

A MULTIMODAL TELECONFERENCING SYSTEM USING HANDS-FREE VOICE CONTROL

D. A. Berkley, J. L. Flanagan, K. L. Shipley and L. R. Rabiner

AT&T Bell Laboratories
Information Principles Research Laboratory
Murray Hill, NJ 07974-0636, USA

ABSTRACT

This talk describes the design and implementation of a digital teleconferencing system that seamlessly integrates a number of speech and image processing technologies together with data transmission capability. The goal is to provide a variety of sophisticated communication features that are easy to learn and fairly natural to use. The system is called HuMaNet, for Human/Machine Network. The system is controlled totally and interactively in a hands-free manner using spoken commands.

The HuMaNet system combines the technologies of speech coding, speech recognition, text-to-speech synthesis, and talker verification with autodirective microphone arrays, image compression, data and hypertext management to provide high-quality audio, image and video conferencing over basic-rate ISDN (Integrated Services Digital Network). The present public-switched transport capacity provides "2B+D", or two 64 kbits/sec circuit-switched channels (2B), and one 16 kbits/sec packet-switched channel (D).

For teleconferencing, the audio can be coded using a wideband (7 kHz) speech coder at 32 kbps, and the video is coded at 128 kbps using a commercial video conferencing codec. Image compression is used to transmit still images at rates of 64–128 kbps (using 2D-subband coding). Adaptive camera control, for group teleconferencing, can be achieved using the speech detection capability of the autodirective microphone array, thereby automatically and rapidly repositioning the camera for new talkers in a group teleconference.

1. INTRODUCTION

The promise of modern telecommunications is to provide seamless, high quality communications between people (or groups of people) and machines, anywhere, anytime, and at a reasonable price. We are still a long way from being able to implement and build a system which delivers on this promise. However, in the past several years, each of the technologies which is fundamental to this vision of telecommunications has made substantial advances; so much so that, within the laboratory, we have been able to build and use (on a day-to-day basis) an experimental teleconferencing system that integrates many of the principles of modern telecommunications. We call this experimental system the HuMaNet (Human/Machine Network) system because it embodies many of the fundamental ideas inherent in seamless, high quality communications between Humans and Machines. This experimental system has been evolving over the last five years. It is the goal of this paper to review the piece part technologies which are integrated into the HuMaNet system, and show how they work together synergistically to create a habitable system for multimodal teleconferencing.

Before describing the individual technologies in HuMaNet, it is worthwhile reviewing the basic concepts which led to HuMaNet more than 5 years ago. In the late 1980's progress in several key technology areas had advanced to the point where it seemed reasonable to ask the following question: "was there a simple, or at least a reasonable, way to integrate a variety of 'piece part technologies' (all undergoing continuous as well as discontinuous change) into an effective, human-engineered, computer terminal or workstation". The proposal that ensued was based on the following premises:

- humans needed communications, computation, and networking (access to information) to cope effectively in modern telecommunications environments

- pervasive, switched, digital connectivity, in combination with sophisticated terminals and peripherals could serve the telecommunications needs of real users
- although modern computers and workstations were quite capable of providing the telecommunications services that people needed, they were difficult for humans, en masse, to use.

Based on these premises, the HuMaNet project [1, 2, 3] was created as a research vehicle to learn how complex communications systems and computers could be made easier for humans to use. In particular, HuMaNet focused on the area of networking and teleconferencing, with a strong emphasis on human-machine communications using interfaces with "human-like" sensory capabilities, primarily *sight* and *sound*. A long term goal was to provide a communications environment with 3D spatial realism in sight (video) and sound (audio) – both in capture and in projection. A secondary goal was to provide natural language voice communication and control of the entire HuMaNet system, thereby obviating the need for any keyboard or handwriting interface.

