A Taxonomy of Multimodal Interaction in the Human Information Processing System

A Report of the ESPRIT PROJECT 8579 MIAMI
– WP 1 –

February, 1995

Written by
L. Schomaker*, J. Nijtmans (NICI)
A. Camurri, F. Lavagetto, P. Morasso (DIST)
C. Benoît, T. Guiard-Marigny, B. Le Goff,
J. Robert-Ribes, A. Adjoudani (ICP)
I. Defée (RIIT)
S. Münch* (UKA)
K. Hartung*, J. Blauert (RUB)

*Editors

Abstract

This document has been prepared in the ESPRIT BRA No. 8579, Multimodal Integration for Advanced Multimedia Interfaces *— in the following referred to as* MIAMI *— in order to serve as a basis for future work. The basic terms which will be used in MIAMI will be defined and an overview on man-machine-interfaces will be given. The term "taxonomy" is used in the following sense, adapted from [217]: "1: the study of the general principles of scientific classification: SYSTEMATICS; 2: CLASSIFICATION; specif: orderly classification of plants and animals according to their presumed natural relationships"; but instead of plants and animals, we attempt to classify input and output modalities.*

Contents

Chapter 1

Introduction

In this chapter we will introduce our underlying model of human-computer interaction which will influence the whole structure of this document as well as our research in MIAMI. The next step will be the definition of basic terms (more specific terms will follow in later sections) and a statement of what will be included and excluded in the project, respectively. Some 'philosophical' considerations will follow in section 1.2, and finally we will give an overview on the structure of this report.

1.1 Definitions of Basic Terms

One thing most publications on *multimodal systems* have in common is each author's own usage of basic terms on this topic. Therefore, we want to state in the beginning of this report what our basic model looks like, how we are going to use several terms within this document, and which levels of perception and control we are considering in MIAMI.

1.1.1 The basic model for human-computer interaction

In order to depict a taxonomy of multimodal human-computer interaction we will have to clarify a number of concepts and issues. The first assumption is that there are minimally two separate *agents* involved, one human and one machine. They are physically separated, but are able to exchange information through a number of information channels. As schematically shown in figure 1.1, we will make the following definitions.

There are two basic processes involved on the side of the human user: *Perception* and *Control*. Note that we take the perspective of the human process throughout this document. With respect to the *Perceptive process*, we can make a distinction between:

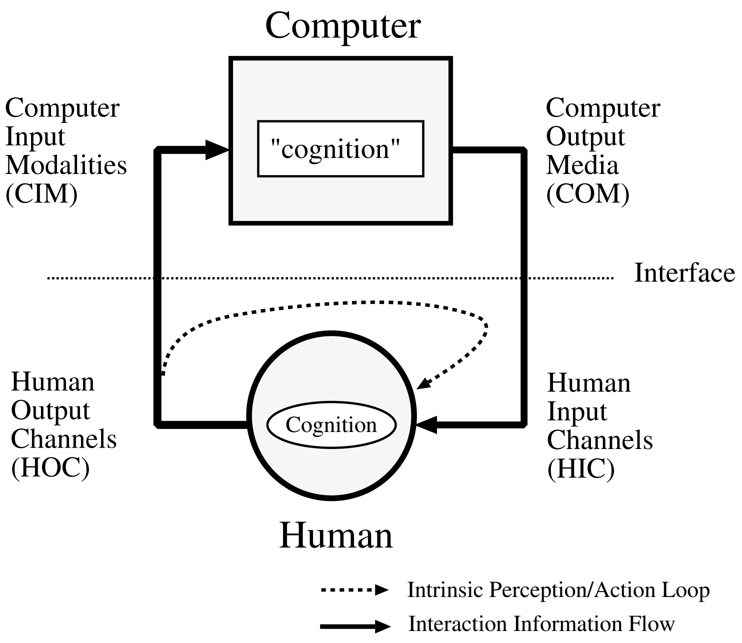

Figure 1.1: *A model for the identification of basic processes in human-computer interaction. Note that in fact two loops exists: the intrinsic feedback as in eye-hand coordination, and the extrinsic loop, imposed by the computer.*

- Human Input Channels (HIC) and
- Computer Output Media (COM)

Within the *Control process*, we can make a distinction between:

- Human Output Channels (HOC) and
- Computer Input Modalities (CIM)

Then, within both of the agents, a cognitive or computational component can be identified, which processes the incoming input information and prepares the output. Also, at this intermediate cognitive level, intentional parameters will influence the processing, either implicitly, such as by design, in the case of non-intelligent agents, or explicitly, as in humans or in more sophisticated agents containing an explicit representation of goals and "beliefs". With respect to the machine, it should be noted that here the design is known, whereas for the human cognitive apparatus, the architecture must be inferred and cannot be observed directly.

Instead of the word modality at the human input side, and the word media at the human output side, we have chosen for the word *channel*, which also allows for a more clear distinction between the abbreviations (HOC $\rightarrow$ CIM $\rightarrow$ COM $\rightarrow$ HIC $\rightarrow$ HOC $\rightarrow \ldots$) which can also be pronounced as:

- HOC: human output to computer
- CIM: computer input from man
- COM: computer output to man
- HIC: human input from computer

The two halves of the loop will be further investigated in sections 2, *Perception*, and 3, *Control and Manipulation*, respectively. The interaction process between the operator (man) and the computer (machine) will be considered in section 4, *Interaction*, whereas the internal processing is described in 5, Cognition.

1.1.2 Levels of observation

For all the four components listed in 1.1.1, a number of epistemological levels of observation can be defined, relevant to both human and machine:

- Physical/Physiological
- Information Theoretical
- Cognitive
- Intentional

At the Physical/Physiological level of observation, device characteristics may be identified, such as resolution (temporal, spatial), delay time characteristics, physical measures such as dB, Lux, force etc. Also analog bandwidth pertains to this level of observation. Examples are the bandwidth of the auditory channels, the bandwidth of the motor output channels (HOC and COM), etc. The system characteristics and process phenomena at this level provide the basic constraints and capabilities in human-computer interaction. Apart from the unimodal constraints, an interesting phenomenon at this level is the improved perception for a given modality under multimodal conditions.

At the Information Theoretical observation level, digital bandwidth and digital communication concepts are relevant, describing the pure informational characteristics of the channels (e. g., entropy). At this level of observation, issues of data compression, but also of optimal search processes play a role. An example is the number of bits/s needed to faithfully transmit the recorded images of a human talking face over a COM$\rightarrow$HIC channel.

At the Cognitive level, *representational* and *procedural* aspects have to be made explicit, in *syntax* and *semantics*. Furthermore involved are all those components of pattern recognition processes which exceed the level of the basic signal features, and require information of a more abstract nature. An example is the use of linguistic "top-down" information in the disambiguation of speech, handwriting, or the use of high-level information in the interpretation of facial movements. Other typical topics at the cognitive level are the notion of *short-term memory* with its limited capacity in the human: 7 ± 2 conceptual items [223, 43]. And, last but not least, an essential property of the Cognitive level is the process of *learning*. Because of the fact that the Cognitive level is the mediating level between low-level processing and the high level of intention, to be clarified in the next paragraph, it is at the core of the MIAMI project.

At the Intentional level, the *goals and beliefs* of the involved agents have to be made explicit. It is still one of the problems of current computer science and AI that this level is often ignored, whereas it is essential to the human user. When asked what he/she is currently doing on the computer, the answer of the user will most likely not be on the level of the tools (e. g., "I'm pressing this button"), but on the level of goals and intentions ("I'm looking for this nice picture in the Louvre homepage"). Typically, much of the research in human-computer interaction involved experimental conditions of goalless behavior. At best, goals artificially imposed on the user subjects are used. However, there is a fundamental difference between "playful browsing activity" and "goal-oriented search in a financial database under time pressure". Such a difference at the intentional level will have a large influence on low-level measures such as movement times, reaction times, and most probably even on perceptual acuity. Another area where the disregard of goal-oriented behavior of the human users leads to serious problems is in the automatic recognition of HOC data. In most pattern recognition methods, training of the algorithms is needed, using a database of human-produced data. However, using isolated words which are collected outside a "natural" application context for the training of speech or handwriting recognizers, will yield data which is not generalizable to the type of human output (HOC) data produced in an application context involving realistic goals. The problem of intention becomes also clearly apparent in the case of teleoperating applications, where the goals of the user must be continuously inferred from action by the computer.

1.1.3 (Multi-) Modality

According to [67], *modality* is defined as follows:

> "*Perception* via one of the three *perception-channels*. You can distinguish the three modalities: *visual*, *auditive*, and *tactile* (physiology of senses)."

visual: concerned with, used in seeing (comp. against *optical*)

auditive: related to the sense of hearing (comp. against *acoustical*)

tactile: experienced by the sense of touch

haptic: most authors are using "*tactile*" and "*haptic*" as synonyms. However, in [120] tactile as perception-modality is distinguished from haptic as output manner.

optics/optical: *optics* is the theory on light as well as the infrared and ultraviolet radiation. Thus the attribute "*optical*" refers to physical quantities and laws rather than to physiological ones. (comp. against *visual*)

acoustics/acoustical: *acoustics* is the theory on vibrations and oscillations in elastic mediums, especially of sound, its generation, spreading, and its reception. Thus the attribute "*acoustical*" refers to physical rather than to physiological quantities and laws. (comp. against *auditive*)

The author of [67] defines only three *modalities* and associates them with three of the human senses. Although they will be the three ones considered in MIAMI, there are some more senses as defined by physiology:

Sensory perception	Sense organ	Modality
Sense of sight	Eyes	Visual
Sense of hearing	Ears	Auditive
Sense of touch	Skin	Tactile
Sense of smell	Nose	Olfactory
Sense of taste	Tongue	Gustatory
Sense of balance	Organ of equilibrium	Vestibular

Table 1.1: *Different senses and their corresponding modalities (taken from: [311])*

In our opinion, the *sense of smell* and the *sense of taste* are not very interesting for our concerns (see 2.1.1 for a deeper discussion). However, the *sense of balance* seems to become more and more interesting with respect to virtual reality environments (see 4.4). Presently it is already used in flight simulators for example.

Whenever more than two of these modalities are involved, we will speak of *multimodality*. To be more precise, in some cases we will also use the term *bimodal* (or *bimodality*) to denote the usage of exactly two different modalities. In this sense, every human-computer interaction has to be considered as multimodal, because the user *looks* at the monitor, *types* in some commands or *moves* the mouse (or some other device) and *clicks* at certain positions, *hears* the reaction (beeps, key clicks, etc.) and so on.

Therefore, in MIAMI our understanding of multimodality is restricted to those interactions which comprise more than one modality on either the input (i. e., perception) or the output (i. e., control) side of the loop *and* the use of more than one device on either side. Thus, the combination of, e. g., visual, auditive, and tactile feedback which is experienced by typing on a keyboard is explicitly excluded, whereas the combination of visual and auditive output produced by the monitor and a loudspeaker when an error occurred is a 'real' multimodal (or — in this case — bimodal) event.

1.1.4 Multimodal vs. multimedia vs. virtual reality system

Notice the following paragraph in [243]:

> **"Multimodal System: A Definition**
> In the general sense, a multimodal system supports communication with the user through different modalities such as voice, gesture, and typing. Literally, 'multi' refers to 'more than one' and the term 'modal' may cover the notion of 'modality' as well as that of 'mode'.
>
> - Modality refers to the type of *communication channel* used to convey or acquire information. It also covers the way an idea is expressed or perceived, or the manner an action is performed.
> - Mode refers to a state that determines the way information is interpreted to extract or convey meaning.
>
> In a communication act, whether it be between humans or between a computer system and a user, both the modality and the mode come into play. The modality defines the type of data exchanged whereas the mode determines the context in which the data is interpreted. Thus, if we take a system-centered

> view, multimodality is the capacity of the system to communicate with a user along different types of communication channels and to extract and convey meaning automatically. We observe that both *multimedia* and *multimodal* systems use multiple communication channels. But in addition, a multimodal system is able to automatically model the content of the information at a high level of abstraction. A multimodal system strives for meaning."

