Multimedia: Technology Dimensions and Challenges

Nikil Jayant
Bryan D. Ackland
Victor B. Lawrence
Lawrence R. Rabiner

Multimedia services are made possible by a host of underlying technologies. These include the processing of speech, audio, image and video signals, and handwritten data, as well as the high-quality transmission of audiovisual messages and data information. Audiovisual signal processing incorporates the subtechnologies of coding, synthesis, and recognition, and the technologies that support acoustical and optical transducers. Synchronous processing of component signals in real time is a particularly important challenge. The communication technologies used to transmit multimedia services include wired and wireless modems, circuit-switched and packet-switched networks, and technologies such as simultaneous voice and data for seamless integration of multimedia messages. All these technologies depend, for their efficiency and pervasiveness, on the low-cost processing power provided by submicron very large-scale integration (VLSI). This paper describes the current capabilities in these technologies and the future challenges, in terms of quantitative metrics and tradeoffs.

Introduction

The promise of the communications revolution in which we are all living is to provide ubiquitous access to information, anywhere, anytime, at a reasonable price. Although no one has a clear vision of where we will end up, or even how we will get there, it is increasingly clear that the revolution will be built on a series of advances in multimedia products and services, and the associated transmission and networking of both "real-time" and non-real-time data streams. This paper, which defines several technological dimensions of multimedia products and services, setting the stage for the rest of the papers in this special issue, also describes and quantifies the current capabilities of the base technologies of audiovisual signal processing and the associated data networking and transmission. In addition, it summarizes the challenges that must be faced in the less easily quantified dimensions of technology integration and user-interface design.

A Formal Definition of Multimedia.

Information is inherently multimodal, and humans process it effectively, efficiently, and simultaneously in several dimensions. The *multi*ple media that define the cornerstones of modern communication technology are *speech, audio, image, video*, and *data*. Speech is a special but important subset of audio, as is music. Machine printed and handwritten documents are significant subsets of still images. Important classes of video communication signals include head-and-shoulders video and closeup views of human faces. Each of these audiovisual signals is discussed in this paper. Examples not covered in this paper are the physical signals of tactile information, sign language sequences, and medical images.

The class of data signals differs from audiovisual signals in that no "real-time" transmission constraints exist. Data signals can tolerate transmission delays of

Table I. Digital audio formats.

Format	Sampling rate (kHz)	Bandwidth (kHz)	Frequency range (Hz)	Bit rate before compression (kb/s)
Telephony	8.0	3.0	(200-3,200)	128
Teleconferencing	16.0	7.0	(50-7,000)	256
Compact disk (CD)	44.1	20.0	(20-20,000)	1,410
Digital audio tape (DAT)	48.0	20.0	(20-20,000)	1,536

Table II. Digital television formats.

Format	Spatio-temporal resolution	Sampling rate (MHz)	Bit rates before compression (Mb/s)
CIF	360 x 288 x 30	3	36
CCIR	720 x 576 x 30	11	132
HDTV	1,280 x 720 x 60	55	660

hundreds of milliseconds or more without degrading their value or ultimate utility. However, for data signals, one usually expects perfect or very high levels of fidelity in representation and transmission. Audiovisual signals, on the other hand, are useful sources of information and entertainment, even when the fidelity of the signal is less than perfect.

The next section of this paper describes audiovisual signal processing and its coding, synthesis, and recognition components. The section on "Communication Technologies" includes subsections on wired and wireless modems, circuit-switched and packet-switched networks, and technologies for seamless message integration, such as integrated services digital network (ISDN), asynchronous transfer mode (ATM), and simultaneous voice and data (SVD). (See Panel 1 for definitions of abbreviations, acronyms, and terms.) "Digital Signal Processing and Computing" deals with the overarching issues of signal processing and computing. Finally, the section entitled "Summary" introduces issues of technology integration, software, and user-interface design.

Audiovisual Signal Processing: Coding, Synthesis, and Recognition

Coding, synthesis, and recognition are three key areas of interest in audiovisual signal processing. The purpose of *coding* is to achieve a compact (compressed) digital representation of the signal for economies in transmission or storage. *Synthesis*, on the other hand, focuses on creating spoken or pictorial information starting from text, rather than from human speech or a real image. The customer for both coded and synthesized audiovisual information is the *human*. In recognition the focus is on machine (computer) understanding of the information content of the signal, usually to help complete some task.

Each area for processing audiovisual signals is often synergistic with other areas. For example, both coding and synthesis are used to build voice response systems, while synthesis and recognition provide dialogue systems for voice control of machines. Recognition and coding provide complementary information in very low bit-rate coding, as in the example of face location in a head-and-shoulders videophone scene, followed by more careful coding of facial features.

Signal Compression: Coding of Speech, Audio, and Image Signals. The four fundamental parameters of coding are *bit rate*, *signal quality*, *processing delay*, and *complexity* of implementation. The overarching goal of coding is to decrease the bit rate while maintaining a specified level of signal quality. In general, it is also necessary to maintain specified levels of processing

Panel 1. Abbreviations, Acronyms, and Terms

ADPCM—adaptive differential pulse code modulation, a coding technique applied within the network for compressed voice/data transmission
ADSL—asymmetric digital subscriber line
ASIC—application-specific integrated circuit
ASPEC—audio spectrum coding
ATM—asynchronous transfer mode
AVP—advanced VLSI packaging
bpp—bit per pixel
CAP—carrierless amplitude and phase
CATV—cable television
CCIR—Consultative Committee on International Radio, part of the ITU. It deals with technical, operating, and tariff questions.
CCITT—International Telegraph and Telephone Consultative Committee, currently known as the ITU
CD—compact disk
CELP—code-excited linear prediction, a method of speech coding that combines linear predictive coding with vector quantization of the excitation signal
CIF—common intermediate format
codec—coder-decoder
CPU—central processing unit
CTIA—Cellular Technology Industry Association (U.S./North America)
DAT—digital audio tape
DCT—digital carrier trunk
DMT—digital multitone
DSP—digital signal processor
EMI—electromagnetic interference
FCC—Federal Communications Commission
FDDI—fiber distributed data interface
FDM—frequency division multiplexing
FIR—finite impulse response
FTTC—fiber to the curb
GSM—Global System for Mobile Communications (previously Groupe Spéciale Mobile)
HDSL—high-speed digital subscriber line
HDTV—high-definition television
HFC—hybrid fiber-coax
HMM—hidden Markov model
IMTV—interactive multimedia television
ISDN—integrated services digital network
ISO—International Organization for Standardization
ITU-T—International Telecommunications Union–Telecommunications Standardization Sector, the portion of the ITU that has jurisdiction over speech coding standards
JBIG—Joint Binary Images Experts Group
JPEG—Joint Photographic Experts Group
LAN—local area network

(continued on next page)

delay and implementation complexity.

Bit rate. Bit rate is generally measured in *bits per second* or *bits per sample*. The number of bits per second is simply the product of the sampling rate (measured in hertz or pixels per second) and the average number of bits per sample used in the quantizing system of the coder. Tables I and II define typical bandwidths and sampling rates in audiovisual communications. The sampling rate is at least twice the bandwidth, as per the Nyquist theorem. The tables also define the bit rates of corresponding signals *prior to* compression, assuming 16 bits per sample for uncompressed audio and 12 bits per sample for uncompressed color pictures. (For audio from a compact disk [CD] or digital audio tape [DAT], coding of a stereo signal is assumed, where the left and right channels are each coded independently and at the same rate.)

Bit rates after compression are generally much lower than the numbers shown in Tables I and II. Several coding systems also use variable-rate (ideally constant-quality) coding of different parts of the nonstationary audiovisual signal. Variable rate coding is particularly matched to packetized transmission.

