Graphical Enhancements for Voice Only Conference Calls

R. Alex Colburn, Michael F. Cohen, Steven M. Drucker, Scott LeeTiernan and Anoop Gupta

October 1, 2001

Technical Report
MSR-TR-2001-95

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Graphical Enhancements for Voice Only Conference Calls

R. Alex Colburn, Michael F. Cohen, Steven M. Drucker, Scott LeeTiernan and Anoop Gupta
Collaboration and Multimedia Systems Group
Microsoft Research
One Microsoft Way
Redmond, WA 98052
{alexco, mcohen, sdrucker, anoop}@microsoft.com

ABSTRACT

We present two very low bandwidth graphically enhanced interfaces for small group voice communications. One interface presents static images of the participants that highlight when one is speaking. The other interface utilizes three-dimensional avatars that can be quickly created. Eleven groups of 4 or 5 people were presented with each enhanced interface as well as conducting a live conversation and a voice only conversation. Experiments show that both graphically enhanced interfaces improve the understandability of conversations, particular with respect to impressions that others in the group could express themselves more easily, knowing who is talking, and when to speak. Little difference was found between the two graphical interfaces. Analysis of voice tracks also revealed differences between interfaces in the length and number of medium duration silences.

Keywords
Teleconference, Avatar, Group Communication, User Study

1 INTRODUCTION

Most of us have some experience with telephone conference calls. Although they save one tremendous time and money, compared to meeting in a common place, the voice-only interactions often make it hard to understand who is talking and who is listening when one talks. This problem becomes worse when the number of people in the conference starts becoming large (more than 3) and/or when there are relative strangers involved in the call – our ability to associate voice with individual person becomes harder. In addition, the non-verbal cues that occur naturally when a small group sits around a table are missing and thus even knowing when it is a good time to talk can be difficult.

Video conferencing systems arose as a means to overcome some of these difficulties but such systems exhibit their own problems. These range from the lack of eye contact to the expense of outfitting special rooms, to the latency of communication, and bandwidth considerations for moving video content over distances. In contrast, voice-conferencing over telephone, or voice over IP, coupled with computer mediated document-sharing has begun to appear. Companies like WebEX and Placeware are having very high adoption rates. A recent study showed that use of voice-conferencing plus NetMeeting grew over 10-fold at Boeing corporation from Jan-98 to Jan-01 [16].

This paper reports on experiments to assess very low bandwidth solutions for enhancing voice-only conference calls with graphics interfaces. We are interested in how non-collocated (each in their own office) small meeting conversations can be improved. We focus here solely on the synchronous conversational aspects of a meeting and purposefully do not address issues surrounding additional means of communication such as whiteboards and/or document sharing.

In work environments users have ready availability of the desktop where such graphics can be shown, and even in mobile situations, PDAs and cell phones have screens that can be leveraged. We focus on two interfaces depicted in Figure 1: 1) A simple icon-based interface that consists of photographs of participants and names – the icon and name of a person lights up in others screens when they speak; 2) A more complex avatar-based interface depicting three-dimensional graphical characters (avatars) in a virtual setting – the avatar is personalized, its mouth and head moves when a person talks, the name above it lights up, and all avatars simulate simple motion rules based on who is talking.

We compare these two enhancements to the voice-only condition and live-interaction where all participants are present face to face. We consider medium sized groups of 4 to 5 participants and compare subjective perceptions and analyze individual voice tracks. Practical experiences with a system to rapidly create three-dimensional facial representations of the participants are described.

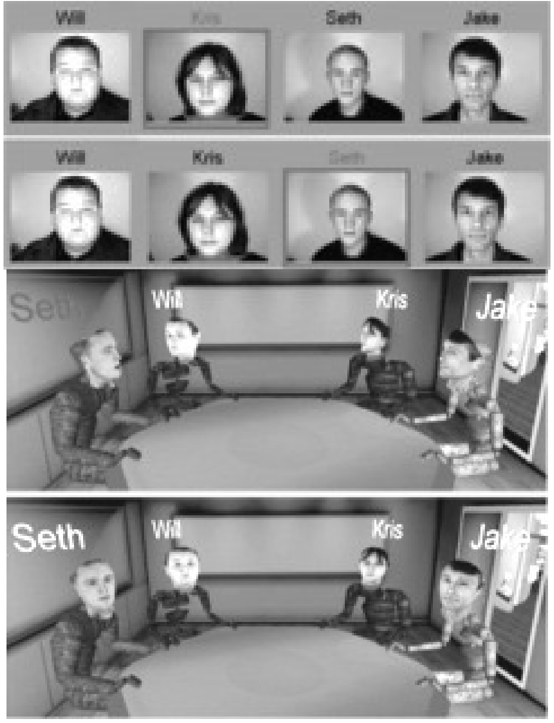


Figure 1: The two images on top depict the static icon *interface at two moments in a conversation when Kris and then Seth is speaking. The two images at bottom show the* avatar *interface at two moments when Seth is speaking and then when "you" are speaking. In both interfaces, the user is not shown.(Note: highlights in red do not show well in greyscale.)*

Findings indicate that having some graphical interface to enhance a voice only exchange can have a significant effect in the subjective measures (e.g., satisfaction, understandability, knowledge of who's talking) of small group conversations. Given the modest technology requirements of our interventions, they can be readily adopted on current hardware to improve user experiences of voice-conference calls.

