Video Telephony

Joel S. Angiolillo
Harry E. Blanchard
Edmond W. Israelski

Although the transmission of video images over telephone lines has a long history, new technology, particularly video compression techniques, now makes video telephony affordable. The technology is practical, and it can be successful if systems maintain high quality and ease of use. The results of behavioral science studies of visual communication indicate that adding video to voice calls increases the effectiveness of human communication. Human factors and behavioral science techniques can evaluate the quality of visual communication systems and make such systems easy for customers to use.

Introduction

The announcement of the AT&T VideoPhone 2500™ in January 1992 began a new era for AT&T. Although skeptics of video telephony continue to point to the false starts of the past, there is ample evidence that this technology is now mature. AT&T is poised to offer a variety of visual communication products and services that are affordable and meet customer needs.

Video telephony begins by adding a video dimension to audio telephone service — simple person-to-person communication, which includes the transmission of the parties' images as well as their voices. The next addition to this service, already a feature of many teleconferencing systems, is the transmission of hard-copy graphics. From there, systems expand to transmit stored and live electronic graphics. Products will also be able to transmit auxiliary video input, such as recorded video segments from a videocassette recorder (VCR). (Panel 1 defines acronyms, abbreviations, and terms.) Live electronic graphics make it possible to offer a shared "blackboard" that both parties can manipulate. A display screen can be used for screen-based control of the video device, display of text messages, and access to information services. At this point, video telephony merges with multimedia computing.[1]

In this paper, we review the history of previous attempts to introduce video telephony, and why this may be the beginning of the video telephony revolution. Will customers see a need for video communication over the telephone? Human factors and other behavioral research studies have assembled evidence indicating what users will and will not gain from having a video dimension added to their telephone conversations. Given that the technology is mature, and that customers would find it useful, we then encounter two other factors important to the success of video: the quality of the transmitted image and how easy the equipment is to use.

The techniques of behavioral sciences can be used to assess the quality of compressed video images and to choose among various techniques and tradeoffs in video compression. Using human factors methods to design the video telephone will ensure that it will be as easy to use as the voice telephone is today. We review several crucial design factors that pose problems for usability of video telephony. Finally, we project which user needs may be fulfilled by the future technology of video communication.

History of Video Telephony

The origins of video telephony can be traced back 65 years to a historic, one-way full-motion video call in 1927 from then-Secretary of Commerce Herbert Hoover in Washington, D.C., to AT&T executives in New York City.[2,3] The frame rate was 18 frames per second. This technology became the foundation of commercial television, first

Panel 1. Abbreviations, Acronyms, and Terms

CCITT — International Telegraph and Telephone Consultative Committee
CIF — common intermediate format
codec — coder-decoder
CRT — cathode ray tube
fax — facsimile
FCC — Federal Communications Commission
ISDN — Integrated Systems Digital Network
IQ — intelligence quotient
kb/s — kilobit per second
Mb/s — megabit per second
MHz — megahertz
ms — millisecond
NTSC — National Television Systems Committee
PC — personal computer
QCIF — quarter common intermediate format
VCR — video-cassette recorder

introduced in 1936. Panel 2 shows many historic events in the evolution of desktop and group video conferencing, up to and including the AT&T Group Video System partnership with PictureTel Corporation, and the introduction of the AT&T VideoPhone 2500.

The earlier failures of Picturephone and Picturephone Meeting Service are now better understood. After 65 years, all the elements needed by AT&T to enter the age of video telephony are in place:

- Bandwidth is cheaper and more readily available, e.g., Integrated Services Digital Network (ISDN), and switched 56 kilobit per second (kb/s) and 384 kb/s services.
- Video compression technology is more advanced and significantly improved.
- Prices for equipment and transmission have become affordable for many customers.
- Saving time and travel expenses is more important than ever as today's businesses compete in a global economy.
- People are more comfortable with video and other higher technologies as a result of large market penetrations of VCRs, camcorders, and personal computers (PCs).
- Worldwide International Telegraph and Telephone Consultative Committee (CCITT) video coding standards (P×64 standards)[4] allow different types of vendor equipment now in place to work together.
- There is a growing universe to call using video telephony: As of 1993, thousands of video conferencing systems will have been installed in businesses all over the world.

Because these seven elements are now in place, many telecommunication marketing consultants and experts predict that a video marketing window will begin to open in 1993. Group video conference vendors are plentiful, and a few are making a profit today. Desktop video vendors are gearing up for a large market window that many expect to open by 1995.

What Does Sight Add to Sound?

Now that we have the technology needed to deliver two-way visual communication successfully and economically to the home and office, we are ready to face the more fundamental question: Who wants it? Since 1876, when the first telephone call was placed, people have gradually become comfortable with voice-only communication. The telephone has become an indispensable tool in the office and a necessity in the home. It works, is easy to use, and gets the job done. On the whole, we seem satisfied with this ubiquitous device. Is there a desire to add sight to sound? Some have argued that customers do not want two-way visual communication over the telephone network at any price.[5] What will it add to our personal communication?

Before answering this question, we must dispel a popular myth. A new technology does not always replace its older, sister technology. Television did not replace radio. The "limitation" of radio is also its advantage, namely, it does not demand your visual attention. Will new video phones replace our common telephones? Probably not. Sometimes it is nice to be anonymous. Often we do not want to worry about how we look. Occasionally, we want to wander around as we talk on the phone.

