and distracted driving scenarios, both of which are commonly encountered. The conference call serves as an ideal starting point to demonstrate how a considerate agent can achieve its scenario goals, and even improve human-human communication, in spite of operating on and sharing the same constrained audio medium as the other participants on the conference call. By identifying and addressing five commonly cited problems, we were able to evaluate different aspects of the agent's actions as depicted in the architecture using the actuation module. These took the form of advisory and assistive actions, and demonstrated the effectiveness of simple triggers and carefully crafted responses in addressing these problems.

The second scenario, i.e. distracted driving, allows the application of the framework to a broader multimodal domain with potentially far-reaching consequences when a system is inconsiderate. Here the focus is on mediating the timing of the agent's actions based on the driver's cognitive load, as depicted in the architecture using the gateway module. Pausing or delaying notifications when the driver is in a high workload was shown to improve their driving performance.

These design solutions emphasizes the utility of a considerate response that balances the pursuit of various system goals with the need to minimize interaction breakdown. For example, in conference calls, developing a language of short, unobtrusive non-judgemental responses was pivotal to its success in that constrained medium. So was building a blackboard system to prioritize and arbitrate between knowledge sources, and to schedule timely responses. The scheduling aims to reduce interruptions through awareness of total message volume and message repletion. With regards to the problems of dominance and extraneous noise, the system encourages the offending users to check themselves. By engaging with them in a considerate manner, the system is eliciting a similar considerate response. This is the give-and-take in communication; a common sense knowledge of how actions can beget desired reactions, and is part of the larger social contract implicit in any interaction.

We further demonstrated how short audio cues like earcons and auditory icons can appropriately provide feedback to stymie the disorienting effects of technology mediation. We show that earcons can improve accuracy on speaker identification, and is comparable to 2D spatialization of speakers. Auditory icons, like keyboard typing and mouse clicking, can act as feedback to reassure participants about the presence of others on the line. We also showed that using metaphorical prompts (intonations) to announce events like entry and exit of participants reduces errors when compared to speech prompts.

The idea of using short communication footprints has become widely popular in the form of notifications. As computing devices become ubiquitous and proactive, we have come to depend on these to receive information and communicate with the world. Notifications, due to their efficiency and perceived safety, have become the primary mode of interaction with the user. Prior research has shown, however, that interacting with these devices, i.e. noticing, attending and responding to messages can be disruptive, and can even have safety implications for tasks such as driving. To improve upon this, the distracted driving scenario was chosen as the second exemplar application in which to evaluate considerate systems. We developed a technology that can know when its appropriate to engage the user based on their cognitive load. We showed that the autonomous mediation of notifications significantly improved participant task performance. Cognitive load assessment is a rich area for exploration, and we hope to inspire other researchers to use our data set to further evaluate models of dynamic task load estimation. By enabling computers to interact appropriately and considerately, we can pave the way for future social computing scenarios.

Society is likely to witness an increasing number of computing systems that make a foray into situated social settings. The reality is that the more interactive these systems are, the greater is the need for them to follow the social contracts of interaction. Like humans, these systems will have to make space for those they are interacting with. They will have to introduce themselves and be available if and when others recognize them. They will have to keep an account of whether their point was made, and realize when further attempts to make their point might be unnecessary or even counter productive. They will have to support the conversations that the people they are around with are engaged in.

If social intelligence plays a big part in how we interact, then this is an area in which technology products are found wanting. If endowing systems with social intelligence is the next logical phase, then considerate systems is a first step towards that goal. Considerate systems is a call to go beyond building more efficient and simple to use systems, to systems that can support and adapt their behaviors to the environment. Considerate system stresses on situational awareness and appropriate response to the situation. This work attempts to do much simpler, and less critical tasks; only attending to the nature of when and how to introduce a communication successfully. This can be based on a tractable common sense knowledge for different application spaces onto which we can tack on more computational units for deeper understanding and affect, and more nuanced behaviors. Social intelligence is a full spectrum that systems might one day occupy as dictated by their goals and duties. This work focuses on ensuring that computing systems are not off this spectrum to begin with.

Intelligent interfaces can focus on various aspects of knowledge, including context and models for user, task, and system. Affective computing focuses on recognizing and modifying system understanding of users based on their affective stance. We present yet another paradigm that allows a system to socially orient itself towards the goals and intentions of the user, and aims to be considerate in its stance. In specific, we challenged ourselves to create an intelligent system which assists human collaboration using the same communication channel and modality as humans do. We demonstrate the valuable influence an agent can have in such settings through better crafting of social cues and responses. We show how careful choice of its *syntax* — its sound and placement (like long utterances), *semantics* — its direct and indirect relationship to the conversational channel (such as adding a speech response on top of a conversational channel), and *timing* — mediating them with respect to the user's cognitive load, can deeply affect its goal of supporting social interactions.

The choices of how and when a considerate agent should intrude on a communication channel is shown here to be delicate but tractable. We are excited about the possibility in the utility of such considerate agents across other interactive scenarios that would benefit from a system's ability to regulate and coordinate social feedback. As it weaves itself into the very fabric of society, the imperative now is for technology that celebrates situational awareness, and appropriateness.