The piece part technologies that were included in the first implementation of HuMaNet included:

- coding of telephone and wideband (7 kHz bandwidth) speech [4]
- synthesis of speech from an arbitrary ascii text (in English) [5, 6, 7]
- speech recognition of connected word sequences with robust rejection of speech not containing "control phrases"
- still image coding using perceptual sub-band image coding methods
- "smart" microphone technology in the form of an electronically steerable array microphone for locating the dominant talker in a room and steering the microphone to that talker (this array microphone was used in the speech recognition module which provided eyes-free, hands-free control of the HuMaNet system) [8, 9]
- high quality teleconferencing microphones and loudspeakers for group teleconferencing
- ISDN networking with 2 circuit switched B channels (64 kbps/channel) and 1 packet switched D channel (16 kbps) [10]
- a hypertext database of text and images [11].

Over time we have added several new piece part technologies (as well as having made substantial improvements to existing technologies), including:

- speaker verification for secure access to databases [12, 13]
- video conferencing using a commercial ISDN codec [14]
- acoustic echo cancellation (AEC) for hands-free voice communications [15, 16]
- automatic algorithms for control and tracking of video conferencing cameras based on tracking of voices by the array microphone [17]
- fully automatic control of all telephony functions including dialing out for voice, data and/or video calls.

Initially HuMaNet was implemented as a 2-node system between adjoining rooms at Bell Labs in Murray Hill. Subsequently we have added 2 additional nodes, one at Murray Hill, and one at the CAIP Center in Rutgers. Communications with the CAIP node is via a 45 Mbps XUNET-II connection [18].

In the following sections we describe the HuMaNet architecture and implementation (Section II), we review the ways in which we use each

of the "piece-part" technologies (Section III), and discuss what lessons we have learned from building and using HuMaNet over the past 5 years (Section IV).

2. THE HuMaNet ARCHITECTURE

Figure 1 shows a functional diagram of the HuMaNet architecture. The complete system is managed by seven 386 PC platforms and one Unix[®] workstation, each such platform (or workstation) having digital signal processing capability appropriate to its assigned function. The machines are locally networked on Ethernet (10 Mbps), and the HuMaNet nodes are connected by public-switched, dialed-up, ISDN from an AT&T ESS5 central-office switch, or via a 45 Mbps connection on the XUNET-II network in the case of the node at Rutgers.

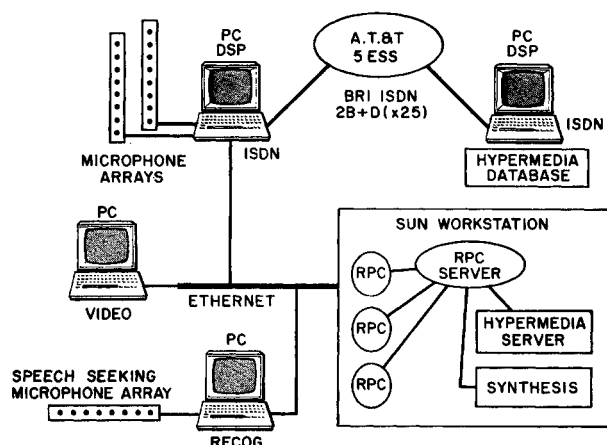


Figure 1

The individual PC's within each HuMaNet system are used in support of the piece-part technologies as follows:

- one PC is used (along with a specialized array processor) [19] for control of a speech recognizer capable of recognizing connected word sequences
- one PC is used to control the autodirective microphone array which provides both the speech input for the recognizer and the camera control positioning information
- one PC is used as a controller for the HuMaNet physical environment (e.g. lighting)
- one PC is used to access and display a local database of vugraphs
- one PC is used to encode the audio signals from the teleconferencing microphone arrays
- one PC is used to decode the audio signals received from the teleconferencing microphone arrays in a second HuMaNet node or other remote telephone connection
- one PC is used to access and display color images from an image database.

The Unix workstation serves several purposes, including acting as the host processor for text-to-speech synthesis, the interface to the speaker verification processor (an Alliant computer), the hypermedia server for requests for text, speech, and images, and the server for handling requests for video, audio, recognition, etc. using the standard Remote Procedure Call (RPC) capability of the Unix operating system.