On the other hand, clearly we are a species that seeks out the visual. Although television has not replaced radio for music, it certainly has for news, drama, and education. Consumers are demanding more and better graphics in the PC world, even for activities that do not require them, in spite of the cost premium. Magazines, newspapers, and books are becoming more visual and

Panel 2. History of Video Telephony

1927	First one-way video phone call — Herbert Hoover in Washington, D.C. to AT&T in New York City (18 frames/sec)
1930	First two-way video phone call — from AT&T, 195 Broadway, to AT&T Bell Laboratories, West St. (Iconaphone)
1936	First TV broadcast by British Broadcast System
1941	Commercial broadcast TV Introduced — Black and white TV over channels 2 to 13
1953	Color TV system approved by the FCC — NTSC system
1956	Visiphone experiments at Bell Labs: Stamp-sized picture every 2 seconds over regular phone lines
1960s	Other non-Bell System video experiments and trials: Stromberg-Carlson (Vistaphone) GTE (Pictel) Trials (England, France, and Sweden PTTs)
1964	Mod I Picturephone introduced World's Fair in N.Y. to Disneyland in California 3 cities (N.Y., Chicago, Washington) — connected via public video booths Ladybird Johnson, Bell and Watson grandchildren on inaugural call
1965	Union Carbide Mod I Trial (N.Y. to Chicago) — used 35 Mod I sets
1967	Mod II Picturephone unveiled — (1-MHz analog, black and white)
1969	Mod II Picturephone Westinghouse Trial completed — N.Y. to Pittsburgh, 40 sets Jules Molnar, Bell Labs' VP, champions technology to AT&T CEO H. I. Rhomnes
1970	(July) Pittsburgh — First commercial Picturephone Service (max. 32 sets) Rate: $160/month and .25/minute N.Y. Public Utility Commission rejects N.Y. Picturephone tariff bid
1971	(April) Chicago Picturephone Service started 389 sets, 53 customers $50 month/line, $25/mo/set .15/minute — Exchange service within Chicago Loop area .75/minute to Oakbrook suburb C&P Telephone (Intercom pockets of Picturephone Service) at hospitals, schools, etc.
1972	John DeButts succeeds Rhomnes as AT&T CEO Executive Policy Committee appoints Ed Goldstein as Picturephone "Czar" AVP to address the market for video telephony
1973	Picturephone Service killed after $500 million investment First AT&T marketing organization created in wake of failure (Visual Communications Services Product Management Organization created under Joe Horzepa, Director)
1974	Picturephone Meeting Service (PMS) started (Intercity Visual Conferencing Service) Picturephone desktop segment trials begun, e.g., law, advertising, health care Public PMS rooms in 12 cities and large internal AT&T trials begun Some private PMS room trials (Department of Energy) and Arthur Anderson (under special FCC Tariff — "See While You Talk")
1980s	ISDN introduced
1990s	AT&T Video Business Units created — Global Business Video Systems and Services Consumer Video Services
1991	P×64 video standards approved by CCITT AT&T Group Video offering announced
1992	AT&T VideoPhone 2500™ introduced ISDN Residence Video telephone trial begins

less verbal. The market for image communication (e.g., fax, multimedia, video conferencing) is growing several times faster than the markets for either voice or data.

The fact that we sometimes *like* visual stimulation seems to be beyond contention. The more interesting question is, "What does a visual channel *add* to voice communication?" The answer to this question is obvious for people with impaired speech or hearing. The visual channel makes communication over distance easier and more natural. Making communication easier for the hard-of-hearing will be an important, perhaps the most important, role of visual communication in the coming years; however, our objective is to examine the role of the visual channel as a *supplement* to voice communication.

The Importance of Nonverbal Information. There are a number of ways in which the visual channel can contribute to the voice channel. It can:

- Add information not present in the voice channel (for example, quietly shaking one's head or pointing to an object in the environment)
- Provide redundant information (a laugh and a smile, together)
- Add emphasis, punctuation, or phrasing to the verbal message
- Create a sense of presence that may encourage the speaker and the listener to be more attentive
- Enhance one's memory of a conversation.

Extensive systems of classifying the nonverbal actions that accompany speech have been proposed.[6,7] Most such systems try to capture a handful of important communication needs:[8]

- *Turn-taking* — Whether the speaker is ready to give up the floor, and whether a listener wants the floor.
- *Feedback* — What the listener thinks of what the speaker has said.
- *General attentiveness* — How intently the speaker and listener are concentrating.
- *Emotional states* — The mood of the speaker and listener, in particular, concerning the discourse itself.
- *Speaker status* — How the speaker perceives his or her relationship to the audience.
- *Information about upcoming topics* — Preparing the listener for a change in topic, return to a previous topic, or relationship between topics.

Kendon adds to this list that nonverbal actions, in particular, gestures, lend interest to the speaker, and therefore hold the listener's attention.[9]

Sources of Nonverbal Messaging. The three most important nonverbal indicators are our eyes and face, our hands, and our appearance.

The eyes and face. The face is the source of the most subtle nonverbal messages, especially the mouth, eyes, and eyebrows. We are remarkably adept at determining what someone is looking at by observing his or her eyes. If we know where Jane is looking, we know what she is interested in, and together with her expression, we can guess what she thinks about what she is observing. The eyes are also a window into our emotional state. When we are in a contemplative or reflective mood, we often look up. When we are analyzing an argument, we often look down. And, of course, when we are tired, we close our eyes. Perhaps the most fundamental nonverbal messages communicated with our eyes are our level of interest in the speaker and whether we are in agreement, disagreement, or neutral.

Speakers and listeners watch the eyes closely to determine when to begin or end talking. When a speaker is ready to relinquish the floor, he or she will often gaze directly at the listener. If the listener is ready to take the floor, he or she will momentarily look away.[10] We are scarcely aware of this complex dance of the eyes. If we focus our attention on them, we become immediately entangled in our own self-consciousness.

Gestures. Gestures are the next most expressive form of nonverbal communication. Gesturing is universal, although some speakers use more gestures than others. Gestures are coordinated with and used to reinforce speech by helping the listener parse it. They are also used to facilitate turn taking. Occasionally, they are used to illustrate shapes or physical actions. McNeill[7] argues that because gestures are not part of a formal, culturally defined system, they reveal more about our thoughts than do our words.