In MIAMI, we decided to agree with this perspective. Especially the common usage of a *multimedia* system, which is often merely a PC with a sound card, a loudspeaker, and a CD ROM (if the hardware is considered) or a kind of hypertext/hypermedia environment with a combination of text, video, and sound (if the software is considered) is *not* what we want to address in this project. Such a kind of explanation of the term *multimedia* is given by Steinmetz:

> "*Multimedia*, from the user's point of view, means that information can also be represented as audio signals or moved images."
>
> [322, page 1] (translation by the authors)

A definition given by Buxton states the following:

> "[...] 'Multimedia' focuses on the *medium* or *technology* rather than the *application* or *user*."
>
> [195, page 2]

The second aspect, namely to concentrate on the user, is what we want to do in this project. The first aspect, the concentration on the application, is characteristic for research performed in the VR domain. Therefore, a distinction between a *multimodal* and a *Virtual Reality (VR)* system has to be made. A 'minimal' definition of a VR system is provided by Gigante:

> "The illusion of participation *in* a synthetic environment rather than external observation of such an environment. VR relies on three-dimensional (3D), stereoscopic, head-tracked displays, hand/body tracking and binaural sound. VR is an immersive, multisensory experience."
>
> [86, page 3]

In this project, we understand that the main difference is the intention behind both research directions: VR aims at the imitation of reality for establishing immersive audio-visual illusions, whereas multimodality attempts to enhance the throughput and the naturalism of man-machine communication. The audio-visual illusion of VR is just a trick for triggering the natural synergy among sensory-motor channels which is apparent in our brain but is not the only possibility of efficient exploitation of the parallelism and associative nature of human perceptual-motor processes. From this point of view, VR research is a subset of multimodality research. For example, in a following section we suggest two views of multimodal systems, *computer-as-tool* and *computer-as-dialogue-partner*, which are a further alternatives to the computer-as-audiovisual-illusion which is typical of VR systems. A discussion of VR will follow in section 4.4.

Having in mind our basic model as well as Nigay and Coutaz' statement that "A multimodal system strives for meaning.", one aspect that distinguishes our research from both, multimedia and VR research, is the concentration on the internal processing steps, known as *cognition*. Whereas synchronization of different modalities (or media) plays a key role in a multimedia system, this won't be enough for a multimodal system. The fusion of modalities as well as of cognitive processes has to be considered in order to find the meaning or the intention of the user's actions (see also 1.1.2 and 1.2.2). This aspect of multimodality will be further investigated in section 5.

1.1.5 Communication channels

Charwat defines communication channel as follows:

> "In general, a *communication channel* is a connection between one sending source and one receiving sink. In the human body a communication channel is formed by the sense organs, the nerve tracts, the cerebrum, and the muscles. In case of competing sensations at different sense organs there is one channel preferred.
> There is no total independence between the communication channels. For example: the capacity of received information coded with combined visual and auditive signals does not reach the sum of the single channels' capacities."
>
> [67]

The following figure shows *estimated capacities* of human information reception, processing, and output summarized over all communication channels [311]:

$$\begin{array}{ccccc} \text{input} & \longrightarrow & \text{processing} & \longrightarrow & \text{output} \\ 10^9 \text{ bit/s} & & 10^1 - 10^2 \text{ bit/s} & & 10^7 \text{ bit/s} \end{array}$$

Some *channel capacities* are mentioned in [67][1]:

[1]Unfortunately, Charwat does not describe the experiments leading to this numbers!

Physical quantity	Channel capacity: Distinguishable steps	Channel capacity: Channel width [bits]
Musical pitch	≈ 6	2.6
Musical volume	≈ 5	2.3
Estimate extent of squares	≈ 5	2.3
Color saturation	≈ 9	3.1
Length of lines	$\approx 6 - 8$	2.6 – 3.0
Angle of lines	$\approx 7 - 10$	2.8 – 3.3
Vibration intensity	≈ 4	2.0
Vibration duration	≈ 5	2.3

Table 1.2: *Capacities of different human input channels (taken from [67])*

Most probably, these channel capacities originated from the Weber, Fechner, and Stevens laws of psychophysics. Therefore, their work will be reviewed here very briefly.

The Psychophysical Law Building upon the classical works of Weber and Fechner in the last century about relative perceptual thresholds, which culminated in the model of the logarithmic law, a better fit to the experimental data for a number of sensory channels has been demonstrated by Stevens with a psychophysical power law [325]:

$$\psi = k \cdot S^m, \tag{1.1}$$

where ψ is the subjective magnitude, S is the physical stimulus intensity and m is a modality-dependent exponent. Table 1.3 summarizes some values of the exponent estimated by Stevens:

Uncertainty and discrimination Information theory has been particularly useful in dealing with experiments on absolute judgment. In these experiments, the observed is considered as the communication channel and the question asked is: “What is the amount of information that the observer can transmit in an absolute-judgment task?” Or, more specifically, “What is the channel capacity of the perceiver with respect to the number of items that can be absolutely discriminated?” Table 1.4 summarizes some of the classical results on the matter.

Details of the experiments are explained in [81].

Modality	Exponent
Brightness	0.5
Loudness	0.6
Vibration	0.95
Duration	1.1
Temperature	1.6
Heaviness	1.45
Electric shock	3.5

Table 1.3: *Exponents for the psychophysical law (taken from [325])*

Stimulus dimension	Investigator	Channel capacity [bits]
Brightness	[97]	1.7
Duration	[236]	2.8
Hue	[66]	3.6
Loudness	[116]	2.3
Pitch	[274]	2.5
Vibration intensity	(from [81])	1.6

Table 1.4: *Amount of information in absolute judgements of several stimulus dimensions (taken from [81])*

1.2 Additional Notes and Caveats

1.2.1 An extra information loop?

The model proposed in 1.1.1 has the advantage of providing the basic, isolated concepts. However, it should be noted that the proposed schematism obscures the fact that perception and motor control in the human are tightly coupled. The clearest example is vision, in which a coherent internal image of the outside world is reconstructed from a series of controlled eye movements (saccades) and fixations, directed at details in the environment. The eye musculature plays an essential role here: The image would fade within seconds if these muscles were paralyzed. Also, processes like continuous eye-hand coordination require an intrinsic Output→Input feedback loop in the human, which shunts the human-machine main information pathway (figure 1.1). This issue will be exposed in the coming chapters, notably chapter 4 on *Interaction*.

1.2.2 Notes with respect to intention

A recurring theme in theorizing about human-computer interaction is the duality of the following two views on the computer:

1. Computer-as-Tool

2. Computer-as-Dialogue-Partner

Ad 1. In this view, the intention (more simply, the goal-directed behavior) is assumed to be present in the human user, whereas the machine is a passive tool, transparent to a high degree, and merely supportive with respect to the human goals. This is in fact the status of most current applications, and is also the acclaimed basic philosophy by well-known researchers in the field of human-computer interaction (e. g., [310]). In fact, a trend is present from earlier anthropomorphizing error messages and dialogue fragments, towards more neutral, tool-like behavior. Initial experimental designs of the now well-known teller machines displayed anthropomorphic announcements like "How are you today, can I help you?", a method which in practice has been quickly replaced by more formal, brief texts and menus.

Ad 2. Alternatively, there is a trend, in which goal-oriented and intentional behavior is explicitly introduced in so-called *intelligent agents*[2], which behave according to a logic containing "beliefs" about the world and about the goals of the user. Synergistic with this trend is the quickly growing capability with which human-like output (speech, faces, movement) can be more or less faithfully produced by computers. Critics of this latter approach pejoratively call it "animism" and point out that user-friendly design is much more helped by giving the user the idea that he/she is fully in control of the situation, rather than having to negotiate with a potentially uncooperative anthropomorphic servant. However, the anthropomorphic dialogue may also be compelling, and the metaphors of "partner", "colleague" or "coach" may be used to make clear to the user what are the characteristics of the dialogue [344], in terms of protocol, attitude, and etiquette.

Rather than deciding *a priori* for either one of these basic interface philosophies, different types of interaction and different application conditions must be defined within this project which are suited for either the *Computer-as-Tool* or the *Computer-as-Dialogue-Partner* approach, possibly also combining aspects of both approaches where possible. As an example: Whereas it may be acceptable to a user to organize the planning of daily activities in a dialogue with an anthropomorphical partner, it is equally unacceptable

[2](Intelligent) Agents are also the topic of section 5.2

to use a human-style dialogue in detailed low-level aspects of text editing or computer programming.

Within the project of MIAMI the concepts introduced here will be substantially refined.

1.2.3 Topics which had to be *excluded* from this report

Due to limitations in time and other resources, not all aspects of multimodality and multimedia systems could be included in this taxonomy report. And, of course, it is also not possible to enumerate all these topics that had to be excluded. Nevertheless, we want to name some of them in order to clarify what you will find in the following sections (see the next section) and what you will look for in vain.

Excluded topics are, among others:

- Data storage aspects
- Compression algorithms
- Operating systems
- Basic software tools
- Communication protocols
- Database systems

Although we will use most of these things, we are not going to develop anything like that.

1.3 Structure of the Document

First we will focus our attention on the human perceptual process, its input channels (HIC) and the characteristics of computer output media (COM) (Chapter 2).

Then, we will approach the process of human control and manipulation, dwelling on the characteristics of the human output channels (HOC) and the computer input modalities (CIM) (Chapter 3).

In the next chapter (4) the issue of the interaction process will be dealt with, addressing aspects of input/output coupling and temporal and spatial synchronization.

Gradually moving from basic processes to more abstract levels, we will describe aspects of cognition and learning, its representational architecture in machine and man, as far as possible, and on the issue of interacting agents (Chapter 5).

Finally, in the last chapter, a number of scenarios, or rather "dreams" will be depicted, elucidating fundamental aspects put forward in this taxonomy.

Chapter 2

Perception

In this chapter we will deal with perception, i. e., the process of transforming sensorial information to higher-level representations which can be used in associative processes (memory access) and cognitive processes such as reasoning. In the perceptual process, two basic components can be identified as described earlier (figure 1.1, right part). First there is the human sensory system with its typical characteristics and constraints, providing the "human input channel" (HIC) processing functions. Second, there are those components of a computer system which provide a multidimensional signal for the human user: "computer output media" (COM).

According to our basic model introduced in section 1, we use the term *perception* to describe the communication from a machine to a human. First, we will show the different kinds of human input channels through which perception becomes possible and give a short overview on these channels at the neurophysiological level (2.1). The second part of this section (2.2) deals with devices and methods used by computers to address these human senses. Finally, in 2.3 we will present our results of bi- and multimodal machine-to-man communication.

2.1 Human Input Channels

In this introductory section on *perception*, we will review the concept of *modality* from the neurobiological point of view [153, 308], gradually narrowing the scope to those modalities relevant to research within MIAMI.

As remarked by Shepherd [308], the notion of sensory modality can be traced back to the 1830s, in particular to the monumental "Handbook of Human Physiology", published in Berlin by Johannes Muller, who promulgated the "law of specific nerve energies". This states that we are aware not of objects themselves but of signals about them transmitted through our nerves, and that there are different kinds of nerves, each nerve having its own "specific nerve energy". In particular, Muller adopted the five primary senses that Aristotle had recognized: seeing, hearing, touch, smell, taste. The specific nerve energy, according to Muller, represented the sensory modality that each type of nerve transmitted.

The modern notion, beyond a great degree of terminological confusion, is not very much different: we recognize that there are specific receptor cells, tuned to be sensitive to different forms of physical energy in the environment and that they serve as stimuli for the receptor cells. A table in Shepherd's book illustrates the point. The table can be simplified and re-written in our framework as follows (Table 2.1).

Sensory modality	Form of energy	Receptor organ	Receptor cell
Chemical (internal)			
blood oxygen	O_2 tension	carotid body	nerve endings
glucose	carbohydrate oxidation	hypothalamus	gluco-receptors
pH (cerebrospinal fluid)	ions	medulla	ventricle cells
Chemical (external)			
taste	ions & molecules	tongue & pharynx	taste bud cells
smell	molecules	nose	olfactory receptors
Somatic senses			
touch	mechanical	skin	nerve terminals
pressure	mechanical	skin & deep tissue	encapsulated nerve endings
temperature	thermal	skin, hypothalamus	peripheral & central
pain	various	skin & various organs	nerve terminals
Muscle sense, kinesthesia			
muscle stretch	mechanical	muscle spindles	nerve terminals
muscle tension	mechanical	tendon organs	nerve terminals
joint position	mechanical	joint capsule & ligaments	nerve terminals
Sense of balance			
linear acceleration	mechanical	sacculus/utriculus	hair cells
angular acceleration	mechanical	semicircular canal	hair cells
Hearing			
	mechanical	cochlea	hair cells
Vision			
	light	retina	photoreceptors

Table 2.1: *An overview of input channels at the neurophysiological level*

The different sensory modalities used by human beings are not processed in isolation. Multimodal areas exist in cortical and sub-cortical areas, such as the posterior parietal cortex (area 5 and 7) and the superior culliculus. The integration of the different channels is essential, among other things, for allowing the brain to reconstruct an internal body model and an internal representation of external Euclidean space [228].

2.1.1 Input modalities

Obviously, not all of these channels/modalities are of the same interest for MIAMI. As outlined in the Technical Annex, the project will mainly address the senses of vision, hearing, and the somatic senses. The remaining question is: Do we consider these senses as "atomic" or do we also take into account the sub-modalities which have been presented partly in table 2.1? We decided to follow the first approach, so in the following paragraphs we will *exclude* some possible input modalities from the project.

