REPORT:

SPEECH PROCESSING FOR AT&T WORKSTATIONS

Authors:
John G. Ackenhusen *is a supervisor in the Speech Processing Department at AT&T Bell Laboratories in Murray Hill, New Jersey, and* ***Syed S. Ali*** *and* ***James G. Josenhans*** *are members of technical staff in that department.* ***John W. Moffett*** *is a supervisor in the Workstation Systems Applications Development Department at AT&T Information Systems in Middletown, New Jersey, and* ***Reuel R. Robertson*** *and* ***Jaime R. Tormos*** *are members of technical staff in that department. Mr. Ackenhusen is responsible for developing efficient algorithms, software, hardware, and silicon for real-time speech processing. He joined AT&T in 1978 and has a B.S. and M.S. in physics, and a B.S.E., M.S.E., and Ph.D. in nuclear engineering, all from The University of Michigan. Mr. Ali is responsible for soft-*
(continued on page 67)

Introduction

Speech processing can exploit the growing overlap of information processing and information transportation typical in many of today's advanced workstations that combine computing and telephony. With speech input and output capabilities added, the ubiquitous telephone becomes a remote terminal for the personal workstation, allowing access to its features from anywhere in the world. For example, features such as voice mail, remote access to text mail using text-to-speech synthesis, and access control using speech recognition can reduce time lost to the unproductive game of telephone tag.

This paper describes the Voice Power speech processor, a speech processing option for the AT&T UNIX® PC. It is a peripheral card (with software) that slides into an expansion slot on the workstation and adds the capability for speech store and playback, speech recognition, and text-to-speech synthesis to the UNIX PC.

The initial offering of application software is also described. This software uses a subset of the hardware's capabilities: speech storage and playback, and text-to-speech synthesis.

The cost of speech processing has been falling with the arrival of efficient algorithms, very large scale integrated circuits, and digital signal processors (DSPs).

The design described here adds a generic speech processing capability in a way that is tightly integrated with the workstation. Besides decreasing the incremental cost of the speech processing, it increases the synergy of speech processing with computing, and allows expansion of the speech processing capability by upgrading software rather than replacing hardware.

Applications

The initial Voice Power product offering on the UNIX PC includes a voice editor, voice mail, telephone answering machine, and remote access. Speech recognition is not used in these initial applications.

Voice Editor. The Voice Power voice editor provides the primary means for creating or hearing voice files on the workstation. It integrates the simplicity of a tape recorder and features that are possible only with digital speech processing.

Because its simple, tape recorder functions—like play, rewind, and record—are already familiar to anyone who has used a recorder or dictating machine, first-time or casual users can use the editor (and entire Voice Power product line) effectively. At the same time, sophisticated users can take advantage of functions like insert, delete, and cursor positioning with the mouse that depend on the digital nature of voice data.

As with screen-based text editors, the voice editor displays the file on the screen and a cursor identifies the current position. The voice appears as a bar along a time axis (Figure 1); at a glance, the user can see the length of the file in minutes and seconds.

The bar consists of both thick and thin sections that identify, respectively, segments of

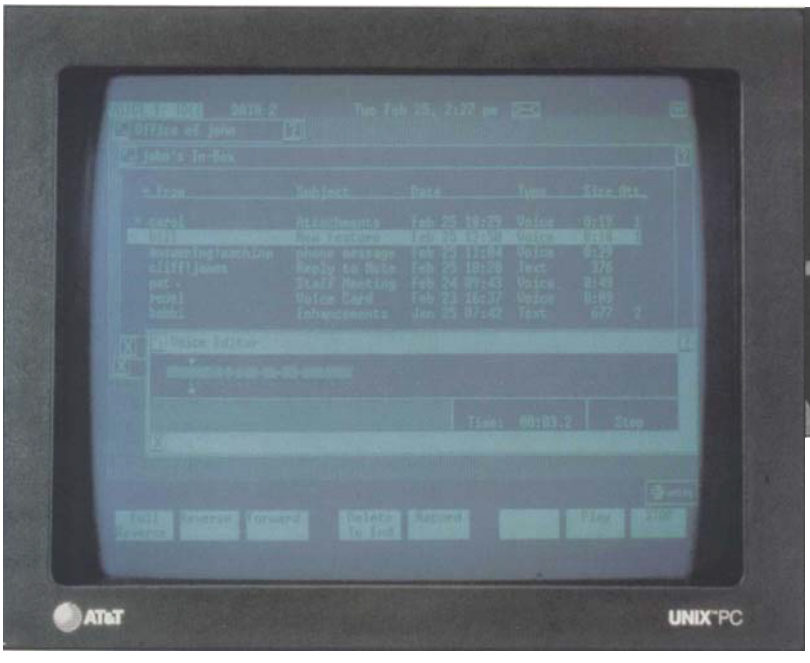

Figure 1. Voice Power voice editor. The thick and thin sections of the bar represent the segments of speech and segments of silence, respectively, in the voice message file. The vertical cursor on the bar identifies the current position in the file. This display also shows a list of the users inbox that contains both voice and text messages.

