Chapter 9

Multimodality

9.1 Overview

James L. Flanagan

Rutgers University, Piscataway, New Jersey, USA

9.1.1 Natural Communication with Machines

The human senses—evolved in primitive times primarily for survival—serve modern man as exquisitely-developed channels for communication and information exchange. Because the sensory modalities are highly learned and natural, we seek to endow machines with the ability to communicate in these terms. Complex machines can thereby be brought to serve human needs easier and more widely. Sensory realism, similar to face-to-face communication among humans, is the long-range objective.

Of the senses, sight and sound have been exploited to the greatest extent for human/machine communication. Technologies for image processing and voice interaction are deploying rapidly. But, understanding of the touch modality is advancing, as tactile and manual interfaces develop. The dimensions of taste and smell are yet to be harnessed broadly. Advocates of Virtual Reality are sure to be at the forefront of research in this direction, as the search for sensory realism progresses.

The human is adept at integrating sensory inputs, and fusing data to meet needs of the moment. Machines, to date, are less able to emulate this ability. This issue is central to current research in multimedia information systems. But, the human ability to process information appears limited to rates that are small in comparison to the transport capacities of modern fiber optic networks.

Experiments (Keidel, 1968; Pierce, 1961; Cherry, 1957) place the human processing capacity for assimilating and reacting to sensory input at the order of 100 bits/sec., or less. But the human's ability to switch and allocate this processing power across modalities seems to reflect a refined mechanism for data fusion. Again, machines do not yet approach this ability.

In the domains of sight, sound, and touch, technologies have developed enough that experimental integration is now being studied. Because the constituent technologies are imperfect, task-specific applications domains represent the most prudent opportunities for realizing synergistic combinations of modalities. Because of performance limitations, careful systems design and human factors analyses are critical. Further, because advances in microelectronics are now providing vast and economical computation, the characteristics of human perception can to greater extent be incorporated in the information processing, resulting in added economies of transmission and storage.

In discharging the duties of an overview, we propose to comment briefly on activities in image, voice and tactile interfaces for human/machine communication. Additionally, we point up issues in data networking, distributed databases, and synergistic integration of multiple modalities in information systems.

9.1.2 Image Compression and Spatially Realistic Displays

Image signals (in contrast to speech) can be of a great variety, ranging from the teeming crowd of a sports contest to a stationary pastoral country scene. Few constraints exist on the source, and most opportunities for efficient digital representation stem from limitations in the ability of the human eye to resolve detail both spatially and temporally. The eye's sensitivity to contrast is greatest for temporal frequencies of about 4 to 8 Hz, and for spatial frequencies of about 1 to 4 cycles/degree (Figure 9.1, adapted from Netravali & Haskel, 1988).

Decomposition of moving images by temporal and spatial filtering into subbands therefore offers opportunity to trade on the eye's acuity in assigning digital representation bits, moment by moment (Podilchuk & Farvardin, 1991; Podilchuk, Jayant, et al., 1990). That is, available transmission capacity is used for those components representing the greatest acuity, hence maintaining the highest perceptual quality for a given transmission rate. Luminance dominates the transmission requirement with chrominance requiring little additional capacity.

Additionally, while video signals are not as *source-constrained* as speech, they

Figure 9.1: Contrast sensitivity of the human eye.

nevertheless tend to be more low-pass in character (that is, the ratio of the frequency of the spectral centroid to the frequency of the upper band edge is typically smaller for video than for speech). As a consequence the signal is amenable to efficient coding by linear prediction, a simple form of which is differential pulse-code modulation (DPCM). Compression advantages over ordinary digital representation by PCM range in the order of 50:1 for television-grade images. It is thus possible, with economical processing, to transmit television grade video over about 1.5 Mbps capacity, or to store the signal efficiently on conventional CD-ROM.

For conferencing purposes, where image complexity typically involves only *head and shoulders*, the compression can be much larger, exceeding 100:1. Quality approximating that of VHS recording can be achieved and coded at less than 0.1 bit/pixel, which permits real-time video transmission over public switched-digital telephone service at 128 kbps.

For high-quality color transmission of still images , sub-band coding permits good representation at about 0.5 bit/pixel, or 125 kbits for an image frame of 500 x 500 pixels.

Spatial realism is frequently important in image display, particularly for interactive use with gesture, pointing or force feedback data gloves. Stereo display can be achieved by helmet fixtures for individual eye images, or by electronically-shuttered glasses that separately present left and right eye scenes. The ideal in spatial realism for image display might be color motion holography, but the understanding does not yet support this.

9.1.3 Speech Processing

The technologies of automatic speech recognition and speech synthesis from text have advanced to the point where rudimentary conversational interaction can be reliably accomplished for well-delimited tasks. For these cases, speech recognizers with vocabularies of a few hundred words can *understand* (in the task-specific sense) natural connected speech of a wide variety of users (speaker independent). A favored method for recognition uses cepstral features to describe the speech signal and hidden Markov model (HMM) classifiers for decisions about sound patterns. As long as the user stays within the bounds of the task (in terms of vocabulary, grammar and semantics), the machine performs usefully, and can generate intelligent and intelligible responses in its synthetic voice (Figure 9.2) (Flanagan, 1992).

Systems of this nature are being deployed commercially to serve call routing in telecommunications, and to provide automated services for travel, ordering, and financial transactions.

The research frontier is in large vocabularies and language models that approach unconstrained natural speech. As vocabulary size increases, it becomes impractical to recognize acoustic patterns of whole words. The pattern recognizer design is driven to analysis of the distinctive sounds of the language, or phonemes (because there are fewer phonemes than words). Lexical information is programmed to estimate whole words. Systems are now in the research laboratory for vocabularies of several tens of thousand words.

A related technology is speaker recognition (to determine who is speaking, not what is said) (Furui, 1989). In particular, speaker verification, or authenticating a claimed identity from measurements of the voice signal, is of strong commercial interest for applications such as electronic funds transfer, access to privileged information, and credit validation.

Coding for efficient voice transmission and storage parallels the objectives of image compression. But, source constraints on the speech signal (i.e., the sounds of a given language produced by the human vocal tract) offer additional opportunities for compression. But, relatively, speech is a broader bandwidth signal than video (i.e., the

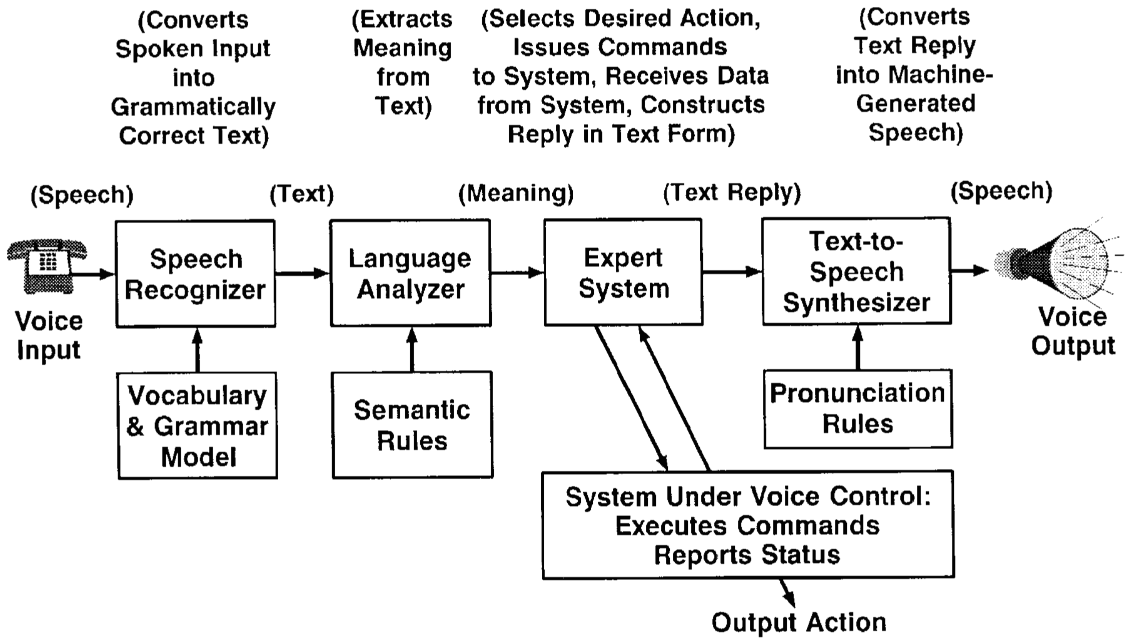

Figure 9.2: Task-specific speech recognition and synthesis in a dialog system.

ratio of centroid frequency to upper band edge is greater). Compression ratios over conventional PCM of the order of 10:1 are becoming possible with good quality. Representation with 1 bit/sample results in an 8 kbps digital representation, and typically utilizes both source and auditory perceptual constraints.

Perceptual coding for wideband audio, such as compact disc quality, is possible through incorporating enough computation in the coder to calculate, moment by moment, auditory masking in frequency and time (Figure 9.3).

A major challenge in speech processing is automatic translation of spoken language. This possibility was demonstrated in concept at an early time by the C&C Laboratory of NEC.[1] More recently, systems have been produced in Japan by ATR for translating among Japanese/English/German, and by AT&T Bell Laboratories and Telefonica de Espana for English/Spanish (Roe, Moreno, et al., 1992).

In all systems to date, vocabularies are restricted to specific task domains, and language

[1] In a major display at Telecom 1983, Geneva, Switzerland, the NEC Corporation provided a concept demonstration of translating telephony. The application scenario was conversation between a railway stationmaster in Japan and a British tourist who had lost her luggage. Real-time, connected speech, translated between Japanese and English, used a delimited vocabulary and *phrase book* grammar.