Bibliography

[1] H. Alm and L. Nilsson, "The effects of a mobile telephone task on driver behavior in a car following situation," *Accident Analysis & Prevention*, vol. 27, pp. 707–715, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/000145759500026V 65

[2] M. Altosaar, R. Vertegaal, C. Sohn, and D. Cheng, "AuraOrb: Social Notification Appliance," in *CHI '06 extended abstracts on Human factors in computing systems*, ser. CHI EA '06. New York, NY, USA: ACM, 2006, pp. 381–386. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1125451.1125533 5

[3] L. Angell, J. Auflick, P. a. Austria, D. Kochhar, L. Tijerina, W. Biever, T. Diptiman, J. Hogsett, and S. Kiger, "Driver Workload Metrics Task 2 Final Report - Appendices," Washington, DC: DOT HS 810 635, Tech. Rep., 2006. [Online]. Available: http://trid.trb.org/view.aspx?id=855927 65

[4] P. Antonenko, F. Paas, R. Grabner, and T. van Gog, "Using electroencephalography to measure cognitive load," *Educational Psychology Review*, vol. 22, no. 4, pp. 425–438, 2010. 69

[5] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, pp. 35–50, 1992. 27

[6] E. Arroyo and T. Selker, "Attention and intention goals can mediate disruption in human-computer interaction," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6947 LNCS, 2011, pp. 454–470. [Online]. Available: http://gti.upf.edu/attention-and-intention-goals-can-mediate-disruption-in-human-computer-inter-computer-interaction/ 5

[7] E. Arroyo, S. Sullivan, and T. Selker, "CarCoach: a polite and effective driving coach," in *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems*, ser.

CHI EA '06, vol. 2. New York, NY, USA: ACM, 2006, pp. 357–362. [Online]. Available: http://doi.acm.org/10.1145/1125451.1125529 5, 16

[8] A. Baddeley, "Working memory: looking back and looking forward," *Nature reviews neuroscience*, vol. 4, no. 10, pp. 829–839, 2003. 65

[9] B. Bailey, J. Konstan, and J. Carlis, "Measuring the effects of interruptions on task performance in the user interface," in *IEEE International Conference on Systems, Man, and Cybernetics, 2000*, vol. 2. IEEE, 2000, pp. 757–762. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=885940 65

[10] S. Bakker, D. Hausen, and T. Selker, *Peripheral Interaction: Challenges and Opportunities for HCI in the Periphery of Attention*. Springer, 2016. 5

[11] P. Balthazard, R. E. Potter, and J. Warren, "Expertise, extraversion and group interaction styles as performance indicators in virtual teams," *ACM SIGMIS Database*, vol. 35, no. 1, p. 41, feb 2004. [Online]. Available: http://doi.acm.org/10.1145/968464.968469 28, 43

[12] J. N. Barkenbus, "Eco-driving: An overlooked climate change initiative," *Energy Policy*, vol. 38, no. 2, pp. 762–769, 2010. 16

[13] A. Battestini, V. Setlur, and T. Sohn, "A Large Scale Study of Text Messaging Use," in *MobileHCI'10, September 7âĂŞ10, 2010, Lisbon, Portugal*. ACM, 2010, pp. 229–238. 65

[14] J. Beatty and B. Lucero-Wagoner, "The pupillary system," *Handbook of psychophysiology*, vol. 2, pp. 142–162, 2000. [Online]. Available: http://www.nrc-iol.org/cores/mialab/fijc/files/2003/090203{_}Pupillary{_}System{_}.pdf 68

[15] M. Beaudouin-Lafon and A. Karsenty, "Transparency and awareness in a real-time groupware system," in *Proceedings of the 5th annual ACM symposium on User interface software and technology - UIST '92*, ser. UIST '92. ACM, 1992, pp. 171–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=142621.142646 32

[16] M. Bekoff, "Social Play Behaviour: Cooperation, Fairness, Trust, and the Evolution of Morality," *Journal of Consciousness Studies*, vol. 8, pp. 81–90, 2001. [Online]. Available: http://www.ingentaconnect.com/content/imp/jcs/2001/00000008/00000002/1075 1

[17] R. D. Berger, S. Akselrod, D. Gordon, and R. J. Cohen, "An efficient algorithm for spectral analysis of heart rate variability," *Biomedical Engineering, IEEE Transactions on*, no. 9, pp. 900–904, 1986. 83

[18] T. Bickmore and J. Cassell, "Social Dialogue with Embodied Conversational Agents," in *Advances in Natural, Multimodal, Dialogue Systems*, ser. Text, Speech and Language Technology, J. C. J. Kuppevelt, L. Dybkjær, and N. O. Bernsen, Eds. Springer Netherlands, 2005, vol. 30, pp. 23–54. [Online]. Available: http://dx.doi.org/10.1007/1-4020-3933-6{_}2 7, 16

[19] M. Blattner, D. Sumikawa, and R. Greenberg, "Earcons and Icons: Their Structure and Common Design Principles," *Human-Computer Interaction*, vol. 4, no. 1, pp. 11–44, mar 1989. [Online]. Available: http://dl.acm.org/citation.cfm?id=1455735.1455736 37

[20] D. Bohus and E. Horvitz, "Computational Models for Multiparty Turn-Taking," Microsoft Research, Tech. Rep., 2010. [Online]. Available: http://research.microsoft.com/en-us/um/people/dbohus/docs/turntaking{_}msr{_}tr.pdf 17

[21] ——, "Facilitating multiparty dialog with gaze, gesture, and speech," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction on - ICMI-MLMI '10*, ser. ICMI-MLMI '10. Beijing, China: ACM, 2010, Conference proceedings (article), p. 1. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1891903.1891910 7

[22] J. R. Brubaker, G. Venolia, and J. C. Tang, "Focusing on shared experiences," in *Proceedings of the Designing Interactive Systems Conference on - DIS '12*, ser. DIS '12. ACM, 2012, p. 96. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2317956.2317973 26

[23] J. K. Caird, K. a. Johnston, C. R. Willness, M. Asbridge, and P. Steel, "A meta-analysis of the effects of texting on driving," *Accident Analysis and Prevention*, vol. 71, pp. 311–318, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000145751400178X 65

[24] J. K. Caird, C. R. Willness, P. Steel, and C. Scialfa, "A meta-analysis of the effects of cell phones on driver performance," *Accident Analysis and Prevention*, vol. 40, no. 4, pp. 1282–1293, jul 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457508000183 77