2 RELATED WORK

There has been extensive examination of the use of audio and video to facilitate group work. Finn [10] provides an excellent compendium of work in the field. He points out that some studies indicate that the use of a video channel has no effect on task performance or user satisfaction. Chapanis et al [7] found that the voice mode makes for the single most important contribution to task completion. [13] also concluded that there is little improvement in performance gained by including any form of video communication channel, and that the audio link often comes close to recreating the face to face situation.

On the other hand, many studies have shown that a visual channel benefits the process, outcome or user experience of the communication such as the work of Olson et al [15]. They examined face-to-face group work compared with remote group work with and without video and found that without video, the satisfaction with the quality of work done is significantly poorer. Since our work is focused on low bandwidth communication, full motion video was not an option. Rather, we examine if satisfaction levels with voice based communication can be made closer to face-to-face meetings. There is also a wide body of work devoted to the effects of transmission delays [12]. We keep transmissions delays to a minimum by using standard teleconferencing systems.

Work by Rodenstein and Donath [17] includes informal studies on using circles with audio drop off based on distance between circles. Graphic changes in the circles helped identify speakers and distance was used to filter the conversation amongst large numbers of participants. Like the work presented here, they found that assistance in speaker identification was an important facet of the system. Attempts have often centered on trying to build systems that support effective eye-gaze, a critical component in face-to-face conversation [1]. For instance, Buxton [3] developed a multi-headed display system for video conferencing which they called the Hydra system, which permitted for important gaze cues. Sellen [19] compared the Hydra System, along with several other video conference applications with face-to-face and voice only conversations. They predicted that for face-to-face and effective video mediated communication, statistical analysis would produce more and shorter turns per session, more equal distribution of turns among speakers, and more simultaneous speech. They did not find as much difference between face-to-face and video mediate conferencing as they expected, though they did find significant differences between audio only and face-to-face conversations. Our avatar interface makes use of eye-gaze behaviors found in real conversations. Other work has examined longer term interactions using systems to link offices together via audio and video systems and support informal awareness [8,20]. The field of conversation analysis provides a framework for comparing the efficacy of different interface modalities based on quantifiable attributes of a conversation. A good reference on this literature is available in Atkinson et al [2]. Connaill and Whitaker looked at the conversation changes under 3 different communication settings [14]. Others [9, 11] have also analyzed the effects that different modalities of communication have on turn taking behavior and on-off patterns often based on a prototypical set of rules described in Sacks et al.

In other recent work, avatars have been suggested as viable proxies for humans in online conversations [4,5,21]. Subtle motions such as appropriate gaze and head-turning behaviors can be important [6]. Our avatar based interface uses a model based on empirical data to support appropriate gaze behaviors based on the auditory channel of the conversation.

3 TECHNOLOGY OVERVIEW

From a technology impact perspective, our goals in providing graphical interfaces to assist in small group voice conferencing were to:

- not rely on new hardware at the user's end, and
- rely only on very low bandwidth solutions.

The first goal is achieved by assuming each participant has a common set of hardware found in most offices. This includes a personal computer with common sound and graphics cards and a telephone. The only additional equipment we used was a telephone headset (although this is strictly not required) and a microphone attached to a sound card in the computer.

The second goal of very low bandwidth precludes capturing and transmitting video. Voice is transmitted with an ordinary telephone. The only extra information transmitted is a one bit wide stream encoding whether a participant is speaking or not (typically already built into analog telephone conferencing systems to achieve "silence suppression"). The talking/not talking determination is made by measuring the energy of the voice signal, and transmitted, over a LAN, with approximately a 100 millisecond delay

3.1 A static image-based interface (icon)

The graphical user interface thus must rely only on who is speaking. The simple image-based interface consisted only of a row of static photo images (collected before the conversations) of the *other* participants in the conversation (i.e., one's own image was not included) with the name in text above. When each participant spoke, the border of the respective image and the text above were highlighted in red.

3.2 An avatar-based interface (avatar)

The more complex graphical interface involved the depiction of three-dimensional characters (or avatars) representing the participants sitting around a virtual table. The appearance of the avatars included a dynamic face model derived by a face capture system [anon]. The graphical representation is unique to each participant, created once, and transmitted once before the conversation. Each avatar face is distinguished by a set of 60 parameters plus a single image texture map. Each face is suspended above a stylized body that does not move in the experiments we conducted.

