TalkBack: a conversational answering machine

Vidya Lakshmipathy, Chris Schmandt, Natalia Marmasse
MIT Media Lab, Cambridge MA, USA
E-mail: {vidya, geek, nmarmas}@media.mit.edu

ABSTRACT

Current asynchronous voice messaging interfaces, like voicemail, fail to take advantage of our conversational skills. TalkBack restores conversational turn-taking to voicemail retrieval by dividing voice messages into smaller sections based on the most significant silent and filled pauses and pausing after each to record a response. The responses are composed into a reply, alternating with snippets of the original message for context. TalkBack is built into a digital picture frame; the recipient touches a picture of the caller to hear each segment of the message in turn. The minimal interface models synchronous interaction and facilitates asynchronous voice messaging. TalkBack can also present a voice-annotated slide show which it receives over the Internet.

KEYWORDS: voicemail, answering machine, computer mediated communication, conversational interface

INTRODUCTION

The most common and expressive [2] setting for language use is face-to-face conversation. It is something that most everyone in the world has some experience doing and, it requires little training. Conversation is both an individual and social process. It is a joint action that requires common ground, shared information that allows for the coordination of meaning and understanding [3]. We have developed many ways to establish common ground in face-to-face conversation. Because the exchange is in real time, we can show understanding with back-channel feedback [15]; we can make references by pointing, gesturing or gazing, or our choice of words, timing, and turn-taking can indicate a continued discussion about a particular subject. We can also interrupt if we wish to speak before it is our turn.

Over the last 50 to 75 years, we've discovered that we no longer need to be face-to-face to communicate in real time. As the telephone has made its way into every house, and now every pocket, we've learned to converse without co-presence. We've established techniques to continue joint actions and establish common ground without facial expression or gesture and only with language. Because the conversation still occurs in real time, we can use back-channel feedback and turn-taking metaphors to establish common ground and have successful communication.

The answering machine added a new dimension to distance communication. Asynchronous communication moved us farther from the familiar face-to-face style, requiring new skills. With voicemail, there is no way to continually ground events over the course of the conversation; the lack of feedback interferes with the normally mutual process of grounding events [3]. In addition to the extra burden required to keep common ground in short term memory, one has to continually remember to check for messages, and often there is an added task of having to respond by calling each person back. While these are all clearly skills we can learn, there might be a cost in the quality or pleasure of communication.

TalkBack bypasses these additional skills by simulating a synchronous conversation in an asynchronous medium. It is an answering machine which breaks incoming voice messages into chunks, and while playing these sequentially, pauses between each to record a response. The recipient can also interrupt and inject a response at any point during playback. The system leverages principles of immediacy and co-presence from conversation to make the interaction simpler and more pleasant and informative for both the message leaver and the message recipient. By embedding an answering machine in a digital picture frame in which photos of the callers represent messages, TalkBack further pursues the face-to-face conversation metaphor.

This paper briefly discusses specific problems with voicemail in its current form and discusses in detail which solutions TalkBack employs. We evaluated TalkBack by having users compare it to voicemail while responding to messages. We continue by outlining the design implications posited by our observations and a glimpse at a revised version of TalkBack based on these implications. Finally we discuss related work that has addressed these and similar problems from different angles.

PROBLEMS WITH VOICEMAIL

A number of factors confound study of the use of stored voice as a communication medium. First, it spans two very different sorts of technologies, *answering machines* (stand alone recording devices, found in domestic settings) and *voicemail* systems, accessed by telephone only and typically (though not exclusively) in business settings. Each of these environments experiences a different mix of voice message genres (e.g. chatty, information gathering, informing, decision making) though there may be some overlap; message type likely influences user interface requirements. Our design of TalkBack was initially motivated by concern for domestic settings, especially among the elderly, but we did not want to limit it to such; the TalkBack solution described in this paper was implemented as “all user interface” and actually is layered on top of a research voicemail system.

This distinction made above is not just limited to where the messages are recorded; in a study of voicemail systems, Rice [13] noted that they may be used primarily as *voice answering* or *voice messaging* systems, depending on whether messages are heard and then discarded, or are annotated, forwarded, and archived. Furthermore, which mode is adopted seems to depend more on community norms than particular user's propensity toward adopting technology [14].

Studies focused on expert users of voicemail have found that there are three main problems experienced when managing voicemail: *scanning*, *information extraction* and *search* [21]. *Scanning* is used to give message priority and for locating saved messages. *Information extraction* is often done by taking notes about a message in order to save important information for future reference. Users also spend a large amount of time *search*ing for archived messages and tracking the status of saved messages. The design of TalkBack focuses specifically on information extraction in the context of formulating a reply to a voicemail. This aspect of the problem of managing voicemail has been addressed with interfaces that allow users to take notes related to the content of the voicemail [19] or allow them to scan a transcript of the message as they listen [20].

Answering machines (or phone-accessed voicemail systems) do not have such rich GUIs, and users are required to either jot down notes or keep the content of the message in memory as they attempt to respond. Voicemail has more recently become a very popular feature for mobile phones. Checking voicemail while mobile and with such a small screen makes it nearly impossible to take notes or view transcripts. As a result, more practical methods of replying to voicemail need to be explored. As is well known [1], memory or recall from memory deteriorates with age, making this task of extracting and remembering information difficult for the elderly. Message recipients must also juggle functionality between listening to a series of messages and then dialing phone numbers, while keeping the message in memory, to reply.