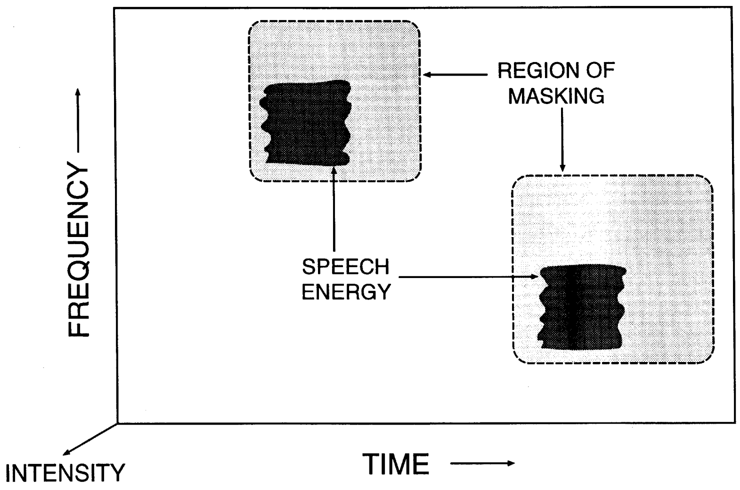

Figure 9.3: Illustration of the time-frequency region surrounding intense, punctuate signals where masking in both time and frequency is effective.

models span limited but usefully-large subsets of natural language.

9.1.4 Tactile Interaction

So far, the sensory dimension of touch has not been applied in human/machine communication to the extent that sight and sound have. This is owing partly to the difficulty of designing tactile transducers capable of representing force and texture in all their subtleties. Nevertheless, systems related to manual input, touch and gesture are receiving active attention in a number of research laboratories (Brooks, Ouh-Young, et al., 1990; Burdea & Coiffet, 1994; Burdea & Zhuang, 1991; Blonder & Boie, 1992; Mariani, Teil, et al., 1992; ICP, 1993). Already, commercial systems for stylus-actuated sketch pad data entry are appearing, and automatic recognition of handwritten characters, substantially constrained, is advancing. Automatic recognition of unrestricted cursive script remains a significant research challenge.

One objective for tactile communication is to allow the human to interact in meaningful ways with computed objects in a virtual environment. Such interaction can be logically combined with speech recognition and synthesis for dialog exchange (Figure 9.4; Burdea & Coiffet, 1994). A step in this direction at the CAIP Center is force feedback applied to a fiber-optic data glove (Figure 9.5; Burdea & Zhuang, 1991). Finger and joint deflections are detected by optical fibers that innervate the glove. Absolute position is sensed by a Polhemus coil on the back wrist. Additionally, single-axis pneumatic

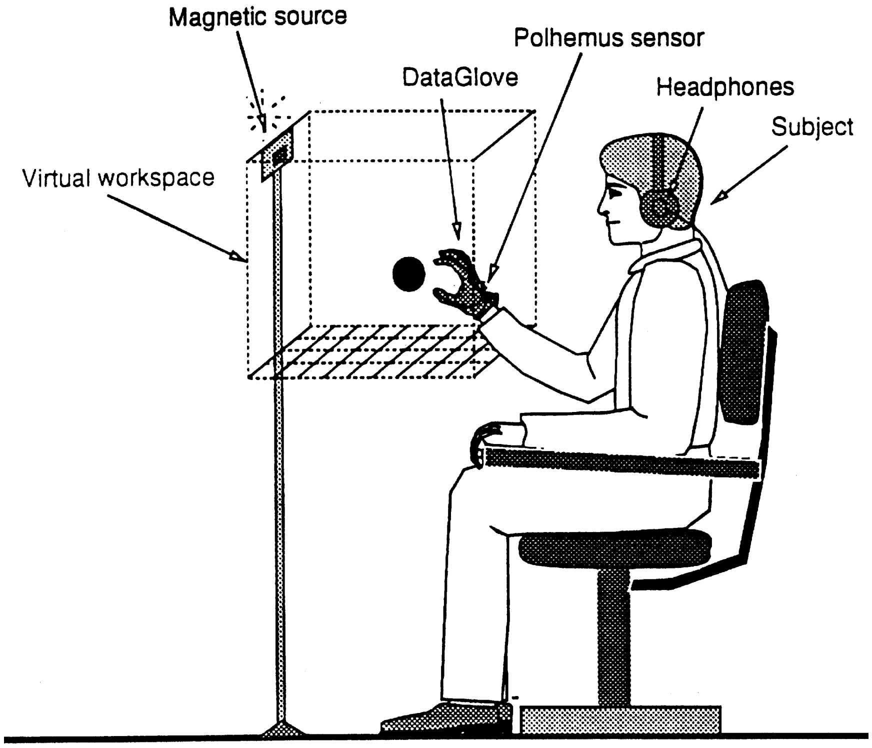

Figure 9.4: With a data glove capable of force feedback, the computer user can use the tactile dimension, as well as those for sight and sound, to interact with virtual objects. A Polhemus coil on the wrist provides hand position information to the computer.

actuators can either apply or sense force at four of the finger tips. While primitive at present, the device allows the user to compute a hypothetical object, put a hand into the data glove and sense the relative position, shape and compliance of the computed object. Research collaboration with the university medical school includes training of endoscopic surgery and physical therapy for injured hands. In contrast, on the amusement side, equipped with glasses for a stereo display, the system offers a challenging game of handball (Figure 9.6). In this case, force feedback permits the user (player) to sense when the ball is grasped, and even to detect the compliance of the ball. More ambitious research looks to devise *smart skins* that can transduce texture in detail.

9.1.5 Data Networking and Distributed Databases

Human/machine communication implies connectivity. In the modern context, this means digital connectivity. And, the eternal challenge in data transport is speed.

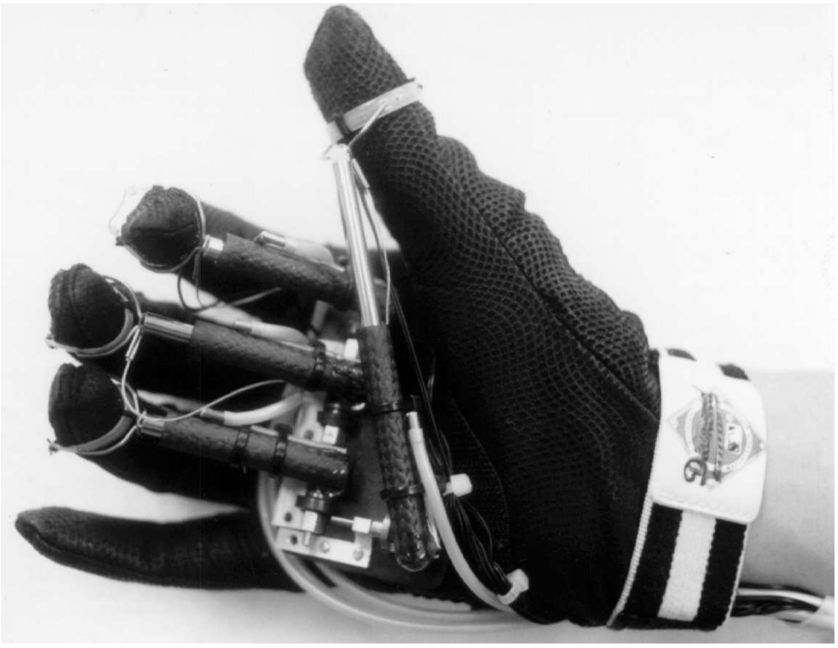

Figure 9.5: CAIP's force feedback transducers for a data glove are single-axis pneumatic thrusters, capable of sensing finger force or, conversely, of applying programmed force sequences to the hand. Joint motion is sensed by optical fibers in the glove, and hand position is measured magnetically by the Polhemus sensor.

Fiber-optic networks, based upon packet-switched Asynchronous Transfer Mode (ATM) technology, are evolving with the aim of serving many simultaneous users with a great variety of information (video, audio, image, text, data). Within some limits, transport capacity can be traded for computation (to provide data compression). Effective data networking must therefore embrace user demands at the terminals, as well as information routing in the network. Particularly in real-time video/audio conferencing, it is important that network loading and traffic congestion be communicated to user terminals, moment by moment, so that compression algorithms can adjust to the available transport capacity without suffering signal interruption through overload.

Research in progress aims to establish protocols and standards for multipoint conferencing over ATM. One experimental system called XUNET (Xperimental University NETwork), spans the U.S. continent (Fraser, Kalmanek, et al., 1992). The network has nodes at several universities, AT&T Bell Laboratories, and several national laboratories (Figure 9.7; Fraser, Kalmanek, et al., 1992). Supported by AT&T Bell Laboratories, Bell Atlantic and the Advanced Research Projects Agency, the network runs presently at DS-3 capacity (45 mbps), with several links already upgraded to 622 mbps. It provides a working testbed for research on multipoint conferencing, shared

Figure 9.6: Using the force feedback data glove, and simulated sound, a virtual handball game is played with the computer. The operator wears glasses to perceive the display in 3-dimensions. Under program control the resilience of the ball can be varied, as well as its dynamic properties. This same facility is being used in medical experiments aimed to train endoscopic surgery and joint palpation.

distributed databases, switching algorithms and queuing strategies. Weekly transcontinental video conferences from laboratory workstations presently are identifying critical issues in network design.

At the same time, public-switched digital telephone transport is becoming pervasive, and is stimulating new work in compression algorithms for video, speech and image. Integrated Services Digital Network (ISDN) is the standardized embodiment, and in its basic-rate form provides two 64 kbps channels (2B channels) and one 16 kbps signaling channel (1D channel).

