On the Application of Multimedia Processing to Telecommunications

Richard Cox, Barry Haskell, Yann LeCun, Behzad Shahraray, and Lawrence Rabiner
Speech and Image Processing Services Research Lab
AT&T Labs-Research, Florham Park/Newman Springs, New Jersey, USA

Abstract

The challenge of multimedia processing is to seamlessly integrate text, sound, image, and video information into a single communications channel, and to do it in a way that provides high quality communications while preserving the ease-of-use and interactivity of conventional telephony. There are a number of technology drivers that are pushing the technology forward, as well as a number of technological problems that must be overcome before multimedia becomes as ubiquitous as voiceband telephony. A key issue with any practical multimedia system has to do with standards that insure connectivity between customers and a range of service providers. Multimedia processing is an area of communications that is rapidly evolving. However, a number of interesting and important multimedia communications applications have evolved over the past several years, and some of these applications will be described in this paper.

1. Introduction

In a very real sense, virtually every individual has had experience with multimedia systems of one type or another. Perhaps the most common multimedia experiences are reading the daily newspaper or watching television. For most of us, when we think about multimedia and the promise for future communications systems, we tend to think about systems that combine video, graphics, animation with special effects (as seen in movies like 'Who Framed Roger Rabbit') and CD quality audio. On a more business oriented scale, we think about creating virtual meeting rooms with 3-dimensional realism in sight and sound, including sharing of whiteboards, computer applications, and perhaps even computer-generated Business Meeting Notes for documenting the meeting in an efficient communications format. Other glamorous applications of multimedia processing include Distance Learning in which we learn and interact with instructors remotely over a broadband communication network, Virtual Library Access in which we instantly have access to all of the published material in the world, in its original form and format, and can browse, display, print, even modify the material instantaneously, and Living Books which supplement the written word and the associated pictures with animations, and hyperlink access to supplementary material.

1.1 The Multimedia Communications Revolution

Modern voice communications networks evolved around the turn of the twentieth century with a focus on creating *Universal Service*, namely the ability to automatically connect any telephone user with any other telephone user, without the need for operator assistance or intervention. This revolutionary goal defined a series of technological problems that had to be solved before the vision became reality, including the invention of the vacuum tube for amplification of telephone signals, mechanical switching to replace the operator consoles that were used in most localities, numbering plans and signaling systems to route calls, etc. The first transcontinental call in the United States was completed in 1915, thereby ushering in the 'modern age of voice communications', an age that has been developed and improved upon for the past 80 or so years.

We are now in the midst of another revolution in communications, one which holds the promise of providing ubiquitous service in multimedia communications. The vision for this revolution is to provide seamless, easy-to-use, high quality, affordable multimedia communications between people and machines, anywhere, and anytime. There are three key aspects of the vision which characterize the changes that will occur in communications once this vision is achieved, namely:

- the basic currency of communications evolves from narrowband voice telephony to seamlessly integrated, high quality, broadband, transmission of multimedia signals;
- the basic access method changes from wireline connections to combinations of wired and wireless, including cable, fiber, cell sites, satellite, and even electrical power lines;
- the basic mode of communications expands from primarily involving people-to-people communications, to include people-to-machine communications.

There are a number of forces that are driving this multimedia revolution, including:

- the evolution of communications networks and data networks into today's modern POTS (Plain Old Telephone Services) network and Packet (including the Internet) networks, with major forces driving these two networks into an integrated structure;
- the increasing availability of (almost unlimited) bandwidth on demand in the office, the home, and eventually on the road, based on the proliferation of high speed data modems, cable modems, hybrid fiber-coax systems, and recently a number of fixed wireless access systems;
- the availability of ubiquitous access to the network via LANs, wireline, and wireless networks providing the promise of anywhere, anytime access;
- the ever increasing amount of memory and computation that can be brought to bear on virtually any communications or computing system—based on Moore's law of doubling the communications and memory capacity of chips every 18 or so months;
- the proliferation of smart terminals, including sophisticated screen phones, digital telephones, multimedia PC's that handle a wide range of text, image, audio, and video signals, "Network Computers" and other low-cost Internet access terminals, and PDAs (Personal Digital Assistants) of all types that are able to access and interact with the network via wired and wireless connections;
- the digitization of virtually all devices including cameras, video capture devices, video playback devices, handwriting terminals, sound capture devices, etc., fueled by the rapid and widespread growth of digital signal processing architectures and algorithms, along with associated standards for plug-and-play as well as interconnection and communications between these digital devices.