Additionally, despite the media richness of computer-mediated communication, voicemail still remains a closed, single-medium system. Although prevalent on mobile devices and in networked environments, it has rarely benefited from the devices and connectivity around it. One of TalkBack’s contributions is to accept and deliver voice messages via the Internet, and to support sender-supplied photos and voice annotated slide shows as messages.

RELATED WORK

TalkBack draws on related work from a number of diverse fields. We are not the first to suggest that human-like attributes may improve computer-mediated communication. Certainly we have seen that people respond to these attributes in computer user interfaces [12]. More specific to communication is the use of “avatars” in chat rooms; there is some evidence that when they perform in a manner which mimics human behavior well, avatars may improve the medium [4]. There is also an extensive literature on gaze in video conferencing systems.

The original “conversational answering machine” was PhoneSlave [17], nearly two decades ago. PhoneSlave used recorded speech and pause-based audio recording to gather responses to questions such as “Who's calling please?”, “What's this in reference to?”, and “At what number can you be reached?”, and later could play each of these snippets back to the PhoneSlave owner, in response to voice commands. PhoneSlave used speech recognition (in lieu of today's telephone caller ID) to try to identify repeat callers, and could deliver personal messages to them when they called back, as well as indicate whether their previous message had been heard.

Part of PhoneSlave's attraction at the time was that voicemail was still new enough that callers were often not facile at leaving messages on a machine; PhoneSlave took complete messages by turning the interaction into a form-filling conversation. The authors believe that most callers would be unwilling to participate in such a routine now, although “Whom may I say is calling?” has been used for call screening in products by Active Voice and Wildfire.

A Japanese project [5] implemented answering machines which would mutter back-channel responses (“hai” in Japanese) to encourage callers to leave longer or more complete messages. The Grunt system [16] presented driving directions over a telephone, pausing between each major route segment and analyzing any user response based on length and pitch contour to decide whether and when to proceed, or offer more explanation.

In the 1990's several research systems used conversational paradigms bordering on natural language input to control live interactive systems over the phone using speech recognition. MailCall[8] emphasized text message retrieval, its successor SpeechActs [22] used more conversational techniques and covered a wider range of applications.

QuietCalls [11] supported live voice interaction over telephones, with one party speaking and the other playing recorded audio snippets, driven by a conversational state model; its similarity to TalkBack lies in that model.

Many systems have implemented graphical user interfaces to control voice messaging systems; two of the more recent and novel ones are Jotmail [19] and Scanmail [20]. But the TalkBack visual component is more influenced by the use of digital images in domestic appliances, such as the Digital Family Portraits project [10] and web-accessed digital picture frame products from Kodak, Ceiva, and others. Digital Family Portraits used a static image while presenting variable data (about the person in the photo) graphically around the frame. TalkBack is meant to be an attractive visual artifact even when it has no stored messages; a changing display indicates messages are waiting and perhaps who left them.

TALKBACK: THE SYSTEM

TalkBack, seen in Figure 1, is a working prototype of a conversational answering machine. A digital picture frame with a touch screen allows the user to control the playback of messages; pictures of the caller indicate new messages. To listen and respond to a message, the user touches the picture corresponding to the message. The message plays in short segments, stopping to allow the listener a turn to record a response for each. When s/he is finished speaking, the device continues to play the next section of the message; this process continues until the entire message is played. The user can also interrupt playback at any time to interject a response.

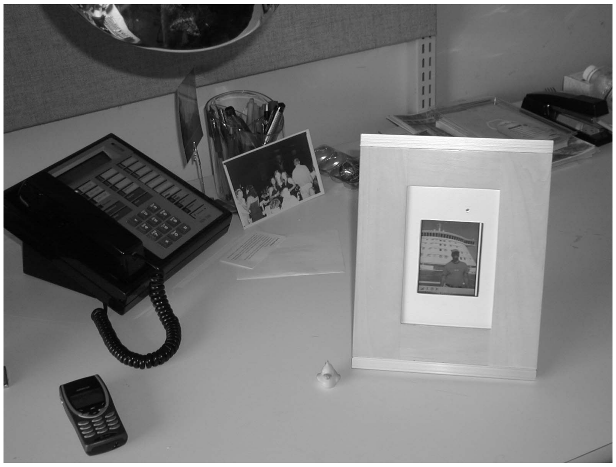

Figure 1. The TalkBack Answering Machine (picture frame on the right).

TalkBack makes replying to messages more conversational. Pausing (with a beep to indicate recording in progress) after significant segments of the message invites a response. Interruption is useful where segmentation failed or in an implementation without segmentation; although the user must first be made aware that interruption is possible. The person who receives the reply is not aware of this process and needs some context to help ground the reply. So, much like an email reply which quotes sections of the original message, responses are aggregated and interspersed with four seconds of the original message, time compressed, as a reply. These replies can be delivered to the original caller by phone or by email as an audio attachment.

The three aspects of TalkBack, display, segmentation with turn-taking, and interruption, are all independent and can be combined in various ways for different implementations. The value of this modularity became clear when evaluating the conversational messaging in a small user study.