Avatar behavior was restricted to head movements. It is computed locally and evolves in real-time during the course of the conversation. Each participant sees a different view of the avatars from their own POV. In fact, there is no technical reason for each participant to see the same interface at all. One might see the simple icon interface displayed, for example, on a palm sized device, one might see the avatar interface and another have voice only. In the experiments that follow, all participants saw the same interface at the same time.

The avatars' behavior is guided by the literature that describes common sequences in the flow of conversations [1]. For example, listeners often look at the speaker. Mutual eye contact is often broken after a few seconds, most often by the speaker [6]. The speaker will often look at the intended main listener. Based on these types of observed behaviors, the avatars act through a stochastic process to make them mimic such behavior. For example, the avatars turn and look at the most recent speaker when listening 90% of the time (this decision is re-evaluated every 3 seconds). A participant is defined as speaking after 1 second of continuous talking to avoid simple affirmations ("uh-huh") from attracting too much attention. Thus, since one's own avatar is not depicted graphically, about a second after one begins to speak, the other avatars will usually turn towards the virtual camera (i.e., towards you). Similarly, a speaking avatar initially turns to face a specific other avatar. We do not know the intended recipient for any comment, so we assume a high probability (70%) that a new speaker addresses the most recent previous speaker. After a few seconds attention may then be shifted randomly to another member of the group.

The avatars also exhibit other behaviors besides turning and looking. When speaking, the mouth moves in a somewhat natural way, although we are not able to mimic real time lip sync. The avatar faces also nod, smile, and frown. These actions are triggered with higher probability immediately after mutual gaze is attained (when addressing a group, try looking directly at someone and you can often illicit a small nod and/or smile). Finally, the participant's name is displayed above the avatar and is highlighted in red when the avatar is speaking.

3.3 Avatar construction

The three-dimensional avatar faces are constructed from a low polygon, generic average face model. Sixty shape deformations represent differences from the norm such as wide/thin head, small/large mouth or nose, eyebrow position and size, etc. Sixty scalar parameters determine how much each difference should be accentuated. This defines the space of all possible face shapes. In addition, a single image acts as a "texture map", the coloration of each point on the face. Since all faces are based on the same base polygons, any motion, such as smiling automatically maps to the new deformed faces representing specific individuals.

The sixty parameters and the texture map are derived through a computer vision process operating on a short video sequence. A user is asked to turn their head from one side to another over the course of approximately 5 seconds. An optimization system adjusts the sixty parameters to best fit the face that was seen. The texture map is then constructed from the video sequence. The entire process takes about three minutes. Once captured, the new avatar is ready to be inserted into the virtual conversations and can act with all the motions predefined for all avatars.

4 EXPERIMENTAL DESIGN

To assess the impact of including a graphical interface in a small group voice-based telephone conference, we assembled 11 groups of 4 or 5 people (a 12th group was intended but at the last minute we were left with only 3 participants). The participants were between 18 and 35, both male and female. They had not met until being assembled for the experiment.

4.1 Avatar construction

After a short welcome, each individual put on a nametag. Participants then sat in front of a camera to collect the short video sequence needed for the avatar construction. In addition to the avatar face, one frame of the video was used as the static image for the simple graphical interface. The entire avatar construction and initial survey took about 25 minutes.

4.2 The conversations

The group was then asked to participate in four 10-minute conversations in sequence. Each conversation was followed by a short survey.

For all groups, the first conversation took place "live" with participants in the same room sitting around a small table. The live conversation was held first to provide a common baseline for all groups and also to allow the groups to get to know each other a bit as most often is the case in real group discussions.

For the three mediated conversations, participants were seated in separate rooms. Voice communication took place over standard phone lines with a small head set with microphone and earpiece. A small microphone also fed the voice stream into a computer. The computer determined the talking/not talking parameter and delivered this one bit on/off stream to the other participants' machines. Individual voice streams were also recorded (48KHz, 16bit mono) for later analysis.

Of the three mediated conversations, one included no additional interface (voice only), one included the highlighting static image (icon), and one included the avatar-based interface (avatar). These were presented in different orders to each group. Given the six possible orders, each ordering was used twice, with one exception.

Tasks were assigned for each conversation. The tasks were presented in the same order for all groups. Thus each interface was used with each task four times. The tasks each involved the selection of three items. These included choosing the three best; a) places to include in a guidebook advising visitors on a day trip to our local city, b) music CD's to take on a long road trip, c) places to go on a workgroup vacation, and d) videos to take along to a wilderness cabin. If the conversations seemed to be concluding before the 10 minutes alloted, an additional related task would be added to keep the discussion going.

