CHAPTER 1

Multimedia Communications

Chapter Overview

The challenge of multimedia communications is to provide services that integrate text, sound, image and video information and to do it in a way that preserves the ease of use and interactivity. At first, the concept of a multimedia communication modeling is described. We present a brief description of elements for multimedia systems. After that, user and network requirements are discussed,. together with the packet transfer concept. Taking into account that Asynchronous Transfer Mode (ATM) uses a fixed-length packet and is suitable for high-speed applications, we describe multimedia requirements. Finally, we give an overview of multimedia terminals.

1.1 Introduction

Multimedia communications is the field referring to the representation, storage, retrieval and dissemination of machine-processable information expressed in multiple media, such as text, image, graphics, speech, audio, video, animation, handwriting and data files. With the advent of high-capacity storage devices, powerful and yet economical computer workstations and high-speed Integrated Services Digital Networks (ISDNs), providing a variety of multimedia communications services is becoming not only technically, but also economically, feasible. In addition, the Broadband ISDN (BISDN) has been given special attention as a next generation communication network infrastructure that will be capable of transmitting full motion pictures and high speed data at 150 and 600 MB/s and voice, as well as data, throughout the world [1.1].

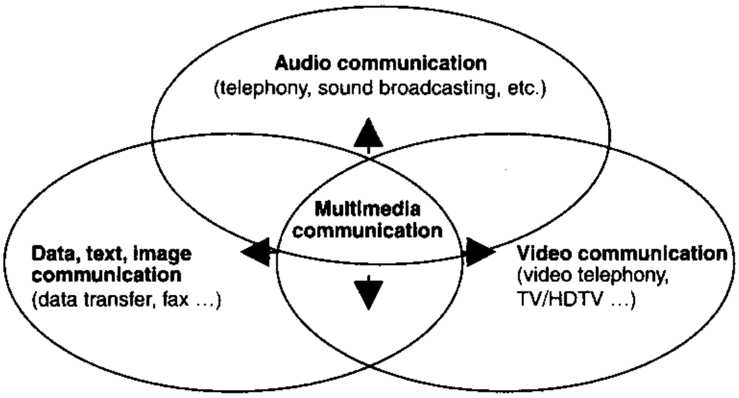

Audio communication
(telephony, sound broadcasting, etc.)
Multimedia communication
Data, text, image communication
(data transfer, fax ...)
Video communication
(video telephony, TV/HDTV ...)

Figure 1.1 Multimedia communication.

Multimedia best suits the human being's complex perception and communicating behavior, as well as the way of acting. Namely, it not only provides communication capabilities and information sharing for people, irrespective of location and time, but it also provides easy and immediate access to widely distributed information banks and information processing centers. Applications in medicine, education, travel, real estate, banking, insurance, administration and publishing are emerging at a fast pace. These applications are characterized by large multimedia documents that must be communicated within very short delays. Computer-controlled cooperative work, whereby a group of users can jointly view, create, edit and discuss multimedia documents, is going to be characteristic of many transactions [1.2]. Some glamorous applications in multimedia processing include distance learning, virtual library access and living books. In distance learning, we learn and interact with instructors remotely across a broadband communication network. Virtual library access means that we instantly have access to all of the published material in the world, in its original form and format, and that we can browse, display, print and even modify the material instantaneously. Living books supplement the written word and the associated pictures with animations, and hyperlinks provide access to supplementary material [1.3, 1.4, 1.5].

Applications that are enabled or enhanced by video are often seen as the primary justification for the development of multimedia networks. Trends toward multimedia communication are represented in Figure 1.1.

Much of the work on packet video has considered a fairly homogenous networking scenario [1.6]. It would be proper if a single type of video service dominated in the networks. However, it is not a valid assumption for the traffic issues. First, video will not constitute a uniform service with easily determined behavior and requirements. Second, video will not share resources with streams of only the same type. This means that multiplexing in the network should be evaluated for a heterogeneous mix of traffic types. In business areas, there is a potential need for various kinds of new communication systems, such as high-speed data networks between geographically distributed LANs, high definition still-picture communication and TV

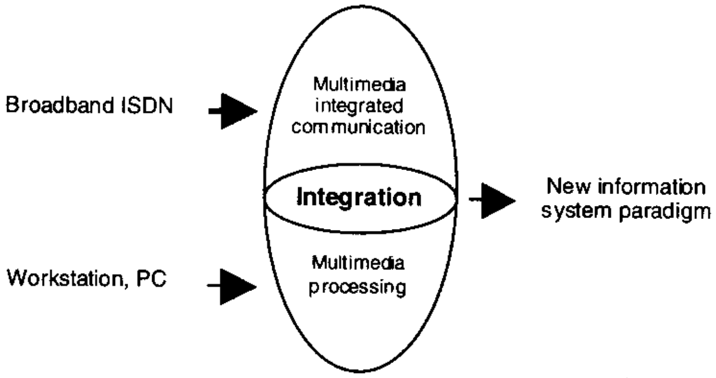

Broadband ISDN
Workstation, PC
Multimedia integrated communication
Integration
Multimedia processing
New information system paradigm

Figure 1.2 New information system paradigm using BISDN and workstations.

conferencing or corporate cable TV services. The new paradigm of the BISDN application system is a result of the integration of multimedia processing by workstations and multimedia communication by BISDN and is shown in Figure 1.2.

It is important to distinguish multimedia material from·what is often referred to as multiple-media material. To illustrate the difference, consider using the application of messaging. Today, messaging consists of several types, including electronic mail (email), which is primarily text messaging, voice mail, image mail, video mail, and hand-written mail often transmitted as a facsimile (fax) document. Each of these messaging types is generally a single medium and is associated with a unique delivery mechanism and a unique repository or mailbox. For convenience, most consumers would like to have all messages delivered to a common repository or mailbox. Hence, you have the concept of multiple media being integrated into a single location.

In networked multimedia applications, various entities typically cooperate in order to provide the real-time guarantees to allow data to be presented at the user interface. These requirements are most often defined in terms of Quality of Service (QoS). QoS is defined as the set of parameters that defines the properties of media streams. We distinguish four layers of QoS: user QoS, application QoS, system QoS and network QoS [1.7]. The user QoS parameters describe requirements for the perception of multimedia data at the user interface. The application QoS parameters describe requirements for the application services, possibly specified in terms of media quality (high end-to-end delay) and media relations (like inter-intrastream synchronization). The system QoS parameters describe requirements on the communications services resulting from the application QoS. These may be specified in terms of both quantitative and qualitative criteria. By quantitative criteria, we mean bits per second or task processing time, while multicast, interstream synchronization, error recovery or ordered delivery of data represent qualitative criteria. The network QoS parameters describe requirements on network services, like network load and network performance. This chapter seeks to provide a brief presentation of multimedia communications including multimedia communication models and elements of multimedia systems, as well as the packet transfer concept. User and network requirements are emphasized together with multimedia in the ATM environment. An outline of the multimedia terminals concludes the chapter.

1.2 Multimedia Communication Model

A multimedia communication model is strongly influenced by the manufacturer-dependent solutions for PCs and workstations, including application software on the one hand and the intelligent network concept on the other [1.8, 1.9, 1.10]. A layered model for future multimedia communication comprises five components:

- Partitioning of complex information objects into distinct information types for the purpose of easier communicating, storing and processing. This comprises data, video or audio and takes into account the integration of different information types.
- Standardization of service components per information type, possibly with several levels of quality per information type.
- Creation of platforms at two levels: a network service platform and a multimedia communication platform. The first level hides the transport networks and network building blocks from an application designer's or user's point of view. The second level provides communication support on the basis of information structure and information exchange building blocks for a large number of applications.
- Definition of generic applications for multiple use in various multimedia environments and different branches meeting common widespread needs.
- Specific applications: electronic shopping, teletraining and remote maintenance, based on special information building blocks and using the network service platform and the multimedia communication platform, as well as including generic applications.

With regard to the capability of the available resources in each case, the multimedia communication applications must be scalable in order to run in a constant manner across different network and terminal types and capabilities.

1.3 Elements of Multimedia Systems

Multimedia systems generally use two key communications modes: person-to-person communications and person-to-machine communications. Figure 1.3 presents key elements of multimedia systems. As can be seen, both of these modes have a lot of commonality, as well as some differences.

In the person-to-person mode shown in Figure 1.3(a), there is a user interface that provides the mechanisms for all users to interact with each other, and there is a transport layer that moves the multimedia signal from one user location to some or all other user locations associated with the communications. The user interface creates the multimedia signal and allows users to interact with the multimedia signal in an easy-to-use manner. The transport layer preserves the qual-

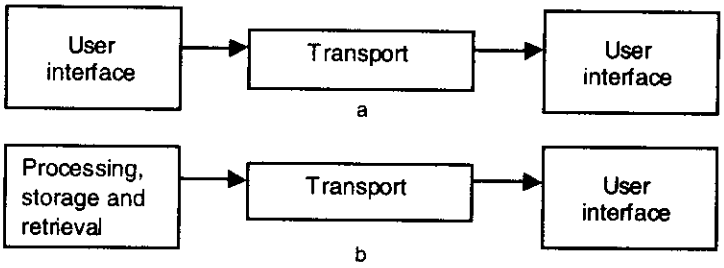


Figure 1.3 Elements of multimedia systems used in (a) person-to-person communications and (b) person-to-machine communications [1.2]. ©1998 IEEE.

ity of the multimedia signals so that all users receive what they perceive to be high-quality signals at each user location.