It is possible to distinguish sub-modalities, according to different sensitivities, e. g. with respect to peak frequency or cut-off frequency. In the eye, there are rod and three types of cone receptors in the retina, each with their specific properties. Similarly, in the motor system, there are the bag and chain fibers in the muscle spindles, etc. However, these very low levels of the infrastructure of the perceptual system can be skipped within the context of MIAMI.

Some of the sensory modalities, however, do not have a cortical representation (sense of balance, chemical senses) or just have a very reduced one (taste) and do not give origin to "conscious perception", whatever is the controversial meaning we attribute to this concept; thus we cannot speak, for them, of "perceptual channels".

The senses of smell and taste might not be very interesting for MIAMI (and human information processing in general). This is not due to the fact that "the corresponding output devices are missing" but to the fact (i) that taste is not a very useful channel of man-machine interaction and (ii) smell has some practical problems. However, it should be emphasized that the sense of smell has great potentialities, particularly if we consider man-machine interaction with mobile robots: in nature, indeed, odors are not only important from the point of view of "chemical" analysis, but also from the navigation point of view, for "marking" the territory and setting "landmarks" which are of great help in path planning. Also, biological memory is linked to odors, most probably because the phylogenetically oldest systems of territory representation are based on the chemical senses. Spatial memory is probably related to the hippo-campus, which is a cortical area in the immediate neighborhood of the olfactory cortex. Some experiments of robotical path planning, following the gradient of some odor, are reported in the literature and it seems reasonable to consider odor as a creative bidirectional channel of communication between man and a moving machine. Unlike sound, olfactory marks have a physical persistence, like visual traces, but can be invisible themselves and may thus be used in parallel to visible traces.

According to [279], there are many different ways to use odors to create a striking sense of presence:

> "The technology of delivering odors is well-developed [343], in trials at Southwest Research Institute. The odors are all Food and Drug Administration approved and delivered at low concentration. The system uses a micro-encapsulation technique that can be dry packaged in cartridges that are safe and easy to handle. Human chemical senses such as taste and smell create particularly salient memories."
>
> [279]

In the context of MIAMI, however, other input modalities are deemed more relevant: vision, hearing, and the somatic senses. Due to the importance of these senses, they will be considered in more detail in the next sections.

2.1.2 Vision

> **vision:** [...] 2b (1): mode of seeing or conceiving; [...] 3a: the act or power of seeing: SIGHT; 3b: the special sense by which the qualities of an object [...] constituting its appearance are perceived and which is mediated by the eye; [...] [217]

As an input modality for information processing, vision plays the most important role. The complexity and sophistication of visual sense is to a large extent beyond our full comprehension but there is a large body of experimental knowledge about the properties of vision collected. Detailed review of all these properties would be voluminous [278], one can say that in general the responses span very considerable dynamic ranges. At the same time the amount information processed is cleverly reduced giving illusion of photographic registration of reality coinciding with fast and precise extraction of very complex information.

From the applications point of view the properties of vision which are basic and important are *light intensity response*, *color response*, *temporal response*, and *spatial responses*.

There are many different levels at which these properties can be studied. At the receptor level, there are retinal receptors sensitive to light intensity and color. However, these raw sensitivities have little in common with perceived light and color and the information form the receptors undergoes numerous transformations making that the visual system is rather sensitive to the changes in light intensity and tries to preserve color constancy in changing illumination. These effects can be quite elaborate, depending also on higher level aspects of visual perception and memory. While the sensitivities of the receptor system is very high, for practical purposes it suffices to assume that it covers 8-10 bit range of amplitudes for the light intensity and each of the primary colors.

The temporal response of the visual system is responsible for many effects like perception of light intensity changes, and rendering of motion. The response can also be very high in specific conditions but in practice it can be considered to be limited to a maximum of 100 Hz for very good motion rendering and few tens of Hz for light intensity change. This is dependent on the type of visual stimulation, distance, lighting conditions and it has as a direct consequence that the frequency of repetition of pictures in TV and cinema is about 50 Hz but this is insufficent for computer displays which require 70-80 Hz or more to eliminate picture flickering effects.

Spatial response of visual system deals with the problems of visual resolution, width of the field of view, and spatial vision. There is a direct and strong impact of these factors on the visual perception. While in normal scenes the resolution to details needs not to be very high (twice the normal TV resolution is considered "high definiton"), in specific situations the eye is very sensitive to the resolution and this is the reason while magazine printing might require a hundred times higher resoltuion than TV. A very important perceptual factor is the width of the field of view. While the center visual field which brings most information is essential, there is a much wider peripheral vision system which has to be activated in order ot increase the perceptual involvement (cinema vs. TV effect). On top of this there is a sophisticated spatial vision system which is partially based on binocular vision and partially on spatial feature extraction from monocular images.

The full visual effect coming from the fusion of visual responses to the different stimulations is rich and integrated to provide optimum performance for very complex scenes. Usually the optimum performance means extremely quick and efficient detection, processing, and recognition of patterns and parameters of the visual scenes.

2.1.3 Hearing

> **hearing:** 1: to perceive or apprehend by the ear; [...] 1: to have the capacity of apprehending sound; [...] 1: the process, function, or power of perceiving sound; specif: the special sense by which noises and tones are received as stimuli; [...] [217]

The main attributes used for describing a hearing event are:

Pitch is the auditory attribute on the basis of which tones may be ordered on a musical scale. Two aspects of the notion pitch can be distinguished in music: one related to the frequency (or fundamental frequency) of a sound which is called *pitch height*, and the other related to its place in a musical scale which is called *pitch chroma*. Pitch heights vary directly with frequency over the range of audible frequencies. This 'dimension' of pitch corresponds to the sensation of 'high' and 'low'. Pitch chroma, on the other hand, embodies the perceptual phenomenon of octave equivalence, by which two sounds separated by an octave (and thus relatively distant in terms of pitch height) are nonetheless perceived as being somehow equivalent. This equivalence is demonstrated by the fact that almost all scale systems in the world in which the notes are named give the same names to notes that are roughly separated by an octave. Thus pitch chroma is organized in a circular fashion, with an octave-equivalent pitches considered to have the same chroma. Chroma perception is limited to the frequency range of musical pitch (50-4000 Hz) [218].

Loudness is the subjective intensity of a sound. Loudness depends mainly on five stimulus variables: intensity, spectral content, time, background, and spatial distribution of sound sources (binaural loudness) [296].

Timbre, also referred to as *sound quality* or *sound color*. The classic negative definition of timbre is: the perceptual attribute of sound that allows a listener to distinguish among sounds that are otherwise equivalent to pitch, loudness, and subjective duration. Contemporary research has begun to decompose this attribute into several perceptual dimensions of a temporal, spectral and spectro-temporal nature [218].

Spatial attributes of a hearing event may be divided into distance and direction:

The perception of a direction of a sound source depends on the differences in the signals between the two ears (*interaural cues*: interaural level difference (ILD) and interaural time difference (ITD)) and the spectral shape of the signal at each ear (*monaural cues*). Interaural and monaural cues are produced by reflections, diffractionsa and damping caused by the body, head, and pinna. The transfer function from a position in space to a position in the ear canal is called head-related transfer function (HRTF). The perception of distance is influenced by changes in the timbre and distance dependencies in the HRTF. In echoic environments the time delay and directions of direct sound and reflections affect the perceived distance.

2.1.4 Somatic senses

> **somatic:** 1: of, relating to, or affecting the body [...]; 2. of or relating to the wall of the body [217]

The main keywords related to somatic senses are *tactile* and *haptic* which are both related to the sense of touch (see 1.1.3). Concerning this sense, there is a lot more than only touch itself. For example, [120] distinguishes five "senses of skin": the sense of *pressure*, of *touch*, of *vibration*, of *cold*, and of *warmth*.

[120] indicates two more senses, called "sense of position" and "sense of force", related to the *proprioceptors*. The proprioceptors are receptors (special nerve-cells receiving stimuli) within the human body. They are attached to muscles, tendons, and joints. They measure for example the activity of muscles, the stressing of tendons, and the angle position of joints. This sense of proprioception is called kinesthesis and [120] calls the accompanying modality *kinesthetical*:

> **Kinesthesis (perception of body movements):** (physiology, psychology) Kinesthesis is the perception that enables one person to perceive movements of the own body. It is based on the fact that movements are reported to the brain (feedback), as there are:
>
> - angle of joints
> - activities of muscles
> - head movements (reported by the vestibular organ within the inner ear)
> - position of the skin, relative to the touched surface
> - movements of the person within the environment (visual kinesthesis)
>
> Kinesthesis supports the perception of the sense organs. If some informations delivered by a sense organ and by kinesthesis are contradictory, the brain will prefer the information coming up from the sense organ.
> Human computer interaction makes use of kinesthesis, e. g. if a key has a perceptible point of pressure or if the hand performs movements with the mouse to position the mouse-pointer on the screen, and so on [67].

For a multimodal system, kinesthesis is not as relevant as it is for a VR system, because aspects like experiencing an outer influence to the sense of balance, e. g. when wearing a head-mounted display (HMD), will not be a major topic here. Therefore, the most relevant somatic sense is the *sense of touch*, which can be addressed by special output devices[1] either with tactile or with force feedback (see 2.2.4).

According to [309], four different types of touch receptors (or *mechanoreceptors*) are present at the human hand[2]. They have different characteristics which will not be addressed within the scope of this report. As tests have shown, the output response of each receptor decreases over time (called *stimulation adaptation*) for a given input stimulus. The *2-point discrimination ability* is also very important. The index finger pulp is able to sense all points with a distance of more than 2 mm whereas in the center of the palm two points that are less than 11 mm apart feel like only one. Other relevant factors are the amplitude and the vibration frequency of the contactor.

2.2 Computer Output Media

The second part of the machine-to-man communication is mainly influenced by the devices and methods that are used to address the senses described in sections 2.1.2 – 2.1.4. Despite the technical limitations that still exist today, there are a number of ways for machines to communicate with their users. Some of them will be presented in the following sections.

2.2.1 Output modalities

Under output modalities or computer output media, we understand the media (devices) and modalities (communication channels) which are used by computers to communicate with humans. As the communication channels for computer output are the same ones as those for human input (see figure 1.1), the following subsections are structured with respect to those in section 2.1. Therefore, we renounce to repeat what has been said there and will instead investigate the devices and methods for the three "main senses" in the following.

2.2.2 Devices and methods for visual output

As stated before (2.1.2), the human visual system has large dynamic ranges. Presently there are many tools available for generation of stimulating the visual sense for many of its properties, with computer displays being the predominant output communication channel. Modern computer display technology can generate stimulations which cover full

[1]Or, in most cases, an input device with additional output features.

[2]These are especially relevant to sense vibrations.

or significant part of practical dynamic ranges for visual properties like color and temporal resolution. While the computer displays are far from being completely 'transparent', they get high marks for their overall perceptual impact. In the near future one can expect full integration of high-quality video and TV-like capabilities with a computer-like contents, integrating and improving the information output from the different sources.

The biggest single deficiency in output devices appears in the spatial vision stimuli, as displays capable of covering a high-resolution wide field of view and especially of 3D visualization are uncommon. The display resolution will be still growing and its aspect ratio may change in the future to the 16:9 from the current 4:3, allowing for greater perceptual involvement. However, stereoscopic and 3D visualization need new solutions to become practical (the difference between the stereoscopic vision and 3D is that in the latter motion parallax is carried on, that is the objects will look differently from different angles of view). The lack of spatial effects has diminishing effect on the sensation of reality and involvement in the content of the scene. A more advanced aspect of this problem is the lack of full immersion in visual scenes, which is being tried to be solved with virtual reality devices. The virtual reality technology in its present form is nonergonomic in the sense that it can be used only for a limited time during one session. Wide-field high-resolution visualization with spatial effect would be optimal from the ergonomic point of view.

2.2.3 Devices and methods for acoustical output

In this section methods for presenting information via acoustical means will be presented. In real life humans are able to get a lot of information via the auditory modality. The communication by the means of speech is practiced every day. Even if sometimes we are not conscious about that, we use a lot of non-speech sounds in daily life to orientate and to get precise knowledge of the state of the environment. When acoustical output is used in technical interfaces the way we use sound every day life is adapted to serve as a powerful tool for information transfer. In the following text the underlying concepts or technologies will be presented. It is not the aim of this text to list different synthesis algorithms or to discuss the underlying hardware problems, although they are limiting the widespread use of some interesting methods. The existing audio hardware found in PC and recent workstations will be presented briefly.

Speech output

If the medium is speech the whole repertoire of learned vocabulary and meaning is exploited. In order to maintain a reasonable intelligibility the bandwidth should not be below three kHz, but higher bandwidth leads to a better quality. In a teleconference environment speech is generated by the participants, compressed, transmitted and uncompressed for presentation. Speech also can be used as interface machine-man interface. Recorded words or sentences can be played to inform the user about the status of the machine. Also information stored in textual form an be transformed into speech by the means of sophisticated synthesis algorithms. Hardware and software for this purpose is commercially available.