Coding standards. Table III summarizes a range of standards for the coding of audiovisual signals. These

LAPM—link access procedure modem
LD-CELP—low-delay code-excited linear prediction. A CELP coder using backward adaptive prediction to reduce delay, used for ITU-T Recommendation G.728.
LPC—linear predictive coding, a method used to remove correlations in the speech signal.
MC-DCT—motion compensation–discrete cosine transform
mflops—millions of floating operations per second
MIMD—multiple instruction–multiple data
mips—millions of instructions per second
MOPS—million operations per second
MPEG—Motion Picture Experts Group
MVP—Multimedia Video Processor
NA—North America
NRZ—nonreturn to zero
NSA—non-service affecting, a network failure classification
NSP—native signal processing
NTSC—National Television Systems Committee
OC-*n*—optical carrier with varying signal rates
ONU—optical network unit
PBX—private branch exchange
PCM—pulse code modulation, another term for digitization of speech. For telephone bandwidth signals, it is usually combined with 8-bit logarithmic quantizers to achieve 64 kb/s.
PE—processing element
POTS—"plain old telephone service"
PSTN—public switched telephone network
QAM—quadrature amplitude modulation
QPSK—quaternary, or quadrature, phase shift keying
RAM—random access memory
RISC—reduced instruction set computer
ROM—read-only memory
RPLPC—regular pulse linear predictive coding
SDV—switched digital video
SG—Study Group
SONET—synchronous optical network
SRAM—static random-access memory
STS—synchronous transfer signal
SVD—simultaneous voice and data
TDM—time division multiplexing
TDMA—time division multiple access
TTI—text to image
TTS—text to speech
TTV—text to video
V. asvd—a simultaneous voice and data scheme approved by ITU-T SG 14
VLSI—very large-scale integration
VSELP—vectorized stochastically excited linear prediction

standards promote digital communication by providing interoperability of coders-decoders (codecs) from different manufacturers. The first group of standards is for telephone bandwidth speech (3 kHz) with applications to network telephony, cellular communications, and secure voice. The second group includes a standard for wideband telephony and teleconferencing (7 kHz) and several standards for storing high-quality audio. The third and fourth groups contain standards for still images and video at several rates.

Signal quality. Given that the ultimate judge of a signal compression system is the human observer, audiovisual signal quality is best described by a subjective criterion. Five-point scales of signal *quality* (or impairment) are widely accepted, and are sometimes supplemented by measurements of *intelligibility*, especially in audio. Measuring subjective speech or image quality in a reliable, repeatable manner is a difficult, often painstaking problem. Likewise, measuring the *composite* quality of an audiovisual signal is at least as difficult, and data on this subject are generally unavailable to date. In the case of image signals, subjective quality is a function of viewing distance. A distance of four times the picture height is a typical assumption for quality measurement and coder design.

Table III. Standards for speech, audio, image, and video coding.

Standards body	Standard	Year	Algorithm	Bit rate	Application
CCITT	G.711	1972	μ-law PCM	64 kb/s	Network telephony
CCITT	G.721	1984	ADPCM	32 kb/s	Network telephony
ITU-T	G.728	1992	LD-CELP	16 kb/s	Network telephony
GSM (Europe)	GSM	1988	RPLPC	13.2 kb/s	Cellular telephony
CTIA (NA)	IS-54	1989	VSELP	8 kb/s	Cellular telephony
NSA	FS1016	1989	CELP	4.8 kb/s	Secure voice
NSA	FS1015	1975	LPC 10E	2.4 kb/s	Secure voice
CCITT	G.722	1984	Subband ADPCM	48-64 kb/s	Teleconferencing
ISO	MPEG-1	1992	Musicam/ASPEC	128-384 kb/s	Two-channel audio
ISO	MPEG-2	1996		320 kb/s	Five-channel audio
ISO	JBIG	1991	Run-length coding	0.05-0.10 bpp	Binary images
ISO	JPEG	1991	DCT	0.25-8.0 bpp	Still images
ISO	MPEG-1,2	1991, 1994	MC-DCT	1-8 Mb/s	Addressable video
CCITT	P x 64	1991	MC-DCT	64-1,536 kb/s	Videoconferencing
FCC	HDTV	1996	MC-DCT	17 Mb/s	Advanced television

Figure 1 shows perceived signal quality as a function of the average number of bits per sample for several signals for rates from 2.0 to 0.25 bits per sample. The horizontal line in the figure represents a high-quality target (quality score of 4.5). For a given number of bits per sample, the perceived quality is highest for video sequences and lowest for wideband audio.

Processing delay. In coding, the processing delay is the sum of delays incurred in the coding and decoding stages of the coding algorithms. At the encoder, delay is introduced in the process of buffering blocks of data to reduce redundancy and thereby to increase efficiency in the signal compression process. Examples of delay-causing algorithms are the block-processing methods of linear transforms and vector quantization. At the decoder, delay is introduced by block-based operations such as inverse transforms, and also by operations such as interpolation to increase the displayed frame rate in video coding.

Another source of delay in communications is transmission delay that stems from the networking of digitized signals. Applications such as telephony and videoconferencing require the lowest possible values of delay—subject to the requirements of the compression algorithm—to provide adequate levels of signal quality. In one-way communications, such as broadcasting, delay issues are less important. To minimize processing delays, however, it is still important to address requirements such as the effects of delays in station switching. In storage applications, encoding delay is essentially irrelevant, as long as the decoding delay is low enough to provide a good quality of service.

Complexity. Complexity is measured both in terms of the arithmetic processing required by the algorithm (typically measured in millions of instructions per second [mips] or millions of floating point operations per second [mflops]) and by its memory requirements (typically measured in kilobytes or megabytes of read-only memory [ROM] or random access memory [RAM]). The use of mips (or mflops) as a complexity measure is particularly appropriate for implementations on general-purpose digital signal processors (DSPs) or microprocessors. In application-specific integrated circuits (ASICs), other metrics

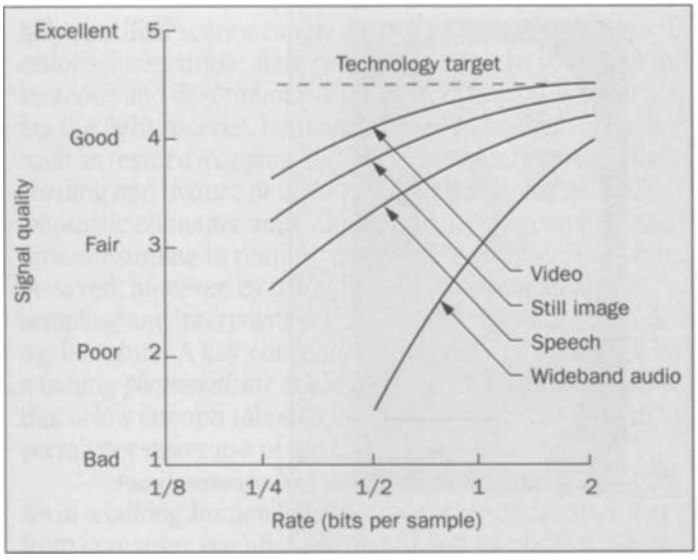

Figure 1. Current capabilities in the coding of audiovisual signals.

of complexity (such as the number of transistors or gates) may be most relevant. Complexity is an important parameter of performance for at least two reasons—the need to minimize cost and the requirement (especially in applications with portable devices) to minimize power dissipation.

In applications such as broadcasting, it is particularly important to minimize complexity at the decoder because decoders greatly outnumber encoders. Complexity of the encoder is a relatively less important, if not insignificant, issue.

The available per-chip complexity (in arithmetic as well as memory) has been increasing exponentially over the last several years, with no saturation in sight at least until the early years of the next century. This will permit the practical use of increasingly sophisticated algorithms for signal coding (as well as other technologies for multimedia).

Coding algorithms. Numerous coding systems have been developed to provide different tradeoffs of delay and complexity. However, there are only two basic principles of signal compression: removing the statistical (or deterministic) redundancies in the audiovisual source signal, and matching the quantizing system to the properties of the human perceptual system. One example of the first principle is using the similarity between consecutive image frames or speech samples in techniques such as differential coding. An example of a simple use of the second principle is the limitation of audio and color image resolutions to 16 and 12 bits, respectively, in the high-quality formats of Tables I and II.

Speech signals have a well-understood universal model of production that permits powerful techniques of redundancy reduction—primarily, linear predictive coding. Likewise, image and video signals have obvious interscan-line and interframe redundancies, especially in the case of videoconferencing scenes with head-and-shoulders inputs of relatively low spatio-temporal activity. In speech as well as image and video coding, the quantizing system can take advantage of a perceptual phenomenon called *noise shaping* to achieve even greater reductions of bit rate for a given level of signal quality. *Masking* is the phenomenon by which a strong stimulus (desired signal) completely covers up a weaker signal (quantizing noise) in its spectral or spatio-temporal vicinity, even if the local signal-to-noise ratio is only modest. In other words, mathematically significant levels of quantizing distortion can be permitted with very low or sometimes no loss of perceived signal quality.