Examples of applications for the person-to-person mode are teleconferencing, videophones, distance learning and shared workspace scenarios. In the person-to-machine mode, shown in Figure 1.3(b), there is again a user interface for interacting with the machine, along with a transport layer for moving the multimedia signal from the storage location to the user. There is also a mechanism for storage and retrieval of multimedia signals that are either created by the user or requested by the user. The storage and retrieval mechanisms involve browsing and searching to find existing multimedia data. Also, these mechanisms involve storage and archiving in order to move user-created multimedia data to the appropriate place for access by others. Examples of applications for the person-to-machine mode include creation and access of business meeting notes, access of broadcast video and document archives from a digital library or other repositories.

1.4 User Requirements

The user needs a multimedia communication system that prepares and presents the information of interest, allows for the dynamic control of applications and provides a natural interface. From a user's point of view, the most important requirements of multimedia communications are the following:

- Fast preparation and presentation of the different information types of interest, taking into account the capabilities of available terminals and services
- Dynamic control of multimedia applications with respect to connection interactions and quality on demand combined with user-friendly human/machine interfaces
- Intelligent support of users taking into consideration their individual capabilities
- Standardization

User requirements in terms of services are defined by the media, the transmission content and the type of communication, as well as the ability to combine the three. On the other hand, multimedia communication services can be classified as being local (interactive or noninteractive), remote noninteractive or remote interactive and can also as be classified as being for residential, business or mobile use. The context in which multimedia services can be used is shown in Table 1.1.

Table 1.1 Context in which multimedia services can be used.

Local		Remote Noninteractive	Remote Interactive
Residential	Leisure (TV) The arts Teaching Games	Broadcasting	Enhanced telephones Videophones Home shopping Games Remote consultation Video on demand
Mobile	Presentation Demonstration	Broadcasting Remote security Monitoring	Project management Contract negotiation
Business	Multimedia presentation Training Database consultation	Teleinformation Teletraining Telesupervision	Video meeting Video conferencing Distance learning Project management Remote security Monitoring Remote diagnostics

Service usage conditions can be defined by their use, place, independence and degree of urgency. Services can be for private or business use. The terminal and services are usually used in the office, the home, the car or a public place. Independence could be defined by the portability of the terminal and its independence of a given infrastructure as perceived by the user. The degree of independence varies from one type of terminal to another. On the other hand, the degree of urgency, from the user's point of view, determines whether the service should be provided in real time or whether an offline service is sufficient.

A number of key requirements are common to the new multimedia services:

- Instant availability
- Real-time information transfer
- Service always online
- Access their services from any terminal (mobile point of delivery)

Whereas traditional voice services already have these characteristics, data services across the Internet (including voice over data) have typically been limited to basic bit transport with no service guarantees, no guaranteed availability and rather fragmented service interruptions.

With new data service emerging, such as Virtual Private Networks (VPNs) and interconnection service between two network service providers, priorities in the data networking domain have to change. In order to resolve and build robust multimedia networks, it is natural that operators will seek to base their data networks on the proven service delivery capability currently deployed in leading-edge voice networks. This will provide the flexibility, functionality and reliability required to meet the new demands of future users. Also, it will enable operators to offer the sophisticated services currently provided for voice in the multimedia domain.

Multimedia applications have several requirements with respect to the service offered to them by the communication system. These requirements depend on the type of the application and on its usage scenario. For instance, a nonconversational application for the retrieval of audio-visual data has different needs than a conversational application for live audio-visual communication (that is, a conferencing tool). The usage scenario influences the criticality of the demands.

1.5 Network Requirements

From the network point of view, the most important requirements of multimedia communications are the following:

- High speed and changing bit rates
- Several virtual connections using the same access
- Synchronization of different information types
- Suitable standardized services and supplementary service supporting multimedia applications

The requirements of applications regarding the communications services can be divided into traffic and functional requirements. The traffic requirements include transmission bandwidth delay and reliability. They depend on the used kind, number and quality of the data streams. The traffic requirements can be satisfied by the use of resource management mechanisms. They establish a relationship between transmitted data and resources and ensure that the audio-visual data is transmitted in a timely manner. For this, during the transmission of data, the information about the resource needs must be available at all nodes participating in the distributed applications, end systems and centers. Hence, resources must be reserved, and states must be created in these nodes, which basically means that a connection is established. The functional requirements are multicast transmission and the ability to define coordinated sets of unidirectional streams.

Current fixed and mobile networks are built on mature architectures with strong traffic management, configuration capabilities, service platforms and well-defined points of interconnection between the networks of different operators. A key requirement is that the same high-quality network services should exist when building integrated networking platforms for voice, data and multimedia services [1.11].

A future multimedia network must be organized to support heavy traffic flows, a wide variety of service mixes and different traffic patterns, both in terms of routing the traffic efficiently and in terms of scaling for overload. The network must adapt quickly to constantly changing traffic conditions. Reliable security features and firewalls must be in place for interworking between the many operators that will be competing in the market as a result of deregulation.

1.6 Packet Transfer Concept

Today's fiber technology offers a transmission capability that can easily handle high-bit rates like those required for video transmission. This leads to the development of networks, which integrate all types of information services. By basing such a network on packet switching, the services (video, voice and data) can be dealt with in a common format. Packet switching is more flexible than circuit switching in that it can emulate the latter while vastly different bit rates can be multiplexed together. In addition, the network's statistically multiplexing of variable rate services may yield a higher use of the channel capacity than what is obtainable with fixed capacity allocation. Many years ago, most of these arguments were verified in a number of projects [1.10, 1.12, 1.13, 1.14].

The project MAGNET II is an integrated network test bed at Columbia University. It was designed and implemented based on requirements of real-time network mangement and control. Switching is based on the concept of packet switching and the multi-class network model. The requirement made on the resource sharing mechanisms is to guarantee the appropriate quality of service for each traffic class [1.12]. However, bandwidth allocation is one of the most important problems in the management of networks that have guaranteed bandwith policy (ATM [1.12], PARIS [1.14]). Thus, a good bandwidth allocation strategy is crucial for packet switching networks.

Compared to circuit switching, packet switching offers dynamic allocation of bandwidths and switching resources, as well as the elimination of channel structure. Packet networks allow integrated service transport. They can carry voice, video and data using the same hardware protocols. Furthermore, packet communication does not require users to allocate a fixed channel or bandwidth before data transmission. Because users send packets only when necessary, and because many users can send packets over the same shared channel, resources in packet networks are used more efficiently than in circuit switched networks. Video signals are especially well suited for packet transmission. Images usually contain regions of high detail and low detail as well as periods of rapid motion and little motion. Thus, effective video coders should be able to produce data streams with variable bit rates that change with local characteristics [1.15, 1.16]. Packet networks can carry Variable Bit Rate (VBR) signals directly. No buffering or rate control feedback is necessary at the transmitter. If a video coder can specify the order in which a network discards data in case of network congestion, the decoder can often suffer less degradation in picture quality when packet loss occurs. Packet networks that offer prioritization of packets give video coders the ability to protect critical information. However, packet networks also provide some difficulties for video coders.

The ATM networks are based on virtual circuit switching. All fixed size packets of a circuit have the fixed route [1.17]. The tasks of packet video transfer across an asynchronous time division multiplexed network or Internet are to code and transfer digital video signal under quality constraints as shown in Figure 1.4. In Internet Protocol (IP) networks, the packets are of variable length, and there is no pre-established route. Therefore, they may arrive out of order at destination. During transfers in ATM and IP networks, delay and some packet loss are unavoidable [1.18, 1.19, 1.20].

Video	Encoder - Transformation - Quantization - Entropy coding - Bit-rate control	Application - data structuring	Network Multiplexing/Routing		Application - Resynchronization	Decoder - Dequantization - Entropy decoding - Inverse transformation - Loss concealment - Postprocessing	Users
			- Overhead (FEC) - Retransmission	- Error detection - Loss detection - Error correction - Erasure correction			

Figure 1.4 Digital video signal transport.

The generic functions of the network in packet transfer from source to user are routing and multiplexing. Routing provides connectivity and does not depend on the information type used in the transfer. Multiplexing determines much of the transfer quality in the network and is highly dependent on the traffic characteristics, the quality requirements and the user's applications.

Statistical multiplexing with quality guarantees is the best choice for video transfer [1.21]. In order to offer probabilistic guarantees, a network must know its current flow of traffic, based on already accepted connections or measurements of the actual network load. New connections are allowed if they can be guaranteed the quality that they request and their characteristics do not risk the quality of already accepted connections or measurements of the actual network load.

1.7 Multimedia Requirements and ATM Networks

One of the biggest factors in the emergence of multimedia computing and processing was the digitization of the telecommunications network, which was completed by the late 1980s. This brought a new era in communications, where digital representations of both signals and data could interact in a common network, and it led to the concepts behind modern data networks and data protocols like ATM [1.22].