Segmentation

In the first version of the TalkBack server, pauses were found by comparing the average magnitude of non-overlapping 200 millisecond windows with a silence threshold. This threshold was initialized to be the average magnitude of the first 200 ms of the recording, which was assumed to be silence. If during recording the average magnitude of any 200 ms window was less than the silence threshold, the silence threshold was reset to that value. The voicemail server normalizes the amplitude of the recording such that the full 8-bit linear scale is utilized (0-255). If the average magnitude of any window was within 12% of the silence threshold, it was considered silence. This range was fine tuned for the recordings being produced by the research voicemail system used in the TalkBack project.

The current version of the TalkBack server had to be made more robust to noise and variable recording levels. First we calibrate to the dynamic range of the sound. We find the silence threshold, i.e. the minimum, by the algorithm described previously. Then we find the overall average magnitude of the entire recording, giving us a measure of the loudness of the speech recorded. The dynamic range is the difference between the overall average and the silence threshold. This is in fact a rough approximation of the dynamic range, but it is approximately correct given that voice messages are mostly speech with relatively few pauses. Second, we look at the average magnitudes of adjacent 200 ms non-overlapping windows. If the difference between these window averages is greater than 10% of the dynamic range, we mark the second window as the beginning of speech if the average magnitudes are increasing, or as the end of speech if decreasing.

To detect filled pauses (e.g. “umm” and “er”) we rely on the fact that they are inordinately long single syllables. TalkBack assumes that any syllable longer than 450 ms is a filled pause; filled pauses may be shorter, but this duration results in a high precision detector. We rely on methods similar to the algorithm described by Mermelstein [9] to detect syllable boundaries. Energy is computed over 10 ms non-overlapping windows, and a syllable begins when energy exceeds a threshold; messages have previously been normalized for energy, so this threshold can be absolute. A syllable ends for one of two reasons. In the simple case, the syllable is terminated by a consonant with significant vocal tract closure, and the energy drops below the same threshold. If closure is incomplete, there is still a drop in energy between vowels, for example in “do you?”. If energy drops to half the peak energy in the preceding portion of the syllable and then rises to twice the minimum after that peak, a new syllable is declared at that intermediate minimum.

Ideally, TalkBack should segment the recorded message into salient, related “chunks”, much akin to text paragraphs. Pauses and filled pauses are useful in that they often reflect thought processing on the part of the talker, and hence shifts of topic or focus. We hypothesized that intonational cues would also be useful, either as additional evidence of topic shift or explicit indication of questions, which might warrant a pause and response.

We computed pitch tracks for a set of 12 voice messages left in one of the author's voicemail boxes before this project began. Messages were screened to be somewhat long and chatty, as opposed to the common “It's me, sorry I missed you, can you call me back?”. We manually aligned pitch, amplitude, and transcription tracks but no consistent intonational cues were detected. This was surprising and, frankly, disappointing. A possible explanation is that the speaking style of a voice message is distant enough from ordinary conversation that intonational cues are weak. It is well known that intonational cues for managing audio playback can vary by genre (conversation, lecture, newscast), so we may be seeing such an effect. If messaging does indeed become more conversational, we may find increase value in intonational cues for segmentation.

Once the silent and filled pauses are found, separate files are created with these pauses as boundaries. Files that are less than 3 seconds are merged with larger chunks so that no segment is too small. This helps assure that each segment has some valuable data and that there are not too many segments per message.

Responding to messages

These voicemail segments are then delivered to the TalkBack client along with the caller ID information. The primary interface to the TalkBack user, the client, is an iPaq placed in a picture frame connected with Wi-Fi to the local area network. The device itself is hidden so that all the user sees are pictures displayed in the frame. Phone numbers known to the system (i.e. friends of the user), are associated with picture files stored locally on the client.

When the user receives a message, s/he sees a new picture displayed in the frame; to listen, s/he just touches the screen. Each section of the message plays in order, giving the listener a chance to record a response between each. The listener does not have to respond to every segment; if the listener chooses not to speak, the system detects the silence and plays the next section. This continues until the entire message has been played. The interface is purposefully simple, allowing only touch controls to begin or end playback of messages.

The recipient of the response receives a small portion of the original message, time-compressed by half, interspersed with the recorded response as shown in the waveform in Figure 2. Time compression is done with the SOLA algorithm [18]. This small portion of the original messages serves to provide a context for the next portion of a response. Many users likened these responses to responses sent by email with the original question interspersed with the response. An example can be heard at [6]. Responses can be delivered via phone or via the Internet as files. Note that because we insert a longer pause after each portion of the response (see Figure 2), if a TalkBack-authored response is delivered to another TalkBack and replied to, segmentation works quite well and the “quoted” audio sections around the second reply will capture the desired snippets of the first reply, not the original message.

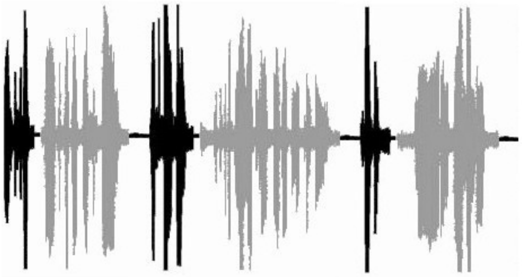

Figure 2. Waveform of message sent back to caller using TalkBack. Each response, seen in grey, is interspersed with a few seconds of the original message time compressed, seen in black.

TalkBack can also be used to display voice-annotated slide shows. As a digital picture frame, it can display pictures in a sequence controlled by the user. These “audio postcards” are composed on a computer and consist of a sequence of picture and wave files.