Appearance. There are two kinds of "appearance" messages, voluntary and involuntary. We cannot control or easily modify the involuntary ones: signs of age, height, weight, a twitch, or the profile of one's nose. On the other hand, we often go out of our way to control the voluntary ones, including the color of our shirt or blouse, or the way we cut our hair. These are the silent messages that we constantly broadcast to the world. For example, putting on glasses will immediately raise your *perceived* intelligence quotient (IQ) by 15 points,[11] although the effect disappears as soon as the person wearing glasses is engaged in conversation.

The Role of Nonverbal Messages. What information is conveyed by this constant flood of nonverbal messages constructed from our clothes, posture, orientation, gaze, facial expression, hand gestures, and eye movements? For 100 years, we have not had the luxury of visual information in our long distance person-to-person communication. The letter is a text-only device and the telephone is a voice-only device. What have we been missing? What will video telephony add to our communication? Researchers have taken a straightforward approach to answering these questions — they have added a visual channel to a conversation to see what happens.

Since 1970, psychologists have conducted many laboratory studies looking for media effect in communication, for example, comparing voice-only communication with voice plus vision communication.[12] The procedure used in these laboratory studies can be generalized as follows: Two subjects, sitting in adjoining rooms, are asked to perform a task that requires both cooperation and communication. For example, one has a telephone book and the other a desire to find a particular restaurant. The type of communication channel between the two is varied. The two channels of interest in this paper are voice-only and voice plus vision.

...when two people can see each other, they can more often agree to a plan of action or a solution to a problem

Although many different types of tasks have been used, they can be divided into two broad categories: *information exchange tasks* and *interpersonal tasks.*

Information exchange tasks. One of most basic measures of communicative success is the *time* it takes to exchange information. Ochsman and Chapanis[13] asked pairs of subjects in separate rooms to perform three different tasks: scheduling a class, diagnosing a problem in an ignition system, and finding a particular part in a bin of similiar-looking parts. The tasks were designed to compel both subjects to work together to solve them. Although different types of communication channels were used in this study, only two are of interest here: voice-only (simulating a telephone) and voice and video (simulating a closed-circuit television).

Because the subjects had enough time to finish the tasks without error, the only measure of performance was the total time to complete the given task. Ochsman and Chapanis' extensive analysis of the results shows that a voice channel is essential for successful performance in an information exchange task, but a visual channel did not add significantly to the audio channel.

Gale[14] had groups of subjects perform three different information exchange tasks: deciding on the best alternative among a set of possibilities, deciding on the quality of a good product manager, and setting up a meeting. Subjects, in small groups, used one of three different media:

- Data sharing (computers linked so the information that one person typed on his or her screen appeared on the screen of each person in the work group)
- Data sharing plus audio
- Data sharing plus audio and video.

Like Ochsman and Chapanis, Gale found that the media choice did not affect the time it took to complete the different tasks.

These two representative studies offer strong evidence that when nonvisual information must be transmitted or negotiated between two or more people, the words we speak are all that we need. We can confidently conclude that the voice-only telephone is a pretty successful device for such cooperative work.

The typical information exchange study is looking for *performance* differences. Some studies have also asked subjects which media they prefer. Generally, they prefer the more "information-rich" channels (channels with vision), even though the addition of vision does not seem to improve performance. One reason might be that channels with vision are rated as higher in "social presence," that is, subjects feel that they know their partners better at the end of the study. Gale[14] reports another interesting finding: subjects *thought* that the voice plus vision channel would save them time in real life, even though his laboratory study did not show that it would.

Interpersonal tasks. Some tasks require more than a simple transfer of information. Argyle[8] writes, "Non-verbal and verbal communication normally play two contrasted roles. Non-verbal communication is used to manage the immediate social relationship — in much the same way as in animals; verbal communication is used to convey information connected with shared tasks and problems." Assuming he is correct, we need to examine the role that nonverbal communication plays in communicating not what we say or hear, but what we *think about* what we say or hear.

Although the results are often task-specific, the general finding from a number of studies[8,15-17] indicates that when two people can see each other, they can more often agree to a plan of action or a solution to a problem. This is especially true when one or both subjects are negotiating from strongly held beliefs, possibly because voice plus vision channel subjects are better able to determine their partner's position on an issue than they are in a voice-only condition.

Ekman[18] coined the term "leakage" to refer to nonverbal indications of our internal states. He presents evidence that nonverbal cues reveal more about our emotional state than do verbal cues, and are harder to suppress.

Even in experiments that did not show media effects, for example, in determining whether or not a partner is lying, subjects in voice plus vision conditions tended to make more confident judgments. Even though they are not any more accurate in determining another's feelings, they *think* they are, and they think their partners are more accurate in deciphering their own feelings.[8] We like to think we are perceptive and convincing even if we are not.

In summary, although information exchange tasks usually do not show a *performance* advantage for vision plus voice compared with voice-only, interpersonal tasks do. The visual channel appears to allow people to pick up nonverbal messages, and it provides a sense of "presence" that encourages commitment to the task.

Thus, behavioral science research suggests that video telephony will be useful to customers, not only because it provides a sense of presence, but also because nonverbal messages enhance interpersonal interaction. Obviously, video communication will add value for customers who have messages that can only be conveyed visually. Given that video telephony will be *useful*, the next step to success is to make it *usable*. Human factors and behavioral research are pivotal in assessing the quality of the transmitted video image and in designing video telephone systems that are easy to use. The next two sections discuss these topics.

Image Quality

For most users of video services and products, the "gold standard," or reference, for video image quality will be the video they see at home. The video signal supplied by North American and Asian broadcasters is currently the National Television Systems Committee (NTSC) system.

For the NTSC analog video system, image quality assessment and measures of user/viewer perception were fairly simple. Broadcasters and others were primarily concerned with the spatial resolution of images, accuracy of color reproduction, and the effects of interference such as video noise. *Spatial resolution* in an image is the level of fine detail often specified by the number of resolvable pixels per line of scanned video. See Table I for a comparison of image quality metrics for various video and other imaging systems (such as motion picture images). Because NTSC signals deliver motion at the highly acceptable frame rate of 29.97 frames per second, it was not necessary to create image quality metrics to measure impairments in the transmission of motion (temporal resolution). Viewers of NTSC or motion picture images are never exposed to any form of motion impairment.

