HUMANET: AN EXPERIMENTAL HUMAN-MACHINE COMMUNICATIONS NETWORK BASED ON ISDN WIDEBAND AUDIO

David A. Berkley and James L. Flanagan

David A. Berkley *is head of the Acoustics Research Department at AT&T Bell Laboratories in Murray Hill, New Jersey.*
James L. Flanagan *was director of the Information Principles Research Laboratory at AT&T Bell Laboratories in Murray Hill until his retirement in August 1990. Mr. Berkley is responsible for research on electroacoustics and acoustic signal processing. He joined the company in 1968 and has both a B.E.E. and a Ph.D. in applied physics from Cornell University (Ithaca, New York). At AT&T Bell Laboratories, Mr. Flanagan has been responsible for research in digital speech processing, acoustics, linguistics, software engineering, image coding, and human-machine communications. He joined the company in 1957 and has a B.S. from Mississippi State University (Starkville) and both an M.S. and*
(continued on page 97)

A human's sensory capacity to assimilate, perceive, and react to information is much smaller than the capacity of modern transport facilities that convey the information. Moreover, the volume of information and the complexity of the terminal used to access the information can overwhelm people. The challenge, then, is to match an information system's capabilities to those of our senses. Recent advances in speech-processing technology have made natural voice—our preferred means for information exchange—feasible for human-machine communications. We describe an experimental network, called *HuMaNet*, that is implemented on commercial ISDN transport. The HuMaNet system uses speech-processing technology to make communications easier and more natural. Spoken commands control the system, which combines image and audio compression, database management, hands-free teleconferencing, and text-to-speech synthesis. Although HuMaNet is only in its initial phase, it has proved a remarkably habitable environment for human control of a complex computer and communications system.

Human-Machine Communication

Consider the telecommunications systems of the future. Over the next decade, switched digital connectivity will spread throughout the network and permit people to access a wealth of information services from terminals in business, home, and school environments.

The transport capacities of the telecommunications network will range from gigabits per second (Gb/s)—for optical fiber—to tens of kilobits per second (kb/s)—for cellular radio and basic-rate ISDN (Integrated Services Digital Network).[1]

Panel 1. Acronyms and Terms

2B+D	ISDN lines: two 64-kb/s circuit-switched channels (2B) and one 16-kb/s packet-switched channel (D)
A/D	analog to digital converter
B channel	64-kb/s circuit-switched channel
CCITT	International Telegraph and Telephone Consultative Committee
CELP	code-excited linear-predictive coding
codec	coder-decoder
D channel	16-kb/s packet-switched channel for carrying data and signaling
DSP	digital signal processor
HMM	hidden Markov model
HuMaNet	human-machine network; an experimental system for studying network-based integration of voice, image, and data transmission and hands-free teleconferencing
ISDN	Integrated Services Digital Network
PC	personal computer
pixel	picture element; one of thousands of dots that form a video image
Q.931	CCITT protocol that specifies the procedures for establishing, maintaining, and clearing connections at the ISDN user-network interface
RPC	remote procedure call
teraFLOP	10^{12} floating-point arithmetic operations

Today's terminals are either telephones, which exclusively allow spoken communications to a distant party, or personal computers (PCs) and workstations, which generally capture words entered on a keyboard or display words on a screen. But in the future, the terminals that people will use will be multifeatured, complex processors for information display and capture, primarily for sight and sound. The new terminals will hear our spoken requests and, in turn, will speak to us, display images, and control other complex telecommunications functions simultaneously.

A Potential Problem. At least two important issues—our sensory capabilities and the complexity of our terminals—attend these advances. Modern transport facilities can convey information[2] much faster than our senses permit us to assimilate, perceive, and react to it. Thus, we may easily be overwhelmed by the volume of information unless intelligent management and display are possible.

Also, as the volume and sophistication of this information grows, the complexity of a user's terminal increases. This includes how to interface with the information source via a terminal, or deal with what is inside or behind the "black box." The complexity can pose a burden for the user who may have no desire or need to become a technical specialist.

A Solution. Ease of use, along with low cost and utility, will decide the mass acceptance and deployment of sophisticated systems. A major challenge, then, is to match the features of an information system to the sensory capabilities of its human users.[3] The terminal must be made intelligent enough to aid a human user by sensing his or her needs of the moment and by displaying information in modalities that are complementary and optimal for human assimilation. (See Figure 1.) An example of such modalities is to display an image with simultaneous voice annotation.

Natural voice is a preferred means for information exchange between people. This normally occurs face to face and hands-free, without either person holding a microphone or telephone handset. Sound (the speaker's voice) is complemented by sight and our other senses (touch, taste, and smell). We associate spatial realism with three-dimensional perception in sight and sound. We can see the other person, and can easily sense position from the direction of his or her voice. Comparable capabilities in machines would greatly simplify and enhance communications between humans and machines.

Recent advances in speech-processing technology, along with new capabilities in hands-free teleconferencing, image compression, database management, switched digital networking, and hypertext systems, offer new dimensions for ease and naturalness in human use of information systems.

All these capabilities are underpinned by advances in microelectronics, which provide great amounts of computational power at low cost.

