REMOTE ACCESS TO VOICE AND TEXT MESSAGES

Chris Schmandt
Research Associate

Media Laboratory
Massachusetts Institute of Technology

3 August 1984

ABSTRACT

The "Voiced Mail" system of M.I.T.'s Architecture Machine Group uses synthesis-by-rule to allow remote access to an electronic message system. A voice storage audio message system is integrated with the text message database, to allow users to send and receive voice replies to text messages and vice versa. Message access is interactive, controlled by Touch-Tones sent by the user, and incorporates many features of conventional on-line mail systems.

INTRODUCTION

A side effect of the growing use of computers for storage and management of a wide variety of textual databases is reliance on CRTs and computer terminals to access that information. Although such hardware may be a common resource in office environments it is often not available elsewhere. This limits access to timely information which may be stored on-line.

The ability of text-to-speech synthesizers to verbalize ASCII text streams and the pervasive nature of the telephone network suggests audio access to such databases over ordinary telephone lines. This paper describes an experiment in such voice access to an on-line electronic mail system. Key issues are the suitability of the database for voice output and human factors concerns with the utility of such output.

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. Text messages are merged with voice messages gathered by a voice mail subsystem in a manner which is transparent to the caller. 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.

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).

VOICED MAIL

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. Messages, both text and voice recordings, are merged, 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 a make a reply (figure 1).

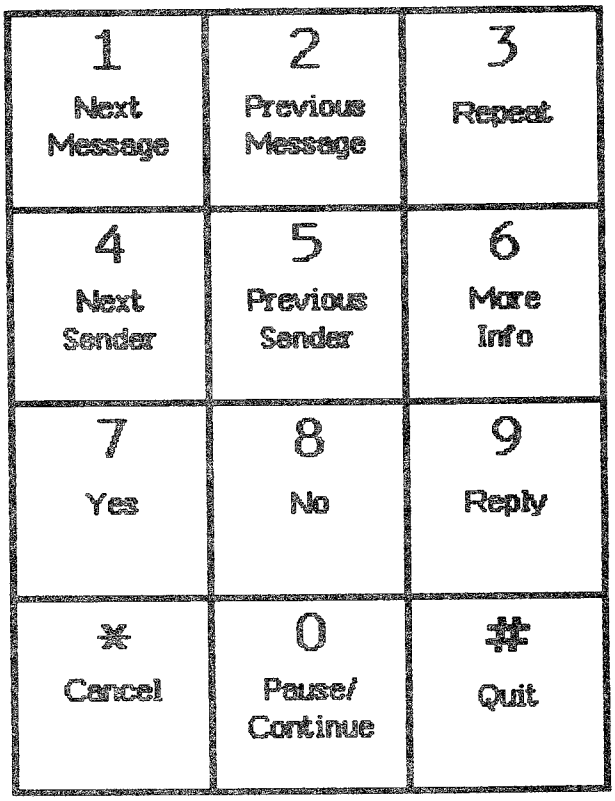

Figure 1. Telephone keypad command layout.

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. The reader may also record a voice reply for any sender, who will then be sent a text message informing them of the reply and containing instructions for telephone access.

Although speech recognition was included in the hardware arrangement (figure 2), it was not used except for demonstration purposes because of poor performance. This can be attributed to the fact that most remote access in the early stages of development were over long distance phone lines, which tend to be particularly noisy, and the fact that the recognizer (NEC DP-100) was not intended for telephone use.

Voiced Mail is an overlay to an existing text message system, and in no way interferes with its operation. Inherent in the design philosophy is the belief that **interaction** with and **organization** of information will be different for screen and voice access. When a caller is identified, their mail file is parsed and translated to Voiced Mail's own internal description; pointers into the mail file indicate the body of each message, but the messages are not presented in the order in which they were received, nor is all the text of the file spoken to the caller.