9.1.6 Integration of Multiple Modalities

Although the technologies for human/machine communication by sight, sound, and touch are, as yet, imperfectly developed, they are, nevertheless, established firmly enough to warrant their use in combination—enabling experimentation on the synergies arising therefrom. Because of the performance limitations, careful design of the

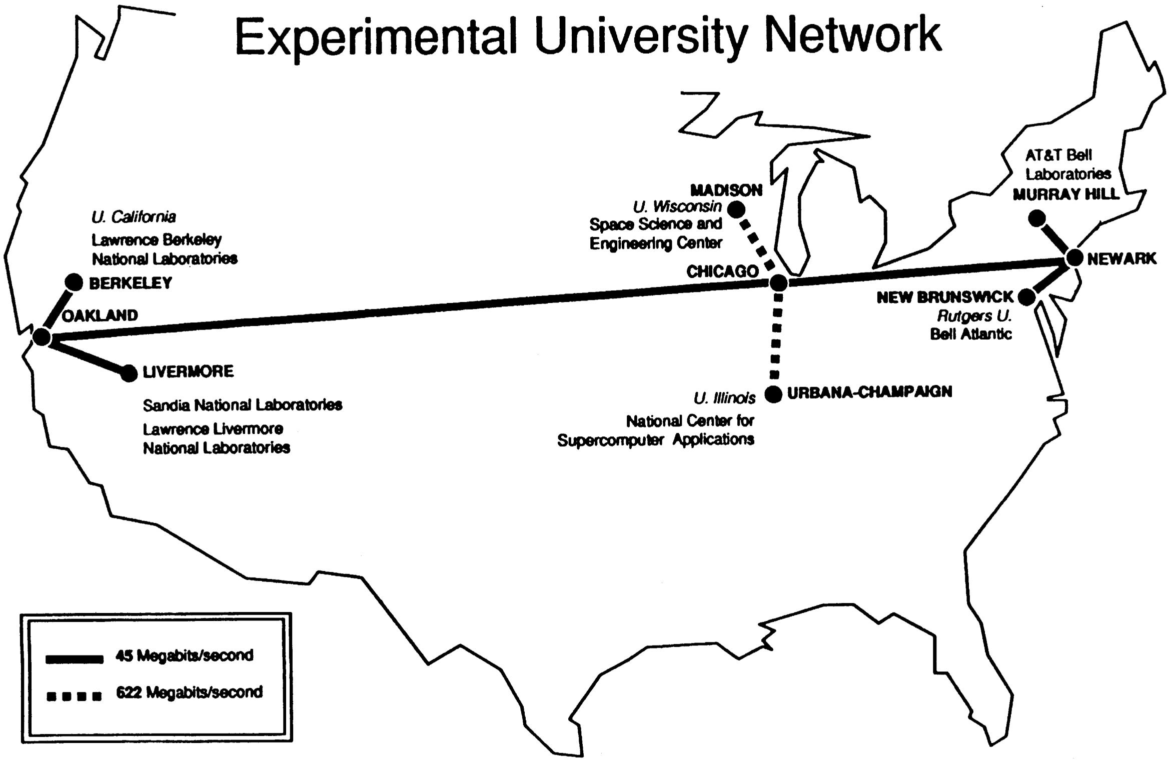

Figure 9.7: Nodes on the Xperimental University NETwork (XUNET).

applications scenario is a prerequisite, as is human factors analysis to determine optimum information burdens for the different modalities in specific circumstances.

Among initial studies on integrating multiple modalities is the HuMaNet system of AT&T Bell Laboratories (Berkley & Flanagan, 1990). This system is designed to support multipoint conferencing over public-switched digital telephone capacity (basic-rate ISDN). System features include: hands-free sound pick up by auto directive microphone arrays; voice-control of call set up, data access and display by limited-vocabulary speech recognition; machine answer-back and cueing by text-to-speech synthesis; remote data access with speaker verification for privileged data, high-quality color still image coding and display at 64 kbps; and wideband stereo voice coding and transmission at 64 kbps. The system uses an a group of networked personal computers, each dedicated to mediate a specific function, resulting in an economical design. The applications scenario is multipoint audio conferencing, aided by image, text and numerical display accessed by voice control from remote databases. Sketch-pad complements, under experimentation, can provide a manual data feature.

Another experimental vehicle for integrating modalities of sight, sound, and touch is a video/audio conferencing system in the CAIP Center (Figure 9.8; Flanagan, 1994). The

Figure 9.8: Experimental video/audio conferencing system in the CAIP Center.

system uses a voice-controlled, near-life-size video display based on the Bell Communications Research video conferencing system. Hands-free sound pickup is accomplished by the same autodirective microphone system as in HuMaNet. The system is interfaced to the AT&T Bell Laboratories fiber-optic network XUNET. Current work centers on communication with HuMaNet. Tactile interaction, gesturing, and handwriting inputs are being examined as conferencing aids, along with automatic face recognition and speaker verification for user authentication. An interesting possibility connected with face recognition includes automatic lip reading to complement speech recognition (Waibel, 1993).

Additionally, inexpensive computation and high-quality electret microphones suggest that major advances might be made in selective sound pick-up under the usually unfavorable acoustic conditions of conference rooms. This is particularly important when speech recognition and verification systems are to be incorporated into the system. Under the aegis of the National Science Foundation, the CAIP Center is examining possibilities for high-quality selective sound capture, bounded in three-dimensions, using large three-dimensional arrays of microphones. Signal processing to produce multiple beamforming (on the sound source and its major multipath images) leads to significant gains (Figures 9.9 and 9.10; Flanagan, Surendran, et al., 1993).

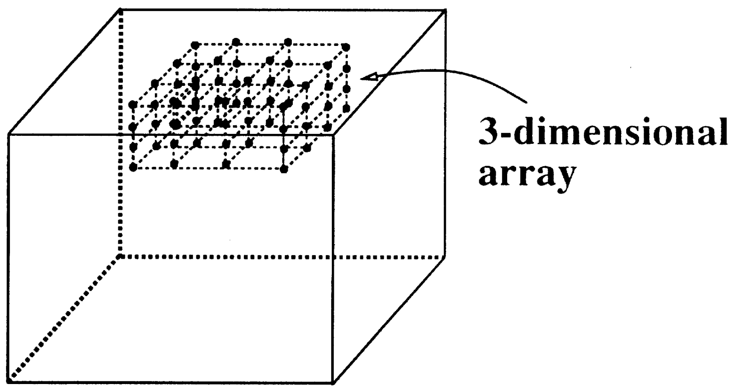

Figure 9.9: Schematic of a cubic 3-dimensional array of microphones for use in conference rooms. The cluster of sensors, which typically is harmonically-nested in space, is positioned as a *chandelier* on the ceiling of the room. Digital processing provides multiple beam forming on the detected sound source and its major images, resulting in mitigation of multipath distortion and interference by noise sources.

Equivalently, matched filtering applied to every sensor of the array provides spatial volume selectivity and mitigates reverberant distortion and interference by competing sound sources (Flanagan, Surendran, et al., 1993).

9.1.7 Future Directions

For quite a few years in the past, algorithmic understanding in processing of human information signals, especially speech, outstripped economies in computing. This is much changed with the explosive progress in microelectronics. Device technologies in the 0.3 μ range, now evolving, promise commercially viable single chip computers capable of a billion operations/sec. This brings closer the possibilities for economical large vocabulary speech recognition and high-definition image coding. Already, we have single chip implementations of low bit-rate speech coding (notably 8 kbps CELP coders for digital cellular telephone, and 16 kbps coders for voice mail coders) which achieve good communications quality. And, we have reliable speaker-independent speech recognizers capable of a few hundred words (Rabiner, 1989; Wilpon, Rabiner, et al., 1990). Even as we approach physical limits for serial processors, the opportunities for massively-parallel

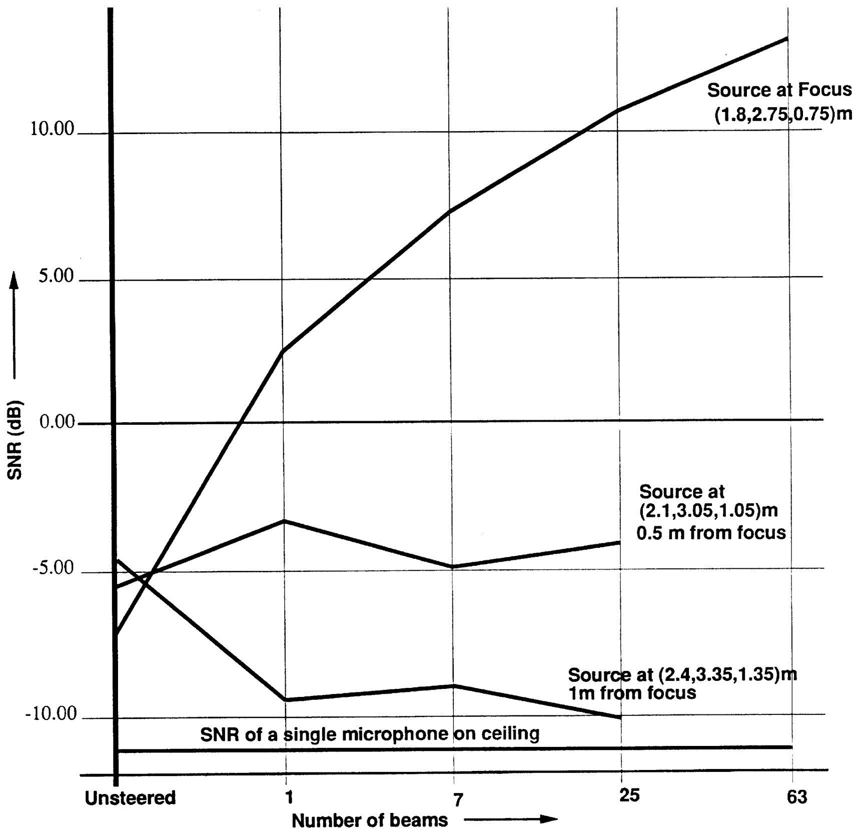

Figure 9.10: Improvements in signal-to-reverberant noise ratio from multiple beam forming on sound source and its major images using a three-dimensional array. The array is 7×7×7 sensors placed at the center of the ceiling of a 7×5×3m room. Comparison is made to the performance of a single microphone.

processing are opening.

This wealth of computation is an additional stimulant to new research in modeling human behavior in interactive communication with machines. To the extent that research can quantify communicative behavior, machines can be made much more helpful if they can understanding the intent of the user and anticipate needs of the moment. Also through such quantification, the machine is enabled to make decisions about optimum sensory modes of information display, thereby matching its information delivery to the sensory capabilities of the human. Automatic recognition of fluent conversational speech, for example, may advance to reliable performance only through good models of spontaneous discourse, with all its vagaries.

For the foreseeable future, success of multimodality systems will depend upon careful design of the applications scenario, taking detailed account of the foibles of the constituent technologies—for sight, sound and touch—and perhaps later taste and smell (Flanagan, 1994). In none of this research does limitation in compute power seem to be the dominant issue. Rather, the challenge is to quantify human behavior for multisensory inputs.