Access to hypermedia databases of text, images, and speech in HuMaNet is both local (via Ethernet to the appropriate PC using an RPC on the Sun workstation), and remote over ISDN. All requests for data (as well as all control aspects of HuMaNet) are made by voice commands, since there is no other viable means of communicating with the system (i.e. there is no keyboard or mouse). Access to privileged data is controlled by a speaker verification "voice lock" procedure provided by a network-based server. Requests for information and all systems communications (e.g. status information) with the user are accomplished by an expert system using the text-to-speech synthesizer. All operations in HuMaNet are eyes-free, hands-free, by virtue of the speech-seeking, autodirective microphone arrays. Because of this feature, the speech recognizer is running all the time looking for voice commands. As such

it is essential that a highly reliable, low false alarm, protocol is used so that normal conversation can occur within a HuMaNet node without triggering the recognizer; at the same time providing immediate and accurate recognition of all spoken commands to the system.

Within each HuMaNet node there is a user terminal with two projection TV's for displaying text, vugraphs, images, and video sequences, (one TV is used for local display, the other for display of remote images) and a stereo sound system for providing high quality audio for teleconferencing. Figure 2 shows a photograph of one terminal of the HuMaNet system. The hands-free speech seeking microphone array is seen on top of the TV monitors. The fixed beam array teleconferencing microphones are seen on the sides of the monitors next to the fixed loudspeakers. The walls of the room are lined with acoustic absorption panels to control the reverberation characteristics of the room.



Figure 2

3. HuMaNet TECHNOLOGIES

The goal of the HuMaNet System is to help humans rapidly identify, locate, display, and assimilate information (in the form of text, data, images, voice, video or any combination of them) needed at the moment. To achieve this goal the HuMaNet system integrates a variety of technologies which, by themselves, are useful in getting a task accomplished; however, when carefully integrated into a unified whole, the synergy of these technologies provides benefits far greater than one would anticipate from the individual technologies. In particular, the technologies integrated into the current version of HuMaNet include the following:

3.1. Speech Coding

For high quality teleconferencing, wide band (7 kHz bandwidth) speech is essential. Although there are known methods for providing high quality speech coding of wideband speech at 16 kbps [20] the complexity of the 16 kbps coding system is such that it was not considered for deployment in HuMaNet. Instead a simpler 32 kbps speech coder was used to perceptually code [21] (using well understood audio criteria) a stereo pair of wideband speech channels and transmit them to another HuMaNet node using a single ISDN B channel.

3.2. Speech Synthesis

For instantaneous system feedback to the user, as well as for spoken accompaniment for image and video sequences, a general purpose text-to-speech synthesis system which can convert an arbitrary text message to an intelligible voice message is required. Over the past 5 years both algorithmic as well as hardware advances have enabled TTS algorithm to migrate from general purpose workstations (as used in HuMaNet) to medium processing power PC's (any 68030/40 or 83386/83486 processor). As a result, with today's technology it is easy to implement TTS and integrate it in with other technologies.

3.3. Speech Recognition

One of the basic tenets of HuMaNet is ease-of-use as embodied in voice commands and control of virtually every aspect of the system. As such there were several requirements that were mandatory for the speech recognition technology, including:

- eyes free, hands free, speech input
- connected word recognition capability for a fixed vocabulary, with a fixed syntax, trained to each user of the system
- excellent rejection of non-command speech (e.g. natural conversation in the room)
- excellent response to all commands
- ease of changing syntax and semantics for control of the system.

These requirements were satisfied using a hidden Markov model (HMM) [22] based connected word recognizer with a small vocabulary (on the order of 80 words) and a constrained syntax such that each valid spoken command was prefaced by one of several key phrases, e.g. “ISDN Control”, or “Room Control”, etc. In this manner the recognizer was able to reject all spoken inputs where the matches to these unusual phrases were not acceptable — i.e. all natural conversations, at the same time providing excellent and immediate response to all commands where the appropriate key phrase was included. A grammar compiler tool (and ultimately a semantics compiler tool) [23] was used to convert the recognition vocabulary and syntax into a finite state network (FSN) control structure for hardware implementation of the HMM recognizer on an AT&T BT-100 array of DSP 32C’s [24].