[25] Y. Cao, F. van der Sluis, M. Theune, R. op den Akker, and A. Nijholt, "Evaluating informative auditory and tactile cues for in-vehicle information systems," in *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '10*. New York, New York, USA: ACM Press, nov 2010, p. 102. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1969773.1969791 67

[26] J. B. Carson, P. E. Tesluk, and J. a. Marrone, "Shared leadership in teams: An investigation of antecedent conditions and performance," *Academy of Management Journal*, vol. 50,

no. 5, pp. 1217–1234, oct 2007. [Online]. Available: http://www.mendeley.com/research/shared-leadership-teams-investigation-antecedent-conditions-performance/ 28

[27] D. Cohen, A. Chandrashekaran, I. Lane, and A. Raux, "The hri-cmu corpus of situated in-car interactions," *Proc. IWSDS*, pp. 201–212, 2014. 63, 66

[28] J. Cohen, ""Kirk here:"," in *INTERACT '93 and CHI '93 conference companion on Human factors in computing systems - CHI '93*, ser. CHI '93. ACM, 1993, pp. 63–64. [Online]. Available: http://portal.acm.org/citation.cfm?doid=259964.260073 27, 31

[29] ——, "Out to Lunch: Further Adventures Monitoring Background Activity," *Proceedings of the Second International Conference on Auditory Display*, pp. 15–20, 1994. [Online]. Available: http://scholar.google.com/scholar?cluster=5090332371982636882{&}hl=en{&}as{_}sdt=2005{&}sciodt=0,5{#}2 27, 31, 37

[30] H. M. Collins, *Tacit and explicit knowledge*, Philadelphia, PA, 2010. [Online]. Available: http://orca.cf.ac.uk/29371/ 4

[31] A. R. a. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle, "Working memory span tasks: A methodological review and user's guide." *Psychonomic bulletin & review*, vol. 12, no. 5, pp. 769–786, oct 2005. [Online]. Available: http://www.springerlink.com/index/10.3758/BF03196772 72

[32] D. Conway, I. Dick, Z. Li, Y. Wang, and F. Chen, "The Effect of Stress on Cognitive Load Measurement," in *Human-Computer Interaction–INTERACT 2013*. Springer, 2013, pp. 659–666. 97

[33] Cooper Joel. M. Nathan Medeiros-Ward and D. L. Strayer, "The impact of eye movements and cognitive workload on lateral position variability in driving," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 55, no. 5, pp. 1001–1014, 2013. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1177/0018720813480177$\delimiter"026E30F$nhttp://hfs.sagepub.com/content/55/5/1001.short 66

[34] E. Cutrell, M. Czerwinski, and E. Horvitz, "Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance," pp. 263–269, 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.26.418 64

[35] M. Czerwinski, E. Horvitz, and S. Wilhite, "A Diary Study of Task Switching and Interruptions," in *CHI '04 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*,

vol. 6. New York, New York, USA: ACM Press, apr 2004, pp. 175–182. [Online]. Available: http://dl.acm.org/citation.cfm?id=985692.985715 64

[36] V. Demberg, E. Kiagia, and A. Sayeed, "Language and cognitive load in a dual task environment," in *Proceedings of the 35th annual meeting of the cognitive science society (cogsci-13)*, 2013. 69

[37] J. M. Dimicco, "Changing Small Group Interaction through Visual Reflections of Social Behavior by," PhD Thesis, Massachusetts Institue of Technology, 2005. [Online]. Available: http://web.media.mit.edu/{~}joanie/thesis/ 34

[38] J. M. DiMicco, A. Pandolfo, and W. Bender, "Influencing group participation with a shared display," in *Proceedings of the 2004 ACM conference on Computer supported cooperative work CSCW 04*. New York, New York, USA: ACM Press, nov 2004, pp. 614–623. [Online]. Available: http://dl.acm.org/citation.cfm?id=1031607.1031713 58

[39] T. Dingler and S. Brewster, "AudioFeeds: a mobile auditory application for monitoring online activities," in *Proc. MM 2010*, ser. MM '10. ACM, 2010, pp. 1067–1070. [Online]. Available: http://dx.doi.org/10.1145/1873951.1874151 30

[40] F. a. Drews, M. Pasupathi, and D. L. Strayer, "Passenger and cell phone conversations in simulated driving." *Journal of experimental psychology. Applied*, vol. 14, pp. 392–400, 2008. [Online]. Available: http://psycnet.apa.org/journals/xap/14/4/392/ 63, 66

[41] E. Durkheim, *The division of labor in society*. Simon and Schuster, 2014. 2

[42] A. Edwards, "Soundtrack: An Auditory Interface for Blind Users," *Human-Computer Interaction*, vol. 4, no. 1, pp. 45–66, mar 1989. [Online]. Available: http://dx.doi.org/10.1207/s15327051hci0401{_}2 30

[43] C. Endres, "Real-time Assessment of Driver Cognitive Load as a prerequisite for the situation-aware Presentation Toolkit PresTK," in *Adjunct Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2012), Portsmouth, New Hampshire, USA*, 2012, pp. 76–79. 96

[44] R. W. Engle, "Working memory capacity as executive attention," *Current Directions in Psychological Science*, vol. 11, no. 1, pp. 19–23, feb 2002. [Online]. Available: http://cdp.sagepub.com/content/11/1/19.short 72

[45] J. Engström, E. Johansson, and J. Östlund, "Effects of visual and cognitive load in real and simulated motorway driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 2, pp. 97–120, 2005. 62, 65

[46] T. Erickson, T. Erickson, W. a. Kellogg, and W. a. Kellogg, "Social translucence: an approach to designing systems that support social processes," *ACM Trans. Comput.-Hum. Interact.*, vol. 7, no. 1, pp. 59–83, mar 2000. [Online]. Available: http://portal.acm.org/citation.cfm?doid=344949.345004 28, 53