9.2 Representations of Space and Time

Gérard Ligozat

LIMSI-CNRS, Orsay, France

Asserting that most human activities requiring intelligence are grounded in space and time is a commonplace remark. In the context of multimodal environments, spatial and temporal information has to be represented, exchanged, and processed between different components using various modes.

In particular, spatial and temporal knowledge may have to be extracted from natural language and further processed, or general knowledge involving spatial and temporal information may have to be expressed using natural language.

This makes apparent the fact that processing spatial and temporal knowledge in natural language draws upon two main domains:

1. Knowledge representation and reasoning about spatial and temporal knowledge, a branch of Artificial Intelligence, including such aspects as qualitative physics.
2. Theoretical and computational linguistics.

In the first domain, the main goal is upon devising formalisms for representing and reasoning about spatial and temporal information in general.

In the second domain, which is of course closely related to the first, and often considered as a subdomain of application, the main focus is upon spatial and temporal information *in* natural language, understanding its content and meaning, and processing it.

Both aspects interact in many applications having to do at the same time with real world data and linguistic data: story understanding, scene description, route description.

In fact, despite the obvious interest and ultimate necessity of considering both spatial and temporal aspects jointly, the state of development of the branch of AI, computational linguistics or formal philosophy dealing with time is much more advanced than that of the corresponding branches which deal with space.

9.2.1 Time and Space in Natural Language

Understanding temporal and spatial information in natural language—or generating natural language to express temporal or spatial meanings, implies: (1) identifying the linguistic elements conveying this information (markers), (2) elucidating its semantic (and pragmatic) content, (3) devising suitable systems of representation to process this content, (4) implementing and using those representations.

Nature and Contents of the Markers for Time

The basic linguistic markers for temporal notions in many languages are verb *tenses* (e.g., preterite, pluperfect in English) and *temporal adverbials* (yesterday, to-morrow, two weeks from now).

However, tenses also express aspectual notions, which are important for understanding the implications of a given utterance: compare *He was crossing the street* and *He crossed the street.* Only the second sentence allows the inference *He reached the other side of the street.*

A basic property of temporal information in natural language is its deictic nature: an event in a past tense means it happened before the time of speech. Reichenbach introduced the time of speech S, time of reference R, and time of event E to explain the difference of meaning between the basic tenses in English.

Another important component is the fact that verbs behave in various ways according to their semantic class (Aktionsart). Variants or elaborations of Vendler's classification (states, activities, accomplishments and achievements) have been in common use.

Systems of Representation

The definition and study by Prior of modal tense logics has resulted in an important body of work in formal logic, with applications in philosophy (the original motivation of Prior's work), the semantics of natural language, and computer science. The simplest versions of tense logics use modal operators $\mathbf{F}$, $\mathbf{G}$, $\mathbf{P}$, $\mathbf{H}$. For instance, $\mathbf{F}p$ means that p *will be true at some future time*, and $\mathbf{G}p$ that p *will be true at all future times*; $\mathbf{P}p$ and $\mathbf{H}p$ have the corresponding interpretations in the past.

Hence tense logic, in analogy to natural language, uses time as an implicit parameter: a formula has to get its temporal reference *from the outside.*

In Artificial Intelligence, both McDermott (1982) and Allen (1983) introduced reified logics for dealing with temporal information. Reification consists in incorporating part of the meta-language into the language: a formula in the object language becomes a propositional term in the new language. For example, p being true during time t might be written $HOLDS(p, t)$. This allows to make distinctions about different ways of *being* *true*, as well as quantification about propositional terms. A recent survey of temporal reasoning in AI is Vila (1994).

Processing Temporal Information

Typically, recent work on temporal information in natural language uses some or all of the preceding tools. A great deal of work is concerned with determining the temporal value of a given sentence. Good examples are Webber (1988) and Moens and Steedman (1988).

Nature and Contents of the Markers for Space

Primary linguistic markers of space are spatial prepositions (*in, on, under, below, behind, above*) and verbs of motion (*arrive, cross*). The seminal work by Herkovits (1986) showed that prepositions cannot be analyzed in purely geometric terms, and that their interpretation depends heavily on contextual factors. Following initial work by Vandeloise (1986), Vieu, Aurnague and Briffault developed general theories of spatial reference and interpretations of prepositions in French (see references in COSIT, 1993). It appears that spatial information also involves:

- Deictic aspects.
- Functional elements (e.g., the typical functions of objects, such as containment for a cup).
- The physical nature of objects (some objects can stick under a table for instance).
- Pragmatic considerations.

Developing systems for dealing with spatial information is best understood in the larger context of spatial reasoning in AI.

9.2.2 Implementation Issues

A general computational framework for expressing temporal information is in terms of binary constraint networks: Temporal information is represented by a network whose nodes are temporal entities (e.g., intervals), and information about binary relations between entities is represented by labels on the arcs.

In Allen's approach, such a qualitative network will represent a finite set of intervals, and labels on the arcs will be disjunctions of the thirteen basic relations (representing incomplete knowledge). Basic computational problems will be:

1. Determining whether a given network is coherent, i.e., describes at least one feasible scenario.
2. Finding all scenarios compatible with a given network.
3. For a given network, answering the previous questions in case new intervals or constraints are added.

The first two problems are NP-hard in the full algebra of intervals, whereas they are polynomial in the case of time points. Recent results of Nebel and Bürckert (1993) identify a maximal tractable subset.

Most algorithms in this framework are variants of the constraint propagation method first introduced in this context by Allen.

Binary constraint networks also are a suitable representation for representing quantitative constraints between time points. A case in point are *time maps* used by Dean and McDermott (see Vila, 1994).

9.2.3 Future Directions

A promising direction of research in the domain of temporal information in natural language is concerned with the integration of *the textual, or discourse level*. Two basic aspects are:

1. Temporal anaphora: in a given sentence, part of the indexes of reference (S, R, E) are determined by other sentences in the text.
2. Temporal structure: this has to do with determining general principles for the temporal structure of discourse. Nakhimovsky (1988) and Lascarides and Asher (1993) are recent examples.

In the spatial domain, two directions of interest are:

1. The development of an interdisciplinary, cognitively motivated field of research on spatial relations (Mark & Frank, 1991; Frank, Campari, et al., 1992; COSIT,

1993). This combines results from cognitive psychology on the perception and processing of spatial information, research in Artificial Intelligence on the representation of spatial information, as well as research on geographic information systems, which has to deal with spatial information both as numeric information (pixels) and symbolic information, e.g., maps, diagrams.

Typical applications include the generation of route descriptions in natural language, the maintenance and querying of spatial databases, the interpretation and generation of maps describing spatio-temporal phenomena.

2. A trend towards the development of general formalisms for spatial reasoning, including:

   - a critical examination of the parameters of existing proposals, such as the nature of the spatial objects considered (abstract regions, regions in 2-D space, points, vectors, rectangles; oriented or non-oriented objects), their physical properties, the dimension and nature of the ambient space, the presence of global orientations and frames of reference, and the types of relations considered (topological, alignment, directional) (Ligozat, 1993).
   - importation of temporal techniques (understood as 1-D reasoning techniques) into the spatial domain;
   - a development of logical tools (either predicate logics, or modal logics) and an investigation of their formal properties (Aurnague & Vieu, 1993; Cohn, 1993).

9.3 Text and Images

Wolfgang Wahlster

Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, Germany

Text and images are ubiquitous in human communication and there are deep connections between the use of these two modalities. Whereas humans easily become experts for mapping from images to text (e.g., a radio reporter describing a soccer game) or from text to images (e.g., a cartoonist transforming a story into a comic strip), such complex transformations are a great challenge for computer systems. In multimodal communication humans utilize a combination of text and images (e.g., illustrated books, sub-titles for animations) taking advantage of both the individual strength of each communication mode and the fact that both modes can be employed in parallel. Allowing the two modalities to refer to and depend upon each other is a key to the richness of multimodal communication. Recently, a new generation of intelligent multimodal human-computer interfaces has emerged with the ability to interpret some forms of multimodal input and to generate coordinated multimodal output.

9.3.1 From Images to Text

Over the past years, researchers have begun to explore how to translate visual information into natural language (McKevitt, 1994). Starting from a sequence of digitized video frames, a vision system constructs a geometrical representation of the observed scene, including the type and location of all visible objects on a discrete time scale. Then spatial relations between the recognized objects and motion events are extracted and condensed into hypothesized plans and plan interactions between the observed agents. These conceptual structures are finally mapped onto natural language constructs including spatial prepositions, motion verbs, temporal adverbs or conjunctions, and causal clauses. This means in terms of reference semantics, that explicit links between sensory data and natural language expressions are established by a bottom-up process.

While early systems like HAM-ANS (Wahlster, Marburger, et al., 1983), LandScan (Bajcsy, Joshi, et al., 1985), and NAOS (Neumann, 1989) generated retrospective natural language scene descriptions after the processing of the complete image sequence, current systems like VITRA (Wahlster, 1989) aim at an incremental analysis of the visual input to produce simultaneous narration. VITRA incrementally generates reports about real-world traffic scenes or short video clips of soccer matches. The most challenging open question in this research field is a tighter coordination of perception and language production by integrating the current bottom-up cascaded architectures with top-down and expectation-driven processing of images, such that text production can influence the low-level vision processing, e.g., by focusing on particular objects and by providing active control of the vision sensors.

A great practical advantage of natural language image description is the possibility of the application-specific selection of varying degrees of condensation of visual information. There are many promising applications in medical technology, remote sensing, traffic control and other surveillance tasks.

9.3.2 From Text to Images

Only a small number of researchers have dealt with the inverse direction, the generation of images from natural language text. The work in this area of natural language processing has shown how a physically based semantics of motion verbs and locative prepositions can be seen as conveying spatial, kinematic and temporal constraints, thereby enabling a system to create an animated graphical simulation of events described by natural language utterances.

The AnimNL project (Badler, Phillips, et al., 1993) aims to enable people to use natural language instructions as high-level specifications to guide animated human figures through a task. The system is able to interpret simple instructional texts in terms of intentions that the simulated agent should adopt, desired constraints on the agent's behavior and expectations about what will happen in the animation. In the ANTLIMA system (Schirra & Stopp, 1993) the generation of animations from text is based on the assumption that the descriptions always refer to the most typical case of a spatial relation or motion. Typicality potential fields are used to characterize the default distribution for the location and velocity of objects, the duration of events, and the temporal relation between events. In ANTLIMA and the SPRINT system (Yamada, Yamamoto, et al., 1992) all objects in the described scene are moved to a position with maximal typicality using a hill-climbing algorithm. If the location of an object is described by several spatial predications holding simultaneously, the algebraic average of the corresponding typicality distributions is used to compute the position of the object in the animation.

