"Roger that!" — The Value of Adding Social Feedback in Audio-mediated Communications

Rahul Rajan[1], Joey Hsiao[2], Deven Lahoti[3], Ted Selker[1]
[1]Carnegie Mellon University, Pittsburgh, PA, U.S.A.
[2]National Taiwan University, Taipei City, Taiwan
[3]Independent

rahul.rajan@sv.cmu.edu, ted.selker@sv.cmu.edu, hideysmk@gmail.com
deven.lahoti@gmail.com

Abstract. Losing track of who is in a conversation, and what is being said, is always a problem especially on audio-only conference calls. This paper investigates how domain-independent social feedback can support such interactions, and improve communication, through the use of audio cues. In particular, we show how an agent can improve people's ability to accurately identify and distinguish between speakers, reassure users about the presence of other collaborators on the line, and announce events like entry & exit with minimum impact on users cognitive ability.

Keywords: Audio-mediated, Conference calls, Considerate, Social feedback

1 Introduction

Communication is a type of social action [1]. 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 [2]. 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 [3], whereas desktop videoconferences are more popular in personal settings [4]. Even so, the reasons for users choice and preference are nuanced and complex, involving multiple tradeoffs related to intrusion, amplification of inattention, and mobility. While there has been significant work and progress to preserve social

cues in video communications [5], audio-mediated technologies have not received the same share of attention, and depend largely on visual cues like participant lists to buttress communications [6].

In this work, we explore how social cues can be restored on the audio channel, while addressing some of its most often cited drawbacks [7, 2, 6]. These are predominantly social in nature, and focus on the process of interaction. These include 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 feed- back and feedforward techniques 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 work well in spite of arguably have the lowest bandwidth for natural synchronous communications between multiple people. The difficulty in improving these communications with an agent further loading this narrow channel should then be maximally hard, making it an ideal place to demonstrate the utility of an agent being considerate and appropriate [8].

A second reason to experiment with augmenting the audio space is that many discussions and meetings involve documents and physical artifacts that occupy users 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.

Thirdly, people can perceive multiple audio channels simultaneously, and do so with considerable ease, especially while listening to music. 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" [9]. Thus, the audio channel can be used as an effective mode of transmitting background information (e.g., [10, 11]). 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., [12–14]).

To evaluate feedback techniques for improving communications, we choose the multiparty audio conference call with a shared-screen as our setting. Previous work demonstrates that considerate agent cues can reduce distraction, and help conference call participants equalize contributions [15]. This paper moves further to show that considerate agents like CAMEO can more generally make call participants aware of others on the line and allow them to focus better on the conversation. In particular, we focus on speaker identification, audio presence, and entry & exit announcements. To evaluate speaker identification methods, we had participants listen to a pre-recorded five-person conference call and an-

swer identity-related questions. We found that while speaker identification is hard, audio augmentations or spatially arranging the different speakers, could aid with identification success. We then show that by feed-forwarding simple audio cues, users were more assured, and less distracted about the presence of other collaborators. Finally, we had participants engage in a memory game while subjecting them to three kinds of entry/exit prompts. We show that by making a more natural, and less syntactic utterance, participants made fewer errors on average.

2 Related Work

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.

2.1 Auditory Interface Design

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 focussed class of audio cues, which are carefully designed to support a semantic link with the object they represent, making them easier to associate. Furthermore, sounds can be perceptually mapped to the events they indicate using symbolic, metaphorical and iconic methods.

Soundtrack [16] was one of the first auditory interfaces to use earcons and synthetic speech. More recently, Rigas et al. [12] 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 [17] 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 [18] 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. [19] 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 [20] explored how audio can be used to monitor social network activity, and PULSE [21] 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.

2.2 Activity Coordination in a Distributed Setting

SoundShark [22] 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 [23]. 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 [24] 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 [25] 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 persons presence is crucial to the outcome.

2.3 Shared Awareness in a Distributed Setting

The Environmental Audio Reminders (EAR) system [10] 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 [13] 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.

The OutToLunch system [11] 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 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 [26]. In Thunderwire [27], 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 [14], 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.

3 AUDIO DESIGN FOR CAMEO

Synthesizing prior work on audio signals, we experiment with feedback techniques to support the social process of communication. Specifically, we incorporate them in the design of three CAMEO features — Speaker Identification, Audio Presence, and Entry/Exit announcements [15]. We focus on the functionality and design of assistive audio cues for each of these features.