speech and segments of silence. This helps the user to perceive the relationship between what appears on the screen and what is heard. It also simplifies many tasks such as repeating or deleting a phrase, inserting a sentence, and positioning the cursor within the voice file.

Voice Mail. Voice mail is designed to provide a sophisticated voice-communication capability between UNIX PCs that are equipped with Voice Power. Voice files are transmitted between machines just like text electronic mail. This package also provides all the functionality of the UNIX PC's standard text-mail package, but with fully integrated text and voice mail. As such, it is compatible with a wide range of UNIX system text-mail packages available on computers from AT&T and other vendors.

If a user is already familiar with electronic mail, then voice mail is an easy transition. Complete symmetry is maintained between text and voice mail, and the voice editor provides access to the voice data. Users of text mail will often find the process of sending voice mail is much easier. And because the speech is of high quality, the sender's voice is recognizable, adding a personal touch to the message.

Sending mail. To send a message, the user first fills in a *mailing envelope* with the recipient's electronic-mail address, subject of the message, and type of message desired. If the type *Text* is selected, the user's text editor is invoked for creating the text message. If *Voice* is selected, the voice editor is invoked, and the user can record a voice message.

After creating a message of either type, the user is given options: add attachments, such as text, voice, spreadsheets, or graphics; re-edit the message; or change the mailing envelope. Finally, the user sends the message.

Once the user sends a message, voice mail operates unattended in the background. It makes sure the message gets delivered, retries if needed, and, on request, gives the user the status of any message. If the message was sent to several people, voice mail ensures that all the deliveries are made.

Receiving mail. When a voice mail user receives a message, a mail icon appears at the top of the workstation's screen.

Messages are delivered to the work-

station's electronic *inbox*. For each message, the inbox shows (Figure 1): who sent the message; subject; date and time sent; message type (text or voice); length (bytes for text, and minutes and seconds for voice); and how many attachments, if any, are included.

The user can select messages in any order, and the correct editor (voice or text) is invoked automatically. Any message may be forwarded, with or without editing, to another UNIX PC user either directly or as an attachment to a new mail message.

Telephone Answering Machine. The Voice Power option comes with a telephone answering machine that can run unattended in the background, leaving the UNIX PC free to do other work. This also means a user doesn't have to remember to activate the answering machine every time he or she leaves the office.

To an outside caller, the Voice Power answering machine sounds like a conventional answering machine. But to the workstation user, it provides a much larger set of features. Of course, it supports high-end, conventional answering machine features such as voice activated recording, message indicator, variable ring count, call screening (with an optional speaker), remote message retrieval, and remote recording of the greeting message. Enhancements include variable ring count; date and time stamped messages; and a user-set, message length limit (up to four minutes).

Digital speech processing enables the Voice Power answering machine to provide features that are not possible on conventional answering machines. They include both interactive and automatic message forwarding, multiple greeting message storage, and remote access.

Message forwarding. With interactive message forwarding, when a user receives a phone message that someone else could handle better, the user doesn't have to take notes to pass the message along. Instead, the user can insert an introduction to the message in his or her own voice and then forward the annotated message to another Voice Power user.

Another useful business feature is automatic message forwarding. If the user is away and someone else should handle all phone messages, they can be forwarded automatically to another Voice Power user. This feature also helps users who work out of two offices, because all phone messages can be directed to a user's current office.

The forward feature can be enabled or disabled remotely from a terminal or another UNIX PC.

Stored greeting. With multiple greeting message storage, the user can keep several greeting messages on file and activate only the appropriate one.

For example, most of the time, a standard greeting message can be used when the user is unavailable. But when the user is on vacation, a vacation greeting message could replace it, and another greeting can be used when the user is away for an extended period.

The user does not need to re-record these messages at each change. On returning to the office, the user just reactivates the standard greeting, and no recording is required.