There is an expanding range of exciting applications for these methods like advanced simulation, entertainment, animation and computer aided design (CAD) systems.

9.3.3 Integrating Text and Images in Multimodal Systems

Whereas mapping images to text is a process of abstraction, mapping text to images is a process of concretion (Figure 9.11). However, in many situations the appropriate level of detail can only be achieved by a combination of text and images.

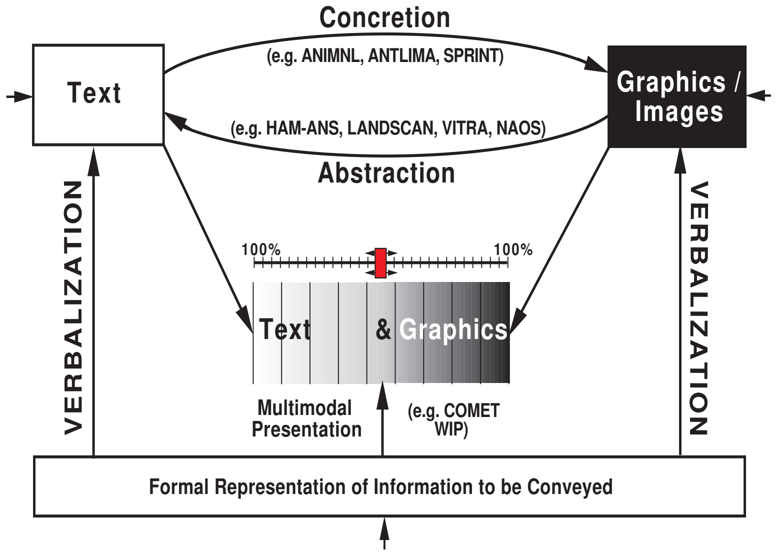

Figure 9.11: Generating and Transforming Presentations in Different Modes and Media.

A new generation of intelligent multimodal systems (Maybury, 1993) goes beyond the standard canned text, predesigned graphics and prerecorded images and sounds typically found in commercial multimedia systems of today. A basic principle underlying these so-called intellimedia systems is that the various constituents of a multimodal communication should be generated on the fly from a common representation of what is to be conveyed without using any preplanned text or images. It is an important goal of such systems not simply to merge the verbalization and visualization results of a text generator and a graphics generator, but to carefully coordinate them in such a way that they generate a multiplicative improvement in communication capabilities. Such multimodal presentation systems are highly adaptive, since all presentation decisions are postponed until runtime. The quest for adaptation is based on the fact that it is impossible to anticipate the needs and requirements of each potential user in an infinite number of presentation situations.

The most advanced multimodal presentation systems, that generate text illustrated by 3-D graphics and animations, are COMET (Feiner & McKeown, 1993) and WIP

(Wahlster, André, et al., 1993). COMET generates directions for maintenance and repair of a portable radio and WIP designs multimodal explanations in German and English on using an espresso-machine, assembling a lawn-mower, or installing a modem.

Intelligent multimodal presentation systems include a number of key processes: content planning (determining what information should be presented in a given situation), mode selection (apportioning the selected information to text and graphics), presentation design (determining how text and graphics can be used to communicate the selected information), and coordination (resolving conflicts and maintaining consistency between text and graphics).

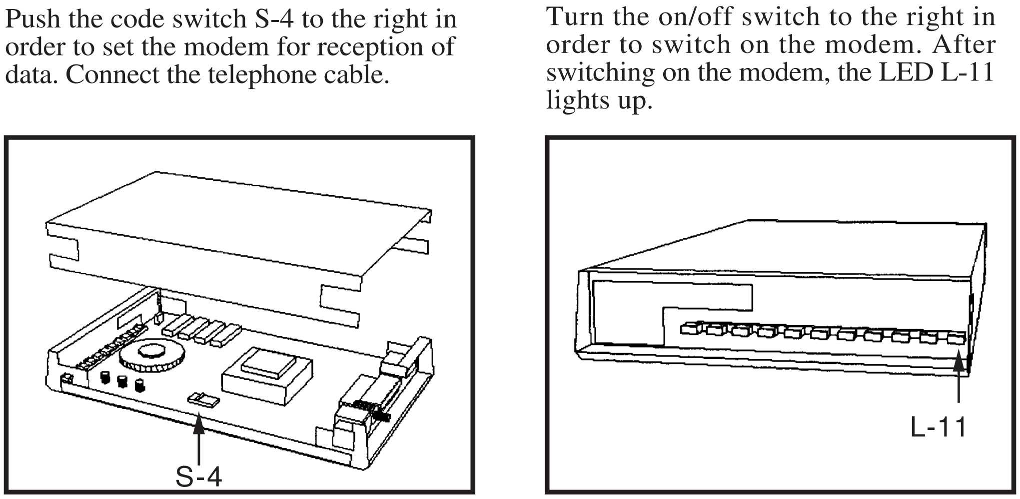

Figure 9.12: A text-picture combination generated by the WIP-System.

An important synergistic use of multimodality in systems generating text-picture combinations is the disambiguation of referring expressions. An accompanying picture often makes clear what the intended object of a referring expression is. For example, a technical name for an object unknown to the user may be introduced by clearly singling out the intended object in the accompanying illustration (Figure 9.12). In addition, WIP and COMET can generate cross-modal expressions like, "The on/off switch is shown in the upper left part of the picture," to establish referential relationships of representations in one modality to representations in another modality.

The research so far has shown that it is possible to adapt many of the fundamental concepts developed to date in computational linguistics in such a way that they become useful for text-picture combinations as well. In particular, semantic and pragmatic concepts like communicative acts, coherence, focus, reference, discourse model, user model, implicature, anaphora, rhetorical relations and scope ambiguity take on an extended meaning in the context of multimodal communication.

9.3.4 Future Directions

Areas which require further investigation include the question how to reason about multiple modes so that the system becomes able to block false implicatures and to ensure that the generated text-picture combination is unambiguous, the role of layout as a rhetorical force, influencing the intentional and attential state of the viewer, the integration of facial animation and speech of the presentation agent, and the formalization of design knowledge for creating interactive presentations.

Key applications for intellimedia systems are multimodal helpware, information retrieval and analysis, authoring, training, monitoring, and decision support.

9.4 Modality Integration: Speech and Gesture

Yacine Bellik

LIMSI-CNRS, Orsay, France

Speech and gestures are the expression means which are the most used in communication between human beings. Learning of their use begins with the first years of life. Therefore they should be the modalities to be privileged in communicating with computers (Hauptmann & McAvinney, 1993). Compared to speech, research that aims to integrate gesture as an expression mean (not only as an object manipulation mean) in Human-Computer Interaction (HCI) has recently began. These works have been launched thanks to the appearance of new devices, in particular datagloves which allow us to know about the hand configuration (flexing angles of fingers) at any moment and to follow its position into the 3D space.

Multimodality aims not only at making several modalities cohabit in an interactive system, but especially at making them cooperate together (Coutaz, Nigay, et al., 1993; Salisbury, 1990) (for instance, if the user wants to move an object using a speech recognition system and a touch screen as in Figure 9.13, he has just to say *put that there* while pointing at the object and at its new position; Bolt, 1980).

In human communication, the use of speech and gestures is completely coordinated. Unfortunately, and at the opposite of human communication means, the devices used to interact with computers have not been designed at all to cooperate.

For instance, the difference between time responses of devices can be very large (a speech recognition system needs more time to recognize a word than a touch screen driver to compute the point coordinates relative to a pointing gesture). This implies that the system receives an information stream in an order which does not correspond to the real chronological order of user's actions (like a sentence in which words have been mixed up). Consequently, this can lead to bad interpretations of user statements.

The fusion of information issued from speech and gesture constitutes a major problem. Which criteria should we use to decide the fusion of an information with another one, and at what abstraction level should this fusion be done? On the one hand, a fusion at a lexical level allows for designing generic multimodal interface tools, though fusion errors may occur. On the other hand, a fusion at a semantic level is more robust because it exploits many more criteria, but it is in general application-dependent. It is also important to handle possible semantic conflicts between speech and gesture and to exploit information redundancy when it occurs.

Time is an important factor in interfaces which integrate speech and gesture (Bellik,

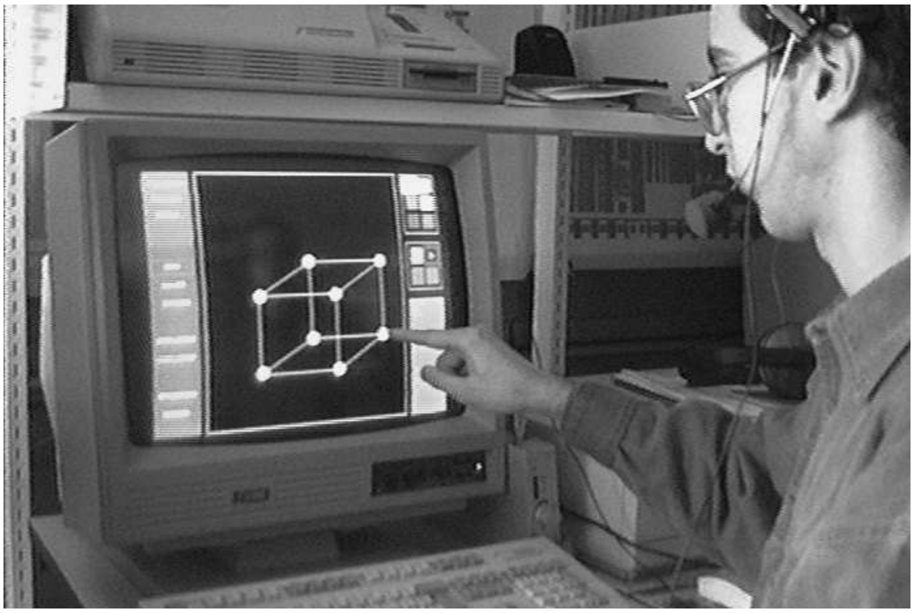

Figure 9.13: Working with a multimodal interface including speech and gesture. The user speaks while pointing on the touch screen to manipulate the objects. The time correlation of pointing gestures and spoken utterances is important to determine the meaning of his action.