The most common form of perceptually tuned coding occurs where carefully selected components in a frequency-transform (short-time signal spectrum) are either coarsely quantized or even discarded. The greatest advances in perceptual coding have been in the compression of unrestricted audio. Unlike speech or videophone signals, unrestricted audio has no universal model or framework for redundancy reduction. In compressing these signals, the approach is to reduce redundancy as much as possible using classic techniques of prediction and transform coding, and to shift much of the total burden to the perceptual (noise-shaping) side of the game. The gains of perceptual coding are greatest when the bit rate is just high enough to provide perceptually lossless coding, implying a noise profile that is exactly at the perceptual threshold. The gains are much less in the so-called suprathreshold points encountered in very low bit rate coding.

Applications of signal compression. At bit rates below 10 kb/s, applications such as secure voice, cellular radio, voice mail, and image phone are practical options. Between 10 and 20 kb/s, applications in network telephony, audio conferencing, and videotelephony can be

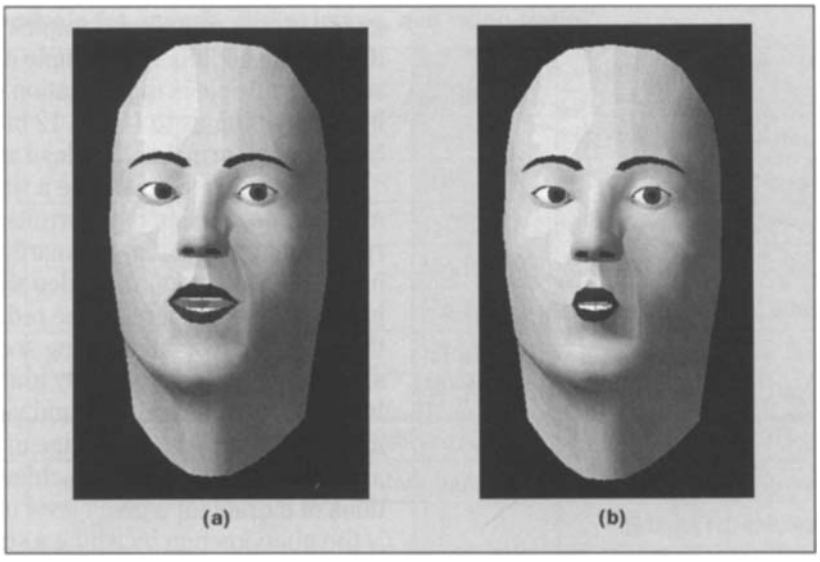

Figure 2. Photo of a talking face. Synthesis of (a) the sound | *a* | and (b) the sound | *i* |.

supported. Between 20 and 100 kb/s, several other audiovisual applications emerge, including slide show graphics, high-resolution facsimile, and music preview and broadcasting. Videoconferencing attains high quality at rates between 100 and 500 kb/s, movies on demand at about 1 to 5 Mb/s, and high-definition television (HDTV) at about 20 Mb/s.

Speech and Image Synthesis. Technologies that let a machine *speak* and *write* (or create a picture) often start from only a textual or mathematical description of the underlying message. By eliminating the need for a human to produce sounds and images, nearly arbitrary levels of flexibility and scope can be attained, producing various combinations of elementary audiovisual data. Key metrics in the technology are the *naturalness* and *intelligibility* of the resulting audiovisual signal, and the complexity of implementation (measured again in terms of the arithmetic effort and memory requirement).

Text-to-speech synthesis. Text-to-speech (TTS) systems can "read aloud," in that they convert text messages to intelligible speech based on both linguistic analyses of the text and the acoustic knowledge of the production of speech sounds (phonemes, syllables, words, phrases), within the context of the desired sentences. The linguistic processing relies heavily on dictionaries of word pronunciations and on the rules for handling words not included in the dictionaries. The acoustic signal processing relies heavily on spectral models of sounds that are easily modified by word context. Finally, processes for generating appropriate word durations and intonation (pitch) are crucial to making the resulting synthetic speech intelligible and to having some reasonable degree of naturalness.

TTS systems can be implemented on modest microprocessors (such as the 386) with about 10 mips of computation. Storage of units, tables, and dictionaries often requires about 10 MB of memory for research systems with maximal modularity and flexibility. However, only about 1 to 2 MB of high-speed memory is required for a range of hardware implementations.

Although TTS systems can maintain high intelligibility for a broad range of applications, making the resulting speech sound natural is still a major challenge.

Image synthesis. As in text-to-speech synthesis, *text-to-image* (TTI) and *text-to-video* (TTV) technologies convert textual descriptions of an image or a sequence of images into synthetic versions, achieving the conversion with the intervention of disciplines beyond signal processing. By using computational geometry and, in particular, techniques of ray tracing, optically accurate depictions of care-

fully specified scenes can be obtained. These textual specifications, for example, may include the size of a room and the locations and descriptions of a few main objects in it, including the light sources. By using applied mathematical tools such as texture mapping and two-dimensional fractals, the shading and texture of the scenes can be enhanced. Fully automatic computer animations are generally complex and time-consuming to render. Significant amounts of time can be saved, however, by using the principles of signal subsampling and interpolation that are well known in the coding literature. A key challenge in TTI and TTV syntheses is attaining *photo-realistic* quality with a computation effort that is low enough (about a few hundred mips or less) to permit pervasive use of the technology.

Face synthesis. Not surprisingly, automatic synthesis of a talking human face has received special attention from computer scientists, in an attempt to create a *personal interface* (personal assistant or agent) for human-machine interaction. In this technology, the modulation of facial features and, to a smaller extent, global head motions is governed by textual information.

Using a large number of polygons (on the order of a few hundred), scientists have made significant advances in developing wire-frame models of the human face, with a flexibility that permits adaptation to specific expressions (such as anger or curiosity). In fact, these models have also synthesized a moving image, where control parameters for facial-feature changes are derived strictly from the printed text. Given that characteristics such as text-derived intonation and emphasis are already available in text-to-speech synthesis, the natural next step is to use these characteristics to modulate the face image, and thus to produce a *talking face*. Incorporating global movements of the synthetic head in horizontal and vertical directions adds further sophistication.

In initial demonstrations, a talking face has indeed been a more compelling interface than a talking computer, and the levels of intelligibility are good enough even to lip-read the synthetic face (see Figure 2). The challenge is to increase naturalness significantly, while reducing the complexity of implementation.

Machine Recognition of Speech and Image Signals. A number of technologies provide a machine with the ability to *hear*, to *see*, and, more importantly, to *understand* audiovisual information. There are, however, key trade-offs in the process. Among these are the correctness (or error rate) in understanding the task, the size and complexity of the audiovisual vocabulary that needs to be understood, and the robustness and performance in light of uncertainties such as the variations in input speech and speaker, or variations in input printed text and handwriting.

Speech recognition. The challenge of speech recognition is to be able to transcribe arbitrary spoken input with exactness. Such a challenge is often far too demanding, and even unnecessary from the point of view of communication. Hence, a subproblem within speech recognition, often referred to as speech understanding, is to extract the meaning from an arbitrary spoken input. Other problems include speaker verification and word spotting.

Speech recognition is generally implemented as a system based on pattern recognition technology—namely with a pattern training/pattern matching paradigm. Recognition systems differ in their patterns (phrases, whole words, subword units, etc.), in the representation of each pattern (statistical models, such as hidden Markov models [HMMs], and templates), and also in the techniques used to match arbitrary inputs to sequences of patterns.