INTERACTION OBSERVATIONS

TalkBack went through multiple iterative designs, based on informal evaluation by the authors and other members of our group and feedback from many visitors. At this point it was ready for more controlled observations. Although TalkBack is a working prototype, its emphasis is on message response; it is more functional than an answering machine but we did not implement enough other features to allow it to directly replace the voicemail systems our prospective subjects were used to. For the sake of controlling a shorter exposure, we wanted subjects to hear a limited set of messages. However, what sort of message we recorded, and even how it was spoken, would likely influence the outcome of the experiment; a short “are you free for dinner tonight” is unlikely to trigger conversational behavior, while if we ask many questions with pauses between each we could accentuate it.

We also considered trying to obtain both voice messages and photos of people whom the subjects were emotionally close to. We were not able to do so while still maintaining any control for message content, and we also wanted subjects to have not previously heard the test messages. So we observed subjects responding to unknown callers, which likely decreased their attraction to the photograph.

We decided this initial study would expose a small number of subjects to two message genres, “chatty” and “information gathering”, under two conditions: TalkBack and a conventional voicemail system. Although the difficulty of remembering phone numbers and then dialing back after listening to messages is a common complaint of voicemail users, we factored this out by having the voicemail condition play the incoming message and immediately record a response because it might be seen as giving an unfair advantage to TalkBack.

Two other aspects of TalkBack also merit evaluation. Since its hypothesized superiority to voicemail depends on its ability to evoke conversational behaviors, it will be only as successful as its ability to segment the incoming messaging into semantically salient chunks. We also need to consider the effectiveness of TalkBack’s reply format and the content quality of the replies. To accomplish this, we had two additional subjects (the “evaluators”) record all the test messages, and later evaluate the responses received from the main set of subjects.

We decided to not evaluate the “voice postcard” feature of TalkBack because it is somewhat off the theme of conversational messaging, the postcard creation software is quite preliminary, and there was no strong technology or interface for comparison. In interviews, however, our subjects made many comments about the photographic interface, and the sudden emergence of camera-equipped mobile phones suggests an area for future study. We return to this subject later.

Segmentation

We were primarily interested in whether TalkBack would result in a “better” message, how TalkBack users found the user interface and responding process, and how well the TalkBack responses were received by the senders of the original message. But a prior question is whether TalkBack's message segmentation scheme is effective; if it chooses inappropriate locations at which to pause and record, it is less likely to evoke conversational behaviors on the part of message recipients.

We transcribed eight voicemail messages. In the first evaluation, one of the authors, who had not heard the messages, marked the transcript with locations of topic shift or clearly defined questions. In some cases, two adjacent sentence boundaries were marked as equivalent, such as “Are you free for lunch? I'd really like to see you. My holiday was very memorable...” Semantically, the middle sentence could be chunked with either the first or third sentence, so we would consider a TalkBack segmentation at either (or both) locations to be acceptable.

We ran the same messages through TalkBack, and on another copy of the transcript noted all TalkBack segmentations. A third party, not associated with the project, then compared the two transcripts. He considered each segmentation in context, using his judgment to determine whether the segments were more or less the same. For example, if TalkBack segmented after the first sentence in the above example, and the manual segmenter instead opted for the end of the second sentence, they could be considered equivalent.

In a second evaluation, a colleague unfamiliar with the work listened to each of the voicemails and manually paused the playback when he wanted to respond. These points were noted on a transcript of the voicemail and were again compared to the TalkBack segmentations.

73% of the topic shifts found by the author and by our colleague in the second evaluation were spotted by the TalkBack segmentation algorithm. The TalkBack segmentation algorithm was detecting many more pauses than the listeners wanted to respond to, however. 40% of the topic shifts found by the TalkBack algorithm were not interesting to the author and 42% were uninteresting to the second evaluator.

This suggests that the algorithm was looking for pauses much smaller than what legitimately indicated a topic shift. This might suggest that tuning the segmentation algorithm to detect larger pauses might reduce the number of extra pauses found by the TalkBack system or perhaps the minimum message length of three seconds is still too short. In practice, this might not be too much of a problem for the user because he can choose not to respond to a particular segment. Additionally, we note that the hand segmenter often found it difficult to precisely define a topic boundary, as we often use filler phrases to link disjoint thoughts. Also, the person who scored the correlation between the two transcripts disagreed with the manual segmentation in places, suggesting that the numbers quoted are rather soft.

Use of TalkBack

In order to explore the use of the TalkBack interface, we asked colleagues and friends unfamiliar with the project to listen to voice messages and leave responses, comparing TalkBack to conventional voicemail replies.

Two special subjects, whom we call evaluators, each left four voicemail messages. Two of the messages were “informational”; these messages were left for strangers and required that the message leaver ask a specific number of questions about specific things relating to the scenario they were given. The other two messages were “chatty” messages, left for friends, chatting and inquiring about specific personal matters.

Eight other subjects (between the ages of 16-35) were separately given a brief demonstration of TalkBack and had time to get familiar with the interface. They then heard two messages of each genre (informational and chatty), one recorded by each evaluator, in each of the two test conditions (TalkBack and regular voicemail); that is, a total of eight messages. Subjects responded to each message immediately after or while hearing it. After leaving their responses, the subjects rated both conditions with a series of questions on a five-point scale, and then participated in a short structured interview with one of the authors. The order of the messages varied between the subjects in order to counter-balance any biasing effects users might have experienced when listening to one type of message or using a particular system first.