1995). It is one of the basic criterion necessary (but not sufficient) for the fusion process and it allows for reconstituting the real chronological order of information. So it is necessary to assign dates (timestamps) to all messages (words, gestures, etc.) produced by the user.

It is also important to take into account the characteristics of each modality (Bernsen, 1993) and their technological constraints. For instance, operations which require high security should be assigned to the modalities which present lower error recognition risks, or should demand redundancy to reduce these risks. It can be necessary to define a multimodal grammar. In a perfect case, this grammar should also take into account other parameters such as the user state, current task, and environment (for instance, a high noise level will prohibit the use of speech).

Future Directions

The effectiveness of a multimodal interface depends in a large part on performances of each modality taken separately. If remarkable progress has been accomplished in speech processing, more efforts should be produced to improve gesture recognition systems, in particular for continuous gestures. Systems with touch feed-back and/or force feed-back which become more and more numerous will allow us to improve the comfort of gesture use, in particular for 3D applications, in the near future.

9.5 Modality Integration: Facial Movement & Speech Recognition

Alan J. Goldschen

Center of Innovative Technology, Herndon, Virginia, USA

9.5.1 Background

A machine should be capable of performing automatic speech recognition through the use of several knowledge sources, analogous, to a certain extent, to those sources that humans use (Erman & Lesser, 1990). Current speech recognizers use only acoustic information from the speaker, and in noisy environments often use secondary knowledge sources such as a grammar and prosody. One source of secondary information that has been primarily been ignored is optical information (from the face and in particular the oral-cavity region of a speaker), that often has information redundant with the acoustic information, and is often not corrupted by the processes that cause the acoustical noise (Silsbee, 1993). In noisy environments, humans rely on a combination of speech (acoustical) and visual (optical) sources, and this combination improves the signal-to-noise ratio by a gain of 10 to 12 dB (Brooke, 1990). Analogously, machine recognition should improve when combining the acoustical source with an optical source that contains information from the facial region such as gestures, expressions, head-position, eyebrows, eyes, ears, mouth, teeth, tongue, cheeks, jaw, neck, and hair (Pelachaud, Badler, et al., 1994). Human facial expressions provide information about emotion (anger, surprise), truthfulness, temperament (hostility), and personality (shyness) (Ekman, Huang, et al., 1993). Furthermore, human speech production and facial expression are inherently linked by a synchrony phenomenon, where changes often occur simultaneously with speech and facial movements (Pelachaud, Badler, et al., 1994; Condon & Osgton, 1971). An eye blink movement may occur at the beginning or end of a word, while oral-cavity movements may cease at the end of a sentence.

In human speech perception experiments, the optical information is complementary to the acoustic information because many of the phones that are said to be close to each other acoustically are very distant from each other visually (Summerfield, 1987). Visually similar phones such as /p/, /b/, /m/ form a *viseme*, which is specific oral-cavity movements that corresponds to a phone (Fisher, 1968). It appears that the consonant phone-to-viseme mapping is many-to-one (Finn, 1986; Goldschen, 1993) and the vowel phone-to-viseme mapping is nearly one-to-one (Goldschen, 1993). For example, the phone /p/ appears visually similar to the phones /b/ and /m/ and at a signal-to-noise ratio of zero /p/ is acoustically similar to the phones /t/, /k/, /f/, /th/, and /s/ (Summerfield, 1987). Using both sources of information, humans (or machines) can determine the phone /p/. However, this fusion of acoustical and optical sources does sometimes cause humans to perceive a phone different from either the acoustically or optically presented phone, and is known as the *McGurk effect* (McGurk & MacDonald, 1976). In general, the perception of speech in noise improves greatly when presented with acoustical and optical sources because of the complementarity of the sources.

9.5.2 Systems

Some speech researchers are developing systems that use the complementary acoustical and optical sources of information to improve their acoustic recognizers, especially in noisy environments. These systems primarily focus on integrating optical information from the oral-cavity region of a speaker (automatic lipreading) with acoustic information. The acoustic source often consists of a sequence of vectors containing, or some variation of, linear predictive coefficients or filter bank coefficients (Rabiner & Schafer, 1978; Deller, Proakis, et al., 1993). The optical source consists of a sequence of vectors containing static oral-cavity features such as the area, perimeter, height, and width of the oral-cavity (Petajan, 1984; Petajan, Bischoff, et al., 1988), jaw opening (Stork, Wolff, et al., 1992), lip rounding and number of regions or blobs in the oral-cavity (Goldschen, 1993; Garcia, Goldschen, et al., 1992; Goldschen, Garcia, et al., 1994). Other researchers model the dynamic movements of the oral cavity using derivatives (Goldschen, 1993; Smith, 1989; Nishida, 1986), surface learning (Bregler, Omohundro, et al., 1994), deformable templates (Hennecke, Prasad, et al., 1994; Rao & Mersereau, 1994), or optical flow techniques (Pentland & Mase, 1989; Mase & Pentland, 1991).

There have been two basic approaches towards building a system that uses both acoustical and optical information. The first approach uses a *comparator* to merge the two independently recognized acoustical and optical events. This comparator may consist of a set of rules (e.g., if the top two phones from the acoustic recognizer is /t/ or /p/, then choose the one that has a higher ranking from the optical recognizer) (Petajan, Bischoff, et al., 1988) or a fuzzy logic integrator (e.g., provides linear weights associated with the acoustically and optically recognized phones) (Silsbee, 1993; Silsbee, 1994). The second approach performs recognition using a vector that includes both acoustical and optical information, such systems typically use neural networks to combine the optical information with the acoustic to improve the signal-to-noise ratio before phonemic recognition (Yuhas, Goldstein, et al., 1989; Bregler, Omohundro, et al., 1994; Bregler, Hild, et al., 1993; Stork, Wolff, et al., 1992; Silsbee, 1994).

Regardless of the signal-to-noise ratio, most systems perform better using both acoustical and optical sources of information than when using only one source of information (Bregler, Omohundro, et al., 1994; Bregler, Hild, et al., 1993; Mak & Allen, 1994; Petajan, 1984; Petajan, Bischoff, et al., 1988; Silsbee, 1994; Silsbee, 1993; Smith, 1989; Stork, Wolff, et al., 1992; Yuhas, Goldstein, et al., 1989). At a signal-to-noise ratio of zero with a 500-word task Silsbee (1993), achieves word accuracy recognition rates of 38%, 22%, and 58% respectively, using acoustical information, optical information, and both sources of information. Similarly, for a German alphabetical letter recognition task, Bregler, Hild, et al. (1993) achieve a recognition accuracy of 47%, 32%, and 77%, respectively, using acoustical information, optical information, and both sources of information.

9.5.3 Future Directions

In summary, most of the current systems use an optical source containing information from the oral-cavity region of speaker (lipreading) to improve the robustness of the information from the acoustic source. Future systems will likely improve this optical source and use additional features from the facial region.

9.6 Modality Integration: Facial Movement & Speech Synthesis

Christian Benoit,[a] Dominic W. Massaro,[b] & Michael M. Cohen[b]

[a] Universite Stendhal, Grenoble, France
[b] University of California, Santa Cruz, California, USA

There is valuable and effective information afforded by a view of the speaker's face in speech perception and recognition by humans. Visible speech is particularly effective when the auditory speech is degraded, because of noise, bandwidth filtering, or hearing-impairment (Sumby & Pollack, 1954; Erber, 1975; Summerfield, 1979; Massaro, 1987; Benoît, Mohamadi, et al., 1994)

The strong influence of visible speech is not limited to situations with degraded auditory input, however. A perceiver's recognition of an auditory-visual syllable reflects the contribution of both sound and sight. When an auditory syllable */ba/* is dubbed onto a videotape of a speaker saying */ga/*, subjects perceive the speaker to be saying */da/* (McGurk & MacDonald, 1976).

There is thus an evidence that: (1) synthetic faces increase the intelligibility of synthetic speech, (2) but under the condition that facial gestures and speech sounds are coherent. To reach this goal, the articulatory parameters of the facial animation have to be controlled so that it looks like and it sounds like the auditory output is generated by the visual displacements of the articulators. Not only disynchrony or incoherence between the two modalities don't increase speech intelligibility; they might even decrease it.

Most of the existing parametric models of the human face have been developed in the perspective of optimizing the visual rendering of facial expressions (Parke, 1974; Platt & Badler, 1981; Bergeron & Lachapelle, 1985; Waters, 1987; Magnenat-Thalmann, Primeau, et al., 1988; Viaud & Yahia, 1992). Few models have focused on the specific articulation of speech gestures: Saintourens, Tramus, et al. (1990); Benoît, Lallouache, et al. (1992); Henton and Litwinovitz (1994) prestored a limited set of facial images occurring in the natural production of speech in order to synchronize the processes of diphone concatenation and *visemes* display in a text-to-audio-visual speech synthesizer. Ultimately, the coarticulation effects and the transition smoothing are much more naturally simulated by means of parametric models specially controlled for visual speech animation, such as the 3-D lip model developed by Guiard-Marigny, Adjoudani, et al. (1994) or the 3-D model of the whole face adapted to speech control by Cohen and Massaro (1990). Those two models are displayed on Figure 9.14.