Speech recognition systems are often classified as belonging to one of three types:

- *Isolated word/phrase systems,* in which the speech patterns are distinct words or phrases, and the user is expected to speak a single word or phrase from the specified recognition vocabulary to command or control a system (for example, selection of a command from a menu).
- *Connected word systems,* in which the speech patterns are distinct words or phrases (often composed from units smaller than words). The user is expected to speak a fluent sequence of the words and phrases and the recognition system provides a best string match. Such systems are used for voice dialing of telephone numbers, entry of credit card numbers, and applications involving ordering from catalogs.
- *Continuous speech systems,* in which the patterns are subword units from which words and phrases are created using a lexicon of word pronunciations (in terms of the subword units). The user is expected to speak fluent speech that is recognized according to the word lexicon and an associated word grammar that specifies

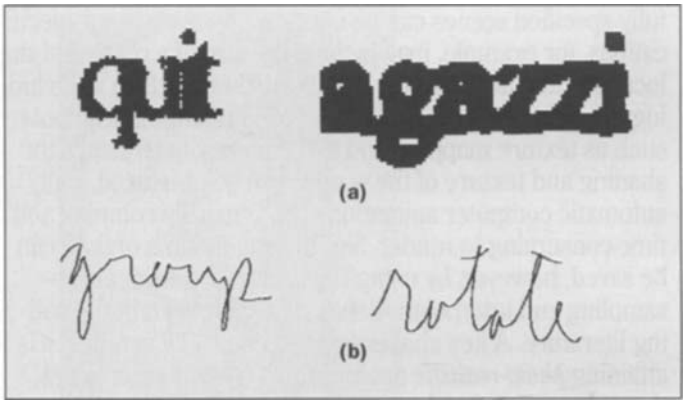

Figure 3. Examples of correctly recognized document samples, including: (a) distorted and blurred text (*quit*, a*gazzi*), and (b) handwritten words (*group*, *rotate*) from a writer outside the training set.

valid sentences in the language. Such systems are used to query databases, for dictation, for form-filling applications, and, ultimately, for language translation.

Current Capabilities in Speech Recognition and Understanding

The current capabilities in speech recognition and understanding can be described by the following speaker-independent performance metrics:

- Word error rates of 0.1 to 0.5 percent for isolated word vocabularies on the order of 10 to 200 words,
- Digit error rates on the order of 0.5 percent for connected digit strings of variable length,
- Word error rates of 2 to 3 percent for understanding of fluent database commands with active vocabularies on the order of 2,000 words, and
- Word error rates on the order of 7 percent for recognition of real speech from the *Wall Street Journal* with a vocabulary of 60,000 words.

Speaker verification. The problem of speaker verification is to decide whether an unknown speech sample was spoken by the individual whose identity is claimed. In speech recognition, the problem is to normalize out, in some sense, the individual speaker and extract the message content of the speech. In speaker verification, the problem is to normalize out the message content and extract information about the individual speaker. The two processes are similar, with some small differences.

For speaker verification, a customer wishing to be verified provides a claimed identity (to access the appropriate stored voice pattern), the spoken phrase suitable to the verification system, and the transaction requested. Depending on the transaction requested, the decision to accept or reject the identity claim is made and provided to the customer by a computer voice response system.

A speaker verification system can make two types of errors: it can reject a true customer (Type I error), or it can accept an imposter (Type II error). The goal of most verification systems is to try to bound Type I errors to less than 0.5 percent, while limiting Type II errors to less than 10 percent. Often, in laboratory testing, performance scores are given for equal rates of Type I and Type II errors. Within the laboratory (a highly controlled environment), equal error rates (of Type I and Type II) as low as 0.5 percent have been obtained, while more realistic field evaluations have yielded equal error rates of from 2 to 4 percent. Clearly, the challenge for speaker verification is to obtain laboratory performance in the field.

Image recognition. To continue the theme of machine recognition, this section focuses on the subfield of document image recognition. The challenge here is to recognize, with useful levels of accuracy, characters or character strings that are machine printed, as well as noisy and highly variable versions of documents that have been faxed or handwritten (see Figure 3). Authentication of a user by means of a sample of the user's signature is another important task. It is well known that authentication procedures can be made arbitrarily close to perfect by combining results from various modalities: for example, speech, face-image, fingerprint, retinal scan, and handwritten signature.

Document recognition is a well-researched area in pattern recognition. Technologies for document recog-

Table IV. ITU Study Groups active in multimedia services over the PSTN.

Standards being developed by the ITU		
Study Group	Question	Recommendation
SG 1	4	PSTN-based telecommunications services
	8	Mobile audiovisual services
	11	Enhanced facsimile services
	12	Message-handling services
	20	Audiovisual multimedia services
SG 8	5	Facsimile apparatus
	10	Audiographic conferencing
	11	Interactive audiovisual services
SG 14	1	Simultaneous voice and data
	1	Capabilities exchange and selection
SG 15	2	Visual telephone systems

nition have also benefited from the science of neural networks and two-dimensional HMMs. As in speech recognition, current capabilities in recognizing complex documents depend on the use of powerful models of context. For static or *off-line* images, these models are ideally two-dimensional. For dynamic images, such as in the task of on-line recognition, one-dimensional models, aided by cues of stroke speed and force, have led to very useful levels of performance.

Current Capabilities in Document Recognition. The current capabilities in document recognition include:

- Font-independent recognition of typewritten or machine-printed characters with a character error rate lower than 0.5 percent,
- Keyword spotting of a small number of words in noisy printed documents with error rates lower than 5 percent,
- Recognition of handwritten digits and letters with character error rates lower than 5 percent,
- Writer-independent recognition of handwritten words with a 32-word vocabulary with error rates of 5 percent, and
- Verification of signatures with equal error rates lower than 5 percent.

Communication Technologies

This section describes the communication technologies used in multimedia services, including the public switched telephone network (PSTN), ISDN, multimedia signals of 1.5 to 7 Mb/s over the local loop, next-generation networks, and private branch exchanges (PBXs) and local area networks (LANs) within customer premises.

Multimedia over the Switched Network. The PSTN is the most widely used medium for transporting voice, voiceband data, and facsimile. It has achieved this status for two reasons—it is nearly ubiquitous, and it is relatively inexpensive to use. Until recently, the bandwidth of PSTN modems was too limited to offer anything but simple data communications, narrowing the PSTN's role to the applications mentioned and using data networks for enhanced services such as videotelephony, simultaneous voice and data, and general multimedia. Now, advances on a number of fronts make delivery of multimedia over the switched network possible.

The principal advances include:

- Development of the V.34 modem standard, which more than *doubles* the available data bandwidth over a PSTN connection to rates on the order of 33.6 kb/s. (The V.34 recommendation was originally adopted in

1994 and specified a maximum data signaling rate of 28.8 kb/s, which was exactly twice the speed of the existing standard, V.32 bis. At the May 1995 meeting of ITU-T SG 14, the V.34 recommendation was extended to include data rates up to 33.6 kb/s.)
- Availability of good quality, low-bit-rate speech coders that compress 7-kHz speech to 16 kb/s and 3-kHz speech to 8 kb/s.
- Availability of powerful DSPs, custom very large scale integration (VLSI), and other processors at reasonable cost, which provide upwards of 50 to 100 mips of processing power.
- Emerging international standards for PSTN videotelephony, audiographic conferencing, and simultaneous voice and data.
- Strong industry interest, especially from computing and communications companies.

Together, they not only make multimedia over the PSTN possible, but by the end of the decade, promise to make it part of our daily lives.

Voiceband modem technology. Ever since the development of the first voiceband modems in the late 1960s, there has been a steady push to drive the maximum data rates higher. In the late 1960s, the Bell 103 modem provided speeds up to 300 b/s. In the late 1970s, it was superseded by the Bell 212A, which quadrupled the speed to 1.2 kb/s. At the time, this was viewed as more than adequate for most data communications needs. As technology improved, 2.4 kb/s became possible, and in the early 1980s, V.22 bis modems became the state of the art. These modems, and their earlier counterparts, used frequency division multiplexing (FDM) to achieve full-duplex communications.

In 1984, when echo cancellation and advanced coding algorithms were introduced, it became possible to support 9.6 kb/s (V.32) over the PSTN. During the remainder of the decade, improvements in the telephone network and in modem technology made speeds of 14.4 kb/s (V.32 bis) practical. In the early 1990s, as a result of the migration to digital networks and the continued sophistication of modulation and coding algorithms, it became possible to communicate at speeds up to 28.8 kb/s (V.34). In fact, this rate was recently extended to 33.6 kb/s by newer and better technology that has been standardized by the ITU SG 14. During the next few years, modem rates will likely increase to about 48 kb/s.

Standards activity over the PSTN. For the last several years, standards bodies, particularly the ITU, have been developing multimedia standards for the PSTN. This work has been driven largely by the availability of technology for individual components such as modems, speech coders, and video coders. The principal activities are shown in Table IV.

The evolving ITU recommendations for PSTN-based multimedia consist of the following elements:
- A V.34 modem (speeds up to 33.6 kb/s),
- A low-bit-rate speech coder at 8 kb/s for 3-kHz bandwidth (and ultimately 16 kb/s for a 7-kHz bandwidth) (Draft Recommendations G.723, G.729, or G.svd),
- A multiplexer (Draft Recommendation H.324 or V.42-based link access procedure modems [LAPMs]),
- Video coding (Draft Recommendation H.326),
- A control channel (Draft Recommendation H.246), and
- Capabilities exchange and selection (Draft Recommendation V.8 bis).