[47] J. E. Fischer, N. Yee, V. Bellotti, N. Good, S. Benford, and C. Greenhalgh, "Effects of content and time of delivery on receptivity to mobile interruptions," in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services - MobileHCI '10*. New York, New York, USA: ACM Press, sep 2010, p. 103. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1851600.1851620 64

[48] J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang, "Predicting human interruptibility with sensors," *ACM Transactions on Computer-Human Interaction*, vol. 12, no. 1, pp. 119–146, mar 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1057237.1057243 67

[49] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 5, 2002. 10

[50] A. G. Francis, M. Mehta, and A. Ram, "Handbook of Research on Synthetic Emotions and Sociable Robotics:," in *International Journal of Synthetic Emotions*, 2010, vol. 1, ch. 08, pp. 391–421. [Online]. Available: http://www.igi-global.com/bookstore/chapter.aspx?TitleId=21518 7

[51] T. K. Fredericks, S. D. Choi, J. Hart, S. E. Butt, and A. Mital, "An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads," *International Journal of Industrial Ergonomics*, vol. 35, no. 12, pp. 1097–1107, 2005. 68

[52] D. Furniss, "Microwave racing." in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, ser. CHI EA '11. New York, NY, USA: ACM, 2011, p. 497. [Online]. Available: http://discovery.ucl.ac.uk/1325733/ 17

[53] W. Gaver, "The SonicFinder: An Interface That Uses Auditory Icons," *Human-Computer Interaction*, vol. 4, no. 1, pp. 67–94, 1989. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327051hci0401{_}3 30

[54] W. W. Gaver, "Sound support for collaboration," in *Proceedings of the Second European Conference on Computer-Supported Collaborative Work*, ser. ECSCW'91. Kluwer Academic Publishers, 1991, pp. 293–308. [Online]. Available: http://dl.acm.org/citation.cfm?id=1241910.1241932 27, 31

[55] W. W. Gaver and R. B. Smith, "Auditory Icons in Large-Scale Collaborative Environments," in *ACM SIGCHI Bulletin*, D. Diaper, D. J. Gilmore, G. Cockton, and B. Shackel, Eds., vol. 23. North-Holland, 1990, p. 96. 30

[56] W. W. Gaver, R. B. Smith, and T. O'Shea, "Effective sounds in complex systems," in *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91*, ser. CHI '91. ACM, 1991, pp. 85–90. [Online]. Available: http://portal.acm.org/citation.cfm?doid=108844.108857 30

[57] W. W. Gibbs, "Considerate computing." pp. 40–47, 2005. 5

[58] D. Gibson and R. F. Bales, *Social Interaction Systems: Theory and Measurement*. Transaction Publishers, 2000, vol. 29. [Online]. Available: http://books.google.com/books?hl=en{&}lr={&}id=JIX{_}OpLIHYcC{&}pgis=1 33

[59] A. Giddens, *Modernity and self-identity: Self and society in the late modern age*. Stanford University Press, 1991. 10

[60] N. J. Goldstein, R. B. Cialdini, and V. Griskevicius, "A room with a viewpoint: Using social norms to motivate environmental conservation in hotels," *Journal of consumer Research*, vol. 35, no. 3, pp. 472–482, 2008. 16

[61] I. Graham, *Understanding computers and cognition: A new foundation for design*. Norwood, NJ: Ablex Publishing Corporation, 1988, vol. 27. [Online]. Available: http://books.google.com/books?hl=en{&}lr={&}id=2sRC8vcDYNEC{&}oi=fnd{&}pg=PR11{&}dq=winograd+flores{&}ots=20pzUinmUj{&}sig=tR7d4ytvbEq2CA9DdG6QriFVuvM 2

[62] J. Gratch, "Editorial," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 1–10, jan 2010. [Online]. Available: http://www.computer.org/portal/web/csdl/abs/trans/ta/2010/01/tta201001toc.htm 3

[63] M. Grist, "Steer: mastering our behaviour through instinct, environment and reason," *London: RSA*, 2010. 10

[64] C. Gutwin, O. Schneider, R. Xiao, and S. Brewster, “Chalk sounds,” in *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, ser. CSCW '11. ACM, 2011, p. 85. [Online]. Available: http://dl.acm.org/citation.cfm?id=1958824.1958838 27, 32

[65] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, “Psycho-physiological measures for assessing cognitive load,” in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 301–310. 68

[66] D. Halbe, “A Qualitative Analysis of Differences in the Features of Telephone and Face-to-Face Conferences,” in *Association for Business Communication*, 2008. [Online]. Available: http://scholar.google.com/scholar?cluster=2377199832961846045{&}hl=en{&}as{_}sdt=0,5{#}1 26

[67] ——, “"Who's there?": Differences in the Features of Telephone and Face-to-Face Conferences,” *Journal of Business Communication*, vol. 49, no. 1, pp. 48–73, 2012. [Online]. Available: http://job.sagepub.com/content/49/1/48.abstract 26

[68] S. Harrison, D. Tatar, and P. Sengers, “The three paradigms of HCI,” in *Alt. Chi. Session at the SIGCHI . . .* , 2007, pp. 1–18. [Online]. Available: http://people.cs.vt.edu/{~}srh/Downloads/HCIJournalTheThreeParadigmsofHCI.pdf 3, 60, 98

[69] E. T. Higgins, “Achieving 'Shared Reality' in the Communication Game: A Social Action That Create; Meaning,” *Journal of Language and Social Psychology*, vol. 11, no. 3, pp. 107–131, sep 1992. [Online]. Available: http://jls.sagepub.com/content/11/3/107.abstract 25

[70] P. J. Hinds and D. E. Bailey, “Out of Sight, Out of Sync: Understanding Conflict in Distributed Teams,” *Organization Science*, vol. 14, pp. 615–632, 2003. [Online]. Available: http://www.jstor.org/stable/4135124 26