A relevant issue, then, is how these separate "piece parts" of technology might be harnessed to benefit people, and what unforeseen synergies might derive from

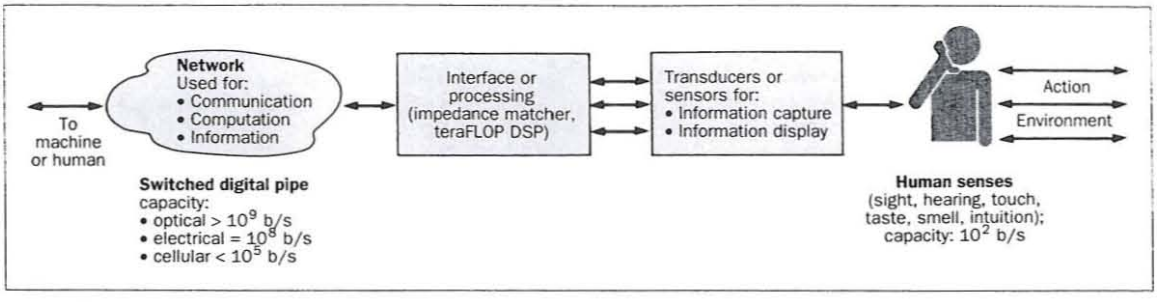

Figure 1. A major issue in human/machine interfaces is matching the features of an information system to human sensory capabilities. Modern transport capacities typically deliver information at rates that exceed any human's ability to assimilate, perceive, and react to information. An interface processor, combined with proper display systems, can, in principle, be used to match raw information coming from remote computers, or other human beings, to the appropriate sensory input of the user. These inputs are, primarily, sight and sound, but others are possible channels as well.

their application in concert.

HuMaNet (for human-machine network) is an experimental system developed at AT&T Bell Laboratories (a division of AT&T) and designed for studying network-based integration of the technologies for voice, image, and data transmission, and hands-free teleconferencing. The transport medium currently used is public-switched, basic-rate ISDN that provides 2B+D, or two 64-kb/s circuit-switched channels (the *2B*) and one 16-kb/s packet-switched channel (the *D*). In this application, the B channels, which are also called *bearer* channels, carry the user's or customer's information (i.e., voice, data, and video signals). The D, or *data*, channel primarily carries the network's control and signaling information but can be used to send some user packet communications as well.

Component Technologies of HuMaNet

The HuMaNet system incorporates a variety of voice, image, and data technologies that are integrated and implemented for real-time operation. These technologies include: speech recognition, speech synthesis, talker verification, speech coding, speech-seeking microphones, image compression, and hypertext and database features.

Speech Recognition. Voice commands are used to control the HuMaNet system and identify information to be retrieved.

For speaker-trained recognition of fluent sentences, the system utilizes hidden Markov model (HMM) technology to provide high performance.[4–9] The system's vocabulary consists of 80 words at present. From these words, over 6 million valid (or useful) phrases and sentences can be spoken and recognized.

Speech Synthesis. The system need not just display words on a screen for user information. Instead, a complete text-to-speech synthesis system also permits spoken output for unrestricted text generated by the HuMaNet system in response to user- and system-control actions.[10–12]

Talker Verification. A phrase-dependent talker-verification system controls access to privileged databases. Before a user can access any of these databases, his or her voice signal must be authenticated.[13,14]

Speech Coding. Perceptually based coding (using auditory criteria) of high-quality speech permits trans-

mission of 7-kHz (kilohertz) bandwidth voice in full stereo over a single ISDN B channel.[15-18]

Autodirective Microphone Arrays. To communicate with the system (or converse with each other), an individual or conference group uses normal voice. Speech-seeking, automatic, beam-steering arrays of microphones make hands-free voice interaction possible. In addition, fixed, multiple-beam arrays provide spatially realistic sound pickup.[19]

Image Compression. The coding techniques determine not just the quality of a displayed image but also the storage resources required and the time needed to transmit the image over a specific transmission system.

In the HuMaNet system, perceptually based subband coding (using visual criteria) of still images provides graphics and picture transmission over one B channel. Typically, the input color images are of television quality and consist of 512 by 512 pixels, quantized using 24 bits to code each pixel. These images are faithfully coded and transmitted at less than 1 bit per pixel, for projection at the receiver.[20-22]

Database Organization. The HuMaNet system has two databases, one local to the system and the other accessed over ISDN. (In the system's current structure, these are the only databases we can access. However, the system is capable of accessing additional databases, if available.) Currently, the databases contain information on the HuMaNet system itself, as well as some examples of more general scenarios, such as the real-estate application described below.

Both databases contain image and text data. The remote database consists of images coded and transmitted as described above. The local database stores image information in a format that is appropriate for the display hardware. The system reads text information "orally" from either database (i.e., text is not displayed on the screen), using text-to-speech synthesis.

Hypertext Features. A person who wants to retrieve material from a database may not know exactly where the information resides in the database nor the storage medium used. Therefore, both databases are organized using hypermedia indexes that are stored separate from the databases.

The indexes consist of the commands that identify the storage medium and location of information in the databases. This allows the index information to be completely isolated from the nature of the material in the database. A user's request for an index entry simply retrieves the appropriate command (or commands), which the system then executes, regardless of the type of storage media or location.