Of course, other standards—such as the Joint Photographic Experts Group (JPEG) for still image and T.30 for facsimile—will be used where applicable, and not all recommendations will be appropriate for every product. In addition, recommendations are being developed for specialized applications. For example, ITU-T SG 14 recently approved V.asvd, a simultaneous voice and data scheme proposed by AT&T, which provides low-delay, communications quality voice, along with a moderate data channel. This recommendation is expected to be used widely in applications like network-based games or telecommunications devices for the hearing impaired.

The availability of international standards for all individual components will facilitate interworking among different manufacturers' products. This factor alone will help accelerate the deployment pace of PSTN-based multimedia products.

Industry activity over the PSTN. During the last eighteen months, interest within the industry has been heightened as companies position themselves to develop and deploy PSTN-based multimedia products. Much of this interest is a result of the anticipated growth in this area and the broad range of applications envisioned. Examples of much discussed applications include:
- Audiographic teleconferencing,
- Collaborative computing,
- Shared whiteboards,

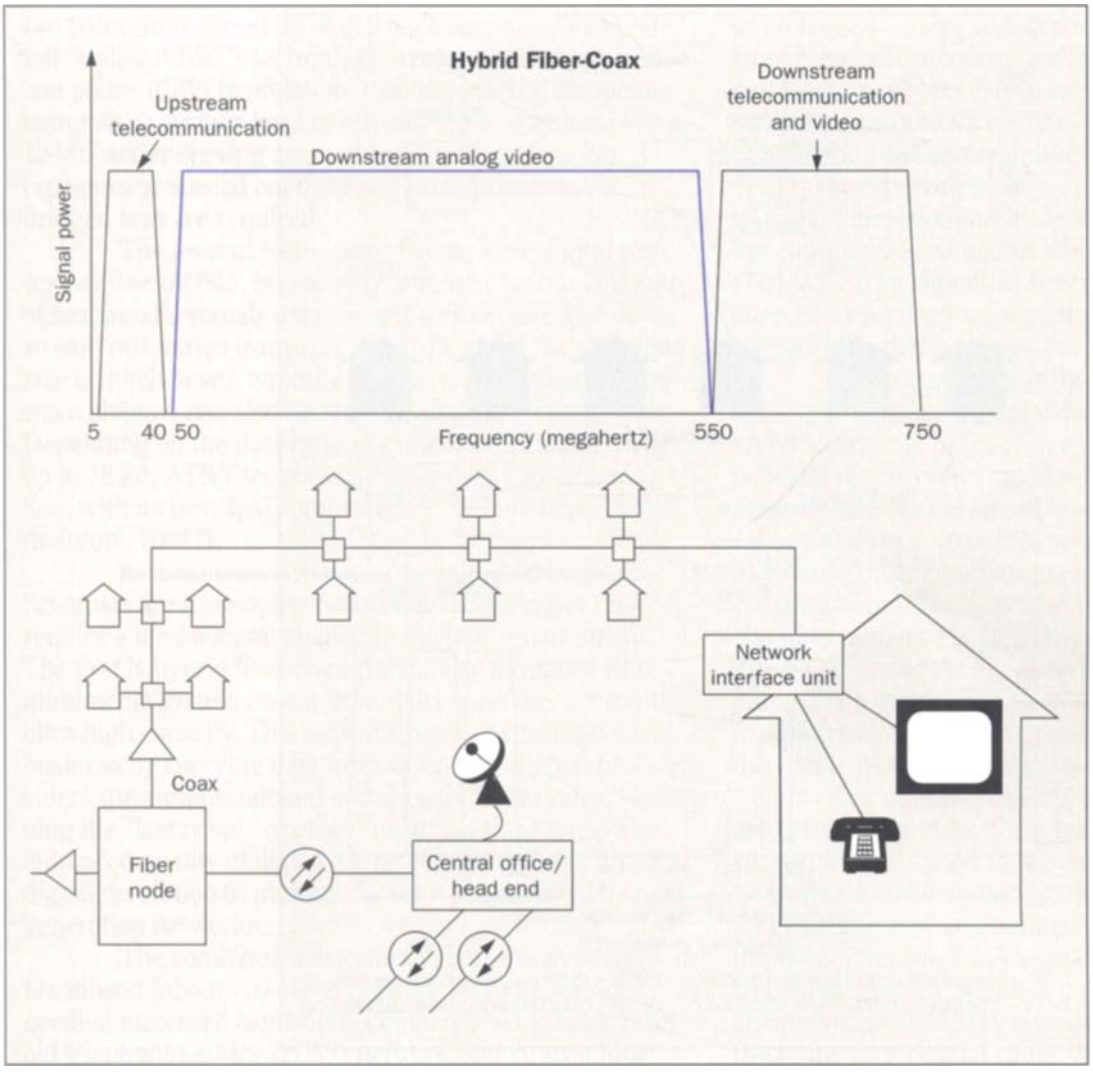

Figure 4. Spectrum utilization in the fiber-coaxial network.

- Remote presentations/telelearning,
- Internet telephony,
- Videotelephony,
- Technical support and customer service,
- Interactive games and transactions,
- Telecommuting, and
- Teleshopping, electronic catalogs.

New applications are being created daily, and companies are anxious to participate in what promises to be an active and rapidly evolving area. Multimedia over the PSTN will potentially be a multibillion dollar industry. Because several standards will be completed and interoperable products will start to appear, 1995 will be viewed as a critical year. Industry partnerships, particularly among computing and communications companies, will be an important factor in accelerating the pace of deployment. With its basic technology (modems, speech coding, image coding, and video processing) for products, as well as the ability to offer network services, AT&T is well positioned in this market.

ISDN: *n* x 64 kb/s. With digital connectivity at *n* x 64 kb/s (*n* = 1 to 24), ISDNs, part of the PSTN, will be tremendously important in the near future for deploying multimedia products and services. In the next five years,

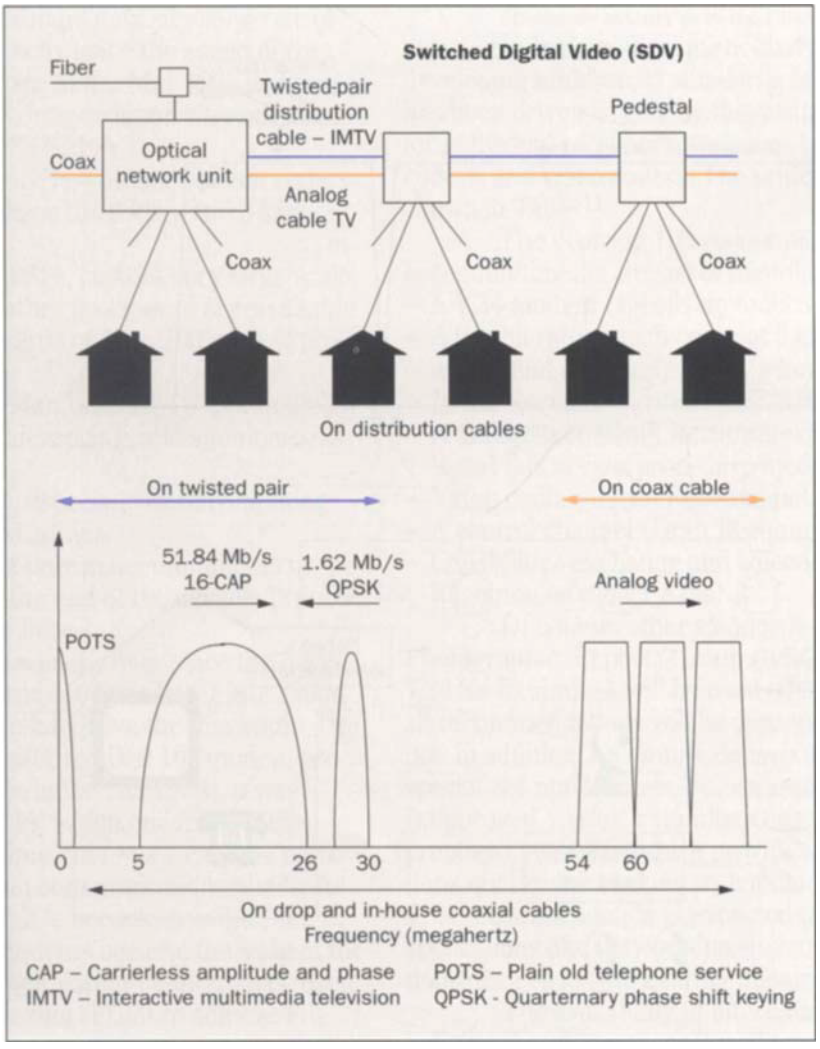

Figure 5. Spectra and bandwidth allocation for on-drop and in-house coaxial cables.