Another interesting application of speech is the use of speech inside of text documents. A lot of text editors offer to append non textual items like pictures and sound documents. Some tools for electronic mail support sound attachments, which might be a speech recording done by the sender. Some text editors allow to make annotations in text. At the position of the annotation an icon will appear. If the icon is activated either a textual or spoken annotation will appear.

Non-speech audio output

For non-speech output four main categories may be identified. Because in the recent literature the same terms are often used in different ways, first the definitions of the terms used in this report will be presented.

Auditory Icon Everyday sounds that convey information about events in the computer or in remote environment by analogy with everyday sound-producing events [118].
Example:

- sound of a trashcan $\longrightarrow$ successful deletion of a file
- machine sound $\longrightarrow$ computer process running

Earcons Tone-based symbol sets, wherein the verbal language is replaced with combinations of pitch and rhythmic structures each of which denotes something in the data or computer environment. They have language-like characteristics of rules and internal references [26].
Abstract, synthetic tones that can be used in structured combinations to create sound messages to represent parts of an interface [42].

Example:

- simple melody for indicating an error status

Sonification Use of data to control a sound generator for the purpose of monitoring and the analysis of data.
Examples:

- Mapping data to pitch, brightness, loudness, or spatial position [162]
- changes in bond market data $\longrightarrow$ changes of brightness of sound

Audification A direct translation of a data waveform to the audible domain for the purpose of monitoring an comprehension.
Example:

- listening to the waveforms of an electroencephalogram, seismogram, radio telescope data [162]

Sound spatialization

An individual sound can be filtered so that it appears to emanate from a specified direction when played by headphones. The filters simulate the distortion of the sound caused by the body, head, and pinna. Today fast DSP (digital signal processing) hardware allows real time filtering and dynamic updating of the filter properties, thus allowing interactive processes to be created for real time spatialization (e. g. allowing head movement and movment of virtual sources). The number of sound sources which can be presented simultaneously is limited by the processing capacity of the DSP-hardware. Presently available hardware like the Motorola 56001 can process two spatial sound sources in real time. Spatial sound is presently used in the auralization of models for room acoustics. Spatial sound gives an enhanced situational awareness and aids in the separation of different data streams ("Cocktail Party Effect"). Therefore, future applications could be in the fields of telerobotic control, air traffic control, data auralization and teleconferences.

Hardware platforms

The different types of computers where multimedia applications will be applied range from personal computers based on x86 family without any audio output hardware in the low price region and RISC-based workstations equipped with audio input and output hardware and digital signal processors designated for audio processing. In the low cost region a lot of suppliers sell additional hardware for audio output. The quality of the audio output ranges from cheap hardware with 8-bit resolution and a sampling frequency of 8 kHz up to cards with two and four audio channels with a 16-bit resolution and up to 48 kHz sampling rate. Most of these sound cards are also equipped with hardware for the synthesis of sound. These built in synthesizers support the GMIDI (General Musical Instrument Digital Interface) which is an extension of the MIDI protocol. In the workstation market most of the workstation like SUN Sparc 20 and Indigo use multimedia codecs which allow sampling rates from 8 kHz to 48 kHz, 8- and 16-bit resolution and two channel audio input and output. With this hardware sampled sound stored in files on the harddisk or CD-ROM or Audio-CDs can be play back. Additional MIDI hardware for sound generation can be connected if the serial port supports a transmission rate of 31.25 KBaud. Depending on the computational power of the CPU or signal processing hardware sound may be manipulated or synthesized on-line. Sound is either presented via the built in loudspeaker or headphones. Loudspeaker may be built in the monitor or in separate boxes. Most hardware realizations only use one speaker but some Apple computers have two speakers built in the monitor for stereo sound output. These overview of existing hardware shows that there is great diversity in the use hardware platforms and which results in big differences in perceived quality and the amount of information which can be transmitted. The future development will lead to sound output with two channels in HiFi-quality with designated hardware for signal synthesis and signal processing.

2.2.4 Devices and methods for tactile/haptic output

Many devices have been developed in the last 30 years in order to address the somatic senses of the human operator, but only few have become widely available. The most probable reason for that is that the devices are either not very useful or really expensive (from US $10,000 up to more than US $1,000,000). By "devices for tactile/haptic output" we mean devices that have been especially designed for this purpose. In some sense, a standard keyboard and mouse do also provide some kind of haptic feedback, namely the socalled *breakaway force* when a key or button, respectively, has been pressed. Although this is important as tests have proven to increase the input rate in the case of a keyboard, we do not consider these devices within this section.

Devices with tactile, haptic, or force output address the somatic senses of the user (see 2.1.4). This can be done by the following methods (taken from [309]):

Pneumatic stimulation This can be achieved by air jets, air pockets, or air rings. Problems arise due to muscular fatigue and the pressure or squeezing effect which means that the ability to sense is temporarily disabled. Another drawback of pneumatic devices is its low bandwidth.

Vibrotactile stimulation Vibrations can either be generated by blunt pins, voice coils, or piezoelectric crystals. These devices seem to be the best ones to address somatic senses because they can be build very small and lightweight and can achieve a high bandwidth.

Electrotactile stimulation Small electrodes are attached to the user's fingers and provide electrical pulses. First results are promising, but further investigation is needed in this area.

Functional neuromuscular stimulation (FMS) In this approach, the stimulation is provided directly to the neuromuscular system of the operator. Although very interesting, this method is definitely not appropriate for the standard user.

Other methods do not address the somatic senses directly. For example, a force-reflecting joystick can be equipped with motors that apply forces in any of two directions. The same method is used for the Exoskeleton. In the following paragraphs, mainly devices with indirect stimulation methods will be described.

Nearly all devices with tactile output have been either developed for graphical or robotic applications[3]. Many different design principles have been investigated, but the optimal solution has not been found yet. Most probably, the increasing number and popularity of Virtual Reality systems will push the development of force feedback devices to a new dimension. In the following, the most popular devices will be reviewed very briefly in chronological order:

The Ultimate Display (1965) Ivan Sutherland described his vision of an "ultimate display" in order to reflect the internal world of the machine as close as possible. He proposed to develop a force reflecting joystick [334].

GROPE (1967 – 1988) In the project GROPE, several "haptic displays" for scientific visualization have been developed in different stages of the project [48]. Starting in 1967, a 2D device for continuous force feedback (similar to an X/Y-plotter) has been developed. In the next stage (1976), a 6D device was used in combination with stereo glasses for operations in 3D space. The device was a kind of master manipulator with force feedback. In the third (and last) phase of GROPE which started in the late 1980s, the hardware has been improved, but the principle layout of the system

[3] An exception which will not be further investigated here is an output device for the blind. Small pins or needles are used to output text in the braille language.

was not changed. The results in the domain of molecular engineering seem to be very promising, i. e. the performance has been increased significantly by the use of haptic feedback.

Joystring (1986) A very special design of a 6D manipulator with force feedback has been realized independently by Agronin [4] and Staudhamer (mentioned in [103], further developed by Feldman). A T-shaped grip is installed inside a box with three strings on each of its three ends[4]. In Agronin's version, the strings' tension is controlled by three linear motors, whereas the second one uses servo motors. No results have been published.

The Exoskeleton (1988) For telerobotic applications, different skeletons which have to be mounted on the operators arm have been built at the JPL [150], the EXOS company [87], and the University of Utah (in [269]). The systems are used as a kind of master-slave combination, and forces are applied by motors at the joints. Unfortunately, these devices are usually very heavy, therefore they can also be used in special applications. EXOS, Inc. has developed a "light" version for the NASA, but this system does not have any force feedback.

The Compact Master Manipulator (1990) The master manipulator that is presented in [148] is based on flight simulator technique. All three translational and rotational axis can be controlled and are equipped with force feedback. Additionally, the thumb, the indexfinger, and the other three fingers control a grip which also applies forces to the user, thus yielding a 9-DOF device with force feedback. Unfortunately, the paper only covers the design of the manipulator but does not contain any results.

A 3D Joystick with Force Feedback (1991) A joystick for positioning tasks in 3D space has been developed by Lauffs [174]. The 3 axis of the stick can be controlled independently, but rotations are not possible. Force feedback has been realized by the use of a pneumatic system, which is very robust but too slow for most applications.

A Force Feedback Device for 2D Positioning Tasks (1991) A device that is very similar to the one developed in GROPE I (see above) has been realized by Fukui and Shimojo [111]. Instead of a knob, the X/Y-recorder is moved with the finger tip. The resulting force will be calculated and sent to the application, and if a collision is detected, one or both axis will be blocked. This device has been developed for contour tracking operations, its advantage is its almost friction and mass free operation.

[4]Obviously, the name "joystring" has been derived from the two words "joystick" and "string".

PUSH (1991) In [131], a **P**neumatic **U**niversal **S**ervo **H**andcontroller is described. It has been developed for the control of industrial robots. By using cardan joints and an orthogonal coordinate system which is placed in the operator's hand, all axis can be controlled independently. The device, which is rather large, can apply forces by pneumatic cylinders. As the 3D joystick described above, it is very slow.

Teletact (1991) A data glove with tactile feedback has been developed by Stone [326] and is used for outputs to the user, whereas a second data glove is used for inputs to the computer. The input glove is equipped with 20 pressure sensors, and the output glove with 20 air pads, controlled by 20 pneumatic pumps. The major drawback of this system is the very low resolution, additionally a force feedback is missing completely. The next generation, Teletact-II, has been equipped with 30 air pads and is available on the market now.

Force Dials (1992) A dial with force feedback has been realized by computer scientists and chemists for simple molecular modeling tasks [127]. The force is controlled by a motor. The main advantage of this device is its low price and its robustness but due to its simplicity it will not be very useful for a multimodal system.

Multimodal Mouse with Tactile and Force Feedback (1993) In [6], an interesting approach to equip a standard input device with output capabilities has been described. A common mouse has been equipped with an electro magnet and a small pin in its left button. This idea is especially appealing because it is cheap and easy to realize. First results are very promising and have shown that the performance in positioning tasks can be increased with this kind of feedback by about 10%.

PHANToM (1993) In his master thesis, Massie developed a 3D input device which can be operated by the finger tip [214][5]. It realizes only translational axis, but it has many advantages compared to other devices, like low friction, low mass, and minimized unbalanced weight. Therefore, even stiffness and textures can be experienced.

A 2D Precision Joystick with Force Feedback (1994) Sutherland's basic idea has been realized by researchers at the University of New Brunswick [10]. The joystick has been made very small in order to achieve a very high precision in its control. Results have shown that the accuracy in a contour modification task can be increased (44%), but the time will increase (64%), too.

[5] In the meantime, Massie founded his own company which manufactures the PHANToM. Its price is US$ 19,500, which is cheap compared to other ones.

2.3 Bi- and Multimodal Perception

2.3.1 Visual-acoustical perception

In biological systems, acoustical and visual input devices and processing systems evolved to a very high degree of sophistication, enabling for receiving and extracting precise information about the current status of environment. Acoustical and visual sensory systems serve also for complex communication tasks which, at least in the case of humans, can carry information complexity of almost any degree. Taken separately, biological functions of acoustical and visual systems are implemented with neural systems showing remarkable organization, which is yet not fully comprehended and understood. Having not a full understanding of each of the systems is making the task of investigating their cooperation very difficult. However, looking at the integrative aspects of acoustical and visual system may in fact lead to a better understanding of their fundamental principles.

It can be seen that in many cases integrative functions of acoustical and visual systems are similar and complementing each other. One of the first tasks of the neural system is to built a representation of environment composed of objects placed in a physical space (practically cartesian) with three spatial dimensions and time. Both visual and acoustical system are extremely well fitted for recovering and building of spatial and temporal representations. They can both represent space by mechanisms like spatial vision and spatial hearing. These mechanisms are separate but their operation is integrated in subtle ways in order to build most probable and consistent representation of the environment. The consistency of the representation brings up interesting theoretical problems. One fundamental problem is the nature of integrated audiovisual representation, if and how it is built on top of the single-modality representation. Another problem is the usual bottom-up versus top-down division in the organization of neural processing, with both organizations participating in a precisely tuned way.

Sensory integration aims for building consistent representation and interpretation of information flowing from the different senses. This integration can deal with various aspects of information and can take place at different levels of processing. As we do not know exactly what are the different aspects and levels of processing, it is hard to devise precise taxonomy of the integration. The taxonomy can be approximately devised by looking into the functions and mechanisms of different sensory systems. One of the basic functions is representation of spatial and temporal properties of the environment. Time and spatial dimensions are basic variables for every representation and they obviously have to be represented consistently. Tuning of the measurements coming from the different senses concerning space and time to built the most consistent representation is the sensory integration process. In spatial representation, space filled with objects is considered, while for time representation we talk about events. One can talk about spatiotemporal objects, taking into account both time and spatial properties of objects. The main problem is how the both systems operation is integrated into a single audiovisual spatiotemporal representation.