The index entries are really nodes in an arbitrarily organized and cross-linked tree. The nodes are named, as are the links that connect the various nodes, and a user can access all nodes (or links) using these names.

Graphical editors are available to aid in the original construction and later modification of the index structure and contents.[23] (Such editing is not done as part of active use of the system.)

Integration. As stated earlier, the HuMaNet system is an experimental vehicle. All the technologies are currently integrated into a voice-interactive, multimedia terminal or workstation, and implemented in two acoustically adjustable conference rooms (Figure 2) in our laboratory at AT&T Bell Laboratories in Murray Hill, New Jersey. One room (the HuMaNet terminal room) is arranged in a living-room decor, and the other (the "remote" room) is arranged as a group conference room. (This issue's cover illustration is a photograph of the HuMaNet terminal room.)

Although the experimental rooms are next to each other in the laboratory building, they communicate only by commercial, public-switched ISDN. They could be arbitrarily remote from one another (assuming ISDN access exists for all sites involved).

The system architecture is designed to be extensible, particularly in terms of higher capacity networking (for example, primary-rate ISDN) and the high-quality motion video that such capacity makes possible. (The primary-rate arrangement for ISDN offers more B-channel

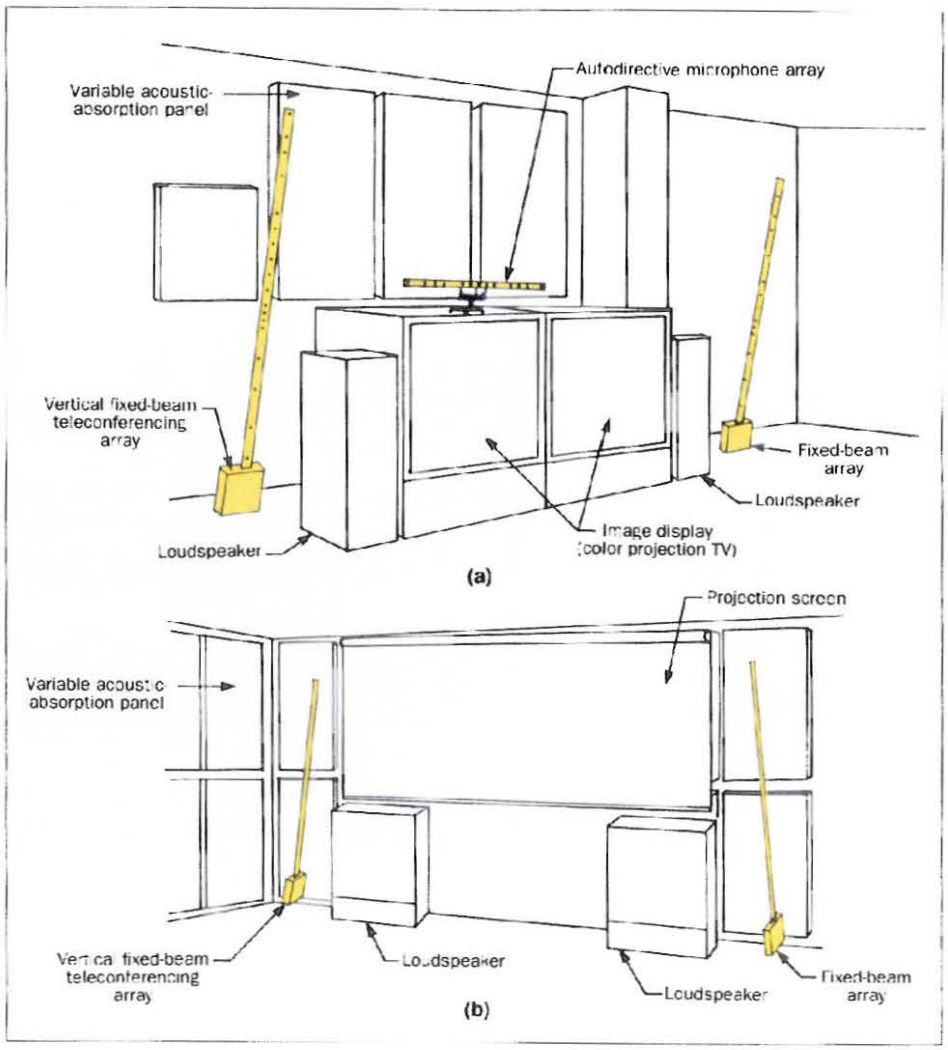

Figure 2. Components in the HuMaNet experimental rooms: (a) HuMaNet terminal or "local" room, and (b) remote conference room. Images can be displayed simultaneously in both rooms. As an image is retrieved and displayed, the system's "voice" reads the related text description. Microphone arrays and stereo sound systems provide auditory spatial realism for conference participants. The speech-seeking microphone (atop the TV monitor in the terminal room) senses the speaker's voice and tracks his or her position, so a speaker need not remain stationary. Because of fixed microphones (for transmission) and the stereo sound system (for reception) in both rooms, listeners correctly perceive the speaker's position. Adjustable acoustic panels on the walls permit experiments on changing room conditions.