ISDN is the only ubiquitous communication fabric likely to be capable of supplying connectivity over 33.6 kb/s. A number of important audio and video coding standards have been designed to use ISDN. Outside the U.S., ISDN deployment is extensive; within the U.S., it is becoming popular as a physical layer for LAN and Internet extension. The *primary rate* ISDN channel (1.536 Mb/s) has been a nice match to audiovisual services such as high-quality videoconferencing, while the *basic rate* ISDN channel (2 x 64 + 16 = 144 kb/s, or the 64 kb/s part of it) has traditionally supported high-quality telephony and audioconferencing. Recent advances in signal compression (CD-like stereo, or high-quality videotelephony at 64 kb/s) will create major new applications in multimedia networking on a single 64-kb/s channel.

Multimedia Signals (1.5 to 7 Mb/s) over the Local Loop. Two technologies that are extensions of ISDN have been developed. The first is the high-speed digital subscriber line (HDSL), intended to carry T1 or E1 rate multimedia signals over the subscriber line between a customer's premises and a telephone company central office or remote terminal. The conservative approach is to use

two pairs (four wires), in which each carries half the rate, full duplex. AT&T's technology is carrierless amplitude and phase (CAP) modulation, with the leading competitor being 2B1Q, or four-level baseband. HDSL operates over a 12-kft carrier serving area using 24-gauge wires. No repeaters or special conditioning such as removal of bridged taps are required.

The second technology, asymmetric digital subscriber line (ADSL), is primarily intended for transmission of multimedia signals over the subscriber line. The downstream rate varies from 1.5 to 7 Mb/s, while the upstream rate is much lower, typically 160 kb/s. A standard analog voice channel can also be carried, all within a single pair. Depending on the data rate, any unloaded pair can carry up to 18 kft. AT&T technology again uses CAP modulation, with its principal competing technology being digital multitone (DMT).

Next-Generation Networks for Multimedia Signals. Scientists are developing two newer technologies that require a fundamental change in the loop environment. The first is hybrid fiber-coax (HFC). The proposed HFC multimedia systems use a "fiber-backbone" network with ultra-high capacity. This network reaches residences and business by carrying data from video information providers through broadband switches to "fiber nodes," spanning the "last couple of miles" using coaxial cables. The enhanced quality of digital TV displays brings the ongoing digital revolution to multimedia services provided by next-generation networks.

The combined networks will become an integrated broadband (about 750-MHz) network for most services needing increased bandwidth, compared to the older plain old telephone service (POTS) network. Ample switching and software control in the newer network will permit symmetric and asymmetric flow of network information. Digital and programmed control (using stored program control in the switching systems) will permit a mixture of customized services whose quality and reliability are high. The fiber-coax network delivers the impact of digital technology, which carries a broader spectrum of multimedia services to the household than plain old cable TV (CATV) services can bring to the public.

Figure 4 shows the spectrum utilization of the combined fiber-coaxial network. This technology carries signals from a head end to a remote terminal over fiber, and from the remote terminal to a small area—as many as 50 homes—using a standard coaxial network. The signals include the usual analog TV signals in their normal frequency slots. Downstream digital signals are typically carried in the frequency band above analog TV. A combination of FDM and time division multiplexing (TDM) provides the broadcasts to the users. In the upstream direction, multiple access techniques, including combined FDM and time division multiple access (TDMA), carry signals in the 5- to 40-MHz band. Our current technology uses quaternary phase shift keying (QPSK) in both directions.

The second new technology is *fiber to the curb* (FTTC) or *switched digital video* (SDV). In this network architecture, the optical fiber goes to a curbside pedestal that serves a small number of homes. At the pedestal, the optical signal is converted into an electrical one and then demultiplexed for delivery to individual homes on copper wiring. Multiplexing and signal conversion functions are also performed for the opposite direction, that is, from the homes to the network. The FTTC system being described uses existing telephone drop wiring or coaxial cable to provide local distribution to interactive multimedia television (IMTV) to the home, as shown in Figure 5.

The downstream and upstream IMTV signals are carried to and from the home on the telephone drop wiring, which is used to provide POTS. This wiring is typically twisted-pair wiring. Standard analog broadcast CATV is delivered to the home on coaxial cable. Inside the home, the two IMTV signals are carried on the coaxial cabling system, which is used for CATV. In another arrangement, the IMTV signals are carried to and from the home on a coaxial cable, which may or may not be the cable that is used for CATV. Typically, the downstream channel operates at the synchronous transfer signal (STS)-1 data rate of 51.84 Mb/s, and the upstream channel from the home operates at a data rate of 1.62 Mb/s. Both channels may carry ATM cells, and the downstream channel uses synchronous optical network (SONET) framing. Figure 5 shows the spectra and bandwidth allocation. With spectral compression technology for digitized video, the downstream data rate is large enough to accommodate two interactive HDTV channels, a dozen video channels with National Television Systems Committee (NTSC) quality, or thirty video channels with VHS quality.

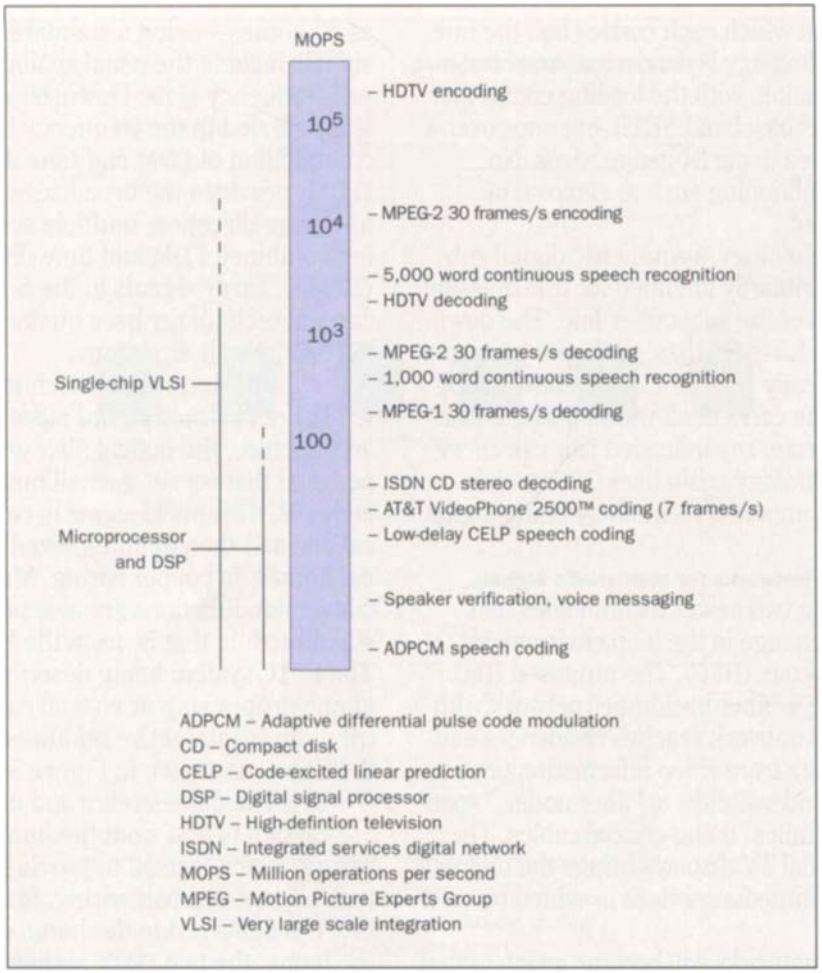

Figure 6. Processing requirements for some speech and image processing applications.

The transmission scheme used for the downstream channel is CAP or quadrature amplitude modulation (QAM); QPSK is used for the upstream channel. A multiple access scheme is also provided in the upstream direction to allow several set-top terminals to be used at the same time in a given house.