A full digital encoding of the analog NTSC signal with no impairments to either spatial or temporal resolution would take at least 90 megabits per second (Mb/s). Practical coders in use by broadcasters operate at around 22 Mb/s with little or no perceptible image impairment. To squeeze the signal down to lower transmission rates requires extensive signal compression to take advantage of known and predicted redundancies of video signals in both the spatial and temporal domains. Unfortunately, compression algorithms that allow a transmission rate below 1.5 Mb/s (the transmission rate of a T1 trunk) cause visible motion impairments, such as jerkiness and/or smearing of objects in motion, which viewers do not experience when they watch NTSC or motion picture images.

Also, electrical noise that causes dots, snow, or streaks in analog video systems can cause major degradations in digital video signals, e.g., large areas of color blocks can appear on the screen as a digital decoder tries to flush out coding errors caused by electrical noise.

Time delays in reproducing the compressed digital image are another new major video impairment. Because of the intensive computations in processing compressed digitally encoded pictures, some algorithms produce time delays of 200 to more than 400 milliseconds (ms). To ensure synchronization of speech with images, the audio signal is delayed. The time delay that results from combining the audio and video signals is another tradeoff to achieving economical (low) video transmission rates. Later in this paper, we describe how the impairments of delay and lip sync affect ease of use for video systems.

In Table I, we compare the image quality metrics for analog NTSC commercial broadcast video and its variations with compressed digital video. Compressed digital video formats follow the new H.320 digital compression standards,[19] issued by the CCITT in 1990. The new

Table I. Image Formats and Quality

Formats	Usable* horizontal lines	Pixels per line	Total pixels per frame	Frames per second	Required bandwidth/ transmission rate
Analog video					
NTSC (Americas, Asia)	338	426	150,000	29.97	4 MHz
PAL (Europe)	411	420	172,000	25.00	5 MHz
VHS	338	280	95,000	29.97	< 4 MHz
Picturephone	175	175	30,500	30.00	1 MHz (6.3 Mb/s)
AT&T VideoPhone 2500™	112	128	15,000	1 - 10	8.2 kb/s
Computer image					
SVGA	1024	768	786,500	60	—
VGA	640	480	307,000	60	—
Motion picture film					
35 mm	(Not a raster-scanned image)		500,000	24	—
16 mm			125,000	24	—
Digital video					
QCIF (P×64)	144	176	25,000	15-30	56 kb/s - 2 Mb/s
CIF (P×64)	288	352	100,000	15-30	56 kb/s - 2 Mb/s
PictureTel, CLI	240	256	61,500	15-30	56 kb/s - 2 Mb/s
HDTV (proposed H4)	806	1920	1,550,000	50	140 Mb/s
MPEG (constrained set)	345	360	124,000	30	1.5 Mb/s and higher

*Eliminates retrace lines and includes the utilization ratio.

standards include the H.261 standard,[4] often called the P×64 standard, where the value of *P* can be any integer between 1 and 32. This standard covers digital video coding from 56 kilobits per second (kb/s) up to 2.048 Mb/s. Within the H.261 standard is the quarter common intermediate format (QCIF), which provides a spatial resolution of 144 lines with 176 color pixels per line. The standard also specifies an optional format, the common intermediate format (CIF), which provides a spatial resolution of 288 lines with 352 pixels per line. For a video equipment manufacturer to be in compliance with the CCITT P×64 standard, it must provide QCIF and, optionally, full CIF. At the present time, no standard objective metrics exist for the subjective motion impairments produced by these video coding systems. Because source material has a direct effect on motion impairments, much human factors work is under way in this area, especially to establish standard test sequences for full-motion video. Subjective image quality standards for compressed digital video pose some challenges. For example, a video system operating at 112 kb/s may be perfectly acceptable for group video conferencing in which there is limited head and shoulder motion, but it may be completely inadequate for showing an action-packed sporting event.

Designing Easy-to-Use Video Telephony

Modern video telephony will enhance communication. Fundamentally, it adds transmission of visual images and graphics, but there is also potential for screen-based control of telephony functions and access to textual information and image libraries. Picturephone also offered the fundamental enhancements of video telephony in the 1960s, but customers did not embrace it. The experience with Picturephone did teach us that one component vital to the success of visual communication is ease of use.[20]

What makes a product or service easy to use? Obviously, customers let us know when a product is easy to use and reaffirm that by purchasing our products and services. We can evaluate ease of use by examining its components: familiarity, transparency, simplicity, predictability, consistency, attractiveness, and adaptation of design to functionality.

Familiarity. The video user interface should use concepts and procedures familiar to users. In many cases, making a video call should be as similar as possible to making a voice call. The same conventions, procedures, and feedback should be used for video calls as for ordinary telephone calls, where these are applicable. For example, video calls should return auditory ring-back and busy tones identical or similar to voice network tones, in addition to any screen messages that may appear.

Transparency. The technology underlying the video interface should be transparent to the user. Users should not be required to provide information or perform operations that could be handled by the terminal or network. For example, users should not have to identify the types of devices they are using or that they want to contact. Just as users make fax-to-fax and voice-to-voice calls in the current network without "informing" the network of their mode of calling, they should not have to inform the network of the type of video device they want to contact.

...one component vital
to the success
of visual communication
is ease of use

Simplicity. Operating video telephony devices and features should be simple and self-evident. Using a video telephone should not require special training. To make a simple video call, a user should be able to either dial a number and be connected to a video mode directly, or press a button (e.g., clearly labeled as "video"). The user should not have to follow a special procedure that would require training or reading a manual before using the device. Users should not be required to dial a separate number or start and end a video call separately.