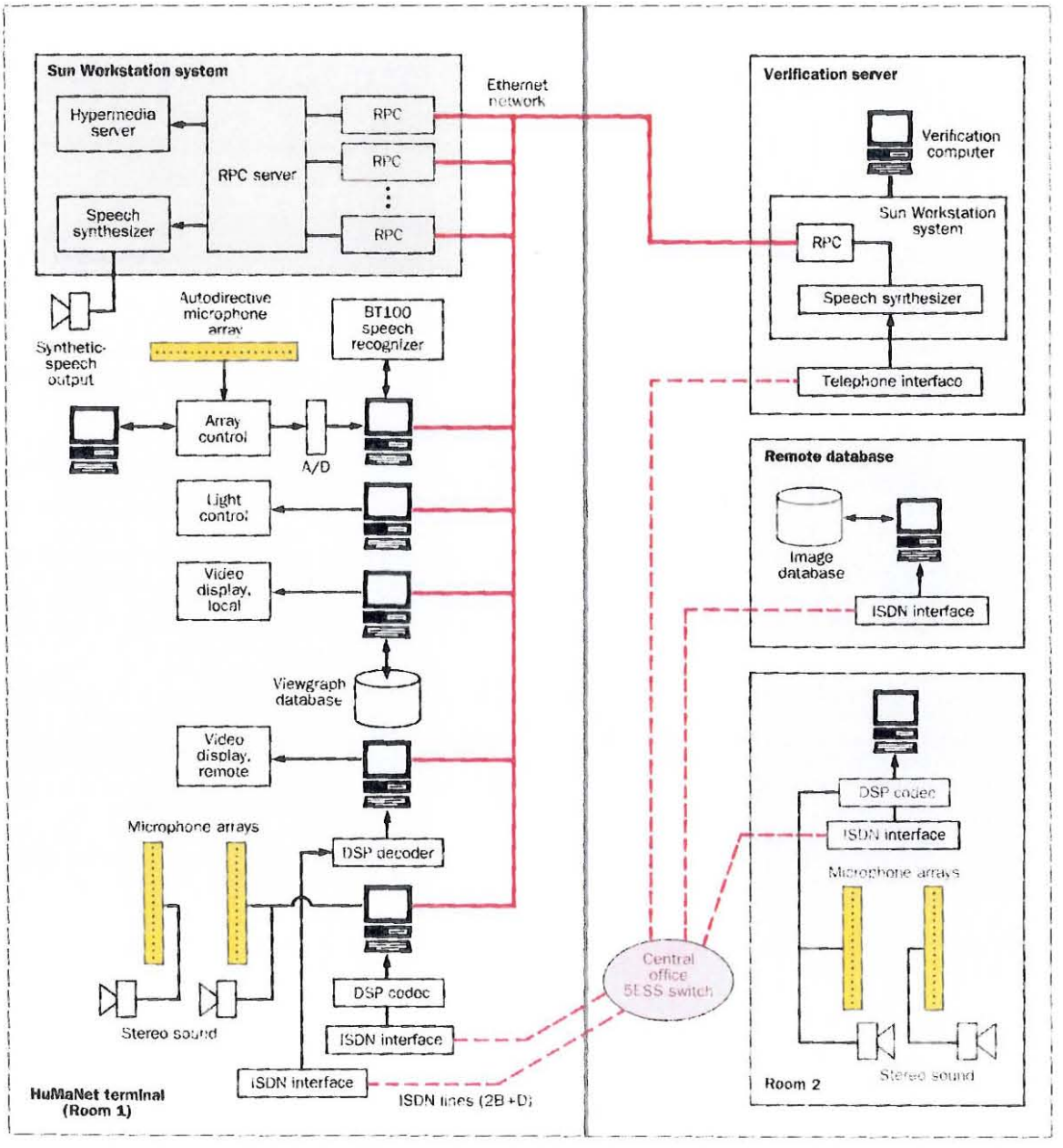
Sun Workstation system
Hypermedia server
RPC server
RPC
RPC
RPC
Speech synthesizer
Ethernet network
Synthetic-speech output
Autodirective microphone array
BT100 speech recognizer
Array control
A/D
Light control
Video display, local
Viewgraph database
Video display, remote
Microphone arrays
DSP decoder
DSP codec
ISDN interface
Stereo sound
ISDN interface
ISDN lines (2B+D)
HuMaNet terminal (Room 1)
Verification server
Verification computer
Sun Workstation system
RPC
Speech synthesizer
Telephone interface
Remote database
Image database
ISDN interface
Central office 5ESS switch
DSP codec
ISDN interface
Microphone arrays
Stereo sound
Room 2

Figure 3. System architecture for HuMaNet. The operator's voice commands control system operation, room environment (such as dimming the lights), teleconferencing, and database access. Individual PCs control various system and environment functions and communicate with the workstation (top left) over a local Ethernet network. Commercial ISDN lines provide the communication paths between the HuMaNet terminal room and any remote database or conference room. When access to a secured database item is requested, the voice password is verified before the ISDN connection to the remote database is established. Text and the compressed images retrieved from this database are sent over ISDN lines to the HuMaNet terminal where they are decoded. As the images are displayed, the corresponding text is read (orally) by the text-to-speech synthesizer. The synthesizer is also used to provide feedback to users about system status and operation.

capacity; typically, 23B+D in North America and Japan, and 30B+D in Europe.)