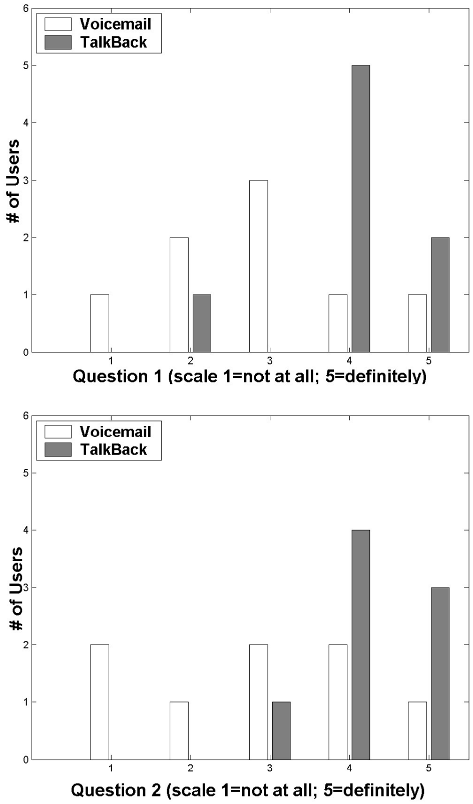

Figure 3. Questions were answered on a scale from 1 to 5; 1 = not at all, 5 = definitely. Q1: *Was the task easy?* Q2: *Did you answer all the questions asked?*

Subjective Evaluation. We were encouraged by the results from the second part of the evaluation. Users seemed to really enjoy using the TalkBack system and found that the conversational style made the system really easy to use. Six out of eight users felt they were "almost in a conversation". And all said they "would definitely use TalkBack" either in the home or office or on a mobile device. The results in Figures 3 and 4 show that more users found the task of responding to voicemail easier with the TalkBack system than with regular voicemail. Although they were not sure that they answered all the questions asked by the caller with voicemail, they were more confident that they answered these questions with TalkBack. Most users felt that they would have said more in a face-to-face conversation than they did on voicemail, however they were undecided whether they would have said more than they did with TalkBack. Users generally seemed to prefer the TalkBack system to voicemail.

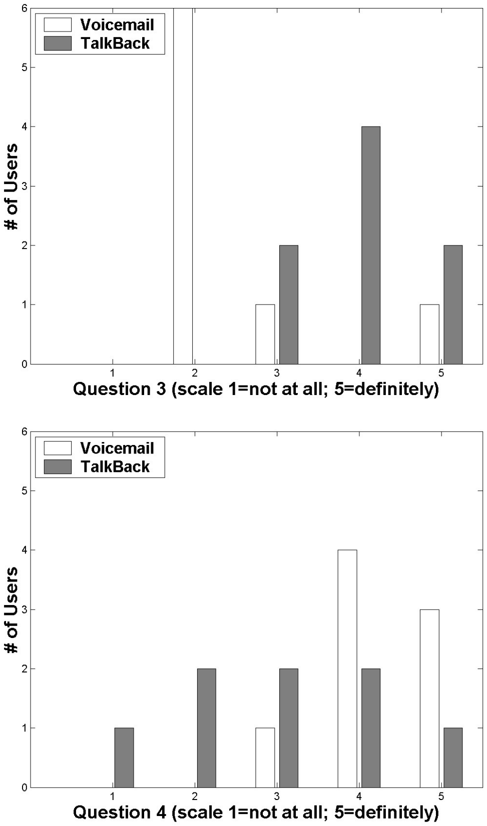

Figure 4. Q3: *Did you like the system?* Q4: *Would you have said more face-to-face?*

Control of Recording. The version of TalkBack tested used pause detection and screen presses to control recording. When playback stopped, TalkBack listened and if the user did not speak after a timeout, no response was logged. If the user began speaking, recording was terminated by silence or by touching the screen. Our users found this timeout to be annoying. Seven of the eight users preferred touching the screen to stop the recording in TalkBack because they did not like being cut off by pause detection if they stopped to think while composing their response. "*I like to have the ability to decide when my thought has been put across.*"

Control of the System. Control of the recording is only one aspect of a broader topic: users generally wanted more control of the system. Six of the eight users mentioned that they would like to have the ability to move backwards and forwards through the sections of the message created by TalkBack. They found the segments to be manageable size chunks. However, they suggested that the ability to jump between sections would give them the greater control they desired. The primary motivation for this control would be to have an overall idea of what the message was about before they chose to respond to certain sections. "*One difficulty with TalkBack was that I didn't have a general plan for the whole message. Given all the information, I may not have answered each individual question in the exact order they were recorded. That said, I didn't have to worry about forgetting details as each segment was manageable as far as the amount of content*." Some users wanted the ability to interrupt the message while it was playing to record a response. However, they felt that if they had an overall idea of the content of the message before they responded, they might not have needed to interrupt.

Use of Pictures. TalkBack's display was designed to be aesthetically pleasing and to be placed on a coffee table or public space rather than hidden in a corner. By making the device a centerpiece, the designers were hoping to make the process of checking for messages more transparent. When new messages arrive, a picture of the caller is displayed on the device indicating that there is a new message; of course this is more attractive when the caller is a friend or family member. If there are no new messages, a default picture that the user selects beforehand is displayed. TalkBack's visual interface was also designed to enhance the feeling of being in a conversation. By displaying a picture of the calling party we hoped that the user would feel almost as if they were talking to someone. Allowing the user to control the pictures that are displayed by default and for each caller also allows a large amount of personalization of the device. This is not dissimilar to people wanting to control the ring tones on their mobile phones.