Today the ATM plays a significant role in realizing the flexibility and economy necessary for multimedia communication [1.23]. The most important ATM capabilities for multimedia requirements are the following:

- Constant, variable or burst-oriented bit streams
- Virtual connections or virtual paths through the subscriber access depending on instantaneous needs with the total capacity of about 150 or 600 MB/s
- Uniform bit rate-independent transmission and switching systems

In addition, the ATM concept aims at a universal network that offers all services using one uniform bit rate and information type-independent access based on harmonized protocols. The ATM protocol is based on the concept of designing a packet protocol that would be appropriate for both real-time signals (such as speech, audio and video) and data. Hence, its packet size is small (53 bytes) for low-latency signal processing, and the header size is small (5 bytes) for high efficiency. This facilitates information packet sequencing and synchronization between various information types within one multimedia application. ATM networks are designed to efficiently handle high volume voice, audio, and video traffic, yet still maintain their effectiveness for bursty data. As demand grows, a universal ATM network will be superior both in performance and in cost to the alternative solution where several specialized networks exist side by side [1.24].

The more standard Transmission Control Protocol/Internet Protocol (TCP/IP) used on the Internet uses significantly larger packets (upwards of 1 to 2 KB packets) for greater efficiency in moving large, bursty and data traffic through the IP network.

In spite of these advantages, ATM has some constraints on multimedia communications. Namely, voice packetization and depacketization result in additional delays that may call for echo compensation measures. Nevertheless, the future multimedia communication system will be created offering a large variety of applications based on an efficient universal ATM network.

1.8 Multimedia Terminals

Every major advance in networking has been preceded by an advance in the user interface that has precipitated the acceptance and growth of the networking advance. For example, the invention of the telephone preceded the growth of switch networks, the invention of the television preceded the growth of TV networks and Cable Television (CATV), the radio telephone led to the cellular network, the PC led to the LAN/WAN network and the browser led to the growth of the Internet and the Web. For the multimedia, new smart terminals need to be created in order to facilitate the displaying, accessing, indexing, browsing and searching of multimedia content in a convenient and easy-to-use manner.

For multimedia systems to achieve the vision of the current communications revolution and to become available to everyone, a number of technological issues must be addressed and put into a framework that leads to integration, ease of use, and high quality outputs. Among the issues that must be addressed are the following:

- The basic techniques for compression and coding the various media that constitute multimedia signals, 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 QoS for multimedia signal and its constituent components
- The basic techniques for accessing the multimedia signals by providing tools that match the user to the machine
- The basic techniques for searching in order to find multimedia sources that provide the desired information or material, or 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 using 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 using text browsing, indexed image browsing, and voice browsing

Multimedia itself denotes the integrated manipulation of at least some information represented as continuous media data, as well as some information encoded as discrete media data (text and graphics). Here, we have the act of capturing, processing, communicating, presenting and/or storing.

Multimedia terminals are needed to retrieve, analyze, store and broadcast the new forms of written, sound and visual content. The architecture of these systems can be defined according to different approaches based on telecommunications data processing and audiovisual technology. By incorporating voice and data as well as still and moving pictures into their communications, business has made functions increasingly sophisticated to improve access to distributed resources and to save valuable time in the decision process. Remote dialog, discussion, information production, maintenance and inspection are now possible from the new multimedia systems at operating costs that are continuing to fall. Existing solutions offer two types of terminals: multifunction office or computer workstations and dedicated equipment, such as enhanced telephone terminals, videophones or shared teleconferencing systems.

Multimedia communication requires powerful terminals: upgraded PCs, desktop workstations or video computers. Today's terminals are enhanced for broadband multimedia applications, for example, PCs by the addition of telecommunication and video-audio capabilities and TV receivers by the addition of intelligence and interactivity. At the same time, High Definition Television (HDTV) is in development, leading the way toward all digital TV. HDTV is a technology driver for memories, image/video processors and flat screens [1.21].

Multimedia terminal equipment also comprises suitable cameras, scanners, printers and mass storage. Special equipment is necessary for editing multimedia information, that is, the creation, alternation and deletion of content and structures. Three-dimensional (3D) display devices and speech recognition systems will further facilitate faster and easier human interaction with multimedia applications or editors.

PCs and workstation architecture are considered for the interconnection of the systems components, based on star configurations and using ATM principles. This could make the integration of all information types easier and could provide the necessary high-bit rates. This concept supports the extension of a PC or workstation into an ATM-oriented desk area network, comprising cameras, printers and other special purpose systems or subsystems offering interfaces to ATM networks.

1.9 Concluding Remarks

The term *multimedia communications* refers to the representation, storage, retrieval and dissemination of computer-processable information expressed in multiple media, such as text, image, graphics, speech, audio, video, animation, handwriting, and data files. Trends toward multimedia communication are audio, video, image, data and text communications.

The key elements of communication models are processing, storage, retrieval, transport and user interface. The user needs a system that prepares and presents the information of interest, allows for the dynamic control of applications and provides a natural interface. From the network point of view, the most important requirements of multimedia communications are high-speed bit rates, synchronization of different information types and suitable standardized services. Multimedia transport based on packet transfer concept has many advantages. The generic functions of network in packet transfer from source to user are routing and multiplexing. Routing provides connectivity and does not depend on information type used in the transfer. Multiplexing determines much of the transfer quality in the network and is highly dependent of the traffic characteristics, the quality requirements and user's applications. The asynchronous transfer mode (ATM) networks are based on virtual circuit switching: All fixed size packets of a circuit have the fixed route. The most important ATM capabilities for multimedia communications are constant, variable or burst-oriented bit streams, virtual connections or virtual paths via the subscriber access depending on instantaneous needs, as well uniform bit rate independent transmission and switching systems.

CHAPTER 2

Audio-Visual Integration

Chapter Overview

In multimedia communication where human speech is involved, audio-visual integration is particularly significant. Not only is it important to consider both verbal and universal information in multimedia communication, but the interaction among different media is also interesting. This chapter explains why multimedia communication is more than simply putting together text, audio, images and video. It reviews a recent trend in multimedia research to exploit the audio-visual interaction and to build the link between audio and video processing. The emphasis is on lip reading, synchronization and tracing, audio-to-visual mapping and bimodal person verification. Some examples cover a broad range of these topics.

2.1 Introduction

Among the possible interactions considering different media types, the interaction between audio and video is the most interesting. A recent trend in multimedia research is to integrate audio and visual processing in order to exploit such interaction. Generally speaking, more interesting research topics can be found when we exploit the interaction among different media types. For example, using speech recognition technology, one can analyze speech waveforms to discover the text that has been spoken. From a sentence of text, a talking-head audio-visual sequence can be generated using computer graphics to animate a facial model and using text-to-speech synthesis to provide synthetic acoustic speech. For person-to-person communication, which is involved in multimedia applications like video telephony and video conferencing, audio and visual interaction is very important. Topics being researched progress from the point of view of audio-visual integration and include automatic lip reading and its use in speech recognition, speech-driven face animation, speech-assisted lip synchronization, facial-feature tracking, audio-visual mapping, bimodal person verification and joint audio-video coding. Some examples cover a broad range, including media interaction, bimodality of human speech, lip reading, lip synchronization, audio-to-visual mapping and bimodal person verification.

2.2 Media Interaction

Integration and interaction among different media types create challenging research topics and new opportunities [2.1].Media interaction is shown in Figure 2.1. As can be seen, media are categorized into three major classes. The first is textual information, the second is audio, including speech and music, and the third represents image and video. The goal of speech recognition is to enable a machine to be able to transcribe spoken inputs literally into individual words, but the goal of spoken language understanding research is to extract meaning from whatever was recognized [2.2, 2.3]. The various Spoken Language Interface (SLI) applications have widely differing requirements for speech recognition and spoken language understanding. Hence, a range of different performance measures on the various systems reflects both the task constraints and the application requirements. Some SLI applications require a speech recognizer to do word-for-word transcription.

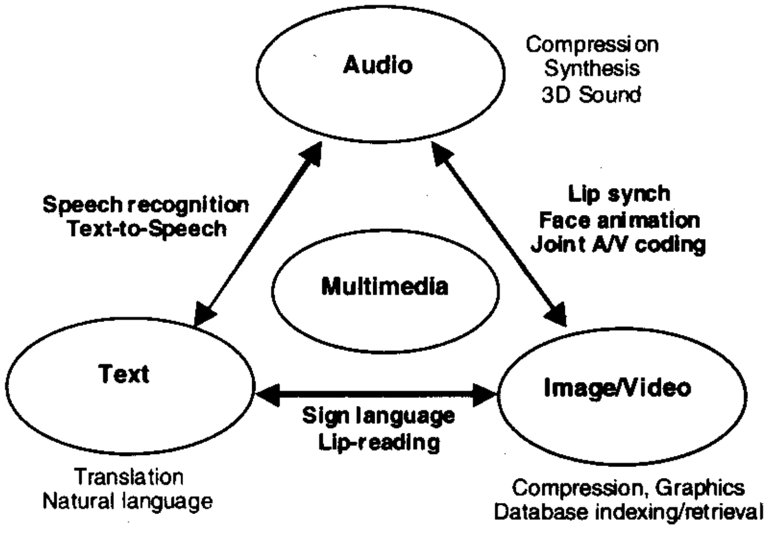


Figure 2.1 Media interaction [2.1]. ©1998 IEEE.