HuMaNet Architecture

The system (Figure 3) is supported by:

- A Sun Workstation® system and file server that provides overall communication and text-to-speech synthesis for the system. (Sun Workstation is a registered trademark of Sun Microsystems, Inc.) However, any UNIX® system-based workstation could fill this role. (UNIX is a registered trademark of UNIX System Laboratories, Inc.)
- Seven 386-class PCs (i.e., PCs that use the Intel 80386 or 80386SX microprocessor). Individual PCs control various elements of the system, such as the local or remote room's environment, access to the local or remote database, and ISDN access. AT&T DSP32C signal processors, adjuncts to the individual PCs,[24] code or decode the speech and image signals for the PC.
- An AT&T BT100 signal processor[25] that recognizes oral commands to the system and discriminates against ordinary conversation.
- Autodirective microphone system. This microphone array "follows" the speaker's voice to provide audio input to the speech-recognition system.
- Stereo sound pickup microphone arrays. These arrays provide audio spatial realism and reduce room reverberation and noise for voice transmission.
- Color projection television, which provides large screens for simultaneous display of information from the local and remote databases. The two rooms use different projection units, appropriate to the environment. (For example, the living-room setting uses large-screen, rear-projection TV monitors. In the conference-room setting, ceiling-mounted units project the image onto flat projection screens at the front of the room.)
- Local Ethernet-network communication between the workstation and the individual PCs in the (local) terminal room.
- ISDN transport using a commercial AT&T 5ESS® switch in a New Jersey Bell Telephone Company central office. These facilities provide remote voice communications and the data link for the remote database.

Users can initiate and carry on conference calls with simultaneous display of images from a remote database (which can be, and is, located completely separate from the conference locations).

Basic Terminal Operation. The easiest way to explain the operation of the HuMaNet terminal is to trace what happens when the person who is controlling the system issues a few specific commands. [Currently, the system can recognize only one controller or operator (recall the speaker-dependent recognition).]

One simple command is "Facility control, lights down." When the person utters this (or any command), the speech-seeking array microphone (the horizontal bar on top of the left-hand TV cabinet in Figure 2a and in the photograph on this issue's cover) locks onto him or her and sends the spoken command to the speech-recognition system. While giving commands, the talker can move freely around the room and even pace back and forth.

The speech recognizer then detects a valid

HuMaNet control phrase. There are three groups of such phrases: "facility control," "database control," and "ISDN control." Each phrase is followed by an appropriate command that controls the local room's environment or database, the remote database, or ISDN connectivity, respectively. By using such a strict format, the speech recognizer is able to operate effectively even when there is normal conversation among conference participants.

Once the command is recognized, the information is forwarded to the Sun Workstation system over the Ethernet network, using a *remote procedure call* (RPC).[26] [An RPC is the protocol used when one device (e.g., the speech recognizer) issues a request to another device (e.g., a workstation) to execute a process. The Sun Workstation system uses an RPC when it instructs one of the PCs to perform a task.] The workstation is responsible for decoding the recognition syntax into valid operation commands. For the local-room environment request, "... lights down," the workstation then forwards the resulting lighting-control command to the PC that contains the dimmer-control hardware, and the room lights lower appropriately.

The architecture is designed so that each PC functions as an individual *technology platform*. Each contains the technology appropriate to its function, allowing easy addition of new hardware, stand-alone testing, and technology transfer. The RPCs "weld" the individual platforms into an integrated multiprocessor system.

Multiprocessor Structure. A more complex command illustrates this point further. For example, the command "database control, image 5 please" provides access to specific remote-database information. (The basic commands have several alternative forms, which are equivalent. "Control" and "controller" are equally acceptable and interchangeable, and the commands may be arbitrarily terminated with a polite "please," if desired.) The grammar is formally specified in a way that makes it easy to construct new recognition scenarios.[27] Appendix A is a formal specification of the grammar.

The Sun Workstation system interprets this command as a request for data contained in the remote database. (The following sequence of events can be traced in Figure 3.) Then, the workstation accesses the appropriate hypermedia index item (here, the node named "5"), which returns a command of the form:

```
#TEXTFILE hfrog.txt
#IMAGEFILE hfrog.y
```

to identify text and a compressed image to be retrieved. These two strings are forwarded to the ISDN-control PC, which sends them—as an ISDN, standard, D-channel packet message—to the remote-database PC. The database PC sends the requested text file, immediately followed by the perceptually compressed image file, back to the ISDN-control PC via a single, 64-kb/s, ISDN B channel. A typical compressed image takes several seconds to transmit, as compared to the 90-seconds transmission time for an uncompressed, 6-Mb (megabit) image.

After the terminal system receives the information, it sends the text information (usually descriptive material related to the image) to the workstation for voice output to the user by text-to-speech synthesis. The image is decoded on a DSP32C board in the ISDN-control PC and presented on the right-hand TV screen shown in Figure 2a (and in this issue's cover photograph). Decoding time for the image is comparable to the transmission time and, with the overlap of the text-to-speech material, images appear in a perceptually prompt manner.