Adding audio cues on to an already overloaded communication channel with multiple speakers provided ample opportunity to be distracting and inconsiderate. We had to try many approaches and various kinds of audio feedback cues to find reasonable candidates to formally evaluate. We chose to work with Apple's GarageBand[1] software. We pre-recorded two ten-minute teleconference calls between five people and used these as the base tracks in Garageband. This was then overlaid with earcons and auditory icons, which included instrument sounds from http://free-loops.com/, and nature sounds from http://www.naturesoundsfor.me/. In the following paragraphs, we summarize the lessons learnt from the broad explorations in the designs of each of these three features.

3.1 Speaker Identity

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 [2, 6]. 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.

We focussed 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 too much crosstalk on the channel.

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. With regards to the duration, a cue that was a second long was distracting, whereas a cue that was 0.25 sec long was too subtle to discriminate from the other cues. Thus, the duration of the cues was set to be 0.5 sec.

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

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

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.

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 [24], 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 second CAMEO feature, i.e. to indicated presence of other collaborators on the line to a user.

3.2 Audio 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 thats picked up from the mouthpiece and instantly introduce 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. There are various methods to present such information visually [6], but the visual channel comes at a 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 [11]. The tracks would play while the user was online, but they overloaded the conversational channel and were, therefore, highly distracting.

To soften the crosstalk, we were inspired by radio shows where people call in to chat with the host. They are usually located in different audio environments, and it's easy to tell them apart. We improvised on this by experimenting with different nature sounds like from a flowing stream, waves landing on the shore of

a beach, and the chirping of birds. These blended so well into the background, though, that it was hard to notice when they stopped or started.

An alternative idea was to employ a roll cqall, 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.

In the final subsection, we consider multiple entry & exit announcements for CAMEO's Entry/Exit feature, which we describe below.

3.3 Entry/Exit

It can be hard to tell when participants get dropped from or reenter a conference call. Some existing conference call systems circumvent this by announcing entry and exit with a lengthy statement, which can be annoying. In this work we explore alternate iconic and metaphoric prompts to announce these events.

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 door open and the door close sounds. We obviated this by appending the sounds to the name using different sequences. 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.

We wanted to compare this to more metaphorical mapping approaches, like using fade and intonations. We modified the TTS in Audacity[2] 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.

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.

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

4 STUDY I: 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.

4.1 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.

4.2 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.

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 1). Overall Accuracy: SEM=(0.024, 0.041); Attempted Accuracy: SEM=(0.046, 0.044); Response Time: SEM=(151.2, 163.7); N=16.

Table 1. Accuracy metrics and average response times for speaker identification with and without 2D spatialization.

	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

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 2). Overall Accuracy: SEM=(0.052, 0.024); Attempted Accuracy: SEM=(0.054, 0.054); Response Time: SEM=(169.0, 111.5); N=16.

Table 2. Accuracy metrics and average response times for speaker identification with and without earcons.

	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

Summary 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 distractive.

5 STUDY II: Audio 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.

5.1 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.

5.2 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 by 37% in the condition where the auditory icons were used ($p<0.05$, 2-tailed t-test, Table 3). There was no significant difference in the attempt rate or error rate. # of Nudges: SEM=(0.72, 0.64); N=10.

Table 3. 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

6 STUDY III: 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).

6.1 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.

Table 4. 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)

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 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[3] 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). 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.

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

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).

6.2 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 5). 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.

Table 5. 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 5. 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.

7 DISCUSSION

The three studies 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.

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, or notice the quiet person in the room when engaged in conversation, especially when they aren't visible. Similarly, keeping track of multiple things while coordinating with others is difficult when there are a lot of distractions in the environment. As interactive systems weave themselves tighter into our social fabric, they need to be designed so as to accommodate such typically complex social situations. 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. We are very excited about being able to add audio to an already loaded channel and improve the perceived and real understanding of the situation.

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.

8 CONCLUSIONS and FUTURE WORK

The work described here focuses on domain-independent social feedback. We show how careful choice of its syntax — its sound and placement (like long utterances), and semantics — its direct and indirect relationship to the conversational channel (such as putting speech on top of a conversational channel), can deeply affect its goal of supporting social interactions. In particular, this paper focuses on how 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 show that using metaphorical prompts (intonations) to announce events like entry and exit of participants reduces errors when compared to speech prompts.

With the proliferation of phones and tablets, more and more interactive systems are being used in social settings. The imperative now is for technology that celebrates situational awareness, and appropriateness. While this work builds a vocabulary of effectors to improve teleconference meetings, we look forward to the opportunities afforded by developing an understanding of how to accommodate system feedback in variety of other social situations. The choices of when and how 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.