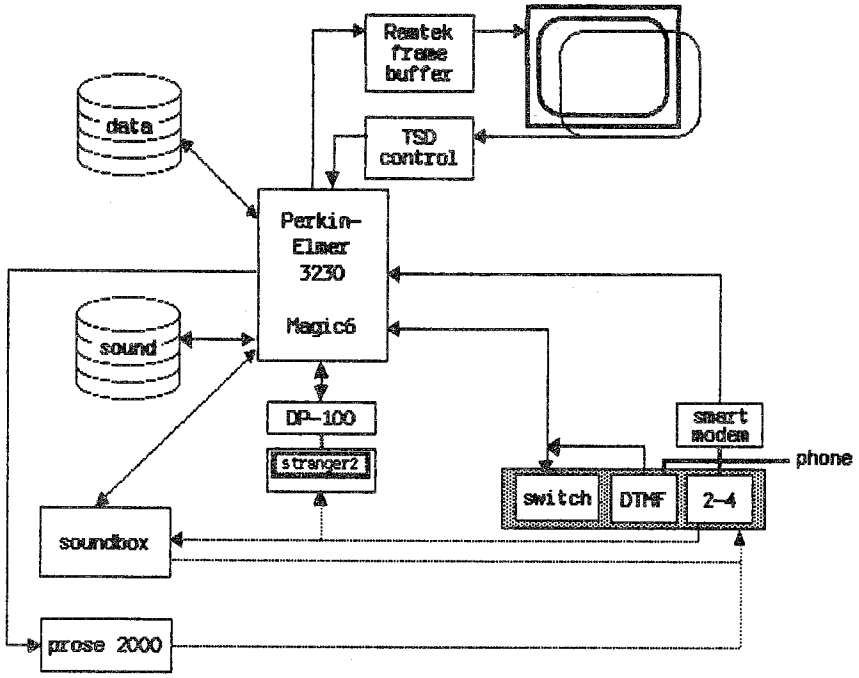

Figure 2. Voiced Mail hardware configuration.

MOTIVATION

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 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.

Many of the limitations in synthetic speech can be overcome, or at least compensated for, by the interface which allows a user to control information access. Only if these limitations are acknowledged can a system robust enough to satisfy users other than dedicated speech researchers be implemented.

SYNTHESIZER LIMITATIONS

Four main problems are encountered with synthetic speech employed for computer generated dialogue: its overall quality, word intelligibility, the speed of speech, and the fact that speech is serial or time sequential in nature.

The primary concern with voice output must be intelligibility, which may prove difficult to quantify (3). Modified Rhyme Tests (4), a common metric, play back a single word, and subjects are asked to identify it from a list of similar sounding words; this is particularly useful for evaluating the correctness of text-to-speech rules. But scores on such a test may not accurately reflect average intelligibility, as one may miss several words of a sentence and still understand the whole. Listeners are very likely to have difficulty understanding unfamiliar words, names, or acronyms.

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). Studies indicate, however, that even though word by word understanding may become fairly high, this takes some effort, such that a listener is less likely to comprehend the meaning of the sentence or paragraph being spoken (6).

Of particular note in applications using synthesis is the necessity for clause and sentence level prosody generation in the synthesizer (7). If voice messages exceed one or two sentences, the monotonous pacing of less sophisticated synthesizers inhibits syntactic perception, which in turn limits one's ability to use syntactic context to improve understanding. The least expensive synthesizers are barely intelligible to previously unexposed listeners.

An important consideration for the use of speech output is its speed. Speech is quite slow as compared to computer terminal baud rates and human reading speeds. In situations requiring rapid response, voice output must be kept terse. In applications necessitating reading a large amount of text (e.g. accessing a database by telephone), as any extraneous information should be removed before it is spoken to the user.

Finally, in applying speech synthesis to applications which formerly used CRT screens, the temporal nature of speech must be kept in mind. Although one may display a number of menu options simultaneously on a screen, they must be spoken sequentially under voice access. If there are a number of choices, the first may well be forgotten before the last is spoken, especially in light of the concentration required for understanding synthetic speech. Such a list of recited options

is necessarily serial, whereas the eye can (and does) skip from place to place on the page or screen scanning for the required information.