ISDN Control. The command "ISDN control, call room two" initiates a procedure that allows a B-channel connection to be established to the called conference room (shown in Figure 2b and on the lower right in Figure 3). To prevent errors in dialing, the system requests confirmation by asking: CONFIRM CALL TO ROOM TWO? Confirmation is desirable because the system can make calls by name (several are included in the recognition syntax), as well as by number. An international dialing error is an undesired expense.

After the HuMaNet system's operator confirms the call (with "okay" or "yes"), the system dials the con-

nection. It announces call progress (using the ISDN Q.931-messaging capability of the D-channel interface), and completes real-time connection between the rooms through DSP32C boards and high-quality codecs on each end.

The resulting completely digital connection to the second conference room uses the vertical fixed-array microphones, to the left and right of the projected images in Figures 2a and 2b. Because these highly directional arrays decouple the paths between the microphone and adjacent loudspeakers, a conference may be carried on without requiring voice switching for stabilization. In this way, we can achieve a two-way, full-duplex, stereo conference with 7-kHz bandwidth over a single B channel, using low-bit-rate coding.

Voice Security. A final feature of the system is voice security. The security system is implemented as a separate security server on the Ethernet network, although the HuMaNet system's basic architecture is followed. Figure 3 shows the connection for verifying voice access to a database item. (The same architecture can be used to provide voice-password access to standard computer systems as well.)

For the HuMaNet terminal, individual database items are designated as secured for designated users, as part of the index entry. If someone attempts to access one of these items (regardless of the command given), the system responds, for example, with: ACCESS TO THIS DATA IS RESTRICTED TO BERKLEY. DO YOU WISH TO PROCEED WITH VOICE PASSWORD VERIFICATION? Once access is confirmed (the user answered "yes" or "okay"), the HuMaNet system handles the entire procedure. It makes the appropriate identity claim to the password system and answers a return ISDN (or standard) call from the voice-password system.

At this point, the voice-password system takes over. It leads the user quickly through the procedure for entering his or her voice-password phrase (using the same array microphone as used for teleconferencing) and determines the password's validity. If the password is accepted, the resulting message allows the HuMaNet terminal to complete the requested database access.

The procedure takes less than 30 seconds and can be set up to allow access to any designated subsets of text and image data.

Notice that the entire procedure follows the overall design philosophy. It hides the technical complexity of the operations from the user as much as possible. Thus, teleconferencing, access to remote databases, and security are made as natural as speaking.

Participants are primarily aware of the smooth progression of images, teleconferencing, and other control operations, without the burden of the complex technological systems that lie behind the HuMaNet terminal.

A Customer Scenario

We designed HuMaNet's architecture to provide a basis for designing application scenarios, as well as for research experiments on the integration of information modalities. The previous discussion did not present a specific scenario. Clearly, the capability for remote-data access, voice security, and teleconferencing fits several possible applications, including real estate, travel services, news and information, and entertainment.

A Real-Estate Application. Real-estate sales is one obvious place where the availability of high-quality images from a remote database could be commercially valuable. The database could have proprietary or privileged aspects, as well as the more usual customer information. Even in its current, preliminary form, this experimental application for a real-estate business is providing new insights about the HuMaNet system's integrated, teleconferencing modalities.

For experimental use, we have stored high-quality images of homes in the HuMaNet system's compressed-image database. (One example appears on the right-hand screen in the photograph on this issue's cover.) In an actual application, such images might be stored in a database anywhere in the country, and the connection to the remote database could be established automatically. The request for the image to be processed

and local decoding of the transmitted image can be accomplished within a few seconds. Our experimental database also stores other images related to the full view of the house, including a full floor plan and high-quality interior shots of various rooms.

Thus, if ISDN connectivity existed, a potential client who currently lives in California could explore real-estate offerings in the New York area, before making the transcontinental trip in person. The simultaneous high-quality teleconferencing capability would allow the client to have detailed discussion with distant agents who were knowledgeable about the area being explored. Multiple areas might be discussed simultaneously, if desired.

Future Directions

Several synergies are apparent from initial experiments with the HuMaNet system. The speech-tracking microphone system, which was designed originally for large-group audioconferencing, permits completely hands-free voice control and verification procedures, thus providing a natural-voice interactive environment for the system.

It has further proved possible to make the interaction "seamless." The substantial technologies of real-time speech recognition and text-to-voice synthesis (as well as image and data compression) are invisible to the system's users. Ease of use is paramount, and human-machine interaction by voice is comfortable and natural.

The receiving room's stereo sound provides spatial realism; it localizes the position and perceptually separates multiple conferees in the transmitting room.[28] The audio bandwidth is broader than that of conventional analog telephone, a differentiating high-quality advantage provided by the ISDN connectivity.

Spatial realism (even three-dimensional projection) in an image is desirable, but the technology does not yet support it. As higher speed switched networking and more sophisticated coding methods become available, the incentive to work toward full-motion color video—and even color holography—will increase as well.