References

1. Higgins, E.T.: Achieving 'shared reality' in the communication game: A social action that creates meaning. Language and Social Psychology 11(3) (September 1992) 107–131
2. Sellen, A.J.: Remote conversations: the effects of mediating talk with technology. HCI 10(4) (December 1995) 401–444
3. Halbe, D.: Whos there?. Business Communication 49(1) (2012) 48–73
4. Brubaker, J.R., Venolia, G., Tang, J.C.: Focusing on shared experiences: moving beyond the camera in video communication. In: Proceedings of the Designing Interactive Systems Conference. DIS '12, ACM (2012) 96–105
5. Nguyen, D.T., Canny, J.: Multiview: improving trust in group video conferencing through spatial faithfulness. In: Proc. CHI 2007. CHI '07, ACM (2007) 1465–1474
6. Yankelovich, N., Walker, W., Roberts, P., Wessler, M., Kaplan, J., Provino, J.: Meeting central: making distributed meetings more effective. In: Proc. CSCW 2004. CSCW '04, ACM (2004) 419–428
7. Tang, J.C., Isaacs, E.: Why do users like video? Proc. CSCW 1992 1(3) (1992) 163–196
8. Selker, T.: Understanding considerate systems – UCS (pronounced: You see us). In: 2010 International Symposium on Collaborative Technologies and Systems, IEEE (2010) 1–12
9. Arons, B.: A Review of The Cocktail Party Effect. The American Voice I/O Society 12 (1992) 35–50
10. Gaver, W.W.: Sound support for collaboration. In: Proc ECSCW 1991. ECSCW'91, Kluwer Academic Publishers (1991) 293–308
11. Cohen, J.: Out to lunch: Further adventures monitoring background activity. In: Proc. ICAD 1994, Santa Fe Institute (1994) 15–20
12. Rigas, D.I., Hopwood, D., Memery, D.: Communicating spatial information via a multimedia-auditory interface. In: Proc. EUROMICRO 1999. Volume 2., IEEE Computer Society (1999) 398 –405 vol.2
13. Cohen, J.: "Kirk here": using genre sounds to monitor background activity. In: Proc CHI 1993. CHI '93, ACM (1993) 63–64
14. Gutwin, C., Schneider, O., Xiao, R., Brewster, S.: Chalk sounds: the effects of dynamic synthesized audio on workspace awareness in distributed groupware. In: Proc. CSCW 2011. CSCW '11, ACM (2011) 85–94

15. Rajan, R., Chen, C., Selker, T.: Considerate Audio MEdiating Oracle (CAMEO): improving human-to-human communications in conference calls. In: Proc. DIS 2012. DIS '12, ACM (2012) 86–95
16. Edwards, A.D.N.: Soundtrack: an auditory interface for blind users. HCI 4(1) (March 1989) 45–66
17. Gaver, W.W.: The SonicFinder: An Interface That Uses Auditory Icons. HCI 4(1) (1989) 67–94
18. Strachan, S., Eslambolchilar, P., Murray-Smith, R., Hughes, S., O'Modhrain, S.: GpsTunes: controlling navigation via audio feedback. In: Proc. MobileHCI 2005. MobileHCI '05, ACM (2005) 275–278
19. Schlienger, C., Conversy, S., Chatty, S., Anquetil, M., Mertz, C.: Improving users' comprehension of changes with animation and sound: an empirical assessment. In: Proc. INTERACT 2007. INTERACT'07, Springer-Verlag (2007) 207–220
20. Dingler, T., Brewster, S.: AudioFeeds: a mobile auditory application for monitoring online activities. In: Proc. MM 2010. MM '10, ACM (2010) 1067–1070
21. McGookin, D., Brewster, S.: PULSE: the design and evaluation of an auditory display to provide a social vibe. In: Proc. CHI 2012. CHI '12, ACM (2012) 1263–1272
22. Gaver, W.W., Smith, R.B.: Auditory icons in large-scale collaborative environments. In: Proc. INTERACT 1990, North-Holland (1990) 735–740
23. Gaver, W.W., Smith, R.B., O'Shea, T.: Effective sounds in complex systems: the ARKOLA simulation. In: Proc. CHI 1991. CHI '91, ACM (1991) 85–90
24. Ramloll, R., Mariani, J.: Do localised auditory cues in group drawing environments matter? In: Proc. ICAD 1998. ICAD'98, British Computer Society (1998) 24
25. McGookin, D., Brewster, S.: An initial investigation into non-visual computer supported collaboration. In: Ext. Abstracts CHI 2007. CHI EA '07, ACM (2007) 2573–2578
26. Beaudouin-Lafon, M., Karsenty, A.: Transparency and awareness in a real-time groupware system. In: Proc. UIST 1992. UIST '92, ACM (1992) 171–180
27. Hindus, D., Ackerman, M.S., Mainwaring, S., Starr, B.: Thunderwire: a field study of an audio-only media space. In: Proc. CSCW 1996. CSCW '96, ACM (1996) 238–247