It is widely recognized that our knowledge about the audiovisual integration is quite limited. This is because of the complexity of the systems involved. Both acoustical and visual systems taken separately have their own sophisticated organizations and integration of their functions is done on top of them. The current status of our understanding can be described as basic experimental level. In [321] there is a review of experimental and physiological facts concerning sensory integration. The topic has been studied mostly by psychological experiments which try to reveal specific properties without providing explanations for the mechanisms responsible. From the application and multimedia point of view these experiments are interesting as a background, but they seem to have a narrow scope depending very much on particular experimental conditions. This stems from the fact that the underlying mechanisms are usually well-hidden as emphasized by Radeau [280]. To uncover the mechanisms, one has to devise experiments putting them in conflict or conditions for cross-modal effects like in [353, 318, 249, 282]. There is one basic issue with these approaches, namely that the results are highly dependent on test signals and conditions.

At present only one rule can be formulated which seems to be in place: the higher the signals interaction and complexity, the more prominent are cross-modal effects between the acoustical and visual system. Relevant taxonomy of interaction and signals is given in section 4.3, Synchronization. Because of this rule, the cross modal effects tend to be strongest in the case of speech [208, 128].

2.3.2 Visual-speech perception

The intrinsic bimodality of speech communication

In 1989, Negroponte [242] predicted that "the emphasis in user interfaces will shift from the direct manipulation of objects on a virtual desktop to the delegation of tasks to three-dimensional, intelligent agents parading across our desks", and that "these agents will be rendered holographically, and we will communicate with them using many channels, including speech and non-speech audio, gesture, and facial expressions." Historically, the talking machine with a human face has been a mystical means to power for charlatans and shamen. In that vein, the first speaking robots were probably the famous statues in ancient Greece temples, whose power as oracles derived from a simple acoustic tube! The statues were inanimate at that time, even though their impressionable listeners attributed a soul (anima) to them, because of their supposed speech competence. If this simple illusion already made them seem alive, how much more powerful would it have been if the statue's faces were animated? One can only wonder how children would perceive Walt Disney's or Tex Avery's cartoon characters if their facial movements were truly coherent with what they are meant to say, or with its dubbing into another language. Of course, these imaginary characters are given so many other extraordinary behavioral qualities that we easily forgive their peculiar mouth gestures. We have even become accustomed to ignoring the asynchrony between Mickey's mouth and his mouse voice.

What about natural speech? When a candidate for the presidency of the United States of America exclaims "Read my lips!", he is not asking his constituency to lip-read him, he is simply using a classical English formula so that his audience must believe him, as if it was written on his lips: If they cannot believe their ears, they can believe their eyes! But even though such expressions are common, people generally underestimate the actual amount of information that is transmitted through the optic channel. Humans produce speech through the actions of several articulators (vocal folds, velum, tongue, lips, jaw, etc.), of which only some are visible. The continuous speech thus produced is not, however, continuously audible: It is also made of significant parts of silence, during voiceless plosives and during pauses, while the speaker makes gestures in order to anticipate the following sound. To sum up, parts of speech movements are only visible, parts are only audible, and parts are not only audible, but also visible. Humans take advantage of the bimodality of speech; from the same source, information is simultaneously transmitted through two channels (the acoustic and the optic flow), and the outputs are integrated by the perceiver. In the following discussion, I will pinpoint the importance of visual intelligibility of speech for normal hearers, and discuss some of the most recent issues in the bimodal aspects of speech production and perception.

Intelligibility of visible speech

It is well known that lip-reading is necessary in order for the hearing impaired to (partially) understand speech, specifically by using the information recoverable from visual speech. But as early as 1935, Cotton [73] stated that "there is an important element of visual hearing in all normal individuals". Even if the auditory modality is the most important for speech perception by normal hearers, the visual modality may allow subjects to better understand speech. Note that visual information, provided by movements of the lips, chin, teeth, cheeks, etc., cannot, in itself, provide normal speech intelligibility. However, a view of the talker's face enhances spectral information that is distorted by background noise. A number of investigators have studied this effect of noise distortion on speech intelligibility according to whether the message is heard only, or heard with the speakers face also provided [330, 241, 21, 93, 95, 331].

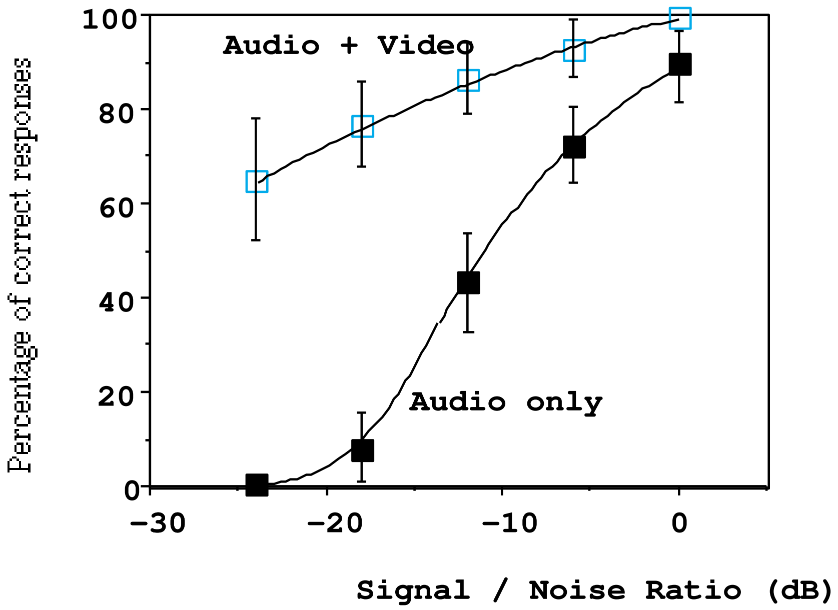

Figure 2.1: *Improved intelligibility of degraded speech through vision of the speakers face. The box indicates the mean, and the whiskers the standard deviation.*

Figure 2.1 replots articulation scores obtained in French by Benoît et al. on 18 nonsense words by 18 normal hearers in two test conditions: audition only and audition plus vision [17]. We observe that vision is basically unnecessary in rather clear acoustic conditions (S/N > 0 dB), whereas seeing the speakers face allows the listener to understand around 12 items out of 18 under highly degraded acoustic conditions (S/N = -24 dB) where the auditory alone message is not understood at all. One may reply that such conditions are seldom found in our everyday lives, only occurring in very noisy environment such as discotheques, in some streets or industrial plants. But using visual speech is not merely a matter of increasing acoustic intelligibility for hearers/viewers: it is also a matter of making it more comprehensible, i. e., easier to understand.

It is well known that information is more easily retained by an audience when transmitted over the television than over the radio. To confirm this, Reisberg et al. [286] reported that passages read from Kant's *Critique of Pure Reason* were better understood by listeners (according to the proportion of correctly repeated words in a shadowing task) when the speakers face was provided to them. Even if people usually do not speak the same way as Emmanuel Kant wrote, this last finding is a clear argument in favor of the general overall improvement of linguistic comprehension through vision. Therefore, it also allows us to better take into consideration the advantage of TtAVS synthesis for the understanding of automatically read messages, assuming that human-machine dialogue will be much more efficient under bimodal presentation of spoken information to the user. An average 11 dB "benefit of lip-reading" was found by MacLeod and Summerfield [199]. This corresponds to the average difference between the lowest signal-to-noise ratios at which test sentences are understood, given presence or absence of visual information. This finding must obviously be tempered by the conditions of visual presentation. Ostberg et al. tested the effects of six sizes of videophone display on the intelligibility of noisy speech [253]. They presented running speech to subjects who where asked to adjust the noise level so that the individual words in the story appeared at the borderline of being intelligible; they observed an increase in the mean benefit of lip-reading from 0.4 to 1.8 dB with the increase in display size. This observation confirms the intuitive idea that the better the visual information, the greater the improvement in intelligibility.

The need for coherence between facial gestures and speech sounds

The main problem researchers have to deal with in the area of speech production and bimodal speech perception (by ear and by eye) is the coherence of the acoustic and the visual signals (see [84, 208, 63], for extended discussions of this phenomenon). I will briefly present experimental results obtained from perceptual studies where various kinds of coherence were not respected: When the auditory and visual information channels have spatial, temporal, or source differences.

Spatial coherence It has been established that either modality influences spatial localization of the source through the other [20]): Subjects who are instructed to point at a visual source of information deviate slightly from it if a competing acoustic source is heard from another spatial position, and conversely, subjects deviate more from the original acoustic source if a competing optical source interferes from another location. In speech, such a capture of the source is well known and widely used by ventriloquists, as the audience is much more attracted by the dummy whose facial gestures are more coherent with what they hear than those of its animator [354]! Even four-to-five month old infants, presented simultaneously with two screens displaying video films of the same human face, are preferentially attracted by a face pronouncing the sounds heard rather than a face pronouncing something else [165]. This demonstrates a very early capacity of humans to identify coherence in the facial gestures and their corresponding acoustic production. This capacity is frequently used by listeners in order to improve the intelligibility of a single person in a conversation group, when the well-known "cocktail party effect" occurs.

Temporal coherence The second problem which arises from the bimodal aspect of speech perception is due to the inherent synchrony between acoustically and optically transmitted information. Dixon and Spitz have experimentally observed that subjects were unable to detect asynchrony between visual and auditory presentation of speech when the acoustic signal was presented less than 130 ms before or 260 ms after the continuous video display of the speakers face [83].[6] Mainly motivated by the applied problem of speech perception through the visiophone (where the unavoidable image coding/decoding process delays the transmission of optical information), recent studies tried to quantify the loss of intelligibility due to delayed visual information.

For example, Smeele and Sittig measured the intelligibility of phonetically balanced lists of nonsense CVC words acoustically degraded by a background interfering prose [315]. They measured a mean intelligibility of 20% in the auditory alone condition of presentation and of 65% in the audio-visual condition. However, if the facial presentation was delayed more than 160 ms after the corresponding audio signal, there was no significant improvement of audio-visual presentation over audio alone. In the other direction, Smeele (personal communication) more recently observed a rather constant intelligibility of around 40% when speech was presented in a range of 320 to 1500 ms after vision. In a similar experiment, Campbell and Dodd had previously discovered that the disambiguation effects of speech-reading on noisy isolated words were observed with durations of up to 1.5 sec desynchrony between seen and heard speech, but they indicated that this benefit occurred whichever modality was leading [54]. On the other hand, Reisberg et al. failed to observe any visual benefit in a shadowing task using the above mentioned text by Kant with modalities desynchronized at 500 ms [286]. These somewhat divergent findings strongly support the idea that audition and vision influence each other in speech perception, even if the extent of the phenomenon is yet unclear (i. e., does it operate on the acoustic feature, the phoneme, the word, the sentence, etc.?) and even if the role of short-term memory, auditory and visual, in their integration remains a mystery. I would simply suggest that the benefit of speech-reading is a function not only of the acoustic degradation, but also of the linguistic complexity of the speech material. The greater the redundancy (from nonsense words to running speech through isolated words), the more the high-level linguistic competence is solicited (in order to take advantage of the lexicon, syntax, semantics, etc., in a top-down process), and the more this cognitive strategy dominates the low-level bottom-up decoding process of speech-reading.

[6] Note that this delay sensitivity is much more accurate in the case of a punctual event, such as a hammer hitting an anvil, where the range is from 75 ms before to 190 ms after.

Source coherence Roughly speaking, the phoneme realizations that are the most easily discriminable by the ears are those which are the most difficult to distinguish by the eyes, and vice versa. For instance, /p/, /b/, and /m/ look alike in many languages, although they obviously sound unlike, and are often grouped together as one viseme. On the other hand, speech recognizers often make confusions between /p/ and /k/, whereas they look very different on the speaker's lips. This implies that a synthetic face can easily improve the intelligibility of a speech synthesizer or of a characters voice in a cartoon if the facial movements are coherent with the acoustic flow that is supposed to be produced by them. If not, any contradictory information processed during the bimodal integration by the viewer/listener may greatly damage the intelligibility of the original message. This dramatic effect can unfortunately result if the movements of the visible articulators are driven by the acoustic flow, e. g., through an acoustic-phonetic decoder. Such a device might involuntarily replicate the well-known McGurk effect [220], where the simultaneous presentation of an acoustic /ba/ and of a visual /ga/ (a predictable decoder error) makes the viewer/listener perceive a /da/! I must emphasize that the McGurk effect is very compelling, as subjects who are well aware of the nature of the stimuli even fall for the illusion. Moreover, Green et al. found little difference in the magnitude of the McGurk effect between subjects for whom the sex of the voice and the face presented were either matched or mismatched [129]. They concluded that the mechanism for integrating speech information from the two modalities is insensitive to certain incompatibilities, even when they are perceptually apparent.