After the fourth conversation and survey, participants were asked to complete a final survey. They were then brought back together to the same room with the experimenters for a closing free discussion about their experiences. The entire experiment took approximately two hours for each group.

5 RESULTS

Below we report the technical, automated voice analysis, and subjective results of the study that was conducted. The technical achievements involve the rapid creation of the avatar faces, speaker detection and real-time transmission of this data, and the resulting dynamic interfaces presented to the participants. Objective measures such as speaker turn lengths and silences are made directly from the voice tracks and compared across interfaces. Subjective measures are derived from the short surveys taken before, between, and after the four sessions.

5.1 Technical Results

The first step in the study involved the creation of a three-dimensional model of the participants heads which were inserted into the avatar-based interface. We created the faces for all members of a 5 person group in a total of approximately 20 minutes.

Participants were later asked if they agreed that the avatars "looked like the other members of the group". All but 6 of 52 subjects "agreed" (5, 6, or 7) on a 7 point Likert scale. When asked if the avatars "acted naturally" the result was quite different. About twice as many subjects disagreed as agreed. Comments indicated that both the strange body shapes and the fact that the heads turned more than expected led to this result. Additionally, unfortunately a bug in the controls resulted in an avatar sometimes getting stuck in the "talking" condition which subjects found disturbing. Typical comments included: "The bodies beneath the heads were really weird.", "They seemed to do some unnatural things at times, like rapidly moving their heads back and forth." and "The faces pulled you in while the bodies distracted.".

5.2 Voice Analysis Results

The voice track for each participant was recorded separately for each of the mediated conversations. Without the human resources to step through and hand code the 40 hours of individual tracks of conversations, we have relied on an automated process. Each voice track was first processed to determine, for each 25 millisecond interval, whether someone was speaking or silent, based on a simple energy measure. We then removed any speaking intervals that lasted only 25 milliseconds, as these mostly likely represent small "clicks" or other noise in the signal. From this we determined distributions of silences for individuals. We also merged all tracks from a group to determine group silences as well. Histograms for both individual and group silences are shown in Figure 2. The log of the number of silences is given for each 100 millisecond length is plotted. Both the individual and group histograms of silence lengths follow a characteristic curve as was observed in [11].

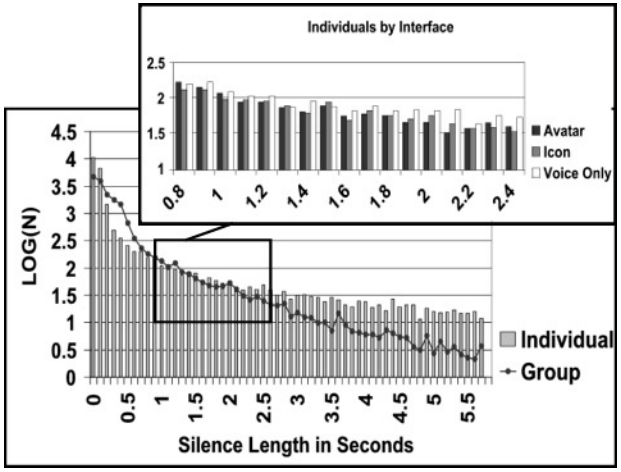


Figure 2: Histograms of individual and group silences. Close-up of individual silences by interface.

Four regimes can be identified; silences of less than 200 milliseconds are mostly gaps between words and sentences. We see more of these silences in the individual statistics since some will be covered by short utterances or noise from others in the group. The 200 to 700 millisecond range contains some gaps between sentences and also rapid turn changes, or turns interspersed with affirmations, laughter, etc. Here we see more occurrences from the group than from individuals since the individuals' silences are broken up by others' affirmations. Between 800 milliseconds and approximately 2500 milliseconds are a range of what we might think of as somewhat uncomfortable silences for groups. For individuals, they may represent restarting a turn after a longer gap for an affirmation (an "uh-huh") from the group that someone is listening. We will examine this region a bit more closely below. Longer silences (>2.5 seconds) quickly become rarer for groups. For individuals they likely indicate having given up the floor to others.

Examining the silences separated by interface type did not yield any significant differences with one exception. The 800 to 2500 millisecond range (the "uh-huh" range) seen in the upper right of the histogram above for individuals exhibits a distinct difference for the voice only interface compared to the avatar and icon interfaces. Using multiple linear regression we analyze the following form,

