A CONVERSATIONAL TELEPHONE MESSAGING SYSTEM

Chris Schmandt and Barry Arons
Architecture Machine Group
Massachusetts Institute of Technology

The *Phone Slave* is a personal, integrated telecommunications management system, combining diverse message functions in a single user interface on a small general purpose computer. This paper will focus on the audio components of that interface (a related publication emphasizes the graphical interface [4]).

The Phone Slave is an intelligent answering machine, conversing with callers to format messages and relaying personal greetings to identified parties. Its owner can access these voice messages as well as electronic mail via speech recognition or Touch-Tones over the phone network. Access to both incoming and outgoing messages, an on-line directory, and autodial features are also provided by a touch-sensitive color monitor.

Conversational Answering Machine

When the machine answers the phone, it quickly engages the caller in a conversation, playing back stored speech segments, while recording the caller's responses. By asking questions such as *"Who's calling?"* and *"At what number can you be reached?"*, the machine formats message segments for its owner. These questions are phrased both for completeness of the message, and to enable response to queries such as *"Who left messages?"* by the owner.

machine: *Hello, Barry's telephone speaking, who's calling please?*
caller: Marc Spehlmann.
m: *What is this in reference to?*
c: Speech Compression.
m: *He's not available at the moment, but he left this message:* <Barry's voice>I'M HAVING LUNCH AT PRIMO'S, I SHOULD BE BACK BY 1.00. *At what number can he reach you?*
c: 258-6681
m: *When will you be there?*
c: I'll be here all night.
m: *Can I take a longer message?*
c: I have the CVSD hardware running on the IBM PC, I'll give you a demo as soon as Dave Chen gets the software running.
m: *I'll be sure he gets that. Thanks for calling. Goodbye.*
c: Bye.

This well timed series of questions guide the caller, who may not be prepared to simply recite a message to a tape recorder. The conversation can take a number of branches to exchange the maximum amount of information between both parties (figure 1).

An adaptive pause detection scheme is used to determine when the caller has finished answering each question. The goal is a snappy response without cutting off those who speak haltingly. The *end of utterance* timeout for pause detection is initially set to 1.25 seconds. If intermediate length pauses of greater than .75 seconds are detected within the speech, this timeout is gradually lengthened to 2 seconds.

It is essential for a segmented playback scheme that the reply to each question be specific and reasonably brief. To counteract callers who ramble or try to answer more than one question at a time, each response has an associated maximum length. If the caller exceeds it, the machine interrupts in a louder voice, politely asks the caller to be precise and repeats the question.

While recording, all reply segments are power normalized to be of approximately the same volume on playback. A poor telephone connection is detected when speech signal levels are not much higher than background noise during pause detection. At this point the machine asks to the caller to speak up and restarts.

A number of possible voices and dialogs were experimented with. While not attempting to deceive callers into believing they are speaking with a person, it is important that they realize this is not a typical answering machine. The machine identifies itself as *"Barry's telephone speaking"* in a pleasant voice which is clearly different from the owner. The owner's voice is heard only to deliver the outgoing message, which is of course changed frequently.

Caller Identification

The answer to the first question, *"Who's calling?"*, is processed by a speech recognizer simultaneous with recording (figure 2). If a match on the voice pattern of a frequent caller is obtained, the conversation branches, with the caller being greeted by name and playing a personal recording for that specific caller.

IEEE Transactions on Consumer Electronics, August 1984, Volume CE-30.

Manuscript received June 11, 1984.

As a backup or possible substitute for speech recognition, the caller may answer the *"Who's calling?"* question by keying in her own phone number with Touch Tones. A familiar caller expects to be greeted by name after identifying herself. If, instead, the machine just asks *"What's this in reference to?"*, she can still key in an ID, at which point the machine apologizes and delivers any personal messages.

This branch of the conversation tree asks whether the caller can be reached at her usual number, informs her if her last message has been heard by the owner, and if not says *"If you'd like to leave a (another) message, I'll record it now, otherwise hang up and I'll tell him you called (again)"*.

The machine encourages participation by providing a variety of options in message type and responding personally to all callers. Most important is the prospect of a specific message with greater content than the generic outgoing "I can't answer my phone right now." A dialog may occur through a series of calls by the owner and a friend, although the parties never connect directly.

Even a previously unknown caller may benefit from this treatment. After a call by an unrecognized person is finished, the digitized voice of their answer to *"Who's calling?"* is used to train a new template in the speech recognizer. On calling back, they will be informed whether the owner has heard their message, receive any personal reply, and be asked if they wish to leave another message.