Predictability. The video telephony user interface should be consistent with user expectations, which evolve from common experience. For instance, video telephones have a "self-view" mode, in which the user can monitor what his or her camera is sending. Some research indicates that users strongly believe that if they see themselves, then the other party cannot see them. Videophones should work consistently with users' expectations or clearly inform users when they can be seen.

Consistency. The same types of video products, services, and systems should work alike from the users' point of view. A user who can operate one system or device should be able to move to a device of the same type and perform *at least* basic functions (such as placing a phone call). User-level interoperability, sometimes referred to as "driveability," is analogous to consistency in automobile design: Although the design of control panels varies widely in different cars, they are enough alike to enable anyone who knows how to drive one car to operate another. For example, if all video telephones had a "video" button that activated video transmission in the same way, then users could refer to that button when moving from device to device or talking to users of other devices.

Attractiveness. The user interface should be attractive and fun to use. Video products and services should not only be easy to use, but enjoyable as well. Although attractiveness is not traditionally considered part of usability, it is vital to a user's motivation to use a product or service, and to his or her subjective satisfaction with a system. For example, personal video telephones will be specialty instruments, not generic utility devices, during their introductory years. Thus, they should have a sleek, high-tech look, created by good industrial design and tailored by customer surveys.

Adaptation of Design to Functionality. AT&T video products and services will cover a wide range of functionality, from basic communication to complex multifunctional information appliances. As complexity increases, the user interface must remain simple to use. Thus, more complex products will require user interfaces that can deal with the increased complexity without intimidating users. An array of buttons or a screen of software buttons may be appropriate for a simple device, but a graphical user interface may be best for a complex one. For example, a video conference device may transmit graphics and images from an attached computer. It would be better to provide control of choice and transmission of images from the computer screen, in graphical format, rather than from an array of buttons on the video control device.

Designing for Ease of Use

How can we ensure that present and future visual communication applications will be easy to use? Involving professional user interface designers from beginning to end in the development process is a proven, successful method.[21–24] User interface designers should be involved in planning and initially defining a product or service, in designing and writing requirements, in directing successive iterations of implementing product requirements, and, after release of the product or service, in collecting customer feedback. Methods employed by user interface designers — task analysis, rapid prototyping, usability testing, and expert review — are described in detail by Day and Boyce[21] and, in this issue, by Opaluch and Tsao.[25]

Crucial Usability Issues for Video Telephony

In this section, we examine usability problems specific to video telephony (as opposed to more common, general aspects, such as ensuring that controls are within the reach of the user, etc.). As user interface designers learn more about video telephones and video services, but before many products and services with incompatible interfaces are available, we can share and standardize the solutions to many of these issues (see Table II).

Table II. Issues in Video Telephony Usability

Usability issue	Design response
Video call setup	• Simple dialing • Intelligent network identifies communication mode (video, voice, graphics) and terminal equipment
Spontaneity of communication	• Prevent call blocking • Portable terminals
Privacy	• Calls begin voice-only • Users have complete control over transmission and reception of video
Self-view	• Provide intuitive control(s) to view transmitted and/or local views • Mirrored for personal self-view, nonmirrored for document views • No surprises: being seen unintentionally while in self-view
Camera control	• Controls to allow natural scanning of other party
Delay and lip sync	• Minimize delays • Preserve lip sync
Physical environment	• Prevent misattribution of environmental problems (e.g., bad lighting) to video hardware

Video Call Setup. From the simplest to the most sophisticated video communication device, the first and most basic function for the user is to dial a video call. Making that call must be as simple as making a voice call in today's voice network. No special prefixes should be needed to identify the call as a video call. The user should not have to perform two-stage dialing, that is, dialing an access number followed by a second number, or having to set up both a voice call and a video call. A good analogy exists with fax calls: setting up a call to a fax machine is identical to setting up a voice call (except that a fax machine exists on both ends). Although the transmission medium is graphics instead of voice, the user making the call follows the same simple, well-learned procedure as in making a voice call.

Despite the seemingly obvious nature of this usability goal, it poses some severe technical challenges. In the near future, the network will probably support a variety of analog and digital video telephones, which may or may not work with one another. Users should not be burdened with identifying the device they are calling from and the device they are calling, yet placing this burden on the network may require extensive changes to databases and risk increasing call setup times to an unacceptable degree. Until video and voice devices work interactively, it may be difficult to provide simple video dialing for all video calls.

The technical challenges increase when the communications network becomes a multimedia network. We envision a network in which voice, video, data, and graphics are exchanged to a wide variety of terminal equipment. This equipment might be voice, video, graphics, or any combination of the three. (Despite the acknowledged need for telecommunication standards, it is likely that some terminals will not be able to communicate with other terminals, at least not without some mediating translation or call to an interworking service center.) The network itself must contain the intelligence to make the interworking between the heterogeneous devices and communication media completely transparent to the user. Users will not accept this new technology if they are required to provide information to the network to place and/or receive a call.

Ensuring Spontaneous Communication. In the past, video conferencing has been anything but spontaneous; rooms had to be scheduled many weeks or months in advance, channels reserved, and, in some systems, conferees had to move offsite to hold a video-mediated conference. The new generation of video conference and personal video telephones promises to bring to video communication the same spontaneity that customers now enjoy with voice. Spontaneity is a key ingredient in the success of video telephony, particularly personal video telephony.

Spontaneity
is a key ingredient
in the success
of video telephony

To ensure that spontaneity is preserved, we must prevent extensive call blocking; the probability of completing a call should not be noticeably different from what customers receive in voice communication. Furthermore, dialing a video call must be simple and easy to complete. Portability of the video terminal equipment is also important to its spontaneity. Initially, business customers are unlikely to provide personal video terminals on every desktop, and small video conferencing systems probably will be moved from conference room to conference room. Video terminals, then, are likely to be shared among people. Careful attention must be paid to the ergonomics of portable terminals. They should be easy to carry, easy to plug in, should not require special training to use, and should enable users to recover quickly from an incorrect setup. Extensive usability testing is essential to determine whether inexperienced users can set up equipment.