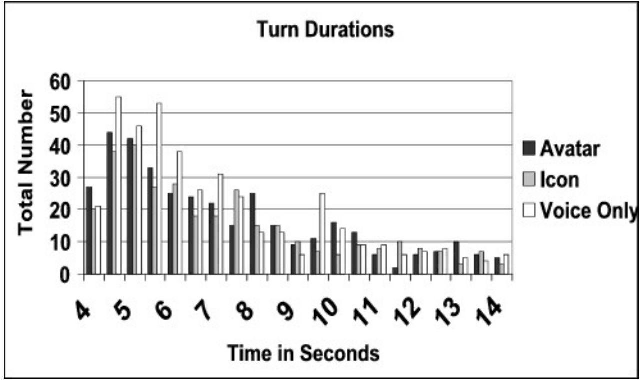
$$\text{Log}(count) = a_0 + a_1\ length + a_2\ avatar + a_3\ icon$$

Where *count* is the number of silences of length *length*, and *avatar* and *icon* are dummy variables indicating whether the silence occurred with each of these interfaces. A zero for both dummy variables indicated it occurred under the voice only interface. The model is highly significant, F[3,53] = 97.9, p <.001, R^2 = .855. The slope, a_1, was the strongest predictor, with the influences of the interfaces also highly significant.

	Coeff.	T	p	β
length	-0.316	-16.54	< .001	-0.892
avatar	-0.095	-3.91	< .001	-0.244
icon	-0.094	-3.86	< .001	-0.241

There are two possible interpretations of this result. One possible inference is that there were more pauses of “uh-huh” length with the voice only interface than with the graphically enhanced interfaces. The results show that a best line drawn through the data is about 0.95 higher for the voice only, which represents an increased count of approximately 25% over the icon and avatar interfaces. This indicates that speakers may have perceived the need to wait for confirmations of someone listening more often. The second explanation is that the length of the pauses in this region are approximately 0.3 seconds longer on average (-.316 / .095). In other words, the line for voice only is not above but rather *to the right* of the other two lines. The latter interpretation suggests there are longer “pregnant pauses” until someone said “uh-huh”.

We also processed the voice data further to extract “turns”. Turns are defined as any sequence over 4 seconds long containing no silences of more than 1.5 seconds. From this we plotted turn lengths and also average turns per minute.



The median turn lengths were 6.8, 7.3, and 6.5 seconds for the avatar, icon and voice only interfaces respectively. An analysis of turn rates (turns/minute) for each participant across interfaces, comparing voice only to the graphical interfaces yields a significant difference F[1,10] = 6.85, p = .026. The voice only interface led to more, shorter turns and similarly more turns per person per minute.

5.3 Survey Results

5.3.1 Per Conversation Surveys

After each of the four conversations all participants completed a 7-item survey concerning their opinion of the conversation. Responses were made on a 7-point Likert scale ranging from “Strongly Agree” to “Strongly Disagree”. Survey responses were analyzed using the within subjects, non-parametric Friedman test to assess differences across interfaces, with specific pair wise comparisons determined by a Wilcoxon signed ranks test using SPSS. Non-parametric tests were used to accommodate the ordinal nature of our response scale. However, mean rather than median differences between ratings of the interfaces are provided to get a sense of overall ratings of the interfaces, as median ratings were frequently 6.0 on the 1 to 7 scale.

Expressing One’s Opinion and Being Heard

Participants did not report any significant differences to questions on how well they were able to express their opinions or how well they thought others understood them across the four interfaces. The mean scores ranged from 6.10 to 6.28 (out of 7) indicating an ease of expressing oneself with all interfaces.

Understanding Who Had What Opinion and How Well Others Were Able To Express Themselves

Participants were asked two questions related to how well others were understood and could express themselves. Highly significant differences were reported across the conversation interfaces for understanding who in their group had what opinions, χ^2(3,N=51) = 19.08, p < .001. The means for live, avatar, icon, and voice only were 6.4, 5.8, 5.9, and 5.1 respectively. All pair wise comparisons were also significant (all p’s < .05) except between the avatar and icon interfaces. Similar differences were found when the participants were asked how well others could express themselves with each interface. The Friedman test resulted in χ^2(3,N=51) = 16.69, p = .001, with means of 6.3, 6.0, 5.9, 5.6. In pair wise comparisons, the live conversation was significantly better than any of the mediated interfaces (p’s < .05). In addition, the avatar vs. voice difference was significant with p = .031.

Enjoyment and Decision Making

Participants did not report any significant differences in levels of enjoyment of the different discussions. All the means were about 6.0. Nor were there significant differences across interfaces in the ease of reaching decisions nor the agreement with the decisions made by the group for each task.

OtherFactors

We also ran parametric analyses to assess whether other factors such as group size (either 4 or 5), or number of females (ranging from 0 to 3) interacted with the different interfaces. While group size was not seen to be a factor, the number of females in the group did show some significant effects. There was a strong interaction between interface and the number of females in a group (F(6,105) = 2.963, p = .01) for the ease with which the group reached decisions. Decisions were easier to make as the number of females increased when using the avatar and icon interfaces, but not at all when conversing live and only marginally when using phone only.