It would have been optimal had we gathered original, new messages, and pictures from every user and had them respond to their own messages with the TalkBack system. Unfortunately, it would be difficult to control the content of messages and prevent the recipient from listening to them before we used them for evaluation. Therefore we used pictures of random people or pictures that were more representative of the content of the message. These types of pictures carry little or no content for the recipient.

Subsequently, we were not surprised to find that users generally had mixed feelings about the use of pictures. Although all users found the device aesthetically pleasing and eye catching, most only noticed the relevance of the picture after hearing the message. Although one user said she "*felt uncomfortable talking to a still picture*," and another felt distracted by the pictures, the rest found pictures to be a good way to indicate who had left messages. Many users felt that this idea would become more powerful as mobile phones with cameras become more widely available. Half of the users suggested that the caller should have control of what picture is displayed whereas the other half felt that the owner of the device should be able to choose which picture or icon should be displayed for each of their friends. They likened this to *buddy icons* in popular instant messaging clients which allow the user to control which icon gets displayed for each member of their buddy list. It is important to note that the users were not told who the callers were and the pictures displayed were of random people and not friends or family of the users. This might have had an adverse effect on their reaction to the pictures.

Six of the eight users mentioned that they used voicemail primarily while "*on the go*" and would have liked to "*have the functionality of TalkBack, maybe without the pictures, on my cell phone*." They liked being able to reply to messages as they heard them and felt this would be even more valuable while mobile. This is a means of usage that the authors had not envisioned for the device but could be quite effective.

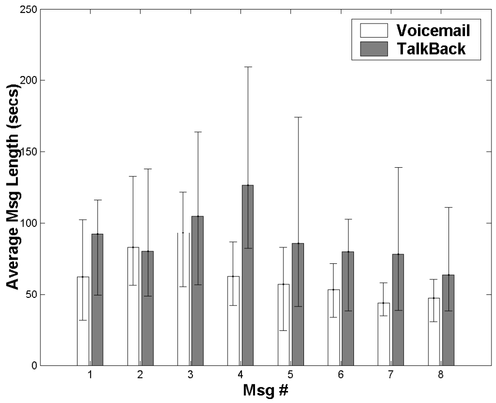

Figure 5. Average response length for each message, with a minimum and maximum bar. TalkBack responses exclude original message.

Response Quality

The final aspect of evaluation was the quality of the responses. We hypothesized that TalkBack would lead to lengthier messages in the "chatty" context, and more complete answers in the "informational" context; we did not know whether length would be positively or negatively correlated with response quality for the informational messages (perhaps simple and direct is better). Because it was difficult to determine good metrics, we relied on the evaluators who recorded the original messages heard by the subjects. The evaluators were more experienced and senior students not affiliated with the project but somewhat familiar with it. The evaluators heard all 32 responses (four from each of the eight subjects) to their original messages, answered several questions about each on a five point scale, and were interviewed at length by one of the authors.

Response Length. As Figure 5 shows, for all but one original message, TalkBack responses were longer, and the longest response for each message was always a TalkBack response. However, it is interesting to note that the differences between message length in the two conditions are greatly affected by the style of the sender (the first four were recorded by evaluator one, and the others by evaluator two). Evaluator one left significantly longer original recordings than evaluator two (mean length 86.964 seconds vs. 53.366 seconds). This helps put in perspective that technology does not overcome differences in style of the human who leaves the original message.

Format of Responses. We were particularly interested in the value of the TalkBack response format. Recall that TalkBack concatenates four seconds of the original speech, time-compressed into two seconds, before each of the response segments, hopefully to provide context when listening to the response. The two evaluators found this method enticing but with some implementation problems. "*TalkBack actually gave me some context for the response but I actually wanted more*."

One evaluator referred to the time compressed segments as "*like a chipmunk*" (although SOLA does not change the pitch), but on further discussion described these snippets as "*too little and too fast*". The second evaluator found that when the snippets were too fast, they were still a good divider between the different response segments. "*Even if the time compressed pieces didn't help, they served as a good breaker*."

These comments point out the limits of iterative design with the designers as subjects. In building and showing TalkBack we became very familiar with test recordings and acclimated to time compressed speech. In a real usage scenario, the evaluators found the inserted snippet duration too short and too fast.

DESIGN IMPLICATIONS

As expected from an exploratory user evaluation, we learned that several aspects of TalkBack require more work, but felt that the conversational answering machine design was validated in large part. In this section we discuss the major design implications which should be incorporated into a refined TalkBack user interface.

Reply recording was the most difficult aspect of the TalkBack user interface; it has been problematic from the start. An initial design paused audio playback between segments, but the user had to touch the screen to start recording. This proved difficult even for the designers of the system because we forgot to touch the screen and simply started speaking at the pauses. We tried using the single LED available on the iPaq for a "recording" indicator, but it is not distinctive enough visually. The second version, tested with several pilot subjects, replaced this method with automatic recording and simple pause detection. At the end of each message segment, the system automatically went into record mode (indicated by the LED and an auditory cue). If the user said nothing, the recording was not saved, and TalkBack played the next portion of the message; this fits nicely with the model of ordinary conversation. But this model has difficulty with termination of recording; if the listener pauses too long to think during a reply, recording can be cut off prematurely. When we made this final silence timeout significantly longer, the conversation lagged with the long pauses and interaction was much less snappy. A third version, tested here, implemented the long trailing pauses to stop recording but also allowed the user to touch the screen to terminate recording if the interaction pace was not quick enough.