[71] D. Hindus, M. S. Ackerman, S. Mainwaring, and B. Starr, “Thunderwire,” in *Proceedings of the 1996 ACM conference on Computer supported cooperative work - CSCW '96*, ser. CSCW '96. ACM, 1996, pp. 238–247. [Online]. Available: http://portal.acm.org/citation.cfm?doid=240080.240262 32

[72] M. Hockenberry, S. Cohen, Z. Ozer, T. Chen, and T. Selker, “Human-Computer Interaction - INTERACT 2005,” in *IFIP International Federation for Information Processing 2005*, M. F. Costabile and F. Paternò, Eds., vol. 3585. INTERACT'05, 2005, pp. 1079–1082. [Online]. Available: http://link.springer.com/10.1007/11555261 37

[73] E. Hoffman, K. a. McCabe, and V. L. Smith, "Behavioral Foundation of Reciprocity: Experimental Economics and Evolutionary Psychology," *Economic Inquiry*, vol. 36, pp. 335–352, 1998. [Online]. Available: http://doi.wiley.com/10.1111/j.1465-7295.1998.tb01719.x 2

[74] L. R. Hoffman, "Applying Experimental Research on Group Problem Solving to Organizations," *The Journal of Applied Behavioral Science*, vol. 15, no. 3, pp. 375–391, jul 1979. [Online]. Available: http://jab.sagepub.com 33

[75] W. J. Horrey, C. D. Wickens, and a. L. Alexander, "The Effects of Head-Up Display Clutter and In-Vehicle Display Separation on Concurrent Driving Performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 47, no. 16. SAGE Publications, 2003, pp. 1880–1884. 66

[76] W. J. Horrey and C. D. Wickens, "Examining the impact of cell phone conversations on driving using meta-analytic techniques." *Human factors*, vol. 48, pp. 196–205, 2006. [Online]. Available: http://hfs.sagepub.com/content/48/1/196.short 65, 67, 77, 78

[77] ——, "In-Vehicle Glance Duration: Distributions, Tails, and Model of Crash Risk," *Transportation Research Record*, vol. 2018, no. 1, pp. 22–28, 2008. 66

[78] W. J. Horrey, C. D. Wickens, and K. P. Consalus, "Modeling drivers' visual attention allocation while interacting with in-vehicle technologies." *Journal of Experimental Psychology: Applied*, vol. 12, no. 2, p. 67, 2006. 77

[79] L. Horstmanshof and M. Power, "Mobile phones, SMS, and relationships," *Humanities & Social Sciences papers*, 2005. [Online]. Available: http://epublications.bond.edu.au/cgi/viewcontent.cgi?article=1076{&}context=hss{_}pubs 65

[80] E. Horvitz and J. Apacible, "Learning and reasoning about interruption," in *Proceedings of the 5th international conference on Multimodal interfaces - ICMI '03*, ser. ICMI '03. New York, NY, USA: ACM, 2003, p. 20. [Online]. Available: http://portal.acm.org/citation.cfm?doid=958432.958440 5

[81] E. Horvitz, J. Apacible, and M. Subramani, "Balancing Awareness and Interruption : Investigation of Notification Deferral Policies," *User Modeling 2005*, pp. 433–437, 2005. [Online]. Available: http://link.springer.com/chapter/10.1007/11527886{_}59 67, 72

[82] E. Horvitz, C. Kadie, T. Paek, and D. Hovel, "Models of attention in computing and communication," *Communications of the ACM*, vol. 46, no. 3, p. 52, mar 2003. [Online]. Available: http://doi.acm.org/10.1145/636772.636798 5, 67

[83] A. Hsu, J. Yang, Y. H. Yilmaz, M. S. Haque, C. Can, and A. E. Blandford, "Persuasive technology for overcoming food cravings and improving snack choices," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 3403–3412. 10

[84] C. Ikehara and M. Crosby, "Assessing Cognitive Load with Physiological Sensors," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE, 2005, pp. 295a—-295a. 68

[85] S. T. Iqbal, P. D. Adamczyk, X. S. Zheng, and B. P. Bailey, "Towards an index of opportunity: understanding changes in mental workload during task execution," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2005, pp. 311–320. 68, 69

[86] S. T. Iqbal and B. P. Bailey, "Oasis," *ACM Transactions on Computer-Human Interaction*, vol. 17, no. 4, pp. 1–28, dec 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1879831.1879833 64, 67

[87] S. T. Iqbal and E. Horvitz, "Disruption and recovery of computing tasks: field study, analysis, and directions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, New York, USA: ACM Press, apr 2007, pp. 677—-686. [Online]. Available: http://dl.acm.org/citation.cfm?id=1240730 67

[88] S. T. Iqbal, E. Horvitz, Y.-c. Ju, and E. Mathews, "Hang on a Sec ! Effects of Proactive Mediation of Phone Conversations While Driving," *Work*, pp. 463–472, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1979008 65, 67

[89] S. T. Iqbal, Y.-C. Ju, and E. Horvitz, "Cars, calls, and cognition: investigating driving and divided attention," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*, pp. 1281–1290, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1753518 66, 72

[90] S. Iqbal and E. Horvitz, "Notifications and awareness: a field study of alert usage and preferences," in *Proc. of ACM CSCW 2010*. New York, New York, USA: ACM Press, feb 2010, pp. 1–4. [Online]. Available: http://portal.acm.org/citation.cfm?id=1718926 64

[91] D. B. Jayagopi, "Computational modeling of face-to-face social interaction," PhD Thesis, Ecole Polytechnique F{\'{e}}d{\'{e}}rale de Lausanne, 2011. 40, 41, 43

[92] D. Jones, "Recent advances in the study of human performance in noise," *Environment International*, vol. 16, no. 4âĂŞ6, pp. 447–458, 1990. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016041209090013V 58