Also, a main effect for the number of females was found for personal satisfaction with the group’s decisions, F(2,35)

= 6.958, p = .003. Groups with only one female reported an average of satisfaction with the group's decisions of 5.85, while with two females the average rose slightly to 6.07, and with three females the average rises again to 6.40. This effect may be due to groups with more females being more agreeable, or to females responding more positively to this item, or to both.

# Female	Live	Avatar	Icon	Voice
1	5.8	4.3	5.4	4.4
2	6.2	5.9	5.4	5.7
3	6.2	6.6	6.6	5.2

Mean ratings for ease of reaching decisions by number of females in the group

Comments about each conversation

Participants also were asked to comment on what was best and worst about each conversation. The questions purposefully did not ask specifically about the technology at this point. Most responses discussed the conversation itself, but a significant fraction mentioned the interfaces, particularly after the avatar and icon mediated conversations. A sampling of comments follows.

Avatar Interface

There were many positive comments about the avatar based interface, such as "The avatar movement and highlight help understand who is speaking.", "With the molded faces, I felt like it was a "richer" conference call….it did feel more like a person-to-person in-the-flesh discussion than the other two methods.", "I felt very comfortable expressing myself with the partial anonymity of the virtual interface.", "I really liked the way the avatars turned to face me when I spoke.", "The 3D figures did make it easier to stay engaged in the telephone conference… it is easy to get distracted and not recognize who is speaking in the standard calls.", and "Having the visual interface with each of the participants was terrific! You felt like you were in the room with them, having a discussion over dinner or something.".

There were also those who really did not like the avatar interface. For example, we received comments such as: "The video screen is distracting...I kept watching the moving heads and not participating!", "I did not like not seeing myself at the table.", "It was creepy wondering if the virtual heads were representing the actual directions that the heads were pointing.", and "I kept using body language, like nodding in agreement, when the others couldn't see me, but seeing the icons moving and talking led me to behave as if they could see me. I was more quiet and used more body language because of it".

Icon Interface

Similarly, there were many positive comments about the icon interface, for example: "I could easily tell who was saying what.", "The visual interface was not as distracting as the one with the computer generated bodies.", "It was easy to figure out who was talking, and if I looked at their photo, I could imagine what they'd look like if I was sitting in the same room with them while they talked.", "Without the highlight I definitely would have missed some people's feelings.", and "Having the colors change when the person spoke was very helpful in tracking the discussion but not as well as the [avatar] interface."

Not surprisingly, this interface also had its detractors. "This interface was very bland -- would be better suited for an office conference meeting as opposed to a group of friends.", "I don't see how the pictures on the screen enhanced the conversation very much.", and "All you had to work with were stills that highlighted and that just didn't satisfy me."

Voice Only Interface

We received less comments directly on the voice only interface as this one was more familiar to the participants. A few that liked this one best commented: "This made it easier to match up who was saying what -- because I could key in on the voices.", "I had the opportunity not to have to pay any attention to the monitor during the discussion.", and "I was not distracted, but remained focused on the conversation."

There were many more negative comments, most related to not knowing who was speaking when, and also more difficulty staying engaged in the conversation, and knowing when to participate. For example, "It was very hard to know who was saying what and it seemed harder to engage everyone in the discussion this time.", "It was harder to stay engaged without the visuals", "It was harder to tell whose opinion was being expressed.", "I interrupted, without meaning to, because it wasn't clear who was speaking or who was getting ready to speak.", and "Interestingly enough I thought it was slightly more difficult to enter into the discussion without something to focus on."

5.3.2 Final Survey

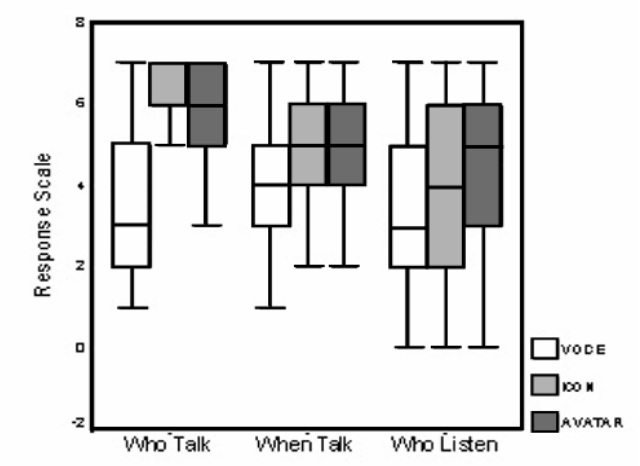
After completing discussions using all four conversation interfaces, participants completed a final survey on which they compared just the three mediated conversation interfaces (avatar, icon, and voice only). Median scores are reported for these analyses when appropriate.