A significant gain in intelligibility due to a coherent animation of a synthetic face has

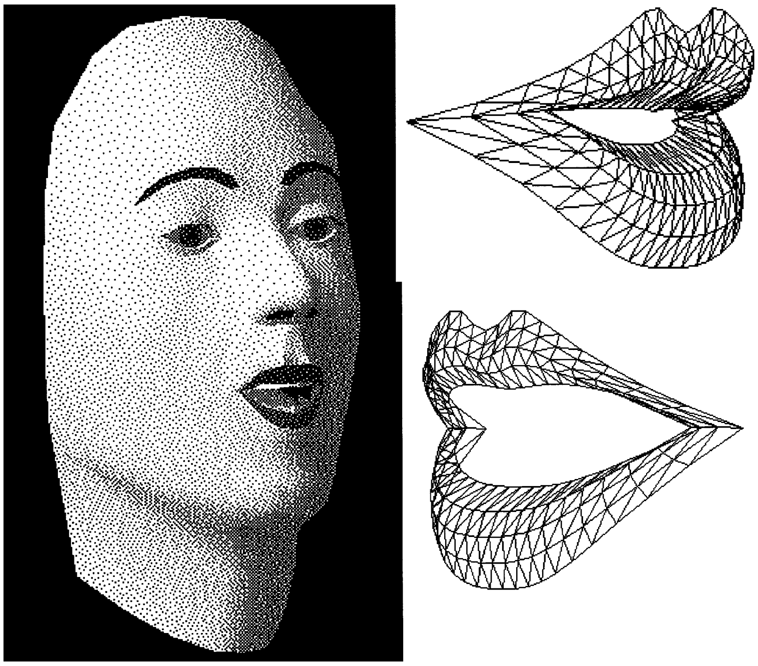

Figure 9.14: Left panel: gouraud shading of the face model originally developed by Parke (1974) and adapted to speech gestures by Cohen and Massaro (1993). A dozen parameters allow the synthetic face to be correctly controlled for speech. Right panel: wireframe structure of the 3-D model of the lips developed by Guiard-Marigny, Adjoudani, et al. (1994). The internal and external contours of the model can take all the possible shapes of natural lips speaking in a neutral expression.

obviously been obtained at the University of California in Santa Cruz by improving the Parke model (Cohen & Massaro, 1993) and then synchronizing it to the MITalk rule-based speech synthesizer (even though no quantitative measurements are yet available). In parallel, intelligibility tests have been carried out at the ICP-Grenoble in order to compare the benefit of seeing the natural face, a synthetic face, or synthetic lips while listening to natural speech under various conditions of acoustic degradation (Goff, Guiard-Marigny, et al., 1994).

Whatever the degradation level, the two thirds of the missing information are compensated by the vision of the entire speaker's face; half is compensated by the vision of a synthetic face controlled through six parameters directly measured on the original speaker's face; a third of the missing information is compensated by the vision of a 3-D model of the lips, controlled only through four of these command parameters (without seeing the teeth, the tongue or the jaw). All these findings support the evidence that technological spin-offs are expected in two main areas of application. On one hand, even though the quality of some text-to-speech synthesizers is now such that simple messages are very intelligible when synthesized in clear acoustic conditions, it is no longer the case when the message is less predictable (proper names, numbers, complex sentences, etc.) or when the speech synthesizer is used in a natural environment (e.g., the telephone network or in public places with background noise.) Then, the display of a synthetic face coherently animated in synchrony with the synthetic speech makes the synthesizer sound more intelligible and look more pleasant and natural. On the other hand, the quality of computer graphics rendering is now such that human faces can be very naturally imitated. Today, the audience no longer accepts all those synthetic actors behaving like if their voice was dubbed from another language. There is thus a strong pressure from the movie and the entertainment industry to overcome the problem of automatizing the lip-synchronization process so that the actors facial gestures look natural.

Future Directions

To conclude, research in the area of visible speech is a fruitful paradigm for psychological inquiry (Massaro, 1987). Video analysis of human faces is a simple investigation technique which allows a better understanding of how speech is produced by humans (Abry & Lallouache, 1991). Face and lip modeling allows the experimenters to manipulate controlled stimuli and to evaluate hypotheses and descriptive parametrizations in terms of visual and bimodal intelligibility of speech. Finally, bimodal integration of facial animation and acoustic synthesis is a fascinating challenge for a better description and comprehension of each language in which this technology is developed. It is also a necessary and promising step towards the realization of autonomous agents in human-machine virtual interfaces.

9.7 Chapter References

Abry, C. and Lallouache, M. T. (1991). Audibility and stability of articulatory movements: Deciphering two experiments on anticipatory rounding in French. In *Proceedings of the 12th International Congress of Phonetic Sciences*, volume 1, pages 220–225, Aix-en-Provence, France.

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Anger, F. D., Güsgen, H. W., and van Benthem, J., editors (1993). *Proceedings of the IJCAI-93 Workshop on Spatial and Temporal Reasoning (W17)*, Chambéry, France.

Asilomar (1994). *Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers*. IEEE.

Aurnague, M. and Vieu, L. (1993). A logical framework for reasoning about space. In Anger, F. D., Güsgen, H. W., and van Benthem, J., editors, *Proceedings of the IJCAI-93 Workshop on Spatial and Temporal Reasoning (W17)*, pages 123–158, Chambéry, France.

Badler, N. I., Phillips, C. B., and Webber, B. L. (1993). *Simulating Humans: Computer Graphics Animation and Control*. Oxford University Press, New York.

Bajcsy, R., Joshi, A., Krotkov, E., and Zwarico, A. (1985). LandScan: A natural language and computer vision system for analyzing aerial images. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 919–921, Los Angeles.

Bellik, Y. (1995). *Interfaces Multimodales: Concepts, Modeles et Architectures*. PhD thesis, Universite d'Orsay, Paris.

Benoît, C., Lallouache, M. T., Mohamadi, T., and Abry, C. (1992). A set of French visemes for visual speech synthesis. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models, and Designs*, pages 485–504. Elsevier Science.

Benoît, C., Mohamadi, T., and Kandel, S. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research*, 37:1195–1203.

Bergeron, P. and Lachapelle, P. (1985). Controlling facial expressions and body movements in the computer generated animated short 'tony de peltrie'. In *SigGraph '85 Tutorial Notes*.

Berkley, D. A. and Flanagan, J. L. (1990). HuMaNet: An experimental human/machine communication network based on ISDN. *AT&T Technical Journal*, 69:87–98.

Bernsen, N. O. (1993). Modality theory: Supporting multimodal interface design. In ERCIM, editor, *Proceedings of the Workshop ERCIM on Human-Computer Interaction*, Nancy.

Blonder, G. E. and Boie, R. A. (1992). Capacitive moments sensing for electronic paper. U.S. Patent 5 113 041.

Bolt, R. A. (1980). Put-that-there: Voice and gesture at the graphic interface. *Computer Graphics*, 14(3):262–270.

Bregler, C., Hild, H., Manke, S., and Waibel, A. (1993). Improving connected letter recognition by lipreading. In *Proceedings of the 1993 International Joint Conference on Speech and Signal Processing*, volume 1, pages 557–560. IEEE.

Bregler, C., Omohundro, S., and Konig, Y. (1994). A hybrid approach to bimodal speech recognition. In *Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers*. IEEE.

Brooke, N. M. (1990). Visible speech signals: Investigating their analysis, synthesis and perception. In Taylor, M. M., Néel, F., and Bouwhuis, D. G., editors, *The Structure of Multimodal Dialogue*. Elsevier Science, Amsterdam.

Brooks, F., Ouh-Young, M., Batter, J., and Jerome, P. (1990). Project GROPE: Haytic displays for scientific visualization. *Computer Graphics*, 24(4):177–185.

Burdea, G. and Coiffet, P. (1994). *Virtual Reality Technology*. John Wiley, New York.

Burdea, G. and Zhuang, J. (1991). Dextrous telerobotics with force feedback. *Robotica*, 9(1 & 2):171–178; 291–298.

Cherry, C. (1957). *On Human Communication*. Wiley, New York.

Cohen, M. M. and Massaro, D. W. (1990). Synthesis of visible speech. *Behaviour Research Methods, Instruments & Computers*, 22(2):260–263.

Cohen, M. M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In Thalmann, N. M. and Thalmann, D., editors, *Models and techniques in computer animation*, pages 139–156. Springer-Verlag, Tokyo.

Cohn, A. (1993). Modal and non-modal qualitative spatial logics. In Anger, F. D., Güsgen, H. W., and van Benthem, J., editors, *Proceedings of the IJCAI-93 Workshop on Spatial and Temporal Reasoning (W17)*, pages 87–92, Chambéry, France.

Condon, W. and Osgton, W. (1971). Speech and body motion synchrony of the speaker-hearer. In Horton, D. and Jenkins, J., editors, *The Perception of Language*, pages 150–184. Academic Press.

COSIT (1993). *Proceedings of the European Conference on Spatial Information Theory (COSIT'93)*, volume 716 of *Lecture Notes in Computer Science*. Springer-Verlag.

Coutaz, J., Nigay, L., and Salber, D. (1993). The MSM framework: A design space for multi-sensori-motor systems. In Bass, L., Gornostaev, J., and Under, C., editors, *Lecture Notes in Computer Science, Selected Papers, EWCHI'93, East-West Human Computer Interaction*, pages 231–241. Springer-Verlag, Moscow.

Deller, John R., J., Proakis, J. G., and Hansen, J. H. (1993). *Discrete-Time Processing of Speech Signals*. MacMillan.

Ekman, P., Huang, T., Sejnowski, T., and Hager, J. (1993). Final report to NSF of the planning workshop on facial expression understanding (July 30 to August 1, 1992). Technical report, University of California, San Francisco.

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, 40:481–492.

Erman, L. and Lesser, V. (1990). The Hearsay-II speech understanding system: A tutorial. In *Readings in Speech Recognition*, pages 235–245. Morgan Kaufmann.

ESCA (1994). *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York. European Speech Communication Association.

Feiner, S. K. and McKeown, K. R. (1993). Automating the generation of coordinated multimedia explanations. In Maybury, M. T., editor, *Intelligent Multimedia Interfaces*, pages 117–138. AAAI Press, Menlo Park, California.

Finn, K. (1986). *An Investigation of Visible Lip Information to be use in Automatic Speech Recognition*. PhD thesis, Georgetown University.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11:796–804.

Flanagan, J. L. (1992). Technologies for multimedia information systems. *Transactions, Institute of Electronics, Information and Communication Engineers*, 75(2):164–178.

Flanagan, J. L. (1994). Technologies for multimedia communications. *Proceedings of the IEEE*, 82(4):590–603.

Flanagan, J. L., Surendran, A. C., and Jan, E. E. (1993). Spatially selective sound capture for speech and audio processing. *Speech Communication*, 13:207–222.