[93] D. Kahneman, *Attention and effort*. Citeseer, 1973. 68

[94] S. Kim, J. Chun, and A. K. Dey, “Sensors Know When to Interrupt You in the Car: Detecting Driver Interruptibility Through Monitoring of Peripheral Interactions,” in *CHI ’15 Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 487–496. 63

[95] T. Kim, A. Chang, L. Holland, and a.S. Pentland, “Meeting mediator: enhancing group collaborationusing sociometric feedback,” in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, ser. CSCW ’08, 2008, pp. 457–466. [Online]. Available: http://dl.acm.org/citation.cfm?id=1460636 28, 58

[96] J. M. Klingner, “Measuring cognitive load during visual tasks by combining pupillometry and eye tracking,” Ph.D. dissertation, Stanford University, 2010. 68, 69

[97] S. Kousidis, C. Kennington, T. Baumann, H. Buschmeier, S. Kopp, and D. Schlangen, “A Multimodal In-Car Dialogue System That Tracks The Driver’s Attention,” in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 26–33. 66

[98] T. T. Kubose, K. Bock, G. S. Dell, S. M. Garnsey, A. F. Kramer, and J. Mayhugh, “The effects of speech production and speech comprehension on simulated driving performance,” *Applied cognitive psychology*, vol. 20, no. 1, pp. 43–63, 2006. 66

[99] J. D. Lee, B. Caven, S. Haake, and T. L. Brown, “Speech-based interaction with in-vehicle computers: the effect of speech-based e-mail on drivers’ attention to the roadway.” *Human factors*, vol. 43, pp. 631–640, 2001. [Online]. Available: http://hfs.sagepub.com/content/43/4/631.short 65

[100] L. Leiva, M. Böhmer, S. Gehring, and A. Krüger, “Back to the app: The Costs of Mobile Appication Interruptions,” *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services - MobileHCI ’12*, pp. 291–294, 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2371574.2371617 64

[101] T. C. Leonard, “Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness,” *Constitutional Political Economy*, vol. 19, no. 4, pp. 356–360, 2008. 10

[102] J. Levy and H. Pashler, “Task prioritisation in multitasking during driving: Opportunity to abort a concurrent task does not insulate braking responses from dual-task slowing,” *Applied Cognitive Psychology*, vol. 22, pp. 507–525, 2008. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/acp.1378/full 67

[103] C. Liu, K. Conn, N. Sarkar, and W. Stone, “Online affect detection and robot behavior adaptation for intervention of children with autism,” *Robotics, IEEE Transactions on*, vol. 24, no. 4, pp. 883–896, 2008. 68

[104] P. Maes, “How to do the Right Thing,” Cambridge, MA, USA, pp. 291–323, 1989. 7

[105] A. Mahr, M. Feld, M. M. Moniri, and R. Math, “The ConTRe (Continuous Tracking and Reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity,” *Adjunct Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications.*, pp. 88–91, 2012. [Online]. Available: http://www.dfki.de/web/forschung/publikationen/renameFileForDownload?filename=ConTRe{_}WS.pdf{&}file{_}id=uploads{_}1846 62, 69, 71

[106] P. Manalavan, A. Samar, M. Schneider, S. Kiesler, and D. Siewiorek, “In-car cell phone use: mitigating risk by signaling remote callers,” *CHI '02 extended abstracts on Human factors in computer systems - CHI '02*, p. 790, 2002. [Online]. Available: http://portal.acm.org/citation.cfm?doid=506443.506599 67

[107] J. Mariani and R. Ramloll, “Do Localised Auditory Cues in Group Drawing Environments Matter?” in *Proc. ICAD 1998*, ser. ICAD'98. British Computer Society, 1998, p. 24. [Online]. Available: http://eprints.lancs.ac.uk/11653/ 31, 36

[108] S. P. Marshall, “Identifying cognitive state from eye metrics,” *Aviation, space, and environmental medicine*, vol. 78, no. Supplement 1, pp. B165—-B175, 2007. 69

[109] T. Matthews, T. Rattenbury, and S. Carter, “Defining, designing, and evaluating peripheral displays: An analysis using activity theory,” *Human–Computer Interaction*, vol. 22, no. 1-2, pp. 221–261, 2007. 6

[110] D. McFarlane, “Comparison of Four Primary Methods for Coordinating the Interruption of People in Human-Computer Interaction,” *Human-Computer Interaction*, vol. 17, no. 1, pp. 63–139, mar 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=1464473.1464475 67

[111] D. McGookin and S. Brewster, “An Initial Investigation into Non-Visual Computer Supported Collaboration,” in *Ext. Abstracts CHI 2007*, ser. CHI EA '07. ACM, 2007, pp. 2573–2578. [Online]. Available: http://eprints.gla.ac.uk/43819/ 31

[112] ——, “Pulse,” in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, ser. CHI '12. ACM, 2012, p. 1263. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2207676.2208580 30

[113] M. McLuhan, *Understanding media: The extensions of man*. MIT press, 1994. 1

[114] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956. 68

[115] L. J. M. Mulder, "Measurement and Analysis-Methods of Heart-Rate and Respiration for Use in Applied Environments," *Biological Psychology*, vol. 34, no. 2, pp. 205–236, 1992. [Online]. Available: {\T1\textless}GotoISI{\T1\textgreater}://A1992KA29300007 68

[116] C. Nass and K. M. Lee, "Does computer-generated speech manifest personality? an experimental test of similarity-attraction," in *Proceedings of the SIGCHI conference on Human factors in computing systems CHI 00*, ser. CHI '00, vol. 2. New York, NY, USA: ACM, 2000, pp. 329–336. [Online]. Available: http://portal.acm.org/citation.cfm?doid=332040.332452 17

[117] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of social issues*, vol. 56, no. 1, pp. 81–103, 2000. 9