Privacy and Control of Video Modes. Customers want to be sure they can control their privacy during transmission of video calls. People want no possibility of being caught unaware. The concern dates back to Picturephone and is clearly evident in recent marketing and human factors research on video telephony. People are also concerned, though to a lesser degree, over when a video picture appears on the video telephone. They want control over whether a stranger can see them.

For AT&T video products and services, we have designed the video telephone user interface so that

- All calls on a video telephone begin as voice-only calls
- Full two-way video communication can only be activated if both parties press a "video" button on their video telephones
- During a video call, each party can control whether he or she can be seen by the other party.

This last function is ensured by a special privacy mode, a button that disables the transmission of video to the other party, but allows the user to continue viewing the picture being sent by the other party (provided, of course, that a picture is being sent). The challenge to designers is to provide a control that users understand will alternate between receive-only video and full two-way video. One option is to provide one "video" button that moves the call from voice to video and activates the privacy (receive-only) mode. The drawback of this method is that it does not allow the user to control reception of video, which would always be controlled by the other party. To solve this problem, the user can press a "privacy" button to activate the privacy mode, while allowing the video button to control whether overall video (sending and receiving) is on or off. For example, Picturephone provided a *disable* button, which sent a pattern to the other party. The AT&T VideoPhone 2500 has a physical shutter over the camera lens, the advantage of which is immediately obvious to the user.

Beginning a call in voice-only mode is the only way to guarantee the privacy that customers require. However, call setup becomes more difficult. The alternative would be to default to full two-way video mode when a call is connected. This eliminates having to choose a video mode after dialing the call. A compromise design would allow users to preset whether they enter voice-only or video when a call is connected, but this involves supplying the user with more controls or procedures. Perhaps as video becomes more ubiquitous, users will have fewer concerns about their privacy; then, placing a video call can be simplified to dialing and entering full two-way video. However, it seems more likely that users may always be concerned that the dimension of face-to-face contact may diminish privacy.

Self-View. Almost any video communication system should be able to display the video picture generated from the user's own camera. Various video products refer to this as *monitor, self-view, vu-self, preview,* or *local view.* This simple function, however, becomes more complex as the video system becomes more complex, for instance, when a video system has more than one video input, (e.g., from a VCR and/or a document camera), in addition to the camera focused on the person.

Self-view contains at least four subfunctions:

- Local view — Viewing the signal from a local video source, i.e., monitoring the video signal from a camera, VCR, graphics tablet, etc. There may be more than one local view when more than one local video source is active.
- Self-view — The local view from the camera pointed at the user. It is the video "mirror" and the user should be able to distinguish it from other local views.
- Monitor — Viewing the video signal being transmitted to the other party.
- Preview — Viewing a video signal that is not being transmitted, typically to inspect the picture before sending it to the other party. For example, a caller may want to check his or her appearance before starting a video call.

Simple video telephones may have one function with a single label that performs just one or two of the functions mentioned. For example, the AT&T VideoPhone 2500 has a self-view function, which enables the caller to preview and monitor the image generated by his or her camera. More complex systems, such as current teleconferencing systems and the next generation of desktop video telephones, will have a choice of video inputs from several sources, only some of which may be transmitted, and any of which may be viewed at one time on the output screen.

Self-View and Privacy. Should the video signal continue to be transmitted while the user is viewing his or her video signal? The AT&T VideoPhone 2500 continues to transmit video while the user views his or her picture. Market research at AT&T suggests that customers want the option of viewing themselves while sending their own picture to the other party. For example, a customer may want to show off new clothes, monitoring his or her own movements in self-view while the other party looks on. However, human factors research at AT&T Bell Laboratories indicates that users think that when they see themselves on their screen they cannot be seen by anyone else. This impression is so strong that it persists even when users are instructed about how video is transmitted during self-view. Users are concerned about privacy in video communication, and they want to control their video transmission. Allowing video to be transmitted during self-view risks giving users a negative impression about the privacy of video communication.

The challenge to human factors is to retain the functionality of allowing video to be transmitted during self-view, while ensuring that users are intuitively aware of whether they are sending their own picture to the other party. Fortunately, the presentation of the self-view can mitigate this problem. One option is full-screen self-view, i.e., when the picture of the other party is replaced by the picture from the user's own camera. Another option is a system that can display a windowed self-view. Here, pressing self-view pops up a small inset picture, overlaid upon the picture of the other party, containing the user's own picture. According to usability studies conducted by AT&T Bell Laboratories, users who can access the windowed self-view find it much easier to understand that self-view does not automatically imply that their own transmission is suppressed.

The Multiple-View Problem. The next step up in videophone complexity is to have more than one video source, e.g., a camera to transmit the user's image and an auxiliary input to transmit a recording from a VCR or laser disk player. The challenge is to allow users to choose which local view they want to watch, e.g., themselves or the VCR. At the same time, they may independently control which video source is being transmitted to the other party. These controls and displays must be designed so that users intuitively understand what they are seeing and what they are sending at any time.

An additional caution is required with multiple local views. When users are viewing themselves, they automatically react to the self-view as if it were their reflection in a mirror. Users have great difficulty adapting to a non-mirrored image of themselves, because they must relearn firmly established eye-body coordination habits. On the other hand, local views of graphics, text in a document window, or a VCR picture cannot be mirror-imaged; there must be a faithful reproduction of the image being transmitted. This poses some technical challenges. It also suggests that control panels be carefully designed for usability.

The issue of how to present the self-view merges with that of video control when video communication evolves into multimedia and shared work systems. The ultimate video communication device has multiple local video sources that can be viewed, the self-view "mirror," a VCR, a still video source, a graphics generator, a document camera, a camera pointed at a display source (e.g., an overhead display), etc. In addition, there are multiple remote sources, such as the other person's image, the other person's graphics, document camera, etc. There may be multiple remote sources as well, as in video conferencing. Finally, there may be a shared video source, e.g., a common, shared "blackboard" on which all parties to the call may enter text or draw graphics.