This version worked quite well; however, a more complete version of the system would not only allow for pause detection and touch screen termination but also allow the user to touch the screen to interrupt the playback to record a response at any point. This could help compensate for sub-optimal message segmentation, and might help users feel the sense of control they desired.

Although users seemed to want more control over all aspects of TalkBack message playback and management, adding many controls would destroy the simplicity of the TalkBack interface. The screen is small enough that buttons would be hard to see and find. We could, during playback, replace the photo with a control panel, but this might be jarring and would certainly have to be tested.

Control could be greatly enhanced by fast-forward and rewind capabilities, perhaps activated by gesture recognition of touches on the screen. Before playback TalkBack has already segmented the message at more or less significant discourse boundaries, and these make excellent landmarks for intelligent navigation within the message.

The included audio from the original message was clearly too short and played too quickly. Although evaluators suggested that they be able to hear their original message in entirety, we believe this would become tedious. Compressing speech to twice its normal speed, even with an algorithm like SOLA that maintains pitch correctly, seriously degrades intelligibility, and hearing such short snippets, even of one's own speech turned out to be quite difficult. The included audio should be slower and/or longer, and more evaluation is required to determine the proper parameters. As one evaluator pointed out, if more time elapses between leaving the original message and reply generation, more original audio could be included, as the sender may have forgotten more of what was said.

PHONE-BASED TALKBACK

The least convincing aspect of TalkBack was its use of photos, but this might have been expected from our experimental scenarios. We were excited by the subjects' suggestion for using the conversational reply mechanism over mobile phones. To incorporate this idea with several of the users' preferences for more control, we built a second, phoned-based interface to TalkBack. This interface allowed users to respond to the TalkBack message segments using the same turn-taking style of the picture frame interface but it also allowed them to interrupt the message playback to record a response at any time by pressing any key. Recording was terminated by lengthy pause detection or by pressing a key while in record mode. Although this version of TalkBack presents a similar user interface, it was written with a totally different code base, from our previous work on telephone-based information retrieval [17].

We thought it would be useful to receive feedback on this new design from some of the original users of our system. We were able to gather four of our original subjects, two male and two female. Each was asked to respond to four messages, two of each type (chatty and informational), on the TalkBack phone system. These messages were the same messages they had listened to in the phone condition in the previous observation. We assumed that enough time had transpired since they had listened to the messages so they would not remember the original message or their response.

We assumed that by using some of the same subjects our only variable would be the presentation type. However, we caught some of our users in significantly different moods. One was under pressure for time and as a result was very terse with his answers. The other was "having a bad day" and behaved very differently to the whole context. For the two users whose demeanor was more similar to their initial experience with TalkBack, we found that their responses were significantly longer in the TalkBack phone condition than they were in the normal voicemail condition (an increase of 315% for subject 1, and 182% for subject 2). For the four informational messages, these two subjects failed to answer a total of 11 questions (aggregated) with regular voicemail and only missed 3 (out of 32 total questions) on the TalkBack phone system. Although this data is extremely small, it suggests that the TalkBack metaphor may be successful independent of the display.

PHOTOS

We hoped that tagging a voice message with a photo of the caller would promote a feeling of closeness for friends and family, but were unable to evaluate this feature. Our subjects did start to pick up the theme that sometimes they might want to send an image of their choice with the message. The "picture frame" version of TalkBack supports audio annotated slide shows, which are sequences of image and audio files.

We believe an appliance such as this will be a natural destination for "audio postcards" sent from mobile phones and PDAs. We have been experimenting with early generation camera phones; at 320x240 the images are sometimes surprisingly good. One of the authors recently used one on vacation; sample shots can be seen at [7]. Note that the images are much higher resolution than what appear on the phone; they almost beg to be sent to a real display. These shots also illustrate a problem with current phones: the title of each shot is the subject line of the email within which the photo was sent to friends, painstakingly typed with a telephone keypad.

From this limited experience we can say that it feels awkward just sending a photo with no message, but takes longer to type even these short text strings than it does to shoot the image. Nonetheless the approximately 10 friends who received these images thought they were preferable to post cards. So we are working on building a Java application on the phones which can merge voice notes with an image or sequence of images, and send them to a TalkBack device for display.

CONCLUSIONS

One of the authors came up with the idea for TalkBack while thinking about an answering machine for aging, technophobic parents with failing memories. Although they have not been able to test TalkBack, our evaluation with much younger and tech-savvy subjects does lend support to the main underlying design principle of TalkBack, that making asynchronous voice messaging more conversational can make responding to messages more pleasant. There is clearly some evidence, albeit a bit less conclusive, that TalkBack results in "better" responses, though message quality is an elusive concept and varies across message genres, as well as, unexpectedly, the personality of the caller. It is particularly difficult to evaluate messages for social or interpersonal contribution between subjects who are not even friends, and just exchange asynchronous messages. Nonetheless our subjects offered many helpful suggestions and pointed out that the conversational messaging paradigm applies to other settings as well.