3.4. Speaker Verification

In order to provide secure access to database tokens, a speaker verification system was used to verify the claimed identity of the user. Although the computational requirements of speaker verification systems are far less than those of the recognizer, the technology was not implemented in a low cost DSP platform and instead was implemented as a network server from a laboratory computer system. Since speaker verification systems are trained for each new user, the eyes-free, hands-free, environment of HuMaNet posed no special problems for speaker verification.

3.5. Autodirective Microphone Arrays

Microphone arrays play a major role in HuMaNet as they provide the means for eyes-free, hands-free, voice control of the system, as well as for full duplex teleconferencing of speech between HuMaNet nodes. Two separate microphone arrays are used; one as a track-while-scan array that scans the room every 20 msec and finds the most prominent acoustic sound source in the room (hopefully the loudest talker) and electronically steers the microphone beam to point at that sound source if it is a talker; the other as a fixed array beam providing excellent spatial representation of the acoustic sources in the room (with the eventual goal of recreating the 3D sound field from one room to another).

3.6. Image Compression

It is simple to select and display local images in HuMaNet via voice commands. Each original image is a color image with 512×512 pixels, quantized using 24 bits to code each pixel. These high information rates pose no problem for local access and display. When a user wants to access such images remotely, the problem arises in the time it would take to transmit a 6 Mbit image ($512 \times 512 \times 24$ bits) over a single ISDN B channel, i.e. more than 90 seconds without compression. Hence some type of image coding and decoding is essential to display images (and ultimately full video) in HuMaNet. The image coder used was a perceptual subband coder which was able to consistently code images faithfully at rates on the order of 1 bit/pixel, or about 4 seconds to transmit a high quality color image over a single ISDN B channel [25].

3.7. Video Compression/Acoustic Echo Control/Camera Control

A commercial video compression system was selected for the video teleconferencing component of HuMaNet because of its availability and its ability to work over two ISDN B channels. Although the video quality is good for simple group conferencing applications, the major technology problems with video conferencing are proper acoustic echo control and appropriate camera control during a multiperson video conference. To achieve good acoustic echo control, specialized acoustic echo cancellation hardware was integrated in with the video system, providing high quality, lip synchronized audio which greatly improved perceived quality of the audio/video teleconferencing. To achieve good camera control, the direction finding capability of the autodirective microphone system was integrated into the camera control mechanisms of the video system, providing rapid and accurate tracking of the most prominent talker in the video conference.

3.8. Databases

The HuMaNet system has two databases of images and associated text (generally converted to speech using the TTS system and spoken to the user). One database is local to the system, and the other is accessed via ISDN. All database items are organized using hypermedia indexes that are stored separately from the databases. This allows easy access and display of database entries without requiring the user to know explicitly where in the database the desired information is located or what type of information is stored there (image, text, control).

4. HuMaNet OPERATION

The easiest way to illustrate how HuMaNet works is to examine some specific user commands. The commands are segmented into 3 categories, namely “Facility Control”, “Database Control” and “ISDN Control”. Each valid spoken command is preceded by one of these utterances, e.g. the simple command “Facility Control, Lights Down”, or a related synonymous phrase (“Room Control”), causes the recognizer to issue an RPC to the Unix Workstation requesting it to execute a process, namely to turn down the room lights. The Sun Workstation ultimately issues another RPC to the PC that contains the dimmer-control hardware, and ultimately the room lights are lowered appropriately.

An example of a “Database Control” command would be “Database Control, Image 5 Please” which, via access to the hypermedia indexes, issues a command (via an RPC) to the PC controlling the remote image database. The command is sent as a single packet over the ISDN D-channel to the remote database PC. Here it is decoded and the image (with any auxiliary text) is returned on a B-channel. The compressed image is decoded and displayed on the appropriate local TV monitor.