Example 2.1 Sending a textual response to an email message requires having capabilities for voice dictation, entering stock information or ordering from a catalog and entering number sequences or lists of data. For these types of systems, word error rate is an excellent measure of how well the speech recognizer produces a word-for-word transcription of the user's utterance.

The current capabilities in speech recognition and natural language understanding, in terms of word error rates, are summarized in Table 2.1. It can be seen that performance is very good for constrained tasks (digit strings and travel reservations). On the other hand, the word error rate increases rapidly for unconstrained conversational speech. Although methods of adaptation can improve performance by as much as a factor of two, this is still inadequate performance for use in many interesting tasks.

Table 2.1 Word error rates for speech recognition and natural language understanding tasks [2.2].

Corpus	Type	Vocabulary Size	Word Error Rate
Connected digit strings	Spontaneous	10	0.3%
Airline travel information	Spontaneous	2,500	2%
Wall Street Journal	Read text	64,000	8%
Radio (marketplace)	Mixed	64,000	27%
Switchboard	Conversational telephone	10,000	38%
Call home	Conversational telephone	10,000	50%

For some applications, complete word-for-word speech recognition is not required. Instead, tasks can be accomplished successfully even if the machine detects only certain key words or phrases within the speech. For such systems, the job of the machine is to categorize the user's utterances into one of a relatively small set of categories. The category identified is then mapped to an appropriate action or response [2.2].

Example 2.2 As an example, consider the AT&T system How May I Help You (HMIHY) task, in which the goal is to classify the user's natural language spoken input into one of 15 possible categories, such as billing credit, collect call and so forth. After this initial classification is done, the system transfers the caller to a category-specific subsystem, which uses either another artificial agent or a human operator [2.2]. Concept accuracy is a more appropriate measure of performance for this class of tasks than word accuracy. In the HMIHY task, word accuracy is only about 50%, but concept accuracy approaches 87%.

Another set of applications of speech recognition technology is the so-called spoken language understanding systems where the user is unconstrained in terms of what can be spoken and in what manner, but is highly constrained in terms of the context in which the machine is queried.

Example 2.3 An example of this type of application includes AT&T's CHRONUS system for air-travel information [2.4] and a number of prototype railway information systems. As in

HMIHY (Example 2.2), results show speech-understanding error rates of 6 to 10%, despite recognition error rates of 20 to 23%. These results demonstrate how a powerful language model can achieve high understanding performance despite imperfect Automatic Speech Recognition (ASR) technology.

A good example of using A/V interaction for human speech communication is lip reading, also referred to as speech reading. Human lip reading is widely used by hearing-impaired persons for speech understanding and automated lip reading.

Lip synchronization is one of the most important issues in videotelephony and videoconferencing. A typical situation in videoconfe encing equipment is when the frame rate is not adequate for lip synchronization perception. One solution is to extract information from the acoustic signal, which determines the corresponding mouth movements, and then to process the speaker's mouth image accordingly to achieve lip synchronization. It is also possible to warp the acoustic signal to make it sound synchronized with the person's mouth movement. This approach is very useful in non-real-time applications, such as dubbing in a studio.

Researchers have tried to produce visual speech from auditory speech, that is, to generate speech-driven talking heads [2.5, 2.6, 2.7]. The major applications of this technique include human-computer interfaces, computer-aided instruction, cartoon animation, videogames and multimedia telephony for the hearing impaired.

The improvement in multimedia interaction can be obtained by using joint audio-video processing compared to the situation where audio and video are processed independently. Audio-visual interaction is very important in multimedia communication where human speech is involved because of its bimodal nature.

2.3 Bimodality of Human Speech

Due to the bimodality in speech perception audio-visual interaction becomes an important design factor for multimode communication systems, such as videotelephony and video conferencing [2.8, 2.9].

The bimodal nature of human speech perception was demonstrated by McGurk and MacDonald [2.10]. When humans are presented with conflicting audio and visual stimuli, the perceived sound may not exist in either modality.

Example 2.4 When a person "hears" the sound /ba/, but "sees" the speaker saying /ga/, the person may not perceive either /ga/ or /ba/. Instead, what is perceived is something close to /da/. Some other examples of audio-visual combinations are shown in Table 2.2. This shows that the speech that is perceived by a person depends not only on acoustic cues, but also on visual cues such as lip movements.

Psychologists have shown that the reverse McGurk effect also exists, that is, the results of visual speech perception can be affected by the dubbed audio speech [2.11]. The McGurk effect is also robust to a variety of different conditions [2.12]. The same answers are obtained to the conflicting stimuli in cases when there are timing mismatches between the stimuli or even when the face of a male speaker is combined with the voice of female speaker [2.13].

Table 2.2 Examples of McGurk effect [2.1].

Audio +	Visual →	Perceived
ba	ga	da
pa	ga	ta
ma	ga	na

Speech production is bimodal in nature, together with speech perception. Human speech is produced by the vibration of the vocal cord and the configuration of the vocal tract (articulatory organs, including the nasal cavity, tongue, teeth, velum and lips). These articulatory organs, together with the muscles that generate facial expressions, produce speech. There is an inherent relationship between the acoustic and visible speech because some of these articulators are visible.

The basic unit of acoustic speech is called a *phoneme* [2.2]. Similarly, in the visual domain, the basic unit of mouth movements is called a *viseme* [2.14]. A viseme is the smallest visibly distinguishable unit of speech. Many acoustic sounds are visually ambiguous. These sounds are grouped into the same class that represents a viseme. There is, therefore, a many-to-one mapping between phonemes and visemes.

Example 2.5 The /p/, /b/ and /m/ phonemes are all produced by a closed mouth shape and are visually indistinguishable. Therefore, they form one viseme group. Similarly, /f/ and /v/ both belong to the same viseme group that represents a mouth of which the upper teeth are touching the lower lip.

Example 2.6 Visemes for English consonants can be grouped into nine distinct groups as shown in Table 2.3. A number of visemes are shown in Figure 2.2 [2.1]. The top three are associated with consonants, and the bottom three are associated with vowels.

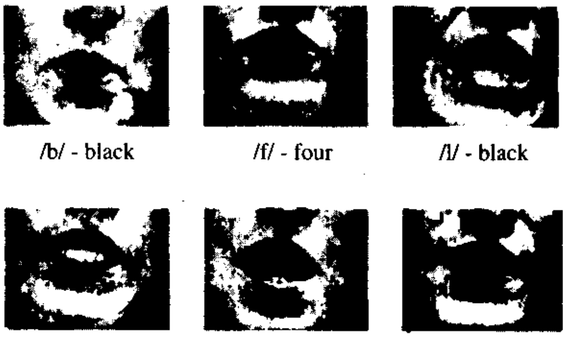
Table 2.3 Viseme groups for English consonants [2.1] ©1998 IEEE.

Number	Viseme Groups
1.	f, v
2.	th, dh
3.	s, z
4.	sh, zh
5.	p, b, m

Table 2.3 Viseme groups for English consonants [2.1]. ©1998 IEEE. (Continued)

Number	Viseme Groups
6.	w
7.	r
8.	g, k, n, t, d, y
9.	l

Example 2.7 Instead of a still image, a viseme can be a sequence of several images that capture the movements of the mouth. This is especially true for some vowels. For example, the viseme /ow/ represents the movement of the mouth from a position close to /o/ to a position close to /w/. Therefore, to illustrate some visemes, we would need to use video sequences. However, most visemes can be approximated by stationary images.



/b/ - black /f/ - four /l/ - black

/iy/ - three /ah/ - mom /eh/ - seven

Figure 2.2 Example visemes [2.1]. ©1998 IEEE.

Example 2.8 Most of the vowels are distinguishable both in acoustic and in visual modality [2.15]. The sounds /p/, /t/ and /k/ are very similar. The confusion sets in the auditory modality are usually distinguishable in the visual modality. The sounds /p/ and /k/ can be easily distinguished by the visual cue of a closed mouth versus an open mouth.

2.4 Lip Reading

A person skilled in lip reading is able to infer the meaning of spoken sentences by looking at the configuration and the motion of visible articulators of the speaker, such as the tongue, lips, teeth and so forth. Knowledge of the positions of these articulators provides information about the content of the acoustic speech signal. Because all of the articulators are not visible, this information may sometimes be incomplete. By combining the visual content with lexical, syntactic, semantic, and programmatic information, people can learn to understand spoken language by observation of the movements of a speaker's face.

Lip reading performance depends on a number of factors. Viewing conditions may affect the quality of the visual information. Poor lighting may make it difficult to judge the shape of the mouth or to detect the teeth or tongue. Likewise, as the speaker and listener move further apart, it becomes more difficult to view important visual cues. Factors such as viewing angle can also affect recognition. Neely [2.16] found that a frontal view of the speaker led to higher recognition rates than an angled or profile view. Lip reading performance can also be improved through training [2.17].

Example 2.9 Before training, a person could confuse an /s/ or /z/ with a /th/ or /dh/. After training, these confusions can be eliminated. Coarticulation can also affect lip-reading performance. It is the process by which one sound affects the production of neighboring sounds. The place of articulation may not be fixed, but may depend on context [2.18].