In summary, the experimental HuMaNet system demonstrates how the technologies of voice, image, data, computing, and teleconferencing can be combined in a sophisticated information system to achieve operation that is natural and easy for people to use.

Acknowledgments

The HuMaNet project is the result of technical contributions from every department of our laboratory. About 25 people were intimately involved in its creation; without them, there would have been nothing for us to write about. We also want to acknowledge the unflagging cooperation and support of all managers in our laboratory in this exceptional integration effort.

In addition, we particularly want to recognize the contributions of K. L. Shipley, who was responsible for most of the HuMaNet system's underlying control software, and A. H. Koenig, who was responsible for project management and ISDN interfaces for HuMaNet.

References

1. R. T. Roca, "ISDN Architecture," *AT&T Technical Journal*, Vol. 65, No. 1, January/February 1986, pp. 4-17.
2. W. D. Keidel, "Information Processing by Sensory Modalities in Man," *Cybernetic Problems in Bionics*, H. L. Oestreicher and D. R. Moore (eds.), Gordon and Breach, New York, 1968, pp. 277-300.
3. J. L. Flanagan, "New benefits from information and communication technologies," *Royal Swedish Academy of Engineering Science, Stockholm, Sweden*, May 1985, L. M. Ericsson prize lecture. Also *Ericsson Review*, Vol. 62, No. 3, 1985, pp. 108-113.
4. L. R. Rabiner, "A tutorial on hidden Markov models and its application to speech recognition," *IEEE Proceedings*, Vol. 77, No. 2, February 1989, pp. 257-286.
5. J. G. Wilpon, L. R. Rabiner, C-H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, to be published, November 1990.
6. L. R. Rabiner, B. S. Atal, and J. L. Flanagan, "Current methods of digital speech processing," *Selected Topics in Signal Processing*, S. Hyakin (ed.), Prentice Hall, Englewood Cliffs, New Jersey, pp. 112-132, 1989.
7. A. L. Gorin, D. B. Roe, and A. G. Greenberg, "On the complexity of pattern recognition algorithms on a tree structured parallel

computer," *Signal Processing, Part 1: Signal Processing Theory*, The IMA Volumes in Mathematics and Its Applications, L. Auslander, T. Railath, and S. Mitter (eds.), Springer-Verlag, New York, 1990, pp. 95-116.

8. A. L. Gorin and D. B. Roe, "Parallel level-building on a tree machine," *ICASSP '88, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, New York, April 11 to 14, 1988, Vol. I, IEEE, New York, 1988, pp. 295-298.
9. J. G. Wilpon, R. P. Mikkilineni, D. B. Roe, and S. Gokcen, "Speech Recognition: From the Laboratory to the Real World," *AT&T Technical Journal*, Vol. 69, No. 5, September/October 1990, pp. 14-24.
10. J. Olive and M. Y. Liberman, "Text to speech work at Bell Laboratories: An overview," *Journal of the Acoustical Society of America*, Vol. 78, Supplement 1, 1985, p. S6.
11. J. P. Olive, "Using digital signal processors for real-time synthesis of acoustic signals," *Journal of the Acoustical Society of America*, Vol. 71S, April 1982, p. S101.
12. J. Hirschberg, S. A. Riederer, J. E. Rowley, and A. K. Syrdal, "Voice Response Systems: Technologies and Applications," *AT&T Technical Journal*, Vol. 69, No. 5, September/October 1990, pp. 42-51.
13. F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-36, June 1988, pp. 871-879.
14. A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *ICASSP '86: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, April 7 to 11, 1986, Vol. II, IEEE, New York, 1986, pp. 873-876.
15. B. S. Atal, "A model of LPC excitation in terms of eigenvectors in the autocorrelation matrix of the impulse response of the LPC filter," *ICASSP '89, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, May 23 to 26, 1989, Vol. I, IEEE, New York, 1989, pp. 45-48.
16. B. S. Atal, R. V. Cox, and P. Kroon, "Spectral quantization and interpolation for CELP coders," *ICASSP '89, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, May 23 to 26, 1989, Vol. I, IEEE, New York, 1989, pp. 69-72.
17. J-H. Chen and N. S. Jayant, "Speech coding with time-varying bit allocations to excitation and LPC parameters," *ICASSP '89, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, May 23 to 26, 1989, Vol. I, IEEE, New York, 1989, pp. 65-68.
18. N. S. Jayant, V. B. Lawrence, and D. Prezas, "Coding of Speech and Wideband Audio," *AT&T Technical Journal*, Vol. 69, No. 5, September/October 1990, pp. 25-41.
19. J. L. Flanagan, J. D. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of the Acoustical Society of America*, Vol. 78, No. 5, November 1985, pp. 1508-1518.
20. C. I. Podilchuk, N. S. Jayant, and P. W. Noll, "Sparse-vector codebooks for the quantization of non-dominant sub-bands in image coding," *ICASSP '90, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, New Mexico, April 3 to 6, 1990, Vol. 2, IEEE, New York, 1990, pp. 2101-2104.
21. R. J. Safranek, K. MacKay, N. S. Jayant, and T. Kim, "Image coding based on selective quantization of reconstruction noise in the dominant sub-band," *ICASSP '88, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, New York, April 11 to 14, 1988, Vol. II, IEEE, New York, 1988, pp. 765-768.
22. V. Ramamoorthy and N. S. Jayant, "High quality image coding with a model-testing vector quantizer and a human visual system model," *ICASSP '88, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, New York, April 11 to 14, 1988, Vol. II, IEEE, New York, 1988, pp. 1164-1167.
23. J. Puttress and N. Guimaraes, "The Toolkit Approach to Hypermedia," submitted to *Proceedings of the European Conference on Hypertext*, Paris, France, November 27, 1990.
24. M. L. Fuccio, R. N. Gadenz, C. J. Garen, J. M. Huser, B. Ng, S. P. Pekarich, and K. D. Ulery, "The DSP32C: AT&T's Second-Generation Floating-Point Digital Signal Processor," *IEEE Micro*, December 1988, pp. 30-48.
25. AT&T, *AT&T DSP Parallel Processor, BT100 Product Brief*, AT&T Federal Systems, Greensboro, North Carolina, 1988.
26. *Network Programming*, Programming Manual, Part Number 800-1779-10, Revision A, Sun Microsystems, Inc., Mountain View, California, May 1988, pp. 57-91.
27. M. K. Brown and J. G. Wilpon, "Automatic Generation of Lexical and Grammatical Constraints for Speech Recognition," *ICASSP '90, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Albuquerque, New Mexico, April 3 to 6, 1990, Vol. 2, IEEE, New York, 1990, pp. 733-736.
28. D. A. Berkley, "Hearing in rooms," *Directional Hearing*, W. Yost and G. Gourevitch (eds.), Springer-Verlag, New York, 1987, pp. 249-260.