Finally the “ISDN Control” command of the form “ISDN Control, Call Room 2” first confirms the recognized command (to prevent errors in dialing), then dials the connection and sets up a full duplex audio (or audio/video) connection between the two HuMaNet nodes.

Over the past 5 years the HuMaNet System has grown and evolved to a full audio/video teleconferencing system which seamlessly integrates voice processing, image processing, networking, and teleconferencing. We have studied the individual technologies which go into HuMaNet extensively and made continuous improvements in the way in which HuMaNet works. Our ultimate goal is to understand how to make an audio/video teleconference between individuals, or groups of people, be as useful as face-to-face meetings between these same people. We have not yet achieved that goal since high quality video conferencing (with good echo control and effective camera motion) has only recently been introduced into the system. What we have learned is that it is possible to integrate a range of technologies into a relatively user friendly system, so that the technologies work together in a synergistic manner to enhance the user interface and make the overall system seem to be more powerful than the component technologies.

HuMaNet is continuously evolving. There are long term challenges to each of the component technologies that must be met before HuMaNet (or any other variant of this system) can be maximally effective. Among these challenges are the following:

Speech Recognition – the ultimate system will allow machines to handle unlimited vocabularies and unrestricted natural grammar with the goal of providing a natural language speech front end to the system.

Speech Synthesis – the ultimate system will provide natural sounding speech for unrestricted text, and be able to speak in any language, with any accent, for any gender.

Speaker Verification – the challenge in speaker verification is to be able to reliably identify a speaker (and similarly to reliably reject imposters) based on short segments of unconstrained speech from the speaker.

Image Compression – the challenge in image compression is to provide “perceptually lossless” images at a rate of about 0.25 bits/pixel so that a typical 512×512 color image can be transmitted over a single ISDN B channel in under 1 second.

Video Compression – one challenge in video compression is to compress a conference room image to 128 kbps (2B channels) while maintaining video realism and interactive communication. To achieve this goal will require high quality video compression, life-size color video displays, interactive capabilities for pointing, displaying multiple windows, controlling windows, and displaying and working on documents of all types.

Sound Pickup and Display – the challenge here is to provide wide-band high-quality pickup and reproduction of spatially realistic speech material, in a natural full-duplex conversational mode, while eliminating echoes, pickup of spurious noise and room reverberation.

5. SUMMARY

The HuMaNet System has been an on-going experimental facility for integrating technology piece parts into a working teleconferencing system. Over the past 5 years a great deal of effort has gone into developing the piece parts technology, and optimizing the interplay of voice processing, image processing, networking, and teleconferencing. Our long term goal is to quantify the synergies arising from the combined and simultaneous information modalities, and learn how to provide seamless communications capabilities using this type of system.