1.2 Technology Aspects of Multimedia Systems

In order for multimedia systems to achieve the vision of the current communications revolution, and become available to everyone, much as POTS service is now available to all telephony customers, a number of technological issues must be addressed and put into a framework that leads to seamless integration, ease-of-use, and high quality outputs. Among the issues that must be addressed are the following:

- the basic techniques for compressing and coding the various media that constitute the multimedia signal, including the signal processing algorithms, the associated standards, and the issues involved with transmission of these media in real communications systems;
- the basic techniques for organizing, storing, and retrieving multimedia signals, including both downloading and streaming techniques, layering of signals to match characteristics of the network and the display terminal, and issues involved with defining a basic Quality-of-Service (QOS) for the multimedia signal and its constituent components
- the basic techniques for accessing the multimedia signals by providing tools that match the user to the machine, such as by using 'natural' spoken language queries, through the use of media conversion tools to convert between media, through the use of agents that monitor the multimedia sessions and provide assistance in all phases of access and utilization;
- the basic techniques for searching in order to find multimedia sources that provide the desired information or material—these searching methods, which in essence are based on machine intelligence, provide the interface between the network and the human user, and provide methods for searching via text requests, image matching methods, and speech queries;
- the basic techniques for browsing individual multimedia documents and libraries in order to take advantage of human intelligence to find desired material via text browsing, indexed image browsing, and voice browsing.

2. Compression of Multimedia Signals

Table 1 shows the signal characteristics and the resulting uncompressed bit rate necessary to support the storage and transmission of speech, audio, image, and video signals with high quality. The table has separate sections for each of these signals, since their characteristics are very different in terms of frequency range of interest, sampling grids, etc.

It can be seen that from Table 1a, that for narrowband speech, a bit rate of 128 kb/s is required without any form of coding or compression—i.e., twice the rate used in ordinary POTS telephony. For wideband speech, a bit rate of 256 kb/s is required for the uncompressed signal, and for 2-channel stereo quality CD (Compact Disk) audio, a bit rate of 1.41 Mb/s is required for the uncompressed signal. We will see later in this section that narrowband speech can be compressed to about 4 kb/s (a 30-to-1 compression rate), wideband speech can be compressed to about 16 kb/s (a 15-to-1 compression rate) and CD audio can be compressed to 64 kb/s (a 22-to-1 compression rate) while still preserving the quality of the original signal.

Table 1b shows the uncompressed size needed for bilevel (FAX) and color still images. It can be seen that an ordinary FAX of an 8 ½ by 11 inch document, scanned at 200 dpi (dots per inch), has an uncompressed size of 3.74 Mb, whereas color images (displayed on a computer screen) at VGA resolution require 2.46 Mb, and high resolution XVGA color images require 18.87 Mb for the uncompressed image. It will be shown that most images can be compressed by factors on the order of 100-to-1 (especially text-based FAX documents) without any significant loss in quality.

Speech/Audio Type	Frequency Range	Sampling Rate	Bits/Sample	Uncompressed Bit rate
Narrowband	200-3200 Hz	8 kHz	16	128 kb/s
Wideband	50-7000 Hz	16 kHz	16	256 kb/s
CD Audio	20-20000 Hz	44.1 kHz	16 x 2 channels	1.41 Mb/s

a

Image Type	Pixels per Frame	Bits/Pixel	Uncompressed Size
FAX	1700 x 2200	1	3.74 Mb
VGA	640 x 480	8	2.46 Mb
XVGA	1024 x 768	24	18.87 Mb

b

Video Type	Pixels per Frame	Image Aspect Ratio	Frames per Second	Bits/Pixel	Uncompressed Bit rate
NTSC	480 x 483	4:3	29.97	16	111.2 Mb/s
PAL	576 x 576	4:3	25	16	132.7 Mb/s
CIF	352 x 288	4:3	14.98	12	18.2 Mb/s
QCIF	176 x 144	4:3	9.99	12	3.0 Mb/s
HDTV	1280 x 720	16:9	59.94	12	622.9 Mb/s
HDTV	1920 x 1080	16:9	29.97	12	745.7 Mb/s

c

Table 1 Characteristics and uncompressed bit rates of speech, audio, image and video signals.

Finally, Table 1c shows the necessary bit rates for several video types. For standard television, including the North American NTSC standard and the European PAL standard, the uncompressed bit rates are 111.2 Mb/s (NTSC) and 132.7 Mb/s (PAL). For videoconferencing and videophone applications, smaller format pictures with lower frame rates are standard, leading to the CIF (Common Intermediate Format) and QCIF (Quarter CIF) standards, which have uncompressed bit rates of 18.2 Mb/s and 3.0 Mb/s, respectively. Finally, the digital standard for HDTV (in two standard formats) has requirements for an uncompressed bit rate of between 662.9 and 745.7 Mb/s.