Customer Premises. The most widely used systems for transporting voice and data in the customer premises have been combinations of PBXs and LANs. The PBXs provide ISDN connectivity, and the most popular LAN types—Ethernet and token rings—deliver 10- to16-Mb/s shared by several users.

With the advent of multimedia, it became necessary to develop higher-bandwidth technologies. Two higher-speed Ethernet-like approaches, 100 BASE-T and 100 BASE-VG, have been developed to provide relief for overloaded Ethernet segments. These systems require significant changes to the existing equipment, such as adding PC adapter cards and software drivers. They have also had to struggle with the violation of Federal Communications Commission (FCC) electromagnetic interference (EMI) compliance for networks running at transmission speeds of 100 Mb/s with spectra above the 30-MHz limit. For this reason, 4-pair unshielded twisted-pair Category 3 cables are run to each desktop computer. With increased speeds of up to 100 Mb/s, the unpredictable nature of their multiple access schemes limits these two schemes.

The fiber distributed data interface (FDDI) technology, also a shared-media-type LAN that runs at 100

Mb/s, is similar to token ring technology. It is a token-passing technology with a second ring for backup. FDDI II is also a 100-Mb/s technology; it emerged as ATM technology started to become popular. The attraction of FDDI II was its purported isochronous capability and its ability to deliver voice, video, and data. Unfortunately, this has not been very successful.

The last important LAN technology is ATM. It runs at several speeds that follow the optical carrier standard. Currently, the most popular speeds for ATM backbone applications are 100 Mb/s, which uses the same physical layer as FDDI, 155-Mb/s optical carrier (OC)-3c SONET framing, and 622-Mb/s OC-12 SONET framing. Other important ATM speeds to computer desktops are 51-Mb/s CAP 16 and 25.6-Mb/s nonreturn to zero (NRZ).

Digital Signal Processing and Computing

Many of the audiovisual and communication signal processing techniques described earlier require large amounts of complex computation that must be performed in real time. This real-time requirement sets multimedia processing apart from standard off-line computing tasks. If insufficient computing power is available, a word processor or spreadsheet application will simply run slowly. However, if the computing engine in a multimedia application that includes real-time communication cannot keep up with the required audio or video processing rate, it will not work at all. For example, efficient video compression algorithms have been available for decades, but only recently have we seen real products based on compressed video. Only in the last few years has submicron VLSI advanced to the point where real-time video compression and decompression can be performed in a few silicon chips costing less than $100.

Figure 6 shows the processing requirements for a number of real-time multimedia processing tasks that require high-performance VLSI for their implementation. Some of these, such as adaptive differential pulse code modulation (ADPCM) speech coding, however, require less than ten million operations per second (MOPS) and can be easily implemented in software on today's microprocessors. Others, such as Motion Picture Experts Group (MPEG) video encoding, require hundreds of MOPS and special-purpose VLSI for their operation. Still others, such as HDTV encoding, require more than 10^{11} operations per second and can presently only be performed with large racks of special-purpose hardware.

This wide range in performance, combined with varying price and flexibility requirements, leads to a wide variety of VLSI platforms on which multimedia signal processing operations can be performed. One key choice that must be made is between programmable and dedicated architectures. Programmable designs offer flexibility at the cost of increased silicon area and power for equivalent levels of performance. Dedicated architectures are most efficient in high-performance applications, but they cannot be modified to capture algorithmic enhancements or reused for other real-time processing tasks. VLSI platforms for multimedia processing tend to fall into one of three generic types: general-purpose microprocessors, dedicated VLSI processors, and general-purpose DSPs.

General-Purpose Microprocessors. Because of the increasing density of silicon VLSI, applications that once required dedicated silicon can now run in software on today's general-purpose reduced instruction set computer (RISC) processors. A 75-MHz Pentium* microprocessor, for example, can provide peak processing rates of more than 120 MOPS and, for some applications, can sustain average processing rates of 40 to 60 MOPS. This makes it suitable for many audio and speech-related tasks, some modem applications, and a limited set of video processing operations. A Pentium-based computer can:

- Decode a common intermediate format (CIF)-sized (Table II) image at frame rates up to about 10 frames per second,
- Decode a CD-quality stereo signal that has been compressed by a perceptual audio coder,
- Easily synthesize multiple channels of speech from text, as in a multilingual transaction, and
- Perform simple forms of automatic speech recognition, such as speaker-independent voice dialing with a menu of about 50 names.

Microprocessor-based implementations have a number of advantages. If the microprocessor is already part of the multimedia platform, such as a PC or a workstation, then the hardware is free and the only investment needed is for software. This kind of *native signal processing* (NSP) creates low-cost solutions that are simple to distribute and upgrade. Improvements in signal processing algorithms and modifications to

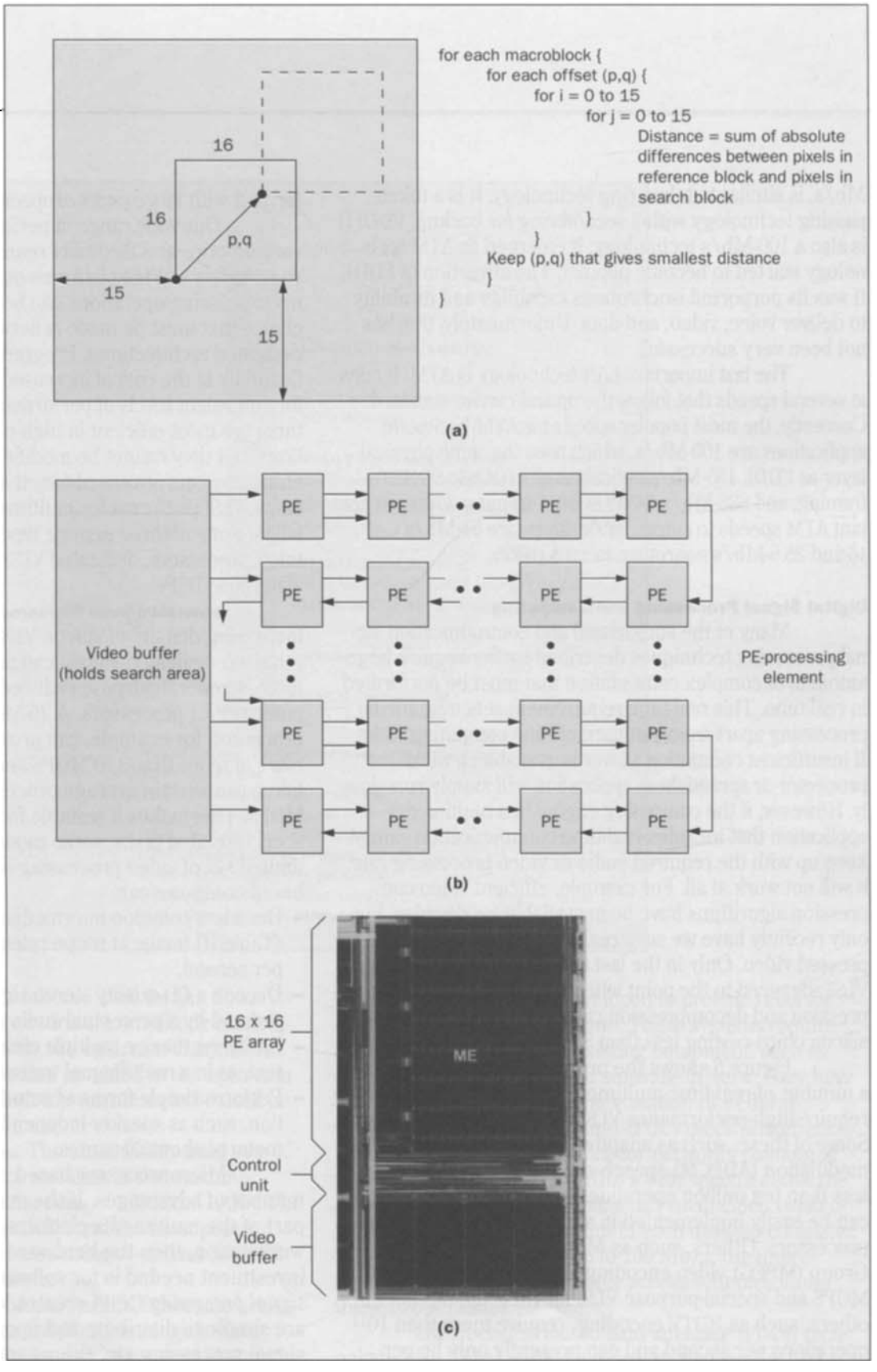

Figure 7. Steps in the design of the AVP 4310 video encoder: mapping from algorithm to regular architecture to custom layout. (a) Algorithm for finding the search block in the previous frame that best matches the reference block in the current frame; (b) architecture comprising 16x16 array of processing elements (PEs) that store reference block pixels and perform computation; and (c) custom layout showing the buffer, PEs, and the control unit.

existing standards are easily incorporated. Also, new generations of microprocessor hardware provide increased real-time performance with no additional hardware modifications.