Example 2.10 The tongue may be in different positions for the /r/ sound in "art" and "arc." The tongue would be more forward in the word "art." This would affect the visibility of the tongue and thus recognition accuracy.

Syllables with a Vowel-Consonant-Vowel (VCV) context were examined by Benguerel and Pichora-Fuller [2.19]. It was found that consonant recognition depended on the vowels that surrounded it. For example, the middle consonant is more difficult to lip read when the vowel is a /u/ as opposed to an /ae/ or /i/. At the same time, the /u/ was the most recognizable of the vowels. This suggests that the /u/ sound is visually dominant. Its appearance is pronounced, and, because of this, there is a recovery period where neighboring sounds may be masked.

Knowledge of the process by which humans extract and incorporate visual information into speech perception can be useful. Also, it is important to know what information is available in the visual channel, what types of parameters humans use to aid recognition, and what means are used to integrate the acoustic and visual information. The acoustic and visual components of the speech signal are not purely redundant. They are complementary as well. Certain speech characteristics that are usually confusable are acoustically distinct, and those characteristics that are acoustically confusable are visually distinct.

Example 2.11 The /p/ and /k/ phonemes have similar acoustic characteristics, but can easily be differentiated by the closing of the mouth. In contrast, the /p/ and /b/ phonemes have similar visual characteristics, but can be acoustically differentiated by voicing.

The widespread beliefs about how humans integrate acoustic and visual information can be classified into early integration and high integration as shown in Figure 2.3. In early integration, the acoustic and visual parameter sets are combined into a larger parameter set. Recognition occurs by finding the word with the template that matches best to the audio-visual parameter sets. On the other hand, in late integration, the audio is compared against an acoustic template for each word, and the video is compared against a visual template for each word. The resulting audio and visual recognition scores are then combined using a certain mechanism.

The recognition systems that have been designed for automated lip reading vary widely and have been used to help answer a number of questions concerning audio-visual recognition. Many systems have been designed to show that speech recognition is possible using only the visual information. Some researchers have done comparisons on a number of visual feature sets in attempts to find those features that yield the best recognition performance. Next, researchers have attempted to integrate these visual-only recognition systems with acoustic recognition systems in order to enhance the accuracy of the acoustic speech-recognition system. A number of studies have examined strategies of early integration, late integration and other novel means for combining audio and video. Some studies have done more investigations into the resiliency of audio-visual recognition systems to varying levels of acoustic noise. Finally, many of the systems in the open literature have attacked different recognition tasks, implementing speaker-dependent and speaker-independent systems and examining isolated vowels, Consonant-Vowel-Consonant (CVC) syllables, isolated words, connected digits and continuous speech [2.1].

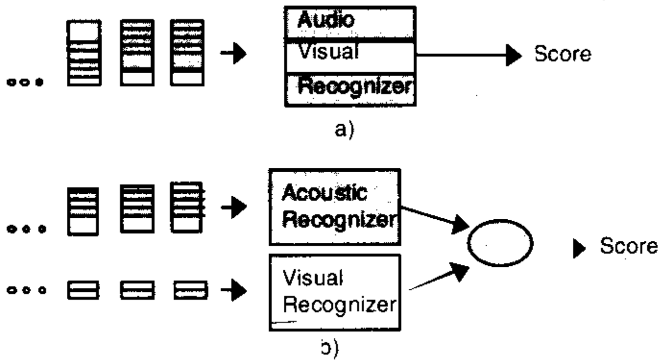


Figure 2.3 Early integration (a) and late integration (b) [2.1]. ©1998 IEEE.

The recognizers that have been designed vary in the acoustic feature sets, visual feature sets, recognition engine and integration strategies. The visual feature sets provide some of the greatest variations between systems. Systems range from performing relatively little processing and having a visual input that consists of a rectangular region of the image to using computer vision techniques to extract a visual feature set to be used for recognition [2.20].

The recognition engine also takes many forms among the various research groups. Some of the earlier recognition systems were based on dynamic time warping. A number of systems have used neural-network architectures, and an increasing number of systems have relied on Hidden Markov Models (HMM) for recognition [2.21].

A number of integration strategies have been proposed. Some of the early systems used either sequential recognition or a rule-based approach. One system converted the visual signal into an acoustic power spectrum and averaged the information sources in this domain. Other systems have employed either early or late integration strategies and compared the differences between the two [2.22, 2.23]. Most of the lip reading systems can be intrusive to the users to some extent. In particular, the user has to remain relatively stationary for the visual analysis system to work well. In Duchnowski et al. [2.24], a modular system was developed to alleviate such constraints. The visual analysis of the system is composed of an automatic face tracker, followed by the lip locator. The capability of face tracking results in a speech recognizer that allows the speaker reasonable freedom of movement. Besides trying to derive the acoustic domain information from visual information, researchers have also tried to produce visual speech from auditory speech, that is, to generate speech-driven talking heads [2.5, 2.6]. In the next section, we study the generation of talking-head images.

2.5 Speech-Driven Talking Heads

The major applications of the talking-heads technique include human-computer interfaces, computer-aided instruction, video games, cartoon animation, and multimedia telephony for the hearing impaired.

Two approaches are used in generating talking-head images: the flipbook method [2.7] and the wireframe model approach, which can be two-dimensional (2D) or (3D) [2.25].

In the flipbook method, a number of mouth images of a person, called key frames, are captured and stored. Each image represents a particular mouth shape, for example a viseme. Then, according to the speech signal, the corresponding mouth images are flipped one by one to the display to form animation. This method results in jerkiness during key-frame transition, especially if the number of key frames is limited. Image warping can be used to create some intermediate frames to make the transition look smoother [2.26]. Image warping is a process whereby one image is distorted through the use of geometric transformations to look like another image. It is useful for producing realistic intermediate images between video frames. To accomplish this, correspondences are made between points in two adjacent video frames. Another similar flipbook approach uses some techniques to enhance the quality of the images [2.1]. Instead of using a small number of mouth images representing the phonemes, this technique uses a long video sequence of a talking person. The sequence contains all the possible mouth shapes of the person. It is analyzed to derive the phonetic transcription. When presented with input audio, matching segments from the original video are found. These segments are concatenated together with an image-processing technique. In order to achieve better synchronization, time warping is also done to these segments.

The wireframe method uses computer-graphics techniques to achieve better realism. A wireframe is composed of a large number of triangular patches that model the shape of a human face. One of the early facial models was developed by Parke [2.27]. Figure 2.4 shows the facial model called Candide developed at Linkoping University [2.28]. A 3D model would contain vertices that correspond to points throughout the head and would allow synthesis of facial images with arbitrary orientations. Because most animation is primarily concerned with the frontal view of the face, and not the back of the head, models are often developed only for a frontal portion of the face.

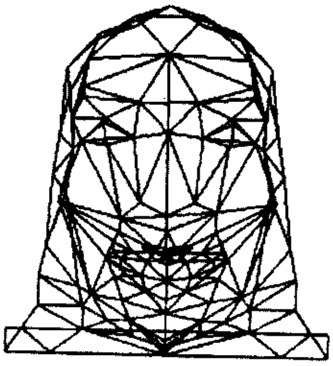


Figure 2.4 The wireframe model Candide [2.1]. ©1998 IEEE.

To synthesize various facial expressions, the Facial Action Coding System (FACS) is often used to generate the required trajectories of the vertices for a large variety of facial expressions and movements. A wireframe model gives only a structural representation of the face. To generate natural looking synthetic faces, a wireframe model must be combined with lighting models that specify how to map the shape and position of the wireframe into intensity when the wireframe is projected onto a 2D image. It is possible to use simple algorithms to synthesize artificial-looking faces together with texture mapping. It is an algorithm that maps pixel values of a 2D image, for example, a photo of a real person, onto patches of a wireframe. The texture from the original image helps create realistic synthetic images.

Generally speaking, there is a duality between the flipbook and the wireframe approaches. The flipbook approach is less computationally intensive, but it requires more data and the right number of images as key frames. The wireframe approach is more computationally intensive, but is needed for texture-mapping purposes. Then, arbitrarily oriented images can be synthesized with the model.

So far, we have discussed only the mechanisms that can be used to create talking heads. Now, we deal with the problem of how to produce the parameters to drive talking heads to make them "say" certain sentences. In [2.5], a 3D wireframe facial model was used to synthesize facial expressions, particularly lip motion. The lip parameters used form an eight-dimensional (8D) vector that includes the position of the upper lip, the position of the chin, and so forth. These parameters are derived by text input or speech input. In the case of text input, a sequence of the 8D feature vector is manually extracted for each phoneme. The input text is then analyzed into a sequence of phonemes, and the corresponding lip feature vectors are used to drive the facial model to produce the lip motion. In the case of speech input, a classifier that derives lip parameters from the input acoustic speech drives the lip feature points [2.2].

A work that focuses on a multimedia telephone for hearing-impaired people is presented by Lavagetto [2.6]. Here, the conversion from speech to lip movements is performed by a number of Time-Delayed Neural Networks (TDNNs). A major advantage of this approach is that TDNNs operate not only the current speech frame, but also its neighbors. Therefore, the estimated lip features can incorporate information from neighboring sounds. This helps model coarticulation effects from speech production.