3. Standards-Based Compression

Over the past two decades, speech and audio coding standards have evolved for network, cellular, and secure telephony applications. Such standards fall into two categories, namely waveform coding, and model-based coding methods. Among the most popular waveform coding methods include:

- PCM (G.711)-pulse code modulation
- ADPCM (G.726, G.727)-adaptive, differential PCM
- Wideband coder (G.722)-2 band ADPCM for 7 kHz bandwidth speech

Among the most popular model-based coding methods include:

- LD-CELP (G.728)-low delay code-excited linear prediction coding
- CS-ACELP (G.729)-conjugate structure, algebraic CELP
- MPC-MLQ (G.723.1)-multi-pulse coding, maximum likelihood quantization
- VSELP (IS-54)-vector sum excitation linear prediction

Coding standards have evolved for FAX including:

- Group 3 and Group 4 FAX for run length coding
- JBIG-1 for pixel prediction based on local neighborhoods
- JBIG-2 for soft pattern matching on segmented regions

Still image coding standards include:

- JPEG for DCT processing, perceptual quantization, and entropy encoding
- JPEG-2000 as a modern multimedia architecture with downloadable software

Finally, video standards include:

- H.261, H.262 and H.263 (p x 64) with motion compensation for interframe coding

- MPEG-1 with specifications for coding, compression and transmission of audio, video, and data in packets
- MPEG-2 with capability of handling multi-channel, multimedia signals over broadband networks
- MPEG-4 an object-based approach to multimedia with independent coding of objects, interactive composition of objects, and the ability to integrate synthetic and natural objects
- MPEG-7 which adds the capability for searching, indexing, and authentication of large databases of multimedia objects.

4. Multimedia Systems

FusionNet Service: A key problem in the delivery of 'on-demand' multimedia communications over the Internet, based on POTS or ISDN access, is that the Internet today cannot guarantee the quality of real-time signals such as speech, audio, and video. The FusionNet service overcomes this problem by using the Internet only to browse and request the video and audio, as well as to control the signal delivery (e.g., via VCR-like controls). FusionNet uses either POTS or ISDN to actually deliver guaranteed Quality of Service (QOS) for real-time transmission of audio and video.

Initial implementations of FusionNet Service required the user to maintain either two POTS lines (one for Internet access, one for the guaranteed QOS audio/video link), or a full ISDN connection (i.e., 2 or more B-channels each with 64 kb/s). The most recent implementation provides the FusionNet Service over a single ISDN B-channel by requiring the ISP (Internet Service Provider) to provide ISDN access equipment that seamlessly merges the guaranteed QOS audio/video signal with normal Internet traffic to and from the user via PPP (Point-to-Point Protocol) over dialed-up ISDN connections. Unless the traffic at the local ISP is very high, this method provides high quality FusionNet Service with a single ISDN B channel. Of course, additional channels can always be ganged together for higher quality service.

Cybrary-the Virtual Library: A key aspect in delivering high quality multimedia is the ability to view printed material in its original (uncompressed) form. This requires a capability of compressing, storing, indexing, and browsing documents stored anywhere. We call a system that provides these capabilities a cyber-library or Cybrary. Such a system essentially allows for virtual presence in a remote archive.

The Cybrary system we have created lets anyone connected to the Internet view any document in the library on any available screen. The key technological innovation which provides this capability is a new standard for document image compression (JBIG2), which makes possible quick page-flipping and browsing. Through the use of advanced OCR (Optical Character Recognition) techniques for translating image text into ASCII characters, the Cybrary system provides full text search, indexing, browsing, and hyperlinking.

Pictorial Transcripts System: A key challenge in multimedia processing is to provide a compact representation of full motion video that can be indexed, stored efficiently, and displayed as a reference or archive. The Pictorial Transcripts System is one proposed solution to this challenge. Essentially the Pictorial Transcripts System provides a complete, albeit condensed, representation of a full motion video sequence, consisting of a carefully selected set of still images matched to (synchronized with) the text version of the audio. The technology has commercial applications in broadcast TV, where a network can automatically convert a closed-captioned broadcast into a Website program in real time, and for business applications where the system could create 'Business Meeting Notes' of meetings, seminars, or conferences, and make these available on the Internet within minutes after the meetings ended.

Pictorial Transcripts automatically analyzes, condenses, and indexes multimedia information from closed captioned video broadcasts and generates web content in real time. The application thus allows selective retrieval of program content, since Pictorial Transcripts generates text and pictorial indices into video and multimedia images. The technical challenges here include finding ways to combine video, speech, and text processing and compression technologies, as well as perform linguistic analyses, computer programming, and systems engineering to index selected video information and transmit it over phone lines.

Using efficient, high-performance algorithms reduces the amount of storage on a one-hour news broadcast from as much as 1 gigabyte to about 1.5 megabytes—i.e., about 1000 to 1 compression ratio. This means that an entire news program can be saved on a single 3 ½" floppy disk.

5. Summary

Multimedia processing is defined as the multiplexing and combining of any number of data streams, where the data streams can represent real-time signals (with their concomitant need for some type of guaranteed Quality of Service), data signals, control signals, conformance testing signals, etc. This has led to the evolution of modern multimedia systems that we believe will become commonplace and widely used in the future.