There are also a number of difficulties associated with "software-only" implementations of these algorithms. Today's microprocessors have been developed primarily to provide high average throughput in traditional computing tasks. Structures like complex multilevel cache (memory) hierarchies give excellent average performance, but they can guarantee only modest performance in a real-time communication application. A cache miss, for example, on a Pentium or PowerPC* processor can cost 10 to 25 clock cycles, instantaneously reducing the performance of the machine to less than 10 MOPS. These processors also associate a large amount of *context*—information about the current state of the process—with each process. In a real-time multitasking application, frequent interrupts and task swapping can degrade performance further.

Another difficulty arises if the microprocessor is being used by some other application while multimedia communication is in operation. The "free cycles" argument assumes that the video or audio processing operation is the only application currently requiring processor cycles. In a complex multimedia application in which communications, data formatting, graphics, and/or number crunching are competing for the same cycles, the situation changes dramatically. Rather than being "free," the operations being performed by the central processing unit (CPU) become expensive compared with more focused hardware implementations.

Dedicated VLSI Architectures. At the other end of the performance/flexibility spectrum are VLSI processors, whose architecture is finely tuned to perform a specific signal processing application. Many DSP algorithms can be described as a set of dataflow operations in which a small number of basic operations are applied to a large body of data in a data-independent, repetitious fashion. These algorithms lend themselves to architectures in which high throughput can be obtained using a combination of data path regularity, specialized arithmetic processors, parallelism, and pipelining. These architectures, in turn, are well suited to custom VLSI implementation in which the layout is structured to capture the necessary data and control flow directly.

Dedicated architectures can provide enormous computing power in a very small silicon area with a relatively low power budget. In the AVP 4310 video encoder, for example, motion estimation is performed by a 16x16 processing element array that provides over 5,000 MOPS in just 35 mm^2. Power dissipation for this unit is a little more than one watt. Figure 7 shows the mapping from regular algorithm to regular parallel architecture and from there to regular layout. Figure 7a shows the algorithm for motion estimation (finding the block in previous frame that best matches a reference block in the current frame). Figure 7b shows an architecture comprising a16x16 array of processing elements that store reference block pixels and perform computation. Figure 7c is the custom layout, showing the buffer, the processing elements (PEs), and the control unit.

Another example is the Net32K chip, a neural net processor used for performing image analysis and segmentation in applications involving document capture and interpretation. This chip provides over 100,000 MOPS (2-bit multiply accumulates) in only 28 mm^2 and dissipates 0.5 watt.

Dedicated VLSI is particularly appropriate for applications in which performance and cost are the driving constraints. An HDTV receiver is a good example. The efficiency of a dedicated solution, however, is offset by the fact that each processor is only capable of performing a very limited range of operations. The AVP motion estimator, for example, is not suitable for high-end image or speech recognition processing. System cost can quickly grow if a special-purpose VLSI processor is required for each new media operation. Some kind of compromise between performance, cost, and reuse is often required.

General-Purpose DSPs. For many applications, general-purpose DSPs provide just such a compromise. DSPs are programmable processors whose architecture and instruction set are optimized to efficiently perform the inner loops of signal processing algorithms. They became popular in the late 1970s as a low-cost means of providing high throughput for speech and communication processing applications—applications in which multiply-add is the dominant arithmetic operation. DSP data paths typically contain heavily pipelined multiplier/accumulators with auto-indexing pointers that keep the pipeline full when implementing opera-

tions such as those found in finite impulse response (FIR) filters, matrix multiplies, and convolution. Silicon die size and, hence, cost is kept low by providing modest instruction decoding and memory support. A $20 DSP chip can typically provide the same arithmetic processing power as a $200 general-purpose microprocessor.

Although DSPs can be programmed to implement a wide range of signal processing tasks, they lack the general-purpose flexibility of a microprocessor. Their instruction sets are poorly matched to applications involving complex data structures, symbolic manipulation, or text processing. Their simple memory architectures rely on expensive high-speed static random-access memory (SRAM), making them unsuitable for applications requiring large amounts of instruction or data memory. They also usually come with minimal compiler support and require a high level of programming skill.

Multimedia applications, particularly those based on video processing, demand much higher levels (hundreds of MOPS) of performance from DSPs. These performance levels can only be obtained through extensive use of parallelism and pipelining. A new class of DSP architectures is emerging in which multiple processors execute separate parallel threads of computation. These are sometimes called multiple instruction–multiple data (MIMD) architectures. One example is the Texas Instruments Multimedia Video Processor (MVP) chip. It consists of a 32-bit master processor and four 64-bit integer DSP cores connected together with a crossbar switch. Split-word instructions allow as many as four operations per processor per clock cycle. At 50 MHz, this single-chip processor provides over 800 million programmable operations per second. These devices are, at present, too expensive to compete effectively with dedicated solutions and too difficult to program to compete effectively with microprocessors. As silicon costs go down, however, and software support for parallel programming improves, these devices will likely provide a very attractive solution to the need for both high performance and flexibility.

Summary

The technologies and tools discussed in earlier sections of this paper form some of the cornerstones of multimedia products and services. Several classes of emerging applications require additional tools, especially user interfaces and application software. Unlike the specific technologies described earlier, the performances of integrated systems are much harder to quantify and measure. However, as several of the papers that follow illustrate, the one-dimensional technologies of the past (such as telephone and facsimile) are being rapidly supplemented by multimedia services involving several signal domains. The integration of multiple modalities has already begun to produce significant, if hard-to-measure, improvements in the quality of service and in the impact of the user interface.

Acknowledgment

The authors wish to thank Dr. Jialin Zhong for the picture of the talking face shown in Figure 2, and Drs. Chinmoy Bose, Oscar Agazzi, and Jianying Hu for the document recognition examples shown in Figure 3.

(Manuscript approved August 1995)

*Trademarks

Pentium is a registered trademark of Intel Corporation.
PowerPC is a registered trademark of International Business Machines, Inc.

Bryan D. Ackland *is head of the VLSI Systems Research Department at AT&T Bell Laboratories in Holmdel, New Jersey. He is responsible for research on VLSI architectures and circuits for high-performance signal processing applications. Mr. Ackland joined AT&T in 1978, after receiving a B.Sc. in physics and a B.E. and Ph.D. in electrical engineering, all from the University of Adelaide, South Australia.*

Nikil Jayant *is head of the Signal Processing Research and Advanced Audio Technology Departments at AT&T Bell Laboratories in Murray Hill, New Jersey. He is responsible for research on audiovisual communications and, in particular, digital audio broadcasting and videotelephony over voiceband modems and wireless multimedia networks. Mr. Jayant received a B.S. in physics and*

mathematics from Mysore University in India, as well as B.E. and Ph.D. degrees in electrical communications from the Indian Institute of Science in Bangalore. He joined AT&T in 1968.

***Victor B. Lawrence** is head of the Advanced Multimedia Communications Department at AT&T Bell Laboratories in Middletown, New Jersey. He is responsible for exploratory development, systems engineering, and human factors for multimedia products and services. Mr. Lawrence joined AT&T in 1974, after receiving a B.Sc. in electrical engineering from London University, England; a D.I.C. from the Imperial College of Science and Technology, London, England; and a Ph.D. in electrical engineering from London University.*

***Lawrence R. Rabiner** is director of the Information Principles Research Laboratory at AT&T Bell Laboratories in Murray Hill, New Jersey. He leads research efforts in several key areas of information sciences: speech and image processing, interactive systems (including handwriting recognition), digital signal processing, and communications. Mr. Rabiner received B.S., M.S., and Ph.D. degrees in electrical engineering, all from the Massachusetts Institute of Technology in Cambridge. He joined AT&T in 1962.*