Remote Access. The Voice Power answering machine supports remote access from any touch-tone phone.

For a remote access session, the user calls the answering machine and depresses any key on the touch-tone dial to interrupt the greeting message. Then, the answering machine prompts the user to enter a password. If

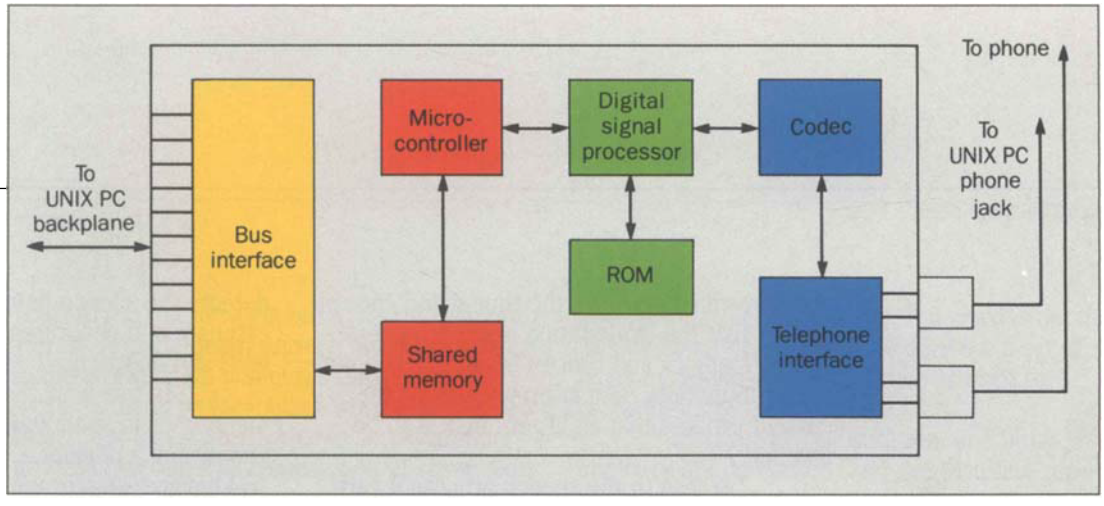

Figure 2. The four functional blocks of the Voice Power speech processor card: telephone interface, signal processor, microcontroller and shared memory, and workstation/card interface.

the password is accepted, the session begins, and the user is told how many phone messages and voice mail messages have been received.

At this point, the user is at the *service selection level* of a two-level user interface. Four services are available: review phone messages, review voice mail messages, record a new greeting, and record a phone message.

After selecting either the phone or voice-mail message review service, the user enters a *message review level*. At this level, the user can listen to or delete messages, and each message is date and time stamped with text-to-speech synthesis.

The touch-tone user interface supports the AT&T Product Family Touch Tone Core Commands.[1] During the remote access session, a help feature is available that can summarize these commands or provide a more detailed explanation of a specific one. Thus, the user can ask how to use the answering machine and ask specific questions rather than general ones.

Hardware

The Voice Power hardware is an advanced speech processing card. It can be considered in four functional blocks (Figure 2): telephone interface, signal processor, microcontroller-shared memory, and workstation/card interface.

Telephone Interface. The telephone interface circuit provides a Model 2500 telephone set, tip/ring interface with ring and touch-tone detection, telephone-set hook status, auxiliary audio in and audio out, relay control of card on/off hook, and battery feed. This circuit allows the speech processor card to answer calls and permits using the telephone as an input and output device. The telephone interface also provides the audio connection to the digital signal processor.

Signal Processor. AT&T's WE® DSP20 digital signal processor does the digital speech processing and supports record and playback, using 16-kb/s subband coding or, optionally, 64-kb/s pulse code modulation coding for toll quality requirements. It also supports speech recognition and text-to-speech synthesis. (We will discuss the speech processing technology and subband coding later.)

Microcontroller and Shared Memory. The speech processor card uses a microcontroller to control and respond to the DSP20 and telephone interface, and interface with the workstation processor. The microcontroller handles the exchange of data between the DSP20 and shared memory, and arbitrates with the workstation for use and control of the shared memory.

Workstation/Card Interface. This functional block, which interfaces with the workstation's expansion card bus, is the only workstation-specific portion of the speech pro-

cessor card. It provides the timing and control signals that the workstation needs to access shared memory and control registers, and generates interrupts. (An interrupt signals the current process that an urgent task is to be handled.)