This is an exciting conclusion, and suggests new dimensions to messaging. Although they remain a staple of consumer electronics, typical answering machines may become obsolete in a world of voicemail where everyone carries a personal mobile phone. But mobile users can gain special benefit from the alternating play and record cycles of TalkBack, due to cognitive load and difficulty taking notes while mobile. The latest phones now include color screens which could display the caller. Recent emergence of camera-equipped phones enables the exchange of images and could support the creation and transmission of voice-annotated slide shows on the telephone

We have deliberately left somewhat vague the question of whether TalkBack is a telephone or internet appliance, as we see value in both sides. It records ordinary phone messages, can be accessed by phone, and can deliver replies back over the telephone network as analog audio. But messages can also be returned as MIME attachments (the TalkBack frame has an IP address), and the annotated slide show feature is one available only over the computer network at the moment.

However technology changes, messaging will still be important. And having learned, as a species, to take turns talking to each other, we can apply this skill to many messaging architectures.

REFERENCES

1. Balota, D. A., Dolan P.O. and Duchek J.M.*: Memory Changes in Healthy Older Adults*, Oxford University Press, 2000.

2. Chalfonte, B.L., Fish R.S. and Kraut R.: 'Expressive Richness: a comparison of speech and text as a media for revision', Proceedings of Human Factors in Computing Systems (CHI), 1991, pp. 21-26.

3. Clark H. H.: *Using Language*, Cambridge University Press, 1996, Great Britain.

4. Garau, M., Slater M., Bee S., and Sasse M.A.: 'The Impact of Eye Gaze on Communication using Humanoid Avatars', *Proceedings of Human Factors in Computing Systems (CHI)*, 2001, pp. 309-316.

5. Gomi, K., Nishino Y., Matsui H, Nakamura F.: 'A Multi-functional Telephone with Conversational Responses and Pause Deletion Recording', IEEE *Transactions on Consumer Electronics,* 1988.

6. http://www.media.mit.edu/~nmarmas/tb_response.wav

7. http://www.media.mit.edu/~nmarmas/arizona/

8. Marx, M. and Schmandt C.: 'MailCall: Message Presentation and Navigation in a Nonvisual Environment'*, Proceedings of Human Factors in Computing Systems (CHI)*, 1996, pp. 165-172.

9. Mermelstein P.: 'Automatic Segmentation of Speech into Syllabic Units', *Journal of the Acoustical Society of America*, vol. 58, no. 4, October 1975, pp. 880-883.

10. Mynatt, E., Rowan J., Craighill S., Jacobs, A.: 'Digital family portraits: Providing peace of mind for extended family members', *Proceedings of Human Factors in Computing Systems (CHI)*, 2001, pp. 333-340.

11. Nelson, L., Bly S and Sokoler T.: 'Quiet Calls: Talking Silently on Mobile Phones', *Proceedings of Human Factors in Computing Systems(CHI)*, 2001, pp.174-181

12. Reeves, B. and Nass C.: *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*, CSLI Publications, 1999, Stanford University.

13. Rice R. E. and Shook D.E.: 'Voice Messaging, Coordination, and Communication', In C. Egido, J. Galegher and R. Kraut, (eds.), *Intellectual Teamwork*, 1990, pp. 327-350.

14. Rice, R. E. and Danowski J. A.: 'Is It Really Just Like a Fancy Answering Machine? Comparing Semantic Networks of Different Types of Voice Mail Users', *Journal of Business Communication*, vol 4, October 1993, pp. 369-397.

15. Schegloff, E.A.): 'Discourse as an interactional achievement: Some uses of `uh huh' and other things that come between sentences', *Analyzing Discourse: Text and Talk*, 1982, pp.71-93.

16. Schmandt C.: 'Employing Voice Back Channels to Facilitate Audio Document Retrieval', *Proceedings of ACM Conference on Office Information Systems (COIS)*, 1988, pp. 213-218.

17. Schmandt, C. and Arons B.: 'Phone Slave: A graphical telecommunications interface', *Proceedings of the Society for Information Display*, 26(1), 1985, pp.79-82.

18. Wayman, J.L., Reinke R.E. Wilson and D.L.: 'High Quality Speech Expansion, Compression, and Noise Filtering Using the SOLA Method of Time Scale Modification', *23rd Asilomar Conference on Signals, Systems, and Computers*, vol. 2, Oct.1989, pp.714-717.

19. Whittaker, S., Davis R., Hirschberg J. and Muller U.: 'Jotmail: a voicemail interface that enables you to see what was said', *Proceedings of Human Factors in Computing Systems (CHI)*, 2000, pp. 89-96.

20. Whittaker, S,. Hirschberg J., Amento B., Stark L., Bacchiani M., Isenhour P., Stead L., Zamchick G., Rosenberg, A.: 'SCANMail: a voicemail interface that makes speech browsable, readable and searchable', *Proceedings of Human Factors in Computing Systems (CHI)*, 2002, pp. 275-282.

21. Whittaker, S., Hirschberg J. and Nakatani, C.H.: 'All talk and all action: strategies for managing voicemail messages', *Proceedings of Human Factors in Computing Systems (CHI)*, 1998, pp. 249-250.

22. Yankelovich, N., Levow G. and Marx M.: 'Designing SpeechActs: issues in speech user interfaces', *Proceedings of Human Factors in Computing Systems (CHI)*, 1995, pp. 369-376.