Biographies (continued)

a Ph.D. from Massachusetts Institute of Technology (Cambridge), all in electrical engineering. Mr. Flanagan is now director of the Center for Computer Aids for Industrial Productivity at Rutgers University in New Brunswick, New Jersey.

Appendix A. HuMaNet Language Specification

This compact specification of a finite-state grammar allows rapid definition of application interfaces and easy changes to existing systems. The meaning of the syntax is fairly self-explanatory. Grouping is provided by parentheses, while the '|' sign is an exclusive OR. Items in all-capital letters (e.g., NONZERO) are defined in terms of other items, as shown.

```
define( N, NULL-ARC )
define( NONZERO, ( 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 ) )
define( ZED, ( zero | oh ) )
define( DIGIT, '( NONZERO | ZED )')
define( O_SILENCE, '( N | silence )')
define( O_COMMA, '( N | silence )')
define( CONFIRMATION, '( yes | no | OK )')
define( CONTROL, '( control | controller )')
define( O_NOISE, '( N | noise )')
define( POLITE, '( N | please )')
define( NET_PREFACE, '( ( teleconference | network | ISDN ) CONTROL O_COMMA )')
define( LOCAL_PREFACE, '( ( facility |room) CONTROL O_COMMA )')
define( REMOTE_PREFACE, '( ( database | computer ) CONTROL O_COMMA )')
define( VOICE_COMMAND, '( ( speech | voice ) recognizer ( on | off ) )')
define( LOCATION, '(
  office | ( ( N | conference ) room ( 1 | 2 ) )
)')
define( EXTENSION, '(
  ( N | extension ) NONZERO DIGIT DIGIT DIGIT
)')
define( NAME, '(
    ( ( David | Dave | N ) Berkley ) |
    ( ( Art | Arthur | N ) Koenig ) |
    ( ( Jim | N ) Flanagan )
  )')
define( ANYBODY, '( NAME | LOCATION | EXTENSION )')
define( NET_COMMAND, '(
    status
    | ( ( drop | disconnect | add ) ( ANYBODY | incoming call ) )
    | ( answer incoming call )
    | ( put ( incoming call | ANYBODY ) on hold )
    | ( ( call | dial ) ANYBODY )
  )')
```

```
define( DATABASE, '(
    ( ( image | text | viewgraph | slide ) ( file | N ) ( number | N )
      NONZERO ( DIGIT | N ) ( ( ( point | dot ) NONZERO ) | N ) )
    | ( ( next | previous ) ( image | slide | viewgraph ) ( N | file ) )
    | ( ( more | additional) information )
  )')

define( ROOM_COMMAND, '(
    status
    | ( ( N | turn ) ( room | N ) lights ( up | down | off | on ) )
    | (
      ( N | turn )
      (
        ( ( slide | viewgraph | video ) ( N | projector ( 1 | 2 | N ) ) )
        | ( ( array | gradient ) microphone )
      )
      ( on | off )
    )
  )')

define( DATA_COMMAND, '(
  status
  | ( ( display | show | N ) ( DATABASE | list of ( image | text ) files ) )
)')

define(STATEMENT, '(
  NET_PREFACE NET_COMMAND POLITE
  | REMOTE_PREFACE DATA_COMMAND POLITE
  | LOCAL_PREFACE ( ROOM_COMMAND | DATA_COMMAND ) POLITE
  | VOICE_COMMAND POLITE
  | CONFIRMATION
  | stop previous command
)')

O_SILENCE STATEMENT * O_SILENCE .
```

(Manuscript received June 11, 1990)