[118] C. I. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Computer-Human Interaction (CHI) Conference: Celebrating Interdependence 1994*, ser. CHI '94. New York, NY, USA: ACM, 1994, pp. 72–78. [Online]. Available: http://doi.acm.org/10.1145/191666.191703 9

[119] D. T. Nguyen and J. Canny, "Multiview," in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, ser. CHI '07. ACM, 2007, p. 1465. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1240624.1240846 26

[120] L. Nunes and M. A. Recarte, "Cognitive demands of hands-free-phone conversation while driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 5, pp. 133–144, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1369847802000128 66

[121] G. Olson and J. Olson, "Distance Matters," *Human-Computer Interaction*, vol. 15, no. 2, pp. 139–178, sep 2000. [Online]. Available: http://dx.doi.org/10.1207/S15327051HCI1523{_}4 26

[122] C. K. L. Or and V. G. Duffy, "Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement," *Occupational Ergonomics*, vol. 7, no. 2, pp. 83–94, 2007. 68

[123] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational psychologist*, vol. 38, no. 1, pp. 63–71, 2003. 68

[124] V. Pejovic and M. Musolesi, "InterruptMe," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*. New York, New York, USA: ACM Press, sep 2014, pp. 897–908. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2632048.2632062 67

[125] M. Pielot, K. Church, and R. de Oliveira, "An In-Situ Study of Mobile Phone Notifications," in *Proc. MobileHCI '14*. New York, New York, USA: ACM Press, sep 2014, pp. 233–242. [Online]. Available: http://dl.acm.org/citation.cfm?id=2628363.2628364 64

[126] R. E. Potter, R. a. Cooke, and P. a. Balthazard, "Virtual team interaction: assessment, consequences, and management," *Team Performance Management*, vol. 6, no. 7/8, pp. 131–137, 2000. [Online]. Available: http://elibrary.ru/item.asp?id=6511275 28

[127] Z. Pousman and J. Stasko, "A taxonomy of ambient information systems: four patterns of design," in *Proceedings of the working conference on Advanced visual interfaces*. ACM, 2006, pp. 67–74. 6

[128] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and brain sciences*, vol. 1, no. 04, pp. 515–526, 1978. 10

[129] R. Rajan, C. Chen, and T. Selker, "Considerate Audio MEdiating Oracle (CAMEO)," in *Proceedings of the Designing Interactive Systems Conference on - DIS '12*, ser. DIS '12. ACM, 2012, p. 86. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2317956.2317972 27

[130] R. Rajan, J. Hsiao, D. Lahoti, and T. Selker, ""Roger that!" - The value of adding social feedback in audio-mediated communications," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8120 LNCS, 2013, pp. 471–488. 27

[131] R. Rajan, T. Selker, and I. Lane, "Task Load Estimation and Mediation Using Psycho-physiological Measures," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, 2016, pp. 48–59. 18

[132] D. a. Redelmeier and R. J. Tibshirani, "Association between cellular-telephone calls and motor vehicle collisions." *The New England journal of medicine*, vol. 336, pp. 453–458, 1997. [Online]. Available: http://www.nejm.org/doi/full/10.1056/nejm199702133360701 65, 66

[133] R. Rienks, A. Nijholt, and P. Barthelmess, "Pro-active meeting assistants: Attention please!" *AI and Society*, vol. 23, no. 2, pp. 213–231, aug 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1403860.1403864 29

[134] D. I. Rigas, D. Hopwood, and D. Memery, "Communicating spatial information via a multimedia-auditory interface," in *Conference Proceedings of the EUROMICRO*, vol. 2, 1999, pp. 398–405. 27, 30

[135] M. Rudary, S. Singh, and M. Pollack, "Adaptive cognitive orthotics: combining reinforcement learning and constraint-based temporal reasoning," in *Procedings of the International conference on Machine learning*. ACM, 2004, pp. 91–98. [Online]. Available: http://dl.acm.org/citation.cfm?id=1015411 53

[136] K. Ryu and R. Myung, "Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic," *International Journal of Industrial Ergonomics*, vol. 35, no. 11, pp. 991–1009, 2005. 68

[137] A. Sahami Shirazi, N. Henze, T. Dingler, M. Pielot, D. Weber, and A. Schmidt, "Large-scale assessment of mobile notifications," *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pp. 3055–3064, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2556288.2557189 64

[138] D. Salvucci, D. Markley, M. Zuber, and D. Brumby, "iPod Distraction: Effects of Portable Music-Player Use on Driver Performance," *Proceedings of the ...*, 2007. [Online]. Available: http://discovery.ucl.ac.uk/125955/ 65

[139] B. D. Sawyer, C. Florida, V. S. Finomore, A. a. Calvo, B. Aerospace, W. Patterson, A. Force, P. a. Hancock, and C. Florida, "Google Glass : A Driver Distraction Cause or Cure ?" *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 56, no. 7, pp. 1307–1321, oct 2014. [Online]. Available: http://hfs.sagepub.com/content/early/2014/10/15/0018720814555723.abstract 66

[140] C. Schlienger, S. Conversy, S. Chatty, M. Anquetil, and C. Mertz, "Improving Users' Comprehension of Changes with Animation and Sound: An Empirical Assessment," in *IFIP Lecture Notes in Computer Science (LNCS)*, ser. INTERACT'07, vol. 4662. Springer-Verlag, 2007, pp. 207–220. [Online]. Available: http://dl.ifip.org/index.php/lncs/article/view/27878 30

[141] T. Selker, "Understanding considerate systems - UCS (pronounced: You See Us)," in *Proceedings - 2011 Conference on Technologies and Applications of Artificial Intelligence, TAAI 2011*. IEEE, 2011, pp. 1–12. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs{_}all.jsp?arnumber=5478532 27, 32