The specific nature of speech coherence between acoustics and optics

Speaking is not the process of uttering a sequence of discrete units. Coarticulation systematically occurs in the transitions between the realizations of phonological units. Anticipation or perseveration across phonetic units of articulator gestures in the vocal tract are well known for their acoustic consequences, i. e., for the differences in allophones of a single phoneme. In French, for instance, the /s/, which is considered a non- rounded consonant, is spread in /si/, but protruded in /sy/, due to regressive assimilation; on the opposite, /i/, which has the phonological status of a spread phoneme, is protruded in /Si/, due to progressive assimilation (which is less frequent). Such differences in the nature of allophones of the same phonemes are auditorily pertinent [65] and visually pertinent [62].

A classic example of anticipation in lip rounding was first given by Benguerel and Cowan who observed an articulatory influence of the /y/ on the first /s/ in /istrstry/ which occurred in the French sequence une sinistre structure [14] (though this has since been revised by Abry & Lallouache [3]). In fact, the longest effect of anticipation is observed during pauses, when no acoustic cues are provided, so that subjects are able to visually identify /y/ an average of 185 ms before it is pronounced, during a 460 ms pause in /i#y/ [64]. In an experiment where subjects had simply to identify the final vowel in /zizi/ or /zizy/ [98], Escudier et al. showed that subjects visually identified the /y/ in /zizy/ from a photo of the speakers face taken at around 80 ms before the time when they were able to auditorily identify it (from gated excerpts of various lengths of the general form /ziz/). They also observed no difference in the time when subjects could identify /i/ or /y/, auditorily or visually, in the transitions /zyzi/ or /zyzy/. This asymmetric phenomenon is due to non- linearities between articulatory gestures and their acoustic consequences [323]. In this example, French speakers can round their lips — and they do so! — before the end of the /i/ in /zizy/ without acoustic consequences, whereas spreading the /y/ too early in /zyzi/ would lead to a mispronunciation and therefore to a misidentification.

To acoustically produce a French /y/, lips have to be rounded so that their interolabial area is less than 0.8 cm^2, above which value it is perceived as /i/ [2]. Lip control is therefore much more constrained for /y/ than for /i/, leading to an anticipation of lip rounding in /i/ → /y/ transitions longer than that of lip spreading in /y/ → /i/ transitions. We see from these observations that coarticulation plays a great role in the possibilities for subjects to process visual information before, or in absence of, acoustic information. This natural asynchrony between the two modes of speech perception depends upon the intrinsic nature of phonetic units, as well as on the speech rate and the individual strategy of the speaker. It is obvious that the increase of intelligibility given by vision to audition relies on it.

The bimodality of speech

In the following two paragraphs, two different aspects of the bimodality of speech will be presented: *synergetic* and *specific* bimodality.

The synergetic bimodality of speech Setting communications parameters at threshold level, Risberg and Lubker observed that when a speaker appeared on a video display, but with the sound turned off, subjects relying on speech-reading correctly perceived 1% of test words [290]. When the subjects could not see the display, but were presented with a low-pass filtered version of the speech sound, they got 6% correct. Presented with the combined information channels, the performance jumped to 45% correctly perceived test words. This observation exemplifies the remarkable synergy of the two modes of speech perception. However, little is known about the process that integrates the cues across modalities, although a variety of approaches and of models to multimodal integration of speech perception have been proposed [208, 332, 40] and tested [211, 213, 128, 210, 291]. Our understanding of this process is still relatively crude, but its study is very active and controversial at present (see the 21 remarks to and in [209]!). For an extensive presentation of the various integration models proposed and tested in literature, see the ICP-MIAMI 94-2 Report by Robert-Ribes.

The specific bimodality of speech To sum up these various psycholinguistic findings: As concerning speaker localization, vision is dominant on audition; for speaker comprehension, vision greatly improves intelligibility, especially when acoustics is degraded and/or the message is complex; this speech-reading benefit generally holds even when the channels are slightly desynchronized; due to articulatory anticipation, the eye often receives information before the ear, and seems to take advantage of it; and finally, as for localization, vision can bias auditory comprehension, as in the McGurk effect. The Motor Theory of speech perception supposes that we have an innate knowledge of how to produce speech [187]. Recently, in a chapter of a book devoted to the reexamination of this theory, Summerfield suggested that the human ability to lipread could also be innate [333]. His assumption allows a partial explanation of the large variability observed in human performance at speech-reading, as this ability seems to be related to the visual performance capacities of the subject [76, 295].

Summerfield also hypothesized that evolutionary pressure could have led to refined auditory abilities for biologically significant sounds, but not for lipreading abilities. Therefore, whatever the innate encoding of speech, whether in an auditory or visual form, an intermediate stage of motor command coding allowing us to perceive speech would provide us not only with the coherence of acoustic and visual signals in a common metric, but also with an improvement in the processing of the speech percept (whose final storage pattern is still an open question). This is my interpretation of the famous formula "Perceiving is acting", recently revised by Viviani and Stucchi [347] into "Perceiving is knowing how to act".

Chapter 3

Control and Manipulation

In this chapter, we will describe the processes of control and manipulation as performed by users of a computer system. Again, two components can be discerned (Figure 1.1, left part). The first component we will approach consists of the "human output channels" (HOC). Theories on motor control will be reviewed briefly. The second component consists of the "computer input modalities" (CIM) that are suitable to be interfaced to the human output activity patterns. In addition, event handling architectures which are relevant for different computer input modalities will be investigated. Finally, our approaches towards bi- and multimodal control schemes will be introduced.

3.1 Human Output Channels

First we will describe some relevant theories on human motor control from a historical perspective. In the Appendix, there will be sections on the HOC channels studied in MIAMI (speech, facial movements, handwriting, gesturing, and object manipulation). There are three theoretical mainstreams in this area.

First, the field of *Cybernetics*, i. e., the sub-field which led to Systems Theory, has exerted a profound influence on our insight of motor control as a process involving target-directed activities and error correction by feedback.

Second, *Cognitive Motor Theory* exposes the need for a description of the representational and computational aspects of motor control.

Third, the *Systems Dynamics Approach* is powerful in explaining a number of motor control phenomena, especially in oscillatory behavior.

3.1.1 Cybernetics: Closed-loop control

Cybernetics is defined as "[...] the study of control and communication in the machine or in the animal [...]" [362]. In the current context, we will use the term cybernetics in a narrower sense, i. e., as referring to the study of control systems. The communication and information-theoretical aspects that were originally partly embedded in cybernetics, are currently studied in a different field, called informatics or computer science. The name cybernetics comes from the Greek word for steersman ($\kappa\upsilon\beta\epsilon\rho\nu\acute{\eta}\tau\eta\varsigma$), and in fact does not have connotations with respect to communication or information. More specifically, even, we will only speak of systems controlling a physical parameter. The following quotation clarifies how Wiener himself envisaged the control problem:

> "[...] what we will is to pick the pencil up. Once we have determined on this, our motion proceeds in such a way that [...] the amount by which the pencil is not yet picked up is decreased at each stage [...]"
>
> [362, page 14]

Since the development of technical servo systems [362], the interest in cybernetics as a paradigm has been increasing and fading in a number of fields, varying from engineering, biology and psychology to economics. Research in cybernetics has led to powerful mathematical tools and theoretical concepts, gathered under the heading of Control Theory.

In theories on motor control, this "Closed loop model" considers the sensorial feedback (tactile and/or acoustic) to be the intrinsic source of control, e. g., of the articulatory activities during speech or handwriting production. The articulatory trajectories are planned on-line depending on the measured feed-backs until the target (spatial or acoustic) is reached. Although this model can explain a number of phenomena in motor control, its basic shortcoming is the fact that propagation delays in the biological system are substantial, necessitating other concepts to explain principles like feedforward, predictive control. Tables 3.1 and 3.2 show a number of typical biological reaction and delay times.

3.1.2 Open-loop models

Cybernetics elegantly describes some existing phenomena in motor control. In speech, pointing and handwriting movements the existence of feedback loops such as the monosynaptic reflex arc and the cortico-cerebellar loops introduce a self-organizing autonomy into the effector system. At the same time however, the concept of feedback is insufficient to explain the corrective properties of motor control in case of absent or delayed sensory

The simple reaction time between seeing a light flash and pushing a button with the index finger (Donders's a-type reaction) is, in adults [154]:	230 ms
The reaction time between seeing a light flash bar and pushing a button with the index finger in a binary decision task is, in adults [172]:	419 ms
The 95 percentile reaction time to brake a car in case of an unexpected obstacle [251]:	1600 ms

Table 3.1: *Typical reaction times*

The time between a passive stretch of an arm muscle and the arrival at the alpha motoneurons in the spine of the afferent neuroelectric burst coming from the muscle spindles is [207]:	23 ms
The time between a discharge of an alpha motoneuron and the peak twitch force of the muscle fibers belonging to the motor unit is [130]:	30–100 ms
The effective duration of a motor unit twitch is (ibid.):	$>$ 60 ms
Estimated time between the occurrence of a motor discharge at the motor cortex and the discharge of hand alpha motoneurons (type A fiber, conduction speed 100m/s) is:	$\pm$ 10 ms
Idem, measured in monkeys [299]:	11 ms
The time between presentation of a light flash and its arrival at the visual cortex [369]:	35 ms

Table 3.2: *Typical delay times*

information. Also, the *origin* of complex patterns like writing is left implicit in a pure cybernetical theory.

Experiments by [23, 24] played an essential role in the paradigmatic shift in which feedback as such was increasingly considered to be inadequate as an general explanation of motor control. It was shown that in fast aiming movements of the head or the arm [349], final targets could be reached in the absence of essential feedback information (visual, vestibular, or proprioceptive feedback). The explanation for this phenomenon that was put forward, and that is still accepted for the greater part today, is that the central nervous system determines in advance of such an aiming movement, the ratios of muscle activation (co-contraction) levels. In this view, the motor apparatus is a combination of tunable mass-spring systems. The role of the existing feedback loops was consequently limited to (1) slow and fine adjustment as in posture control, to (2) adaptation to new or strange postures [350], not internalized by the "programming" system, and (3) and to learning.

More cognitively oriented researchers interpret the "Open loop model" in terms of Computer Science and Artificial Intelligence concepts (like the Turing machine concept) and utilize the metaphor of "brain like a computer". It is assumed that each articulatory activity is controlled through a sequence of instructions [106, 107] (Magno-Caldognetto-and-Croatto-1986,Harris-1987,Keller-1987, Wilson-1987,Wilson-and-Morton-1990). In the extreme (purist) case, the articulatory trajectories are pre-programmed (off-line) point by point and are executed deterministically. However, it is well known that only a limited number of parameters is needed to produce curved trajectories. The impulse response of the biomechanical output system, and the well-timed production of intrinsically primitive electro-myographical burst patterns, leads to smooth trajectories by itself. The details of such a movement do not have to be stored explicitly in the brain. What the neural control system does, in fact, is to represent the inverse transform of the non-linear output system in such a way that (1) non-linearities may be counteracted, and (2) useful properties, like the energy present in a mechanical oscillation, may be exploited. The third group of theories elaborates on this latter idea.

3.1.3 Coordinative structure models

The "Coordinative model" [339, 106, 163, 164, 307] (Scott Kelso 1983,Salzman 1986, 1991,Keller 1990) is based on the metaphor of generalized physics models. Other names referring to the same group of theories are "Synergetics" or "Systems Dynamics" (as applied to motor control). This theoretical direction developedin contrast to cybernetic and algorithmical models. It is based on coordinative structures, that is functional groups of muscles behaving like coordinated units according to rules learnt by training. The articulatory trajectories are the result of the interactive work of coordinative structures aimed at reaching a “qualitative” target in a flexible and adaptive way (Saltzman 1991). Using tools from non-linear systems dynamics, oscillatory behaviors like walking and juggling with balls can be nicely described. However, the parsimony of the approach breaks down in complex patterning tasks. As an example we can take cursive handwriting. In this motor task, discrete action pattern units are concatenated. Even if we would succeed in describing an oscillator configuration for a single letter, or even two letters, how then are the basic action units concatenated? Is this done by an associative process, or by ”buffering”? The same argument holds for speech and manipulation tasks. In order to solve this problem, cognitive representational concepts will be needed in a theory of motor control, eventually.

3.1.4 Relevance of these theories for multimodal interaction

In the context of theorizing on multimodal man-machine interaction, it must be pointed out that phenomena in motor control (HOC) at the physical/physiological level may not and cannot be disconnected from the higher, cognitive and intentional levels. Control problems at first considered intractable, like the degrees of freedom problem (inverse kinetics and kinematics) may become very well tractable if task-level constraints are taken into consideration. A simple example is a 3 df planar, linearly linked arm, for which the mapping from (x, y) to joint angles is an ill-posed problem, unless also the object approach angle for the end effector (the last link) is known for a specific grasp task.