Message Retrieval

Messages are recorded as a series of distinct audio segments, to facilitate message access. The machine may playback individual responses, or a series of responses to indicate who left messages, or the entire content of a single message (figure 3). Local access is by a touch-sensitive display (figure 4), with remote access by speech recognition or DTMF tones.

The owner may access all message components remotely over a phone connection, leave a new personal reply for any caller, or request the time of a call or the caller's phone number from the directory. The machine switches between a *command mode*, during which the owners speech is being interpreted though the recognizer, and a *record mode*, during which replies are being recorded until a significant pause is detected.

owner: Hi this is Barry.
m: *Hi Barry*, you have three new messages, one from MARC SPEHLMANN
o: What's it about?
m: SPEECH COMPRESSION.
o: What's the message?
m: I HAVE THE CVSD HARDWARE RUNNING ON THE IBM PC, I'LL GIVE YOU A DEMO AS SOON AS DAVE CHEN GETS THE SOFTWARE RUNNING.
o: Who left messages?
m: MARK SPEHLMANN, Chris, Walter.
o: Next Message.
m: One from Chris, it's about "Japanese Video Tape."
o: What'd he say?
m: We will re-shoot Phone Slave videotape, in English and Japanese...
o: When was it?
m: Message received yesterday afternoon at three.
o: Take a reply.
m: *Ready to record.*
o: Hi Chris, it should only take about 3 days to get the Phone Slave software to run in Japanese. Lets schedule taping with Shigeru tomorrow.
m: *Stopped recording, mail was sent.*
o: Next Message.
m: One from Walter Bender, this is a very long message, its about "Anti-Aliased Line Routines."
o: Goodbye.
m: Goodnight.

Although speech recognition over telephone lines [1, 3] is improving, accuracy degrades significantly with the noise levels frequently found on trunk lines. As a limited backup, a significant subset of the commands are provided through Touch-Tones. A small but comprehensive set of *single stroke* commands were selected (figure 5), in the belief that this would be more useful than the full set of command implemented with either multiple stroke entries or sub-menus.

Unified Electronic Mail

Electronic mail messages are integrated with voice messages, and may be viewed on the screen, or heard over the phone with a text-to-speech synthesizer. On the prototype system in use in our laboratory, it is quite common to receive both forms from the same person, and they are grouped together appropriately for easy access. This allows text replies to voice messages and vice-versa.

Several limitations of synthetic speech have been addressed. The first is intelligibility, which may be disappointingly low. As a listener is exposed to a particular synthetic speech peripheral, and becomes accustomed to it, misunderstanding errors decrease significantly, much as one improves in ability to understand a regional or foreign accent [5]. Pronunciation is improved through an on-line exception dictionary, translating names, local jargon, or other confusing words into an alternate spelling for correct pronunciation.

A second intelligibility aid is a REPEAT command, which replays text, starting from the previous sentence, at a slower rate. The second invocation of REPEAT spells the sentence in question letter by letter.

Even though word by word understanding may become fairly high with usage, this takes some effort, such that a listener is less likely to comprehend the meaning of the sentence or paragraph being spoken [2]. To avoid clutter in the speech channel and minimize memory demands, header information, such as the date and time of message delivery, is withheld until requested. With similar intent, messages are grouped according to the sender, so all messages from a particular source are played sequentially.

A voice reply to a text message may be taken, in which case mail is sent informing the original sender that a voice message awaits, giving the phone number and an access code. The universal accessibility of the phone network allows speech to be transmitted anywhere, so all sound storage can be local, with no assumptions about remote site capability or message protocols.

Acknowledgments

This work has been funded by grants from Atari, Inc. and NTT, the Nippon Telegraph and Telephone Company. Speech synthesis hardware was supplied by Speech Plus. The authors also wish to thank Marc Spehlmann for his dedicated work building the telephone interface hardware.

References

1. S. E. Levinson, A. E. Rosenberg, J. L. Flanagan. Evaluation of a Word Recognition System Using Syntax Analysis. *The Bell System Technical Journal* 57, 5 (May-June 1978), 1619-1626.

2. P. A. Luce, T. C. Feustel, and D. B. Pisoni. Capacity Demands on Short-Term Memory for Synthetic and Natural Word Lists. *Human Factors 25*, 1 (1983), 17-32.

3. A. E. Rosenberg and C. E. Schmidt. Automatic Recognition of Spoken Spelled Names for Obtaining Directory Listings. *The Bell System Technical Journal 58*, 8 (October 1979), 1797-1823.

4. Christopher Schmandt and Barry Arons. Phone Slave: A Graphical Telecommunication Interface. Digest of Technical Papers, SID International Symposium, 1984.

5. L. M. Slowiaczek and D. B. Pisoni. Effects of Practice on Speech Classification of Natural and Synthetic Speech. *Journal of the Acoustical Society of America 71* (1982).

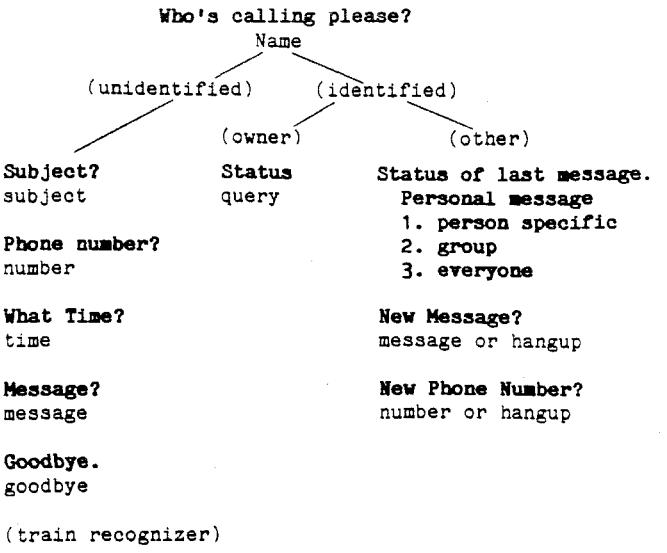

Figure 1: Tree of possible conversations.

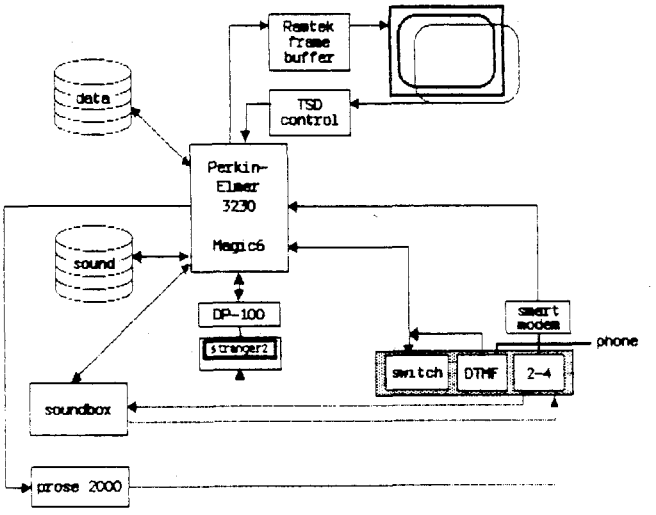

Figure 2: Hardware configuration used in Phone Slave.

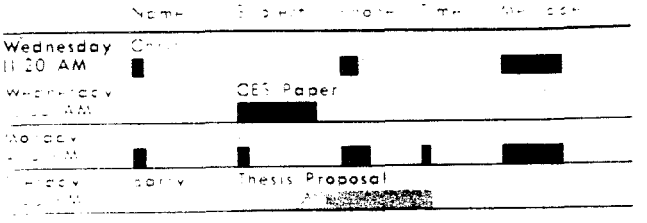

Figure 3: Message screen with text and voice messages.

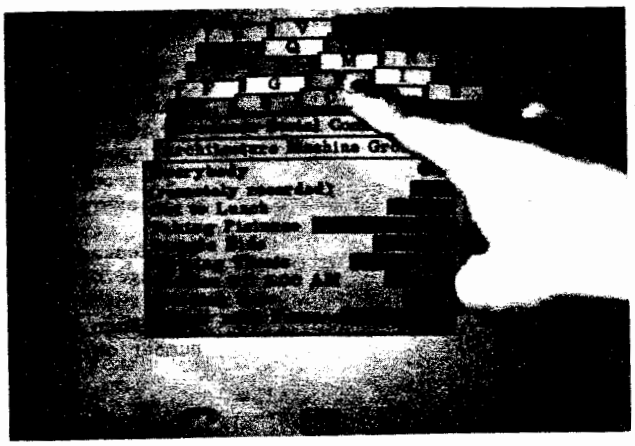

Figure 4: Touch screen access to the on-line directory.

1 Next Message	2 Previous Message	3 Repeat
4 Next Sender	5 Previous Sender	6 More Info
7 Yes	8 No	9 Reply
* Cancel	0 Pause/ Continue	# Quit

Figure 5: Commands available through the keypad.