[142] A. Sellen, "Remote Conversations: The Effects of Mediating Talk With Technology," *Human-Computer Interaction*, vol. 10, no. 4, pp. 401–444, dec 1995. [Online]. Available: http://dx.doi.org/10.1207/s15327051hci1004{_}2 26, 35

[143] Y. Shi, T. Park, N. Ruiz, R. Taib, E. H. C. Choi, and F. Chen, "Galvanic Skin Response (GSR) as an Index of Cognitive Load," in *CHI EA '07 CHI '07 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2007, pp. 2651–2656. 68, 69

[144] D. Skierkowski and R. M. Wood, "To text or not to text? the importance of text messaging among college-aged youth," *Computers in Human Behavior*, vol. 28, pp. 744–756, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563211002676 65

[145] S. Strachan, R. Murray-Smith, and S. O'Modhrain, "gpsTunes - controlling navigation via audio feedback," in *Proc. MobileHCI 2005*, ser. MobileHCI '05. ACM, 2005, pp. 275–278. [Online]. Available: http://eprints.pascal-network.org/archive/00001267/ 30

[146] D. L. Strayer and W. a. Johnston, "Driven to distraction: dual-Task studies of simulated driving and conversing on a cellular telephone." *Psychological science : a journal of the American Psychological Society / APS*, vol. 12, pp. 462–466, 2001. [Online]. Available: http://pss.sagepub.com/content/12/6/462.short 65, 66

[147] D. L. Strayer, J. M. Watson, and F. a. Drews, "Cognitive Distraction While Multitasking in the Automobile," *Psychology of Learning and Motivation - Advances in Research and Theory*, vol. 54, pp. 29–58, 2011. [Online]. Available: https://books.google.com/books?hl=en{&}lr={&}id=6L9bl-JrQD4C{&}oi=fnd{&}pg=PA29{&}dq=related:43tyJe1XuNPF2M:scholar.google.com/{&}ots={_}tBqxXSODo{&}sig=DJcWsqueVxmZMSBPXZjCEluN8QE 62

[148] J. C. Tang, "Why do users like video? Study of multimedia supported collaboration," *Proc. CSCW 1992*, vol. 1, no. 3, pp. 163–196, 1992. [Online]. Available: http://dl.acm.org.ezproxy.uct.ac.za/citation.cfm?id=974914 26

[149] T. Taylor, a. K. Pradhan, G. Divekar, M. Romoser, J. Muttart, R. Gomez, a. Pollatsek, and D. L. Fisher, "The view from the road: The contribution of on-road glance-monitoring technologies to understanding driver behavior," *Accident Analysis and Prevention*, vol. 58, pp. 175–186, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457513000456 65

[150] P. Tchankue, J. Wesson, and D. Vogts, "The Impact of an Adaptive User Interface on Reducing Driver Distraction," in *rd International Conference on Automotive User Interfaces and Interactive*

Vehicular Applications. New York, New York, USA: ACM Press, nov 2011, pp. 87–94. [Online]. Available: http://dl.acm.org/citation.cfm?id=2381416.2381430 96

[151] K. G. Tippey, E. Sivaraj, W.-J. Ardoin, T. Roady, and T. K. Ferris, "Texting while driving using Google Glass: Investigating the combined effect of heads-up display and hands-free input on driving safety and performance," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58, no. 1, pp. 2023–2027, oct 2014. [Online]. Available: http://pro.sagepub.com/content/58/1/2023.abstract 66

[152] a. J. Trevino and J. H. Turner, "Handbook of Sociological Theory," in *Contemporary Sociology*, ser. Handbooks of Sociology and Social Research, J. H. Turner, Ed. Springer US, 2003, vol. 32, ch. 15, p. 282. [Online]. Available: http://www.springerlink.com/content/w3k04713114l7621/ 59

[153] M. van Dooren, J. J. G. G.-J. de Vries, and J. H. Janssen, "Emotional sweating across the body: Comparing 16 different skin conductance measurement locations," *Physiology & Behavior*, vol. 106, no. 2, pp. 298–304, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031938412000613 83

[154] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009. 3, 7, 33, 34

[155] R. T. Warne, "A Primer on Multivariate Analysis of Variance (MANOVA) for Behavioral Scientists," *Practical Assessment, Research & Evaluation*, vol. 19, no. 17, p. 2, 2014. 75, 92

[156] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992. 54

[157] M. Weber, *Economy and society: An outline of interpretive sociology*. Univ of California Press, 1978. 10

[158] M. Weiser, "The Computer for the 21st Century," *Scientific American*, vol. 265, pp. 94–104, 1991. [Online]. Available: http://wiki.daimi.au.dk/pca/{_}files/weiser-orig.pdf 98

[159] M. Weiser and J. S. Brown, "The coming age of calm technology," in *Beyond calculation*. Springer, 1997, pp. 75–85. 6

[160] C. D. Wickens, "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science*, vol. 3, pp. 159–177, 2002. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/14639220210123806 67

[161] ——, "Multiple resources and mental workload," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 3, pp. 449–455, 2008. 65

[162] B. Wilpert, *Organizational behavior.*, 10th ed. South-Western College Pub, 1995, vol. 46. 33

[163] G. F. Wilson, "An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, 2002. 68

[164] N. Yankelovich, N. Yankelovich, W. Walker, W. Walker, P. Roberts, P. Roberts, M. Wessler, M. Wessler, J. Kaplan, J. Kaplan, J. Provino, and J. Provino, "Meeting central: making distributed meetings more effective," *Proceedings of the 2004 {ACM} conference on Computer supported cooperative work*, pp. 419–428, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?id=1031607.1031678 26, 28, 33, 35, 36

[165] S. Young, "Cognitive user interfaces," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 128–140, may 2010. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs{_}all.jsp?arnumber=5447049 7

[166] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell, "Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior," in *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 2008, pp. 322–331. 96