The second observation is related to the consequences of multi-channel HOC. What happens if more human output channels are used in human-computer interaction? Two hypotheses can be compared:

Hypothesis (1): The combination of different human output channels is functionally interesting because it effectively increases the bandwidth of the human→machine channel. Examples are the combination of bimanual teleoperation with added speech control. A well-known example is the typing on a keyboard where key sequences which require alternating left and right hand keying are produced 25% faster (50 ms) [250] than a within-hand keying sequence. The reason is thought to reside in the independent, parallel control of left and right hand by the right and left hemispheres, respectively. It may be hypothesized that other forms of multimodality profit in a similar way, if their neural control is sufficiently independent between the modalities involved.

Hypothesis (2): The alternative hypothesis goes like this: Adding an extra output modality requires more neurocomputational resources and will lead to deteriorated output quality, resulting in a reduced effective bandwidth. Two types of effects are usually observed: (a) a slowing down of all the output processes, and (b) interference errors due to the fact that selective attention cannot be divided between the increased number of output channels. Examples are writing errors due to phonemic interference when speaking at the same time, or the difficulty people may have in combining a complex motor task with speaking such as in simultaneously driving a car and speaking, or playing a musical instrument and speaking at the same time. This type of reasoning is typical for the *cognition-oriented models* of motor control and may provide useful guidelines for experimentation.

At the physical level, however, the combination of output channels may also result in a system which can be described as a set of coupled non-linear oscillators. In the latter case, it may be better to use the *Coordinative Structure Models* to try to explain inter-channel interaction phenomena, rather than trying to explain phenomena on a too high, cognitive, level.

And finally, the effective target area of *Cybernetical, Closed-loop Theories* will be those processes in multimodal motor control which can be classified as tracking behaviour or continuous control (as opposed to discrete selection tasks).

3.2 Computer Input Modalities

The human output channels are capable of producing physical energy patterns varying over time: potential, kinetic and electrophysiological energy. Of these, the resulting force, movement and air pressure signals are the most suitable candidate for the development of transducer devices, such as has been done in practice already: keys, switches, mouse, XY-digitizer, microphone, or teleoperation manipulandum. As regards the electrophysiological signals, the following can be said. From the development of prosthetic devices for disabled persons, it is known that the *peripheral nervous signals* are very difficult to analyse and to interpret in a meaningful way. This is because the basic biomechanical transfer function characteristic is not as yet applied on the measured neuro-electric (ENG) or myoelectric (EMG) signals. Some recent advances have been made, incorporating the use of artificial neural network models to transform the electrophysiological signals, but a number of electrodes must be placed on the human skin, and often needle or axial electrodes are needed to tap the proprioceptive activity, which is usually needed to solve the transfer function equation. As regards the development of electrophysiological transducers, measuring *central nervous system signals*, it can be safely assumed that the actual use of these signals in CIM is still fiction (as is functional neuromuscular stimulation in the case of COM, page 27). Surface EEG electrodes only measure the lumped activity in large brain areas and the measured patterns are seemingly unrelated with implementation details of willful and conscious activity. As an example, the Bereitschafspotential is related to willful activity, in the sense that "something" is going to happen, but it cannot be inferred from the signal what that motor action will be.

Disregarding their impracticality for the sake of argument, implanted electrode arrays into the central nervous system would allow for a more subtle measurement of neural activity, but here the decoding problem is even more difficult than in the case of the electrophysiological signals derived from the *peripheral nervous system*. Therefore, we will concentrate on those types of computer input modalities, where input devices measure the force, movement, and sound patterns which are willfully generated by the human user by means of his skeleto-muscular system, under conscious control.

However, before looking into detail at all kinds of possible input and output interfaces that are possible on a computer, first it makes sense to look at what already has been achieved in modern user-interfaces. The first interactive computers only had a keyboard and a typewriter in the form of a Teletype (Telex) device. Later in the development, a CRT screen became the standard output. But not for all applications the keyboard is the most reasonable way of input. Many systems nowadays have a 'mouse', more or less as a standard feature. In the near future more input and output devices will be connected to computers. Already, the microphone for audio input is emerging as a standard facility on workstations and on the more expensive personal computers. Video cameras will probably be considered as a standard feature in the near future. It is good to consider what kind of operations are needed in many different applications, and what devices can handle these operations effectively.

Input devices can be divided into multiple groups, selected for functionality:

- Pointing devices (mouse, pen tablet, light pen, touch screen, data glove, ...)
- Keyboard (ASCII (Querty, Dvorac), Numeric keypad, Cursor keys, MIDI, ...)
- Sound input (microphone)
- Image input (camera)
- Others (sensors)

In fact it is possible for other devices than pointing devices only to input coordinates. For instance many systems allow the keyboard TAB-key to select input fields in a form, which has the same effect as pointing to the appropriate field.

Image input devices are currently not part of a user interface. They are used to input images, for further processing. Video cameras, however, gradually start being used interactively in videoconferencing applications.

Some input devices supply coordinates indirectly: The position of a joystick indicates movement in some direction, which can be translated into coordinate changes. Quick jumps from one place to another (as is possible with a pen) is difficult with a joystick. Also the arrow or cursor keys can be used to change the position on the cursor (In the GEM window manager, the ALT-key in combination with the cursor keys is used as an alternative for moving the mouse)

In principle, it is also possible to translate two speech parameters (e. g. volume and pitch) into screen coordinates. This might be useful for some applications for disabled people, being able to control the position of a screen object by voice only.

The simplest pointing device supplies two coordinates (x & y) and a button status. Mice might have more buttons and a pen might have a continuous pressure/height parameter. Pens might have additional parameters about the angles they make with the tablet (in x and y directions) and about the rotation of the pen. But these are only likely to be used in special applications.

3.2.1 Keyboards

Five possible groups of operations that can be performed by the keyboard are:

1. Text Input. Typing, mostly Western-language, characters which are displayed directly.
2. Action. Keys like ENTER/ESCAPE have generally the function of accepting/ cancelling the prepared action. Also keys like HELP, UNDO, DEL, BACKSPACE and the Function keys belong to this group.
3. Mode change. These are similar to the Action keys, but only work as on/off or selection switches. Examples are the CAPS-LOCK and the INS-key. In fact the SHIFT, CTRL and ALT-key could also be in this group.
4. Navigation. This is performed by the cursor keys, TAB, HOME, END.
5. Others (e. g., MIDI-keyboards)

Keys sometimes have different functions in different modes, or even in different applications. For instance on most keyboards CTRL-'[' is the same as the ESCAPE key.

The most common keyboard is the Qwerty-keyboard, used on most computers and typewriters. The meaning of a key is given by its label, not by its position. This is also true for the Dvorac keyboard. Even minor variations of key location over different keyboards are very annoying to the user.

There also exist chord-keyboards, in which the simultaneous stroking of a combinations of keys give some predefined output. The function is generally the same as the normal keyboard, only less keys are necessary. Letter order ambiguities in a chord are solved by an internal language model (example: "... [txr] ..." $\rightarrow$ "... xtr ..."), such that entering abbreviations or text in another language is not very well possible. In these cases, the user must fall back on single-key entry.

Another kind of keybord exists for a completely different purpose, but still belongs to the same group of input device: MIDI-keyboards. MIDI (Musical Instrument Digital Interfaces) is a standard for exchanging music- related information. MIDI-keyboards have keys as on a piano or organ. The only labels are the color (black and white), and the further meaning of the key is fully signified by the position. This keyboard often is pressure-sensitive, allowing the volume of each sound to be controlled by the keypress only.

3.2.2 Mice

Mice are the most common input devices for graphical user interfaces (GUIs). They come in different shapes and are equipped with one, two, or three buttons which are used for positioning and selection tasks. Mice can work either mechanically or optically. Within this report, the method of construction is not as interesting as the functionality of the mouse.

Only three of the above mentioned operations are possible with the mouse:

1. *Action* On schreen some action buttons are displayed like OK or CANCEL. Moving the mouse pointer to these buttons and clicking the mouse button starts the desired action.

2. *Mode change* This can be performed by clicking check-boxes or radio-buttons. But also the clicking on windows to select them, can be seen as a mode change.

3. *Navigation* This is the most natural use the mouse is meant for.

   But through software the mouse can do more:

   - Dragging. Selecting an item, keeping the mouse-button down. Moving the mouse to a new locationg and release the mouse button. The effect of this is moving or copying the item to a new location.

- Numerical ('Value') input: By selecting a scroll bar and dragging it, windows can be scrolled. Using sliders, a continuous input value can be given by dragging the slider control.

3.2.3 Pens

For the measurement of the pen-tip position in two dimensions, i. e., the writing plane, a wide variety of techniques exists. Using resistive, pressure sensitive, capacitive, or electromagnetical transducer technology, the position of a pencil can be accurately measured. Sometimes, even acceleration transducers inside the pen are used to measure pen movement [185].

Table 3.3 gives an overview of parameters that may be controlled with a pen.

x, y	position (velocity, acceleration, etc.)
p	pen force ("pressure")
	binary penup/pendown switch analog axial force transducer
z	height
ϕ_x, ϕ_y	angles
Switching	by thresholding of pen force p (above) or with additional button(s) on the pen

Table 3.3: *Parameters controlled by a pen*

Identifying the "inking" condition

An essential issue when using pens as an input device is the detection of the "inking" condition. This is not a trivial problem. In a normal writing implement, such as the ballpoint, the inking process starts when the pen tip touches the paper and the ball starts to roll, smearing the paper surface continually with ink. In electronic pens, we can make a distinction between two dimensions: the height of the pen tip above the writing plane, and the force exerted by the pen on the writing plane after contact is made. If the height of the pen tip is within a threshold, say 1 mm above the paper, this is usually identified as the "pen-near" condition. However, height is not a reliable identifier for inking and leads to annoying visible hooks in the measured trajectory at pen-down and pen-up stages.

A more reliable method is the measurement of the normal force. Direct measurement of the normal force can be done by touch tablets or by force transducers under the writing surface. Another technique is using the axial pen force. The pen may be equipped with a force transducer, which measures the force component exerted in the longitudinal direction of the stylus. Either the transducer is positioned at the end of the stylus within the barrel, or it is integrated within the stylus. If analog axial force is used to identify the “inking” condition, a force threshold must be used, above which we may safely assume that the pen is on the paper. However, analog amplifier drift and mechanical problems like stiction an hysteresis pose a problem. Writers have their typical average pen force (0.5 N – 2.5 N), such that the threshold cannot be fixed to a single small value. In practice, thresholds between 0.15 N and 0.4 N have been used for pen-up/pen-down identification. If necessary, the normal force may be calculated from the axial pen force if the pen angle with respect to the writing surface is known. Pen force is a useful signal for writer identification, but the force variations are mostly unrelated to character shape [302].

However, the most-often used technique for identifying the inking condition is an axial on/off switch inside the barrel. In the case of writing with an electronic pen with ballpoint refill on paper, there may be an inconsistency between the actual ink path on paper and the recorded pen-up/pen-down because the writer did not produce sufficient force to activate the switch. In such a case, pieces of the trajectory are misidentified as a pen-up sequence. Another problem with binary tip switches is that their axial movement is noticeable at each pen-down event, which disturbs the writing, drawing or menu selection movements by the user. Also, axial force transducers which infer axial force from the amount of compression (axial displacement) of the transducer itself suffer from this problem. An acceptable axial displacement at time of contact is 0.2 mm, but larger displacements are unacceptable for serious use. The use of pens with an axial displacement which is larger than 0.2 mm is restricted to ’sketching’-like drawing.

Integrated display and digitizer

The “inking” problems mentioned above are solved by using *Electronic Paper (EP)* which consists of an integrated display (e. g., a Liquid Crystal Display, LCD) and an XY-digitizer. One of the earliest EP devices consisted of a 12”x12”, 1024x1024 pixel plasma device with integrated XY-digitizer [44]. In the case of EP, a non-inking (“untethered”) stylus refill is used for the pen. By providing ’electronic ink’, i. e., a trace of legible pixels directly below the pen tip, writers will automatically adapt the exerted force to produce a clean ink path, due to the immediate graphical feedback. Electronic paper currently has a number of other problems, which have mainly to do with the quality of the LCD screen, the writing surface texture, and the parallax problem. The parallax problem refers to the perceived disparity between pen-tip position and ink pixels, due to the thickness of the display glass or plastic cover. Another problem is the friction of the writing surface. Plain glass, for instance would be much to glitchy for normal use, since writers are used to the higher friction of the paper surface. In writing on paper, the friction is assumed to be 4% of the normal force [85]. For this reason, the writing surface on electronic paper is treated to be both non-glaring and have an acceptable friction. Finally, the resolution and image quality of even the best LCD display is incomparable to the high resolution of the ballpoint ink on white paper. A typical dimensioning is 640 pixels over 192 mm. This yields a pixel width of 0.3 mm, which is three times worse than the average whole-field digitizer accuracy (0.1 mm), and 15 times larger than the local digitizer resolution (0.02 mm). Particularly problematic are unlit monochromatic (black & white) LCD screens. At least 64 gray levels are needed, which also allows for anti-aliasing techniques to be used for the graphical inking process. Currently, color TFT versions of electronic paper are appearing, but the relatively high-level signals of the in-screen transistors injects noise into the position sensing apparatus [312].