One solution to the multiple local view problem is to make the video communication device more like a multiple window system, or graphical user interface, common in computing environments. Multiple views of local and remote images could appear in separate windows, each window bearing a label indicating its origin, which is what the AT&T/NCR Personal Video System Model 70 does. Here, the visual communication device merges with the multimedia computer. However, this is not the only way to deal with the complexity of the multiple view problem. For example, in NTT's ClearBoard[26] prototype, a video link between two or more people is supplemented by a common "blackboard" on which all conferees can draw. The video views, however, are full-screen, and users can draw on a liquid crystal screen on top of the image of the other party. Thus, two people are having a conversation, and making eye contact, with a "transparent glass blackboard" between them.

Camera Control. In live, face-to-face conversation, a speaker's control of his or her view is natural and automatic, mediated by the muscles of the neck and eyes. When face-to-face conversation is simulated over a video link, the naturalness of the user's control over his or her own view of the other person(s) is lost. The basic unnaturalness is that other parties can control your view using their cameras. Most video conferencing systems provide remote controls for the other camera. Such controls need to be designed so that their operation is intuitive. The layout of controls, e.g., arrows, should take the perspective of the user looking at the other party, not the camera's perspective.

Delay and Lip Sync. Video compression promises to make video communication affordable, and therefore a viable commercial enterprise. However, in addition to degradation of picture and motion quality, video compression also delays transmission of the picture. Delays can range between 250 and 500 ms, depending on the relative adjustment between level of picture quality and delay engineered into the video coder/decoder (codec). Given the video delay, lip sync would be destroyed unless the audio is also delayed. Thus, video delays averaging 300 ms accompanied by deliberate audio delay are common in video compression systems. Such delays are readily apparent if a live picture and a delayed picture are presented side-by-side, or if a delayed self-view is presented to the user.

But delays are difficult for inexperienced users to detect when viewing a person over a remote link. What users can detect, however, is an aberration in the normal flow of conversation. This is similar to the problems people encounter in a voice conversation over a satellite link: the other party appears to be unusually slow to respond in the turn-taking of conversation. This can cause users to interrupt, breaking the flow of conversation. Interestingly, users tend to attribute the delay to the person and not to the communication link. This may change with experience. Because enhanced personal contact is one of the principal advantages of using video, this phenomena could be particularly worrisome. As codec technology advances, delays may be decreased to the point where people will not perceive them. Meanwhile, we need to know more about users' reactions to audio and video delays; we also must be aware of the tradeoff between allowing this delay to interrupt conversation, versus destroying lip sync to eliminate audio delay.

The Physical Environment. One of the most difficult factors to control is the physical environment in which users place their video telephone. This is a problem for personal video telephones and portable video conferencing systems. The designers of the personal and portable video system have little or no control over lighting and other environmental conditions in which the video telephone operates. Yet the physical environment can have a profound effect on the perceived operation of the video telephone. Lighting problems can cause a complete failure of the system from the user's point of view. The picture may be dark, purely because of lighting, yet the user has no indication that lighting is the problem. The user may blame the condition on hardware failure, or on mishandling of the controls. Users must be informed about possible problems in the environment, and video telephone design must, if possible, mitigate these problems.

Future Directions

First, we must have a firm understanding of user needs for visual communication. Only then can we think about the *products* and *services* that will serve those needs in the future.

Asymmetrical, or one-way, services such as television are appropriate when one wants to see but not be seen. Television does not allow the user to select the content, other than changing the channels. It is not personal. It does not do a good job of delivering content that will appeal to only a few viewers. Personal videophones and high-bandwidth services of 128 kb/s, 384 kb/s, or 1.5 Mb/s could offer content tailored to individual needs. In the future, "video-on-demand" will allow customers to retrieve every video recording ever made, and "news-on-demand" will present video clips from the most recent news events. One of the biggest applications for asymmetrical services might be anytime, anywhere education and training. Equipment to support these services would require large screens, good quality audio, and upstream signaling, which allows the viewer to request and interact with programming.

...we must have a firm understanding of user needs for visual communication

Person-to-person calling can use smaller screens, simple cameras, and handsets. Group-to-group calling requires larger screens, cameras that pan and zoom, and high-quality speakerphones. In both cases, customers want a service that gets as close as possible to the spontaneity, clarity, and warmth of face-to-face interactions.

Basic video calls will be made from one location to another. However, customers will eventually want one video call to be connected to several locations. On a multipoint *voice* call, the voices can be combined, but on a multipoint *video* call, other approaches are necessary, for example, splitting each viewer's screen into pictures from two or more remote sites, sending the picture of the current talker to each site, and so on.

Some applications will not require motion. For example, a catalog shopping service will require only high-resolution still-frame images. Others will emphasize smooth motion, sacrificing high-resolution imaging, as in a movie delivery service. Still other applications will allow the user to alternate between a still image mode and a motion mode. Business conferencing systems may be one such example.

A variety of products and services will be required to serve these diverse needs. Each will have different design constraints. At the same time, the user should be able to move effortlessly from one video product to another and from one video service to another. Users have complained that although computer companies may have brought high-speed information processing to the consumer, they have delivered products that are difficult to learn and remember, inconsistent, and non-intuitive. The human factors engineers who are designing future video products and services will have to work hard to make them as usable as common, everyday voice telephony.