Which Was Best

A simple question put to participants asked which of the three interfaces was best and worst without asking for reasons. An overwhelming majority of participants ranked the icon and avatar interfaces better than the voice only. The avatar interface received the most "best" ratings but a few more also thought the avatar interface was worst compared to the icon interface. Chi-squared tests confirm that the distribution of rankings represents significant differences from random responses, with p's = .000, .004, .025 for voice only, icon and avatar interfaces.

	Avatar	Icon	Voice
Best	26	21	3
Middle	14	24	14
Worst	11	6	34

Rankings of the three mediated interfaces



Median ratings for how well participants felt they knew who was talking, when to talk, and who was listening, across three conversation interfaces. Shaded portion represents the interquartile range; bars denote minimum and maximum values.

Knowing Who Was Talking

Participants were then asked if it was easy to know who was talking in each interface, again using the 7 point Likert scale. Not surprisingly, the graphically enhanced interfaces showed highly significant improvements $\chi^2(2,N=52) = 53.84$, $p < .001$. Median scores of 6.0, 6.0, and 3.0 for the avatar, icon, and voice only interfaces indicate the level of differences. In pair wise comparisons, voice only was significantly worse than the avatar and icon, $p < 0.05$.

Knowing When to Talk

Perhaps less predictable, there were also significant differences in how well people reported about how well they knew when to talk, $\chi^2(2,N=52) = 25.04$, $p < .001$. Medians of 5.0, 5.0, and 4.0 for avatar, icon, and voice respectively indicated the graphical interfaces helped the perceived flow of conversation. Pair wise, there was no difference between the two enhanced interfaces.

Knowing Who Was Listening

A similar result was found when asked if participants felt they could tell who was listening despite no direct indication in the graphical interfaces. Results were significant again $\chi^2(2,N=45) = 21.29$, $p < .001$, with medians of 5.0, 4.0, and 3.0 for avatar, icon, and voice respectively. Pair wise comparisons again showed an increase with the avatar and icon interfaces over voice (both p's < .001) and no significant distinction between them.

General comments

We also asked participants if they had additional comments at the end of the final survey. Many repeated earlier comments. A few people noted that they would like to have some visual representation of themselves in the avatar and icon interfaces. Others provided an overview of reactions, such as "I didn't realize how much I relied on knowing who was talking until we did the voice only, then I lost track of some of the voices.", "The animated characters made me feel like I'm in a real group meeting, the other ways were like a conference call.", "I concentrated more on the voices and listened better, I think, when it was voice only. But the icons and characters helped me know who was talking.", "The flow of communication--who was speaking--was much clearer in person & almost as good with the animated figures.", "Knowing who was talking seemed to shift cognitive load off of the "who is this" task and strangely seemed to make it easier to understand what was being said".

6 DISCUSSION

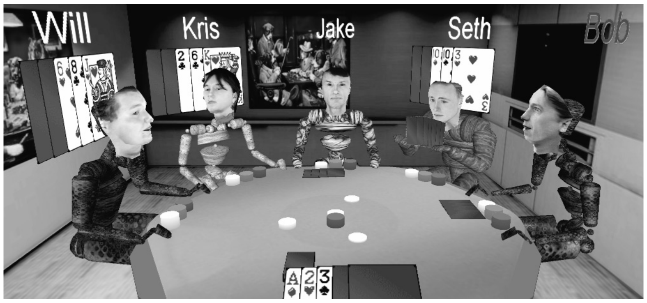
The results clearly indicate that at least subjectively, the participants found benefits in the graphically enhanced interfaces. It is perceived to be easier to understand many aspects of the conversation, who is speaking, when to speak, and even who is listening. Some even reported is being easier to understand *what* is said perhaps due to a lower cognitive load provided by the enhanced interfaces. The differences between the two graphical interfaces are less striking. Many people found the avatar interface the "best" but many also found it distracting. The individual comments provide great hope that future versions that provide better representations of bodies and behaviors will reduce the distraction and provide even more visual cues. Analysis of the voice tracks shows that the conversations have somewhat different dynamics with and without the graphical interfaces with respect to middle-length pauses, 0.8 to 2.5 seconds. We cannot determine whether there were more pauses in the voice only interface or whether the pauses were longer, however, a common hypothesis can be constructed for both. In either case we might surmise there was a greater need to affirm the existence of the rest of the group through seeking affirmations and/or providing others a chance to respond when no graphical interface is provided

7 CONCLUDING REMARKS AND FUTURE WORK

We have shown that by transmitting only a single bit wide stream per person, a voice only small group conference call can be significantly enhanced by a graphical interface. Given the simplicity of adding such systems to current voice only communication, we think those offering voice communication should consider including such graphical interfaces.

There is clearly a lot of work to be done on the avatar based interface. Bodies with natural behavior need to be constructed, as in [4,5].