Signal processing issues

Using current state of the art technology [221, 352], a typical resolution of 0.02 mm and an overall accuracy of 0.1 mm can be achieved. Taking only into account the bandwidth of handwriting movements [338], a Nyquist frequency of 10-14 Hz would indicate a sampling rate of 20-28 Hz. However, since this means that there are only two sample points per stroke in a character, interpolation (e. g. by convolution) is absolutely necessary to recover the shape, and also the delay between samples will be annoying in interactive use. Using higher sampling rates is generally a cheaper solution. The most often used sampling rate is 100 Hz for handwriting, 200 Hz for signature verification, and the minimum for interactive use is probably about 50 Hz, yielding a delay of 20 ms, which happens to be 20% of a stroke movement. Another solution is to use spatial sampling, generating coordinates only if a position change above a given threshold is detected. This is similar to most types of sampling in the mouse. However, the disadvantage of spatial sampling is that it is not a proper time function, such that operations like digital filtering are ill-defined. This drawback is sometimes counteracted by adding time stamps to each detection of position change.

Pen ergonomics

Originally, pens provided with digitizer tablets had a diameter of 10-14 mm and were connected to the tablet by a wire. Such a pen thickness is unusable for handwriting, especially if the pen tip is not visible while writing. From the ergonomical point of view, a good guideline for the pen thickness is a diameter of 8 mm [194]. Fortunately, current technology allows for wireless pens, since the wire was a continual nuisance (curling, risk of connection damage, etc.). In the case of touch-sensitive technology, any object with a sharp tip may be used (as long it does not scratch the surface). In case of electromagnetic tablets, a transponder technique is used. Formerly, either the tablet grid or the pen coil was the transmitter, and the other component was the receiver of a high-frequency carrier (50-150 kHz). The peak of the electromagnetic field strenght corresponds to the (estimated) pen-tip position. With the transponder technique, modern tablets transmit in short bursts and receive the emitted resonant signal from a coil and capacitor circuit which is fitted inside the pen and tuned to the carrier signal. The tablet thus alternates between transmission and reception. Axial pen force can be coded by, e. g., capacitively, de-tuning the circuit in the pen. A disadvantage of the wireless pen in electromagnetic devices, is that it is expensive and may be easily lost, or mixed up with regular pens. For this reason, pen-based Personal Digital Assistents are equipped with touch-sensitive tablets, on which any pen-shaped object may be used.

In addition to the internal "inking" switch, electronic pens may have a barrel switch to the side of the stylus. It is still unclear whether this unnatural addition, as compared to the normal pen-and-paper setup, is ergonomically acceptable. Two problems are present. First, the index finger is normally used for ensuring a stable pen grip. Applying force to the barrel button disturbs this posture-stability function of the index finger. Especially the release movement from the button is incompatible with the normal pen manipulation. The finger must 'find' a free area on the pen after the switching action. The second problem is more on the software and application side. Since there is no standard approach for using the side switch over several application, the user may forget the function of the side switch, much as mouse users often forget the function of the second (and/or third) mouse button.

Other signals

In acceleration-based pens, the pen angle and rotation may be derived from the acceleration signals. These pens are very expensive and usually have a deviant shape and/or size. In electromagnetic systems, the skewness of the sensed electromagnetic field function is indicative of the pen angles with respect to the writing plane [194]. This property is also used in commercial systems to improve the accuracy of the pen-tip position estimation.

3.2.4 Cameras

This device can, dependent on the frame speed, supply single images or moving-image video to the computer. Because the amount of data coming from such a device is large, currently only live display on a screen window, or data compression and storage for later playback are currently possible on the computer in practical settings. For more advanced features much more processing power is necessary than is generally available today. Within the context of MIAMI, the camera will be used for lip reading and videophone experiments.

Current camera technology makes use of optically sensitive charge-coupled devices (CCD). A density of 400k pixels is standard (e. g., 752x582). Optical signal-to-noise ratios may be 50 dB, and a minimum lighting condition of 0.9 lux. More and more, camera parameters like tilt, pan, zoom, focus, and diaphragm may be remotely controlled from the computer.

3.2.5 Microphones

The microphone and the Camera have in common that recording, storage and playback are already useful (in fact the same holds for the pen, too). Because audio has lower demands than video, much more is possible. Some application could be controlled by speech recognition. Or the incoming sound itself could be processed in some way (filtered).

The most useful application of sound in computers will be speech recognition. Several methods already proved to be capable of doing reasonable recognition in real-time. Especially the Hidden Markov Model (HMM) is becoming popular.

3.2.6 3D input devices

Because 3D input devices are mainly used for special applications, they will not be described here in detail. Nevertheless, some devices will be reviewed briefly:

Spaceball A device which provides three translational and three rotational degrees of freedom (DOFs). It can be used in CAD and robotic applications, for instance. Typical representatives are the SPACEMASTER and the SPACEMOUSE (see also section 3.4.4).

Data glove The data glove is a glove which is equipped with a tracking sensor. The (unconstrained) movements of the user's hand in 3D space are sensed and used to control the application. Data gloves are most often used in VR applications. They might be equipped with tactile feedback capabilities, see the Teletact on page 29.

Tracker The tracker is very similar to the data glove, but instead of a glove it has to be held with the hand. Therefore, no feedback can be applied. The most popular one is the PolhemusTracker which is used for in VR applications. Both, data glove and tracker, are often used in combination with head-mounted displays (HMDs).

Joystring See section 2.2.4, page 28 for a discussion of the joystring.

3.2.7 Other input devices

Other input devices which will not be further reviewed within this report include the following:

Dial A dial is an input device with only one rotational DOF which is mainly used in CAD applications to control the view of a model or a scene. Usually, several dials are used to control more DOFs.

Paddle/Slider Paddles and Sliders do also have only one DOF, but in this case it is a translational one. They are very cheap and are mainly used in the entertainment industry.

Trackball The trackball's functionality is comparable with those of the mouse (see section 3.2.2). They are mostly used in portable computers (laptops and notebooks) in order to save the space that is needed to move a mouse.

Joystick Joysticks are mainly used in the entertainment industry. They are used to control cursor movements in 2D and are equipped with one or more buttons.

Digitizer Digitizers are 2D input devices for exact inputs of points, usually from already existing sketches. Their main purpose is to digitize hand-drawings, i.e. to get 2D scenes in a format which can be processed by a CAD program.

Eyetracker Eyetracker are controlled by the movement of the user's eyes which are sensed with infrared light. It is possible for the user to control the cursor by simply looking in the desired direction.

Touchpad/Touchscreen Both devices may be operated by a pen or by the user's fingertips. With the touchscreen, the user can directly point to objects on the screen. Both devices are mainly used for simple pointing and selection tasks.

3.2.8 Generalized input devices

In fact we see a translation of 'user input actions' to 'events'. Clicking a 'cancel'-button with the mouse has exactly the same result as pressing the ESC-key. This suggest that it is useful to design an extendable set of events, independent of the used input device. Such as:

Action events	OK, CANCEL, HELP, DELETE, UNDO
Navigation	Select an object visible on screen
Dragging	Move object to another place
	Copy object to another place
Value input	Typing the value with the keyboard (or use +/-)
	Use a slider

Input from other input devices (pen, speech, external knobs, joysticks, etc.) can all be translated to such events, such that it makes no difference for the application where the input comes from (see also the Meta Device Driver, figure 3.3, on page 70). This is the first step of implementing multimodality. Some examples are:

PEN	Certain gestures are translated to action events (OK, CANCEL)
	Clicking on a screen object means selecting it
	Pointing an object, keeping the pen down, then write a line means moving the object to another place
	Writing characters can be used for direct ASCII-input
SPEECH	Speaking some words mean action (OK, CANCEL)
	Other recognized words can be translated to direct ASCII-input

In order for this translation to be correct, there is a lot of interaction between input device and application. Pressing a mouse button has different results, depending on the coordinates. Translation into actions is only possible if the 'mouse driver' knows which screen area is covered by the appropriate button. Similarly, it is very useful for a speech recognition system to know what words are expected in some appliciation. The schematic becomes as follows:

The arrows in this picture represent different flows of information. The applications perform some initialization (e. g. choice of pen resolution), or supply information to devices when their input is requested. If for instance an application cannot accept sound input, it is not necessary for the audio input system to supply any data. A similar argument holds for XY tablets. In general, the on and off switching of a peripheral sampling process may save CPU time.

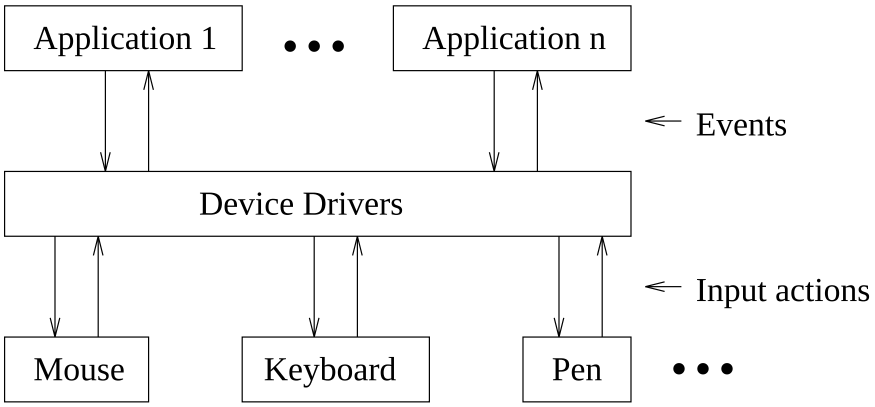

Figure 3.1: *A schematic model for generalized input devices*

A basic function of drivers is to translate data into a hardware-independent form. As an example, mouse-button clicks may be translated into keyboard key presses. Also, spoken words may be processed by a speech recognition system to simulate mouse-button or keyboard-key presses too. As an additional functionality, drivers might translate hardware coordinates (like tablet XY coordinates) into application coordinates (in this example the screen bitmap).

In practice, it is always necessary that there is two-way communication between the applications and the input devices, usually with some drivers in between. These drivers in principle can take any form and degree of software complexity, from simple pass-through interfaces to complex handwriting or speech recognition systems.

Note that the same schematic can be applied to general output devices (COM).

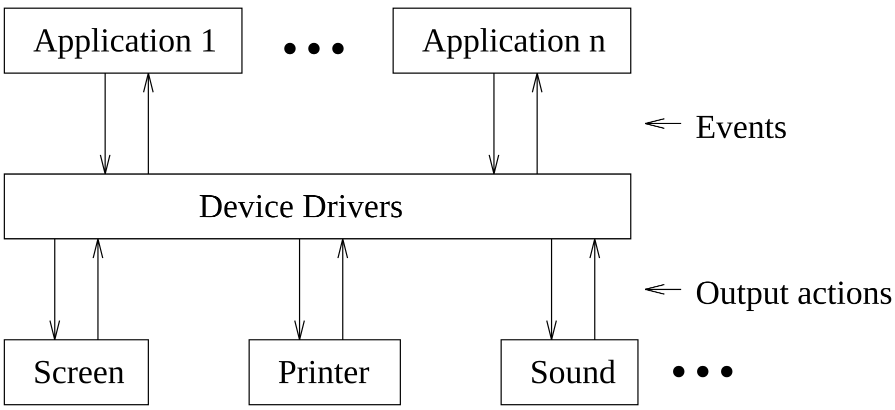

Figure 3.2: *A schematic model for generalized output devices*

3.3 Event Handling Architectures in CIM

Most current state-of-the-art applications are event-driven, for example all applications written in X11 or GEM. Every time something happens (e. g. a key is pressed or the mouse is moved) an event is generated. An event is characterized by a unique identity code and a data structure containing information about this event. Usually, the system maintains a queue of events, in order to ensure that all events are being handled, independent of CPU usage fluctuations. This means that "immediate response but probability of missed events" is traded off against "possibly delayed response but all events are handled". There exist standard procedures for getting an event from the event queue. The application is responsible for performing some action that corresponds with the event. As an example, each application is responsible for its screen content integrity. Applications send eachother events in case they modify the screen content. The receiving application must then decide whether a portion of the screen must be "repainted".

The effect on the application is that long calculations are forbidden, because if events are not processed for a long time, the user cