Speech Synthesis Gives Voiced Access to an Electronic Mail System

MIT's Architecture Machine Group uses Speech Plus's Prose 2000 to read and answer their electronic mail.

Chris Schmandt
Research Associate
Massachusetts Institute of Technology
Cambridge, Mass.

THE "VOICED MAIL" system of MIT's Architecture Machine Group allows users of a research computer system to access their electronic mail through a text-to-speech synthesizer (Speech Plus's Prose 2000) and a touch-tone telephone. In an environment of heavy on-line mail usage, this system has gained acceptance by a community of about 30 subscribers needing to both read and transmit messages.

To use Voiced Mail, a user calls in and gives a unique identifier (home phone number) and a password by touch-tones, much like using an automatic bank teller.

Mail messages are sorted by source, and played sequentially within each group. The caller may interact with the system by touch-tones, to jump to the next message or next sender, obtain more information about the sender, repeat a sentence, or make a reply.

Several types of replies may be generated. The caller may send back an electronic message of the form "I read your message about [subject line] and the answer is *yes*," or *no*, or *please call me at* [a telephone number], which is then keyed in.

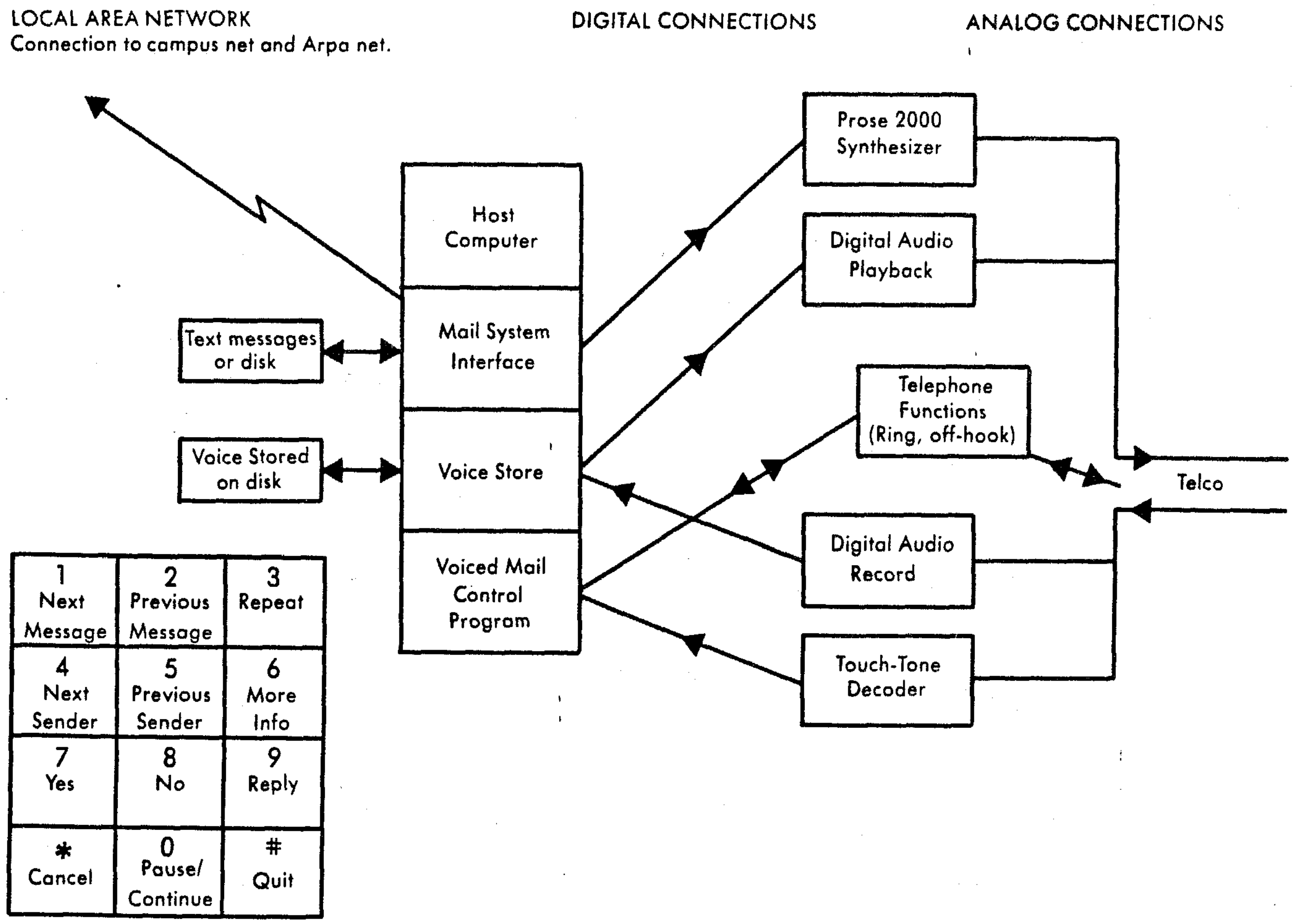
A recent addition allows the reader to record a voice message for the sender, which can be accessed through a voice storage system which transparently integrates recorded and text messages.

Although this system was developed as a component of a wider research activity in telephone access to multi-media message systems, it was found useful enough as an internal utility to be implemented as a stand-alone system and left running 24 hours a day on one of the laboratory's minicomputers. Major portions of Voiced Mail have been incorporated as is into a personalized text and voice message storage and answering machine. [1, 2].

One of the main motivations for electronic mail systems is convenience of access. On-line message systems enable quick and reliable communication between parties or groups in disparate locations and perhaps on differing schedules.

Rapid message delivery and 24-hour availability make it ideal for communication or decision making on issues requiring timely response. To use conventional electronic mail systems, however, one must carry a terminal and

Fig. 1



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

A keypad menu was devised to simplify control functions.

modem to read and send messages. A telephonic text-to-speech interface removes this restraint.

Speech synthesis can gain acceptance as a viable new means of *access* to existing text databases much faster than it can in a totally new environment or application. Users already accustomed to some service seem much more tolerant of the limitations of synthesis when it clearly makes that service more available, while experiments using synthesized speech in completely new situations (e.g., talking supermarket cash registers) have been much less successful. It is important to note that all the users of Voiced Mail were already experienced and regular users of one or more electronic mail systems.

Implementation

Although text-to-speech synthesizers accept streams of ASCII characters just like a terminal, from the user's end their behavior is very different. The user interface must be designed to compensate for this. Both *intelligibility* and *speed* of synthesized speech interfere with data transfer. Speech is necessarily *serial*, which hampers menu presentation and the conceptual framework within which data are presented to the user. With telephone access, a further consideration is that the single voice channel must be *multiplexed for both data and control functions*.

To minimize short-term memory requirements already taxed with synthesized speech [3], and to simplify the conceptual framework of the control functions, a simple, single-keystroke telephone keypad menu was devised (see figure).

The slow, serial nature of speech output mitigated against complicated submenus. By the time the list of options has been recited, one has already

MR. SCHMANDT *received his B.S. in Computer Science and M.S. in Computer Graphics from M.I.T. He is currently a Research Associate there at the Architecture Machine Group, a component of the new Media Laboratory. His work is focused on human-computer interaction, particularly voice input/output in telecommunication systems. Some of the concepts behind the application presented in this paper are currently being developed by Active Voice of Seattle, WA.*

forgotten some of them. The menu chosen provides comprehensive single-key functionality at the price of a few extra features which could not be implemented.

Commands

Grouping message recital by sender rather than by sequential time order was also done to simplify organization.

Commands, such as next or previous, can operate on a single message or a group. A pause/continue command allows speech to be stopped at any moment; it is resumed from the beginning of the current sentence.

A repeat command replays the last sentence more slowly, which makes it more intelligible, and allows the rest of the dialog to continue at a higher speech rate. The second invocation of repeat spells the sentence letter by letter.

It proved important to minimize the amount of information transmitted with each message to speed up the interaction, as users consistently found the system slow. The headers associated with each message were not transmitted (they are often longer than the message body itself), although a "more info" key recites the sender's full name, mailing address, and time of message. The synthesizer warns "This is a very long message" if the body exceeds 200 characters. Such messages are often ignored until the recipient is at a terminal.

Another aid to responsiveness is that the system is always interruptible, an aspect which is vital to its success. If one is reading one's mail from a telephone, one is often in a hurry, and it is important to abort playing a message once it becomes clear that it is irrelevant or cannot be acted upon until one is in one's office. It also allows a convenient adaptability between naive and experienced users. While speaking instructions, such as how to enter a password or a list of menu options, the first digit entered immediately cuts off the explanation. The naive user can receive a lengthy tutorial, while the experienced user just touches in a series of commands and skips all narrative.

User Experience

This system supports a user population of about 30, of which half are occasional and the remainder call in three to ten or more times per week. Many user comments were offered and many suggestions incorporated into successive versions of the mail facility.

All users found they could understand most of their messages most of the time, supporting, informally, the observation that speech understanding increases rapidly after even a short exposure [4].

A significant increase in system usage was noted after it was integrated into the voice storage system; at this point, it became much more reasonable to initiate messages remotely, and of course replies are much richer. Many users prefer to leave a voice message, or leave a much more detailed message, than they would if a secretary has to transcribe it on paper.

Of course, speech access to electronic mail is no substitute for CRT access if one is in one's office. The greatest use is among those on differing schedules, those who travel frequently and members of a software team working split shifts to maximize computer time.

Voice Mail was designed by and for users of an existing mail environment. Its success demonstrates that appropriate user interface techniques as well as incorporation of user feedback and simplicity of command structure can render synthesized speech a viable means of remote access in a society experiencing a growing reliance on telecommunications and speed of information exchange.

REFERENCES

1. **C. Schmandt** and **B. Arons** "A conversational telephone messaging system," in Digest of Technical Papers, IEEE International Conference on Consumer Electronics, 1984.
2. **C. Schmandt** and **B. Arons** "Phone Slave: a graphical telecommunication interface," in Digest of Technical Papers, Society for Information Display International Symposium, 1984.
3. **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.
4. **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.

FOR FURTHER INFORMATION

Contact the Architecture Machine Group, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Mass. 02139.