Access to the speech processor card occurs through a 16-kbyte, memory-mapped input/output space and five status-and-control register bits.

Current Speech Processing Capabilities

The Voice Power speech processor currently provides speech store and playback. It uses the technique of subband coding[2] to store speech in compressed form thus reducing the storage requirements by a factor of four, to 16 kb of storage per second of speech. For playback, the compressed representation is decoded to reconstruct the speech signal.

The subband coding used in the Voice Power speech processor will comply with the AT&T Cross Product Standard.[1] Therefore, this speech processor can handle compressed voice files from other AT&T speech store and playback products, such as the Audix voice mail system.

In subband coding, a filter bank analyzer divides the input speech signal into a set of five subband signals. After an adaptive pulse code modulation (APCM) encoder encodes each band, the bits are multiplexed into a single data stream.

The subband coder employs a silence detector scheme to further improve storage efficiency. When the transmitter (encoder) adaptively detects a lull in the speech, it embeds a signal flag, silence duration, and silence noise-level measurement in its output bit stream. When the receiver (decoder) detects the silence flag, it reads the noise level estimate and generates noise at the encoded level and duration.

In this way, the silence intervals—normal pauses in spontaneous speech—are encoded at a negligible bit rate rather than the 16-kb/s speech storage rate.

The DSP20 on the Voice Power speech processor card executes the subband coder and silence encoding algorithm. Then, the compressed speech is transferred to disk storage in the workstation for later retrieval or transmission via UNIX system mail. The encoder and decoder each require 1 kbyte of DSP program memory.

Future Speech Processing Capabilities

The Voice Power speech processor can support speech recognition and text-to-speech synthesis. These capabilities are under development.

Speech Recognition. The Voice Power speech processor's speech recognition capability permits the automatic identification of an unknown word or phrase from a vocabulary of 50 words or phrases. First, the user or an automatic application compiler must select the words; then, the user trains the recognizer.

The speech processor uses linear predictive coding (LPC) to compute a spectral pattern that represents the unknown word. Then, through the time-alignment process of dynamic time warping,[3] it compares this pattern to LPC-based patterns of words in the vocabulary.

At the start of the recognition process, the DSP20 on the speech processor card computes the LPC spectral representation.[4] During recognition, the card continuously outputs LPC feature vectors that are then

transferred to the workstation and blocked into words. A high-speed, pattern matching process in the workstation compares these unknown word patterns to vocabulary word patterns. To match the test pattern, the pattern matching process stretches or compresses the time axis of each reference pattern to compensate for nonlinear differences in speaking rates.

Although the Voice Power speech processor currently uses speaker-dependent, isolated word recognition, it also has basic processing capability for either speaker-independent or connected-word recognition of a smaller vocabulary.

The isolated word recognizer provides buffering for talk-ahead capability to allow the rapid input of words with minimal pauses. In addition, it can provide syntax-directed recognition. Here, a subset of the recognition vocabulary can represent the only grammatically correct choices at each point in a speech transaction. The words in this subset change as the transaction proceeds, providing a substantial increase in overall vocabulary size.

Text-to-Speech Synthesis. The text-to-speech synthesizer uses a set of rules and table look-ups to convert printed English text into a spoken representation. Thus, it provides an unlimited synthesis vocabulary.[5,6]

The conversion of letters into sounds begins with a preprocessor that expands abbreviations into spelled words and assigns parts of speech to all words. Next, words and punctuation marks are examined to determine phrase groups so that stress and tone patterns can be assigned later. Then, the resulting words are looked up in a dictionary to determine their pronunciation.

The dictionary consists of base word forms, and derivation procedures are used to add endings such as *s* or *ing*. If a word is not in this dictionary and cannot be derived, a set of letter-to-sound conversion rules is used to compute its pronunciation.

Finally, timing, duration, and pitch contours are computed, and the resulting representation of speech sounds is converted into a speech waveform using multipulse linear predictive coding.[7]

In the Voice Power speech processor, the workstation converts text into the pronunciation representation, and the DSP20 on the speech processor card converts the pronunciation into the digital speech waveform.

Future Applications

Two broad categories of future applications for the Voice Power speech processor include personal services (like the answering machine) and voice servers.

Several personal services could build on the base of applications that were provided initially. For example, with text-to-speech synthesis, remote access to the inbox could be extended to include text mail. A voice-annotated text editor could further integrate voice and data. For voice mail delivery to telephones, the UNIX PC could automatically call and deliver messages to a distribution list of telephone numbers.