The interfaces tested were fully "hand-off". The users simply talked. One can also imagine a host of behaviors based on more input from users. For example, users might indicate who to address a particular comment to and their avatar would turn to that person. By indicating a desire for a turn, their avatar may, for example, raise its hand. Clearly, there would be a trade-off between a more intrusive interface and any benefits.



We have also begun developing an avatar interface to serve as the view to a poker game application. Here, not only do the heads turn, but the bodies move to place chips and "take the pot" when the game calls for it. We believe such social settings will be the early adopters of this type interface and we look forward to trying out this and other related applications.

Postscript (Sept. 11, 2001): *My hands shake as I try to find words to describe the horror I have just witnessed. May the world somehow come to see that only in a world where all are able to share its blessings will we be able to put this behind us.*

REFERENCES

1. Argyle, M., and M. Cook. 1976. Gaze and mutual gaze. Cambridge: Cambridge University Press
2. Atkinson, J. M., & Heritage, J. (1984). Structures of social action; Studies in conversation analysis. Cambridge, U.K.: Cambridge University Press.
3. Buxton, W. (1992) Telepresence: Integrating shared task and person spaces. Proceedings of Graphics Interface '92 (pp. 123-129) May 11-15. Vancouver, BC.
4. Cassell, J. and H. Vilhjalmsson. Fully embodied conversational avatars: Making communicative behaviours autonomous. Autonomous Agents and Multi-Agent Systems, 2:45--64, 1999.
5. Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., and Yan, H. "Conversation as a System Framework: Designing Embodied Conversational Agents", in Cassell, J. et al. (eds.), Embodied Conversational Agents. MIT Press, Cambridge, MA, 1999.
6. Colburn, A., Cohen, M. F., and Drucker, S., The Role of Eye Gaze in Avatar Mediated Conversational Interfaces, MSR-TR-2000-81. Microsoft Research, 2000.
7. Chapanis, A. Ochsman, R. Parrish, R. and Weeks., G. (1972). Studies in interactive communication: II. The effects of four communication modes on the behavior of teams during cooperative problem-solving. Human Factors. 14(6). 487-509.
8. Dourish, P, and Bly, S. (1993). Portholes: Supporting awareness in a distributed work group. CHI'93 Human Factors in Computing Systems, 541-547, New York: ACM Press.
9. Egido, C. (1990). Teleconferencing as a technology to support co-operative work: a review of its failures. In J.Galegher, R. Kraut, & C.Egido , Eds. Intellectual Teamwork. Hillsdale, N.J.: Lawrence Erlbaum Press.
10. Finn, K. Introduction, in Video-Mediated Communication, eds. Finn, K., Sellen, A. and Wilbur, S. Lawrence Erlbaum Associates, 1997, pp. 3-22.
11. Jiang, W. and H. Schulzrinne. Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation. In The 9th IEEE International Conference on Computer Communication Networks, 2000.
12. Krauss, R. M., and P. D. Bricker, "Effects of transmission delay and access delay on the efficiency of verbal communication," J. Acoust. Soc. Amer., vol. 41.2, pp. 286--292, 1967.
13. Masoodian, M., Apperley, M., & Frederikson, L. (1994). The Impact of Human-to-Human Communication Modes in CSCW Environments. OZCHI, 1994, Computer Human Interaction Special Interest Group of the Ergonomics Society of Australia, Melbourne, Australia, 193-198.
14. O'Connail, B. and Whittaker, S., "Characterizing, predicting and measuring video mediated communication: A conversational approach," in Video Mediated Communication, K.E. Finn, A.J. Sellen, and S.B. WIlbur, Editors. 1997, Lawrence Erlbaum, Mahwah, NJ. pp. 107-132.
15. Olson, J., Olson, G. and Meader, D. "Face-to-Face Group Work Compared to Remote Group Work With and Without Video" in Video-Mediated Communication, eds. Finn, K. , Sellen, A. and Wilbur, S. Lawrence Erlbaum Associates, 1997. pp.157-172.
16. Poltrock, S., Boeing Corp., personal communication
17. Rodenstein, Roy and Judith S. Donath. (2000) Talking in Circles: Designing A Spatially-Grounded AudioConferencing Environment. In Proceedings of CHI '2000, pp. 81-88.
18. Sacks, H., E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. Language, 50:696-- 735, 1974.
19. Sellen, A. (1995) Remote conversations: The effects of mediating talk with technology, Human Computer Interaction, 10(4). 401-444.
20. Tang, J., Isaacs, E. and Rua, M. Supporting distributed groups with a montage of lightweight interactions. ACM Conference on Computer-Supported Cooperative Work (CSCW'94), Chapel Hill, USA, 1994, p. 23-34.
21. Vertegaal, R., Slagter, R., Van der Veer, G.C., and Nijholt, A. "Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes." ACM CHI 2001 Conference on Human Factors in Computing Systems. Seattle ACM 2000