2.6 Lip Synchronization

One of the most important questions in videotelephony and video conferencing is lip synchronization because human speech perception is bimodal. A typical situation in videoconferencing equipment with bandwidth constraints is when the frame rate is not adequate for lip synchronization perception. The solution can be to extract information from the acoustic signal, which determines the corresponding mouth movement and then process the speaker's mouth image accordingly to achieve lip synchronization.

It is possible to warp the acoustic signal to make it sound synchronized with the person's mouth movement [2.1]. This approach is very useful in non-real-time applications such as dubbing in a studio. In movie production, the dialog is usually recorded in a studio to replace the dialog recorded while filming a scene because the latter has poor quality due to background noises. To ensure lip synchronization, a spectrum analyzer analyzes both the studio audio and the original recording. The results are then input to a processor to find the best time-warping path that is required to modify the time scale of the studio audio to align the original recording. According to the time-warping path, the studio audio is edited pitch synchronously, that is, a period of sound segment being cut out or repeated. Thus, the studio dialog can be made in synchronization with the original lip movement.

One can time warp the video instead of warping the audio. We can warp the image of the speaker to make the lip movement hit the studio dialog. A video codec often skips some frames to meet the bandwidth requirement, which results in a lower frame rate at the decoder. Frame skipping introduces artifacts, such as jerky motion and loss of lip synchronization. To solve this, we can extract information from the speech signal and process the mouth image accordingly to achieve lip synchronization [2.29]. Figure 2.5 represents the block diagram of this approach. In this system, the feature-tracking module analyzes the input frames to find the location and shape of the mouth [2.30]. The audio-to-visual mapping module analyzes the audio signal and produces a sequence of the corresponding mouth shapes that are missing in the low-frame-rate video. Image warping is then applied to the input frames to modify the mouth shape to produce new frames that are to be inserted in the original video. Hence, lip synchronization is achieved in the high-frame-rate output video. Note the interaction between image analysis and speech analysis. The results of image analysis can be used to improve the accuracy of speech recognition, as is done in automatic lip reading. On the other hand, speech information can be used to improve the result of image analysis.

Example 2.12 We can decide whether the mouth is open or closed by speech analysis and then apply different algorithms to locate the lip points. Mouth closeness, for example, during /p/, /b/, /m/ and silence are important perceptual cues for lip synchronization. Therefore, the lip synchronization is good as long as speech analysis and image analysis together detect these closures correctly and image synthesis renders mouth closures precisely.

In addition to the low frame rate, another issue that causes loss of lip synchronization in videoconferencing is the transmission. The transmission delay for video is longer than the audio delay. We can always delay the audio to match the video. The speech-assisted video-processing

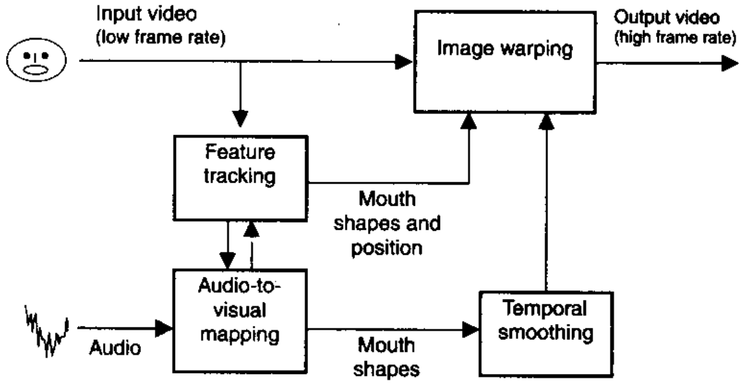


Figure 2.5 Block diagram of information extraction from the speech signal and processing the mouth image to achieve lip synchronization [2.1]. ©1998 IEEE.

technique can solve this problem by warping the mouth image of the speaker to be synchronized with the audio. Therefore, we can actually decrease the overall delay.

The speech-assisted interpolation scheme can be embedded into a video codec. In a typical video codec, such as H.263, some constraint usually exists on the number of frames that can be skipped between two coded frames [2.31]. This is needed in order to prevent too much jerkiness in the motion rendition and also to prevent too much loss of lip synchronization. When a speech-assisted interpolation scheme is in place at the decoder, the encoder can skip more frames than usual. Therefore, more bits can be assigned to each frame that is coded so that image quality and lip synchronization can be improved. For better lip synchronization, frame skipping can be controlled also by the perceptual importance of mouth shapes and ease of speech-assisted interpolation at the decoder. The encoder can avoid skipping frames that are crucial for lip synchronization, for example, frames that contain mouth closures. There are a number of ways to parameterize speech waveforms. For example, linear predictive coefficients and line spectral pairs are often used for coding purposes, and filter-bank outputs are used for recognition purposes. The question is how to analyze the visual signal, that is, the lip movement. Unlike the speech signal, which is one-dimensional (1D), the visual input is a 3D video signal with two spatial dimensions and one temporal dimension. A visual analysis system must convert this sequence of images into a meaningful parameter. We next discuss the enabling technologies for audio-visual research, including lip tracking.

2.7 Lip Tracking

A number ways to parameterize speech waveforms have been developed. As an example, linear predictive coefficients and like-spectral pairs are often used for coding purposes, and filter bank outputs are used for recognition purposes. In connection with this fact, the question often arises as to how to analyze the visual signal, that is, the lip movement. Unlike the speech signal, which is essentially 1D, the visual input is a 3D video signal with two spatial dimensions and one temporal dimension. A visual analysis system must convert this sequence of images into a meaningful parameter. Visual analysis systems can be divided into two major classes. The first classifies the mouth image into one of several categories, for example, into visemes. The other measures parameters or dimensions from the input image, for example, the mouth height and the mouth width. For the first class, vector quantization and neural networks are standard methods for classifying input images into several categories. For these systems, intensity images, Fourier transform coefficients and thresholded binary images are often used as the input. For the second class of image analysis systems, the task is to obtain parameters or dimensions that have some physical significance. For instance, we may want to measure the height between the lips and the width between the corners of the mouth. The next task is to construct a model for the lips and to find the parameters of the model that provide the closest match between the model and the image [2.1].

The system used in Pentajan [2.32] for visual speech recognition took an input image and applied a threshold. The resulting binary images were then analyzed, and parameters, such as the area of the mouth opening, height, width and perimeter length, were extracted to provide an adequate representation of the shape of the mouth. In another system [2.33], the vertical and horizontal projections of both intensity and edge images were used to locate points of interest on the mouth. The distances between these points were successfully used in speech-recognition applications.

The system in Rao and Mersereau [2.30] uses state-embedded deformable templates, a variation of deformable templates that exploits statistical differences in color to track the shape of the lips through successive video frames. Assume that, based on pixel colors, the image can be divided into foreground (pixels within the outer contour of the lips) and background (pixels that are part of the face) regions. The shape of the foreground is modeled by a template composed of two parabolas, as shown in Figure 2.6. This template is specified by the five parameters, a, b, c, d and e. When the parameters change, the shape and position of the template change. This template divides the image into foreground and background regions. Last, we assume that distinct Probability Density Functions (*pdfs*) govern the distribution of pixel colors in the foreground and background. The assumption is valid because the lips and face have different colors.

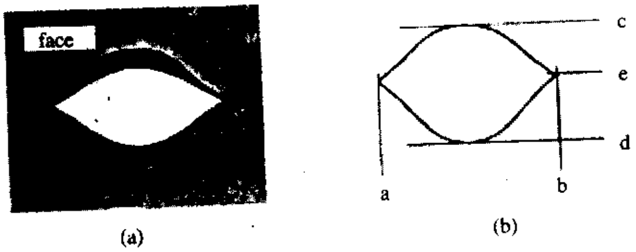


Figure 2.6 Lip tracking: (a) The face image and (b) the template [2.1]. ©1998 IEEE.

The *pdf* for the foreground pixels (the lips and interior of the mouth) and the *pdf* for the background pixels (the face) are first estimated. Based on these two *pdfs*, the joint probability of all pixels in the image can be calculated. Then, the visual analysis system uses a maximization algorithm to find the template that maximizes the joint probability. Sample results from this tracking algorithm are presented in Figure 2.7. Images in the top row show the derived template overlaid on the original image. The bottom images show pixels that are more likely to be part of the foreground.

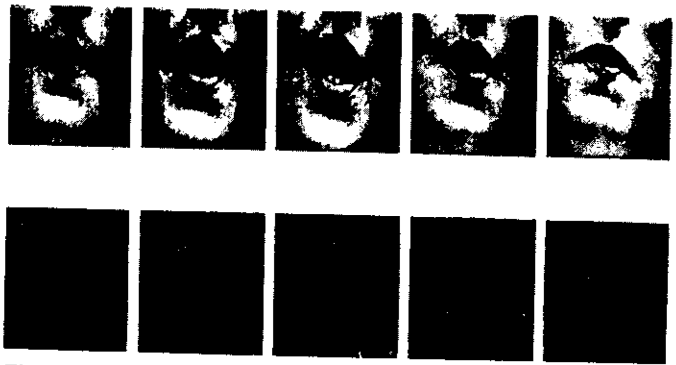


Figure 2.7 Lip tracking results [2.1]. ©1998 IEEE.