References

- [1] Berkley, D. A. and Flanagan, J. L., "HuMaNet: An experimental human/machine communication network based on ISDN," *AT&T Tech. Journ.* **69**, pp. 87-99 (Sept./Oct. 1990).
- [2] Berkley, D. A. and Flanagan, J. L., "Integration of speech recognition, text-to-speech synthesis, and talker verification into a hands-free audio/image teleconferencing system (HuMaNet)," *Proc. 1990 ICSLP 2*, pp. 861-864, Kobe, Japan (Nov. 1990).
- [3] Flanagan, J. L., Berkley, D. A. and Shipley, K. L., "A digital teleconferencing system with integrated modalities for human/machine communications: HuMaNet," *Proc. IEEE/ICASSP '91*, **5**, pp. 3577-3579, Toronto, Canada (May 1991).
- [4] Quackenbush, S. R., "A 7 kHz bandwidth, 32 kbps speech coder for ISDN," *Proc. IEEE/ICASSP*, Toronto, Canada (May 1991).
- [5] Olive, J. and Liberman, M. Y., "Text to speech work at Bell Laboratories: An overview," *J. Acoust. Soc. Amer.* **78S**, pg. S6 (1985).
- [6] Olive, J., "Using digital signal processors for real-time synthesis of acoustic signals," *J. Acoust. Soc. Amer.* **71S**, pg. S101, (1982).
- [7] Sproat, R., Hirschberg, J. and Yarowsky, D., "A corpus-based synthesizer," *Proc. ICSLP*, pp. 563-566, Banff (Oct. 1992).
- [8] Flanagan, J. L., Johnston, J. D., Zahn, R. and Elko, G., "Computer steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.* **78**, pp. 1508-1518 (Nov. 1985).
- [9] Flanagan, J. L., Berkley, D. A., Elko, G. W. and Sondhi, M. M., "Autodirective microphone systems," *Acustica* **73**, pp. 58-71 (Feb. 1991).
- [10] Roca, R. T., "ISDN architecture," *AT&T Tech. Journ.* **65**, pp. 4-17 (Jan./Feb. 1986).
- [11] Puttress, J. J. and Guimaraes, N. M., "An extensible hypermedia toolkit," *Proc. European Conf. on Hypertext* (Nov. 1990).
- [12] Rosenberg, A. E. and Soong, F. K., "Evaluation of a vector quantization talker recognition system in text-independent and text-dependent modes," *Comp., Speech and Language* **22**, pp. 143-157 (1987).
- [13] Rosenberg, A. E., DeLong, J., Lee, C-H., Juang, B-H. and Soong, K., "The use of cohort normalized scores for speaker verification," *Proc. ICSLP '92*, pp. 599-602, Banff (Oct. 1992).
- [14] AT&T Group Video System, Global Business Communications Systems, Basking Ridge, NJ.
- [15] Sondhi, M. M. and Kellerman, W., "Adaptive echo cancellation for speech signals," pp. 327-356 in *Advances in Speech Signal Processing*, Furui, S. and Sondhi, M. M. (Eds.), Marcel Dekker, NY (1991).
- [16] Gay, S. L. and Mammone, R. J., "Fast converging subband acoustic echo cancellation using RAP on the WE DSP16A," *Proc. ICASSP* (1990).
- [17] Berkley, D. A., "Audio for video teleconferencing," *Jour. Acoust. Soc. Am.* **94S**, pg. 1827 (Sept. 1993).
- [18] Fraser, A. G., Kalmanek, C. R., Kaplan, A., Marshall, W., Restrict, R., "XUNET 2: A nationwide testbed in high-speed networking," *Proc. Infocom '92*, pp. 582-589, Florence, Italy (1992).
- [19] Shively, R. R. and Gorin, A. L., "The AT&T BT-100 processor: A systolic processing ensemble," *Proc. Command and Control Res. Symp.*, pp. 196-201 (June 1988).
- [20] Jayant, N. S., Johnston, J. D. and Shoham, Y., "Coding of wide-band speech," *Proc. Eurospeech '91* (Sept. 1991).
- [21] Jayant, N. S., Johnston, J. D. and Safranek, R. J., "Signal coding based on models of human perception," *IEEE Proc.*, pp. 1385-1422 (Oct. 1993).
- [22] Juang, B. H. and Rabiner, L. R., "Issues in using hidden Markov models for speech recognition," pp. 509-553 in *Advances in Speech Processing*, Furui, S. and Sondhi, M. M. (Eds.), Marcel Dekker, NY (1991).
- [23] Brown, M. K. and Wilpon, J. G., "A grammar compiler for connected speech recognition," *IEEE Trans. Sig. Proc.* **39**(1), pp. 17-28 (Jan. 1991).
- [24] Gorin, A. L. and Roe, D. B., "Parallel level-building on a tree machine," *Proc. IEEE Int. Conf. Acous. Speech and Sig. Processing*, pp. 295-298 (Apr. 1988).
- [25] Safranek, R. J., Johnston, J. D., Jayant, N.S. and Podilchuck, C. I., "Perceptual coding of image signals," *Proc. 24th Asilomar Conf.*, pp. 346-350, CA (Nov. 1990).