Frank, A. U., Campari, I., and Formentini, U. (1992). Proceedings of the international conference GIS—from space to territory: Theories and methods of spatio-temporal reasoning. In *Proceedings of the International Conference GIS—From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning*, number 639 in Springer Lecture Notes in Computer Science, Pisa, Italy. Springer-Verlag.

Fraser, A. G., Kalmanek, C. R., Kaplan, A. E., Marshall, W. T., and Restrick, R. C. (1992). XUNET 2: A nationwide testbed in high-speed networking. In *INFOCOM 92*, Florence, Italy.

Furui, S. (1989). *Digital Speech Processing, Synthesis, and Recognition.* Marcel Dekker, New York.

Garcia, O., Goldschen, A., and Petajan, E. (1992). Feature extraction for optical automatic speech recognition or automatic lipreading. Technical Report GWU-IIST-92-32, The George Washington University, Department of Electrical Engineering and Computer Science.

Goff, B. L., Guiard-Marigny, T., Cohen, M., and Benoît, C. (1994). Real-time analysis-synthesis and intelligibility of talking faces. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 53–56, New Paltz, New York. European Speech Communication Association.

Goldschen, A. (1993). *Continuous Automatic Speech Recognition by Lipreading.* PhD thesis, The George Washington University, Washington, DC.

Goldschen, A., Garcia, O., and Petajan, E. (1994). Continuous optical automatic speech recognition. In *Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers*. IEEE.

Guiard-Marigny, T., Adjoudani, A., and Benoît, C. (1994). A 3-D model of the lips for visual speech synthesis. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 49–52, New Paltz, New York. European Speech Communication Association.

Hauptmann, A. G. and McAvinney, P. (1993). Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies*, 38(2):231–249.

Hennecke, M., Prasad, K., and Stork, D. (1994). Using deformable templates to infer visual speech dynamics. In *Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers*. IEEE.

Henton, C. and Litwinovitz, P. (1994). Saying and seeing it with feeling: techniques for synthesizing visible, emotional speech. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 73–76, New Paltz, New York. European Speech Communication Association.

Herkovits, A. (1986). *Language and Cognition.* Cambridge University Press, New York.

ICP (1993). Bulletin de la communication parlée, 2. Université Stendhal, Grenoble, France.

Keidel, W. D. (1968). Information processing by sensory modalities in man. In *Cybernetic Problems in Bionics*, pages 277–300. Gordon and Breach.

Lascarides, A. and Asher, N. (1993). Maintaining knowledge about temporal intervals. *Linguistics and Philosophy*, 16(5):437–493.

Ligozat, G. (1993). Models for qualitative spatial reasoning. In Anger, F. D., Güsgen, H. W., and van Benthem, J., editors, *Proceedings of the IJCAI-93 Workshop on Spatial and Temporal Reasoning (W17)*, pages 35–45, Chambéry, France.

Magnenat-Thalmann, N., Primeau, E., and Thalmann, D. (1988). Abstract muscle action procedures for human face animation. *Visual Computer*, 3:290–297.

Mak, M. W. and Allen, W. G. (1994). Lip-motion analysis for speech segmentation in noise. *Speech Communication*, 14:279–296.

Mariani, J., Teil, D., and Silva, O. D. (1992). Gesture recognition. Technical Report LIMSI Report, Centre National de la Recherche Scientifique, Orsay, France.

Mark, D. M. and Frank, A. U., editors (1991). *Cognitive and Linguistic Aspects of Geographic Space*, Dordrecht. NATO Advanced Studies Institute, Kluwer.

Mase, K. and Pentland, A. (1991). Automatic lipreading by optical flow analysis. *Systems and Computer in Japan*, 22(6):67–76.

Massaro, D. W. (1987). *Speech perception by ear and eye: a paradigm for psychological inquiry.* Lawrence Earlbaum, Hillsdale, New Jersey.

Maybury, M. T., editor (1993). *Intelligent Multimedia Interfaces.* AAAI Press, Menlo Park, California.

McDermott, D. (1982). A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6:101–155.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.

McKevitt, P. (1994). The integration of natural language and vision processing. *Artificial Intelligence Review Journal*, 8:1–3. Special volume.

Moens, M. and Steedman, M. J. (1988). Temporal ontology and temporal reference. *Computational linguistics*, 14(2):15–28.

Nakhimovsky, A. (1988). Aspect, aspectual class, and the temporal structure of narrative. *Computational Linguistics*, 14(2):29–43.

Nebel, B. and Bürckert, H.-J. (1993). Reasoning about temporal relations: A maximal tractable subclass of Allen's interval algebra. Technical Report RR-93-11, DFKI, Saarbrücken, Germany.

Netravali, A. and Haskel, B. (1988). *Digital Pictures*. Plenum Press, New York.

Neumann, B. (1989). Natural language description of time-varying scenes. In Waltz, D., editor, *Semantic Structures*, pages 167–207. Lawrence Earlbaum, Hillsdale, New Jersey.

Nishida (1986). Speech recognition enhancement by lip information. *ACM SIGCHI Bulletin*, 17(4):198–204.

Parke, F. I. (1974). *A parametric model for human faces*. PhD thesis, University of Utah, Department of Computer Sciences.

Pelachaud, C., Badler, N., and Viaud, M.-L. (1994). Final report to NSF of the standards for facial animation workshop. Technical report, University of Pennsylvania, Philadelphia.

Pentland, A. and Mase, K. (1989). Lip reading: Automatic visual recognition of spoken words. Technical Report MIT Media Lab Vision Science Technical Report 117, Massachusetts Institute of Technology.

Petajan, E. (1984). *Automatic Lipreading to Enhance Speech Recognition*. PhD thesis, University of Illinois at Urbana-Champaign.

Petajan, E., Bischoff, B., Bodoff, D., and Brooke, N. M. (1988). An improved automatic lipreading system to enhance speech recognition. *CHI 88*, pages 19–25.

Pierce, J. R. (1961). *Symbols, Signals and Noise*. Harper and Row, New York.

Platt, S. M. and Badler, N. I. (1981). Animating facial expressions. *Computer Graphics*, 15(3):245–252.

Podilchuk, C. and Farvardin, N. (1991). Perceptually based low bit rate video coding. In *Proceedings of the 1991 International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2837–2840, Toronto. Institute of Electrical and Electronic Engineers.

Podilchuk, C., Jayant, N. S., and Noll, P. (1990). Sparse codebooks for the quantization of non-dominant sub-bands in image coding. In *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing*, pages 2101–2104, Albuquerque, New Mexico. Institute of Electrical and Electronic Engineers.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Signal Processing. Prentice-Hall, Englewood Cliffs, New Jersey.

Rao, R. and Mersereau, R. (1994). Lip modeling for visual speech recognition. In *Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers*. IEEE.

Roe, D. B., Moreno, P. J., Sproat, R. W., Pereira, F. C. N., Riley, M. D., and Macarron, A. (1992). A spoken language translator for restricted-domain context-free languages. *Speech Communication*, 11:311–319. System demonstrated by AT&T Bell Labs and Telefonica de Espana, VEST, Worlds Fair Exposition, Barcelona, Spain.

Saintourens, M., Tramus, M. H., Huitric, H., and Nahas, M. (1990). Creation of a synthetic face speaking in real time with a synthetic voice. In Bailly, G. and Benoît, C., editors, *Proceedings of the First ESCA Workshop on Speech Synthesis*, pages 249–252, Autrans, France. European Speech Communication Association.

Salisbury, M. W. (1990). Talk and draw: Bundling speech and graphics. *IEEE Computer*, pages 59–65.

Schirra, J. and Stopp, E. (1993). ANTLIMA—a listener model with mental images. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 175–180, Chambery, France.

Silsbee, P. (1993). *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*. PhD thesis, The University of Texas at Austin.

Silsbee, P. (1994). Sensory integration in audiovisual automatic speech recognition. In *Proceedings of the 28th Asilomar Conference on Signals, Systems and Computers*. IEEE.

Smith, S. (1989). Computer lip reading to augment automatic speech recognition. *Speech Tech*, pages 175–181.

Stork, D., Wolff, G., and Levine, E. (1992). Neural network lipreading system for improved speech recognition. In *Proceedings of the 1992 International Joint Conference on Neural Networks*, Baltimore, Maryland.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26:212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, 36:314–331.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. and Campbell, R., editors, *Hearing by Eye: The Psychology of Lipreading*, pages 3–51. Lawrence Earlbaum, Hillsdale, New Jersey.

Vandeloise, C. (1986). *L'espace en français: sémantique des prépositions spatiales*. Seuil, Paris.

Viaud, M. L. and Yahia, H. (1992). Facial animation with wrinkles. In *Proceedings of the 3rd Workshop on Animation, Eurographic's 92*, Cambridge, England.

Vila, L. (1994). A survey on temporal reasoning in artificial intelligence. *AICOM*, 7(1):832–843.

Wahlster, W. (1989). One word says more than a thousand pictures. on the automatic verbalization of the results of image sequence analysis systems. *Computers and Artificial Intelligence*, 8:479–492.

Wahlster, W., André, E., Finkler, W., Profitlich, H.-J., and Rist, T. (1993). Plan-based integration of natural language and graphics generation. *Artificial Intelligence*, pages 387–427.

Wahlster, W., Marburger, H., Jameson, A., and Busemann, S. (1983). Over-answering yes-no questions: Extended responses in a NL interface to a vision system. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 643–646, Karlsruhe.

Waibel, A. (1993). Multimodal human-computer interaction. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume Plenary, page 39, Berlin. European Speech Communication Association.

Waters, K. (1987). A muscle model for animating three-dimensional facial expression. In *Proceedings of Computer Graphics*, volume 21, pages 17–24.

Webber, B. L. (1988). Tense as discourse anaphor. *Computational linguistics*, 14(2):61–73.

Wilpon, J., Rabiner, L., Lee, C., and Goldman, E. (1990). Automatic recognition of key words in unconstrained speech using hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(11):1870–1878.

Yamada, A., Yamamoto, T., Ikeda, H., Nishida, T., and Doshita, S. (1992). Reconstructing spatial images from natural language texts. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 1279–1283, Nantes, France. ACL.

Yuhas, B., Goldstein, M., and Sejnowski, T. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, pages 65–71.