It has been suggested that geometric features like mouth shapes could be combined with other image-based features like the Karhunen-Loeve Transform (KLT) of the mouth image. In [2.34], the result of the KLT of the gray-scale mouth image to assist the tracing of mouth shapes is used. The result [2.35] directly combined mouth shapes with KLT for lip reading with color video.

Motivated by applications in surveillance and in human computer interfaces, an increasing amount of research has been done on the analysis of human body motion from video. As an example, human body tracking is presented in Figure 2.8, with a) segmentation of two walking human bodies into heads and legs, b) detection of four independently moving humans in an outdoor scene, and c) display of the heads, bodies, hands and feet of the humans in b) [2.36].

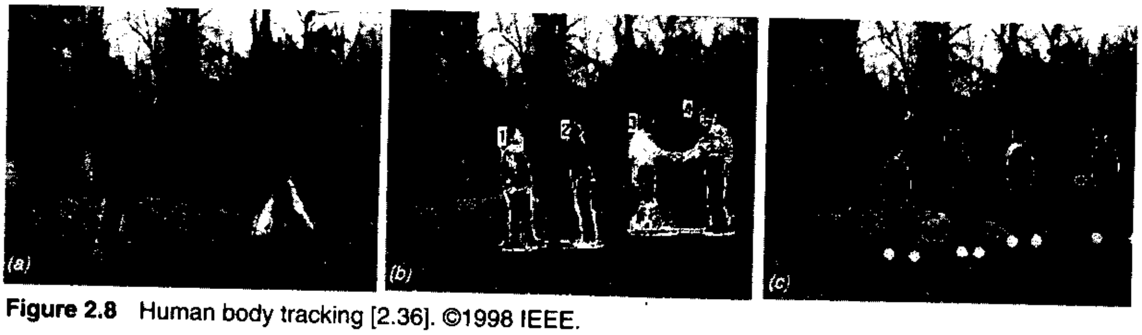


Figure 2.8 Human body tracking [2.36]. ©1998 IEEE.

Based on computer-vision techniques for tracking lip movements of a speaking person, a computer can be trained to understand visual speech [2.37]. One key issue in bimodal speech analysis and synthesis is the establishment of the mapping between acoustic and mouth-shape parameters. It means that, given the acoustic parameters, such as filter-bank coefficients, we need to estimate the corresponding mouth shape and vice versa. Thus, the next section seeks to provide audio-to-visual mapping, or the task of converting acoustic speech to mouth-shape parameters.

2.8 Audio-to-Visual Mapping

The problem for converting acoustic speech to mouth shape parameters can be solved in two different ways. The first one stresses that speech is a linguistic entity. Namely, the speech is first segmented into a sequence of phonemes. Then, each phoneme is mapped to the corresponding viseme. This scheme could be implemented using a complete speech recognizer followed by a lookup table [2.29]. The advantage of this approach is that the acoustic speech signal is explored to the full extent so that all the context information is used, and coarticulations are completely incorporated. Therefore, this approach provides the most precise speech analysis. Also, this approach has a certain amount of computation overhead because we do not really need to recognize the spoken words or sentences in order to achieve audio-to-visual mapping. The construction of the lookup table that maps phonemes to visemes is not trivial. Because a physical relationship exists between the shape of the vocal tract and the sound that is produced, a functional relationship may exist between the speech parameters and the visual parameters set. The conversion problem becomes one of finding the best functional approximation given sets of training data. Many algorithms can be modified to perform this task. These approaches include vector quantization (VQ), neural networks (NNs) and HMMs with Gaussian mixtures [2.5, 2.6, 2.8].

2.8.1 Classification-Based Conversion

An approach to classification-based conversion is given as a block scheme in Figure 2.9. It contains two stages. In the first one, the acoustics must be classified into one of a number of classes. The second stage maps each acoustic class into a corresponding visual output. In the first stage, VQ can be used to divide the acoustic training data into a number of classes. For each acoustic class, the corresponding visual code words are then averaged to produce a visual centroid. Thus, each input acoustic vector would be classified using the optimal acoustic vector quantizer and would then be mapped to the corresponding visual centroid. One shortcoming with this approach is the error that results from averaging visual feature vectors together to form the visual centroids. Another problem invoked by applying the classification-based method is that it does not produce a continuous mapping, but rather produces a distinct number of output levels. A few application examples exist that apply temporal neural models to conversion and/or synchronization including HMMs for audio-to-visual conversation and another example using TDNNs.

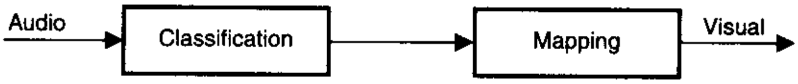


Figure 2.9 Block scheme of a classified-based approach.

2.8.2 HMM for Audio-to-Visual Conversion

HMMs have been used in speech recognition for many years. Although the majority of speech-recognition systems train HMMs on acoustic parameter sets, they can also be used to model the visual parameter sets. Recent multimedia results exploit the audio-visual interaction, which includes speech-assisted lip synchronization and joint audio-video coding [2.38]. The goal of speech-driven facial animation is to synthesize realistic video sequence from acoustic speech.

Chen and Rao [2.38] accomplished the audio-to-visual conversion process with HMMs. The correlation between audio and video was exploited for speech-driven facial animation. One problem addressed is that frame skipping due to limited bandwidth commonly introduces artifacts, such as jerky motion and loss of lip synchronization in talking-head video.

Consider estimating a single visual parameter v from the corresponding multidimensional acoustic parameter a. Defining the combined observation to be

$$O = [a, v]^T \tag{2.1}$$

the audio-to-visual conversion process using HMMs can be treated as a missing data problem. More specifically, a continuous-density HMM was trained with a sequence of O for each word in the vocabulary. In the training phase, the Gaussian mixtures in each state of the HMM are modeling the joint distribution of the audio-visual parameters. When presented with a sequence of acoustic vectors that correspond to a particular word, conversion can be made by using the HMM to segment the sequence using the Viterbi algorithm [2.39]. More exactly, when presented with a sequence of acoustic vectors $\{a\}$ that correspond to a particular word, the maximum likelihood estimator for the associated visual parameter vectors $\{v\}$ is equal to the conditional expectation, which can be derived from the optimal state sequence using the Viterbi algorithm of the HMM.

2.8.3 Audio and Visual Integration for Lip-Reading Applications

Due to the maturity of digital-video technology, it is now feasible to incorporate visual information in the speech-understanding process, that is, lip reading. These new approaches offer effective integration of visually derived information into the state-of-the-art speech-recognition systems so as to gain an improved performance in noise without suffering degraded performance on clean speech.

A complete audio-visual lip reading system can be decomposed into three major components [2.40]:

- Audio-visual information preprocessing (explicit feature extraction from audio and visual data)

- Pattern-recognition strategy (hidden Markov modeling, pattern matching with dynamic or linear time warping and various forms of NNs)
- Integration strategy (decision from audio and visual signal recognition)

2.8.4 Audio-Visual Information Preprocessing

Audio information processing has been well discussed in the speech-recognition literature [2.40]. The digitized speech is commonly sampled at 8 KHz. The sampled speech is pre-emphasized and then blocked and Hamming windowed into frames with a fixed-time interval (say, 32 ms long) and with some overlap (say, 16 ms). For each frame, an N-dimensional feature vector is extracted. Two major types of visual features are useful for lip reading: contour-based and area-based features [2.41, 2.42, 2.43]. The active contour models are a good example of contour-based features, which have been applied to object contours found in many image-analysis problems. Principal Component Analysis (PCA) of a gray-level image matrix, which is a typical area-based method, has been successfully used for principal feature extraction in pattern-recognition problems [2.41, 2.42].

2.8.5 Pattern-Recognition Strategies

Most lip-reading systems used similar pattern-recognition strategies as the traditional speech recognition, such as dynamic time warping and HMMs [2.44, 2.45]. NNs can also be used to convert acoustic parameters into visual parameters. In the training phase, input and output patterns are presented to the network. An algorithm called back propagation can be used to train the network weights. The design choice lies in selecting a suitable topology for the network. The number of hidden layers and the number of nodes per layer can be experimentally determined. A single network can be trained to reproduce all the visual parameters, or networks can be trained with each network estimating a single visual parameter [2.46]. NN architectures have also been extensively explored, such as static-feed forward-back propagation networks [2.47], multistage TDNNs [2.48] and the HMM recognizer with NNs for observation probability calculation [2.49]. One project [2.34] combines the acoustic and visual features for effective lip reading. Instead of using NNs as the temporal sequence classifier, the HMM is adopted, and Multilayer Perceptron (MLP) is used to calculate the observation probabilities {P(phoneIaudio,visual)}. The system combines the ten-order PCA transform coefficients (and/or the deltafeatures) from a gray-level eigen lip matrix (instead of the PCA from the snake points) derived from the video data of the acoustic features from the audio data [2.50]. The discriminatively trained MLP as one of the most popular NN models is used to compute the observation probabilities needed by the Viterbi algorithm. The bimodal hybrid speech recognition system has already been applied to a multispeaker spelling task and to a speaker-independent spontaneous speech-recognition system.