References

1. N. I. Benimoff and M. J. Burns, "Multimedia User Interfaces for Telecommunications Products and Services," *AT&T Technical Journal*, Vol. 72, No. 3, pp. 42–49.
2. E. A. Mainzer, *AT&T Picturephone: The dysfunctionality of a functional structure*, New York University, Spring, 1984.
3. I. Dorros, et al., "The Evolution of PICTUREPHONE service," *Bell Labs Record*, Vol. 47, No. 5, May/June, 1969, pp. 137–141.
4. "Video Codec for Audiovisual Services at P×64 kb/s," *CCITT Recommendation H.261*, International Telegraph and Telephone Consultative Committee, Geneva, Switzerland, 1990.
5. Michael Noll, "VideoPhone: A Flop That Won't Die," *The New York Times*, January 12, 1992, p. 13.
6. R. L. Birdwhistel, *Introduction to Kinesics*, Louisville University Press, Louisville, Kentucky, 1952.
7. D. McNeill, *Hand and Mind*, University of Chicago Press, Chicago, 1992.
8. M. Argyle, "Non-verbal communications in human social interactions," *Non-Verbal Communications*, R. A. Hinde (ed.), Cambridge University Press, New York, New York, 1972.
9. A. Kendon, "Some relationships between body motion and speech: an analysis of an example," *Studies in Dyadic Communication*, A. Siegman and B. Pope (eds.), Pergamon, Elmsford, N.Y., 1972.
10. George A. Miller (ed.), *Communication, Language, and Meaning*, Basic Books, New York, 1973.

11. M. Argyle and R. McHenry, "Do spectacles really affect judgements of intelligence?" *British Journal of Social and Clinical Psychology*, Vol. 10, 1970, pp. 27–29.
12. A. A. Reid, "Comparing Telephone with Face-to-Face Contact," *The Social Impact of the Telephone*, Ithiel de Sola Pool (ed.), MIT Press, Cambridge, Massachusetts, 1972.
13. R. B. Ochsman, and A. Chapanis, "The effects of 10 communication modes on the behavior of teams during co-operative problem-solving," *International Journal of Man-Machine Studies*, Vol. 6, 1974, pp. 576–619.
14. S. Gale, "Adding audio and video to an office environment," *Studies in Computer Supported Cooperative Work*, J. M. Bowers and S. D. Benford (eds.), Elsevier Science Publishers, New York, 1991.
15. H. Wichman, "Effects of isolation and communication on cooperation in a two-person game," *Journal of Personality and Social Psychology*, Vol. 16, 1970, pp. 114–120.
16. J. W. Dorris, G. C. Gentry, and H. H. Kelley, *The effects on bargaining of problem difficulty, mode of interaction, and initial orientations*, University of Massachusetts, Amherst, Massachusetts, 1972.
17. I. E. Morley, and G. M. Stephenson, "Interpersonal and interparty exchange: A laboratory simulation of an industrial negotiation at the plant level," *British Journal of Psychology*, Vol. 60, 1970, pp. 543-545.
18. P. Ekman, "Non-verbal leakage and clues to deception," *Psychiatry*, Vol. 32, 1969, pp. 88-106.
19. "Narrow-Band Visual Telephone Systems and Terminal Equipment," *CCITT Recommendation H.320*, International Telegraph and Telephone Consultative Committee, Geneva, Switzerland, 1990.
20. C. G. Davis, "Getting the picture," *Bell Laboratories Record*, Vol. 47, No. 5, 1969, pp. 143–147.
21. M. C. Day, and S. J. Boyce, "Human factors in human-computer systems design," *Advances in Computers*, M. Yovits, (ed.) Volume 36, Academic Press, San Diego, California, 1993, pp. 333–430.
22. J. D. Gould and C. Lewis, "Designing for usability — Key principles and what designers think," *Communications of the ACM*, Vol. 28, No. 3, 1985, pp. 93–100.
23. J. Grudin and S. E. Poltrock, "User interface design in large corporations: Coordination and communication across disciplines," *Proceedings of ACM CHI '89 Conference on Human Factors in Computing Systems*, Austin, Texas, April 30–May 4, 1989, pp. 197–203.
24. R. Mulligan, M. Altom, and D. Simkin, "User interface design when you *do* know how," *Communications of the ACM*, Vol. 35, No. 9, 1992, pp. 19–22.
25. R. E. Opaluch and Y-C. Tsao, "Ten Ways to Improve Usability Engineering: Designing User Interfaces for Ease of Use," *AT&T Technical Journal*, Vol. 72, No. 3, May/June 1993, pp. 75–88.
26. H. Ishii and M. Kobayashi, "ClearBoard: A seamless medium for shared drawing and conversation with eye contact," *Proceedings of ACM CHI '92 Conference on Human Factors in Computing Systems*, Monterey, California, May 3–7, 1992, pp. 525-532.

(Manuscript approved June 1993)

***Joel S. Angiolillo** is a member of technical staff in the Integrated Systems Digital Network (ISDN) and Video Architecture Department at AT&T Bell Laboratories in Holmdel, New Jersey. He is currently working on making visual communication products easy to use and useful. Mr. Angiolillo joined AT&T in 1982 after receiving a B.A. and M.A. in linguistics at Amherst College, Massachusetts, and a Ph.D. in cognition and communications at the University of Chicago, Illinois.*

***Harry E. Blanchard** is a member of technical staff in the Applications Architecture Department at AT&T Bell Laboratories in Holmdel, New Jersey. He is working on video telephony and multimedia standards and guidelines to develop speech recognition applications. He is also involved in formulating standards in the field of human–computer interaction. Mr. Blanchard received a B.A. in psychology from the University of Connecticut, Storrs, and an A.M. and Ph.D. in experimental psychology from the University of Illinois at Urbana-Champaign. He joined AT&T in 1988.*

***Edmond W. Israelski** is technical manager of the Human Factors Group in the Cross-Platform Customer Solutions Department at AT&T Bell Laboratories in Middletown, New Jersey. He supports the design of user interfaces for customer premises switching systems and their adjuncts. Mr. Israelski received a B.S. in electrical engineering from the New Jersey Institute of Technology, Newark, New Jersey, an M.S. in operations research from Columbia University, New York, New York, and a Ph.D. in engineering and industrial psychology from Stevens Institute of Technology, Hoboken, New Jersey. He joined AT&T in 1970.*