Today, touch-tone commands control remote access but, in the future, speech recognition could be used and allow access from rotary dial telephones. Talker verification can add an additional degree of access security.

Voice servers are an exciting, new applications area for the Voice Power speech processor. A workstation with multiple speech processing cards (up to seven for the UNIX PC) provides a low-cost voice server platform

for applications such as remote access to voice databases, telephone message centers, and voice-based, order entry systems.

Although the Voice Power option is initially being offered for the UNIX PC, it is a generic hardware-software speech processing system. As such, its capabilities can be extended easily to other AT&T workstations (e.g., PC 6300 and PC 6300 PLUS personal computers).

Many applications are possible but close communications with early customers and the marketplace will identify the ones that have the greatest merit. What will be provided next is a Voice Power development package for value added resellers—the tools that they need to implement these and other voice applications. The initial Voice Power offering only hints at the true utility of voice processing.

Acknowledgments

As with any complex projects, many people contributed to the success of the workstation speech processing project. The authors would like to acknowledge the valuable contributions of their colleagues.

W. E. Wetzel was responsible for the innovative design of the voice editor and the UNIX system software, including the speech processor driver. J. P. Hickey developed the speech processing card's firmware and integrated the recognition and text-to-speech subsystems. B. A. Adams added the voice capability to the UNIX PC's electronic mail package.

C. A. Russell provided the electromagnetic interference testing and overall hardware support. E. F. Stefanacci successfully introduced the speech card into the factory, and C. A. Franz provided software support and archiving.

For the text-to-speech synthesizer, H. F. Carbonneau ported the synthesizer software from a minicomputer to the UNIX PC and PC 6300 speech processors. A. Zegarek provided excellent software engineering to convert the text-to-speech synthesizer into an efficient, reliable form, and R. Blessing and L. Granrud provided substantial speed improvements to the speech recognition.

L. F. Rosa coordinated a models program for early versions of the speech processor and hand-built many early prototypes.

References

1. J. G. Josenhans, et al., "Speech Processing Application Standards," *AT&T Technical Journal*, Vol. 65, No. 5, September/October 1986, pp. 23-33.
2. R. E. Crochiere, R. V. Cox, and J. D. Johnson, "Real Time Speech Coding," *IEEE Transactions on Communications*, Vol. COM-30, No. 4, April 1982, pp. 621-634.
3. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-23, February 1975, pp. 67-72.
4. J. G. Ackenhusen and Y. H. Oh, "Single-Chip Implementation of Feature Measurement for LPC-Based Speech Recognition," *AT&T Technical Journal*, Vol. 64, No. 8, October 1985, pp. 1787-1805.
5. J. P. Olive and M. Y. Liberman, "Text to Speech—An Overview," *Journal of the Acoustical Society of America*, Supplement 1, Vol. 78, Fall 1985, pg. S6.
6. Cecil H. Coker, "A Dictionary-Intensive Letter-to-Sound Program," *Journal of the Acoustical Society of America*, Supplement 1, Vol. 78, Fall 1985, pg. S7.
7. B. S. Atal and J. R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Paris, France, 1982, pp. 614-617.

Biographies (continued)

ware and hardware development for real-time speech processing. He joined AT&T in 1979 and has a B.S. and an M.S. in physics from the University of Punjab, Pakistan. Mr. Josenhans is responsible for fitting speech and signal processing algorithms into digital signal processing hardware. He joined AT&T in 1963 and has a B.Sc. in physics from the University of Toledo (Ohio) and a Ph.D. in electrical engineering from Ohio State University. Mr. Moffett is responsible for developing speech processing systems (hardware and software) for AT&T workstations, including the UNIX® PC, PC 6300, and PC 6300 PLUS personal computers. He joined AT&T in 1972 and has a B.S. from Georgia Institute of Technology and an M.S. from Massachusetts Institute of Technology, both in electrical engineering. Mr. Robertson designs and develops speech processing applications and value-added-reseller tools for AT&T workstations. He joined AT&T in 1981, and has a B.S.M.E. from the University of Massachusetts at Amherst and an M.E. in electrical engineering and computer science from Princeton University. Mr. Tormos works on the hardware design and development of AT&T's Voice Power speech processor and an interactive display terminal. He joined AT&T in 1982 and has a B.S. in computer science and an M.E. in electrical engineering from Old Dominion University in Norfolk, Virginia.

(Manuscript received August 22, 1986)