2.8.6 Integration Strategy

The audio and visual features can be combined into one vector before pattern recognition. Then, the decision is solely based on the results of the pattern recognizer. In some lip-reading systems that perform independent visual and audio evaluations, some rule is required to combine the two evaluation scores into a single one. Typical examples include the use of heuristic rules to incorporate knowledge of the relative confusability of phonemes in the evaluation of two modalities. Others used a multiplicative combination of independent evaluation scores for each modality. These postintegration methods possess the advantages of conceptual and implementional simplicity and they also give the user the flexibility to use just one of the subsystems if desired.

Example 2.13 Let us compare the results of the HMM-based method and the NNs based method in considering audio-to-visual mapping. Lip height versus time is observed. Figure 2.10 shows the results of the HMM-based method. The dotted line represents the height variation of the mouth when speaking a particular phrase. The solid line represents the estimation.

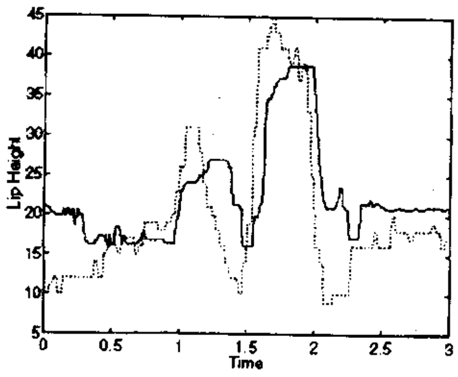


Figure 2.10 The results using HMM method [2.1]. ©1998 IEEE.

Figure 2.11 represents the results of the same phrase using the NN approach. As can be seen, the HMM-based approach gives better approximation to the original waveform.

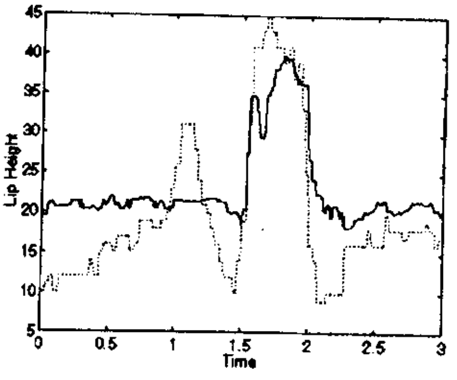


Figure 2.11 The results using NN approach [2.1]. ©1998 IEEE.

2.9 Bimodal Person Verification

Audio-visual interaction can also be used for person verification. Existing methods for person verification are mainly based on either face image or voice [2.1]. Using each single modality has certain limitations in both security and robustness. Using still images alone can be ineffective because it is easy to store and use prerecorded images. Image-only person verification can also suffer from image-coding artifacts and variations in lighting conditions. On the other hand, use of voice only for verification is not reliable because it is possible to rearrange phonemes from a prerecorded speech of a person to synthesize different phrases. In addition, voice-only systems may fail when the acoustic environment is noisy or contains echo, such as in a typical office environment. Joint use of voice and video can solve these problems. By combining these two modalities, we can obtain more secure and more robust person-verification systems.

A number of techniques use lip movement together with acoustic speech to identify or verify a person. PCs with multimedia capabilities (cameras and microphones) make these techniques attractive. During the registration phase, the user says a chosen phrase while the voice and lip movements of the user are recorded into a database. During the verification phase, the user is then asked to read the displayed phrase. The user's voice and video data are then compared with those in the database to verify the user [2.51]. Lip movement has been used mainly for speech recognition and not for speaker verification until recently. Luettin, Thacker and Beet [2.52] showed that lip movement also contained information about a person's identity.

Figure 2.12 demonstrates the time variations of the mouth height of two persons who each say "Hello! How are you?" two times. The lip movements while saying the same phrase vary a great deal from individual to individual, but they stay relatively consistent for the same person. With dynamic time warping [2.2], a technique commonly used in acoustic-based speaker verification, to match the features, the scores of match and no match could diffe. by a factor of more than 40.

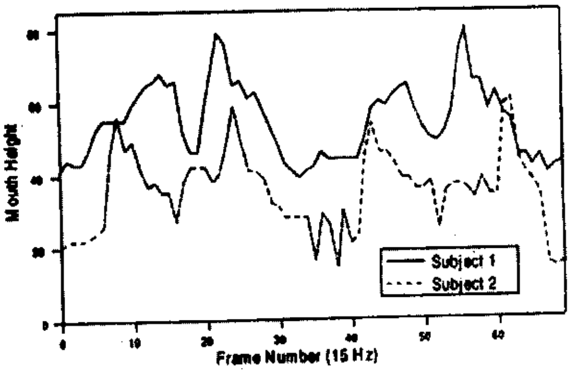


Figure 2.12 Time variation of mouth height [2.1]. ©1998 IEEE.

2.10 Joint Audio-Video Coding

Audio-visual information can be exploited in many ways. The correlation between audio and video can be used to achieve more efficient coding of both audio and video. One example to exploit this correlation is a predictive coding manner. Predictive-coding of video uses information from video frames to help construct an estimate of the current frame. The difference between the original and estimated signals can then be transmitted to allow the receiver to reconstruct the original video frame. This method is useful for removing the temporal redundancy in video. Also, the prediction can be done in a cross-modal manner to explore cross-modal redundancy. The basic idea is that here is information in the acoustic signal that can be used to help predict what the video signal should look like. Because the acoustic data is also transmitted, the receiver is able to reconstruct the video with very little side information. The process of cross-modal prediction is shown in Figure 2.13. This system provides a coding scheme that is scalable to a wide range of bit rates. An acoustic-to-visual mapping module estimates a visual parametric set, such as mouth height and width, given the acoustic data. The image analysis module measures the actual parameter set from the video.

The measured parameter set is compared with the parameter set estimated from the acoustics, and the encoder decides what information must be sent. If the acoustic data lead to a good prediction, no data has to be sent. If the prediction is slightly off, an error signal can be sent. If the prediction is wrong, the measured parameter set can be sent directly. The decision of what information needs to be sent is based on the Rate Distortion (R-D) criteria.

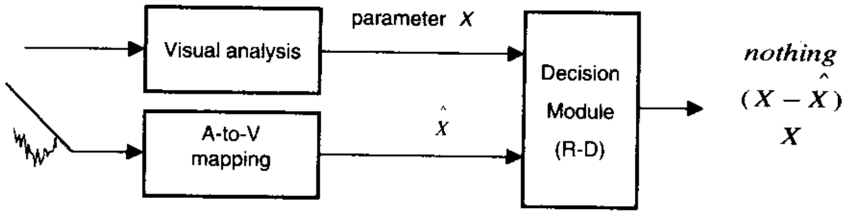


Figure 2.13 Block scheme of the cross-modal predictive coding [2.1]. ©1998 IEEE.

2.11 Concluding Remarks

The joint processing of audio and video provides additional capabilities that are not possible when audio and video are studied separately. Future communication will place a major emphasis on media integration for human communication. Multimedia systems can achieve their potential only when they are truly integrated in three key ways: integration of content, integration of human users and integration with other media systems [2.53]. A prime example of the audio-visual interaction due to the bimodality in speech perception is lip reading. It is not only used by the hearing impaired for speech understanding. In fact, everyone uses lip reading to some extent, in particular in a noisy environment. Based on computer-vision techniques for tracking lip movements of a speaking person, a computer can be trained to understand visual speech. In addition, automatic lip reading has also been used to enhance acoustic speech recognition.

If the frame rate is not adequate for lip-synchronization perception, one solution is to extract the information from the acoustic signal that determines the corresponding mouth movements and then process the speaker's mouth image accordingly to achieve lip synchronization [2.29]. On the other hand, it is also possible to warp the acoustic signal to synchronize with the person's mouth movements. This approach is very useful in non-real-time applications, such as dubbing in a studio. One key issue in bimodal speech analysis and synthesis is the establishment of the mapping between acoustic parameters and mouth shape parameters. In other words, given the acoustic parameters, one needs to estimate the corresponding mouth shape and vice versa. A number of approaches have been proposed for this task that use VQ, NNs and HMMs.

Audio-visual interaction can be exploited in many other ways. One of the characteristic examples is person verification. The correlation between audio and video can be used to achieve more efficient coding of both audio and video [2.5]. Other applications include dubbing of movies, segmentation of image sequences using video and audio signals [2.54], human-computer interfaces and cartoon animation. Also, joint use of audio and video has been applied to multimedia content classification [2.55, 2.56, 2.57, 2.58]. Further discussion of different aspects concerning audio-visual integration in multimedia communication is given in other research [2.39, 2.59, 2.60, 2.61, 2.62, 2.63, 2.64, 2.65, 2.66, 2.67, 2.68, 2.69, 2.70]. Multimedia Signal Processing (MMSP) provides additional capabilities that are not possible when audio and video are studied separately [2.39]. After we break down the artificial boundary between audio/speech and image/video processing, many new research opportunities and applications will arise. Thus, in the next chapter, we deal with the past, present and future of MMSP.