IMPLEMENTATION

To minimize short-term memory loads and simplify the conceptual framework of the control functions, a modeless, single-keystroke telephone keypad menu was devised (see figure 1). The slow, serial nature of speech output mitigates against complicated sub-menus. The menu chosen provides comprehensive single key functionality at the price of a few extra features which could not be implemented. Slightly expanded functionality was provided when speech recognition was used for input.

Grouping message recital by the sender, rather than the normal time sequential order, is also done to simplify organization. It is often the case that a single sender will mail a series of related messages about a common issue, and it makes sense to hear those messages together. Commands, such as **next** or **previous**, can operate on a single message or a group.

It is 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. Likewise, the subject of a long message will be spoken, and the caller will be asked whether to continue, e.g. "This is a very long message. It's about the price of widgets in China. Do you wish to hear it?" A short message will simply be spoken, as it takes less time to hear the message than it would to answer questions about it.

Another aid to responsiveness is that the system is <u>always interruptible</u>, an aspect which is vital to its success. A **pause/continue** command allows speech to be stopped at any moment; it is resumed from the beginning of the current sentence. In fact, any command will be responded to immediately at any time. It is important to be able to abandon playing a long message which is not urgent and hence may be more appropriately read at a terminal.

Interruption 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 keys in a series of commands and skips all narrative.

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. 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.

USER EXPERIENCES

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. Being a small research group, many user comments were offered and many suggestions incorporated into successive versions of the mail facility.

Users generally found the single keystroke menu system adequate and easy to learn. Though some users simply listen to their messages in order (the system will do this even if no keys are pressed), others make common use of the random access and extended function features.

In an early version of Voiced Mail, in which only text replies could be generated, very little use was made of this facility. When voice storage was added, general interest in the system increased significantly. A significant portion of the calls in are primarily for the purpose of leaving a voice message for another party; these messages are generally much richer than those left if a receptionist transcribes a message.

Voiced 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.

ACKNOWLEDGMENTS

Barry Arons and Caren Baker, as graduate and undergraduate students respectively, contributed both software and design input to the Voiced Mail system. Members of the Architecture Machine Group acted as Guinea pigs and in turn offered additional input which is reflected in the system's final design.

Portions of this work were funded by NTT, the Nippon Telegraph and Telephone Public Corporation. Speech Plus assisted this project with
hardware support.

REFERENCES

1. Schmandt, C. and Arons, B
A Conversational Telephone Messaging System.
In Digest of Technical Papers. IEEE International Conference on Consumer Electronics, 1984.

2. Schmandt, C. and Arons, B.
Phone Slave: A Graphical Telecommunication Interface.
In Digest of Technical Papers. SID International Symposium, 1984.

3. Miller, G.A., Heise, G.A. and Licten, W.
The intelligibility of speech as a function of the context of the test materials. Journal of Experimental Psychology 41:329-335, 1951.

4. House, A.S., Williams, C.E., Hecker, M.H.L., and Lryter, K.D
Articulation testing methods: consonantal differentiation with a closed response set.
Journal of the Acoustical Society of America 37:158-166, 1965

5. Slowiaczek, L.M. and Pisoni, D.B.
Effects of practice on speech classification o natural and symthetic speech.
Journal of the Acoustical Society of America 71, 1982.

6. Luce, P.A., Feustel, T.C., and Pisoni, D.B.
Capacity demands short-term memory for symthetic and natural word lists.

Human Factors 25(1):17-32, 1983

7. McPeters, D.L. and Tharp, A.L.
The infuence of rule-generated stress on computer synthesized speech.
International Journal of Man-Machine Studies 20(2):215-226, 1984.

BIOGRAPHY

Chris Schmandt
Research Associate

Media Laboratory
Massachusetts Institute of Technology
Room 9-516
77 Massachusetts Avenue
Cambridge, MA 02139

Mr. Schmandt received his B.S. in Computer Science and his M.S. in Computer Graphics from M.I.T. He has continued there as a Research Associate at the Architecture Machine Group, a component of the new Media Laboratory. His research interests there are focused on interactive systems and human-interface issues, with emphasis on voice interaction and telecommunications. Some of the concepts presented in this paper are being developed by the author in conjunction with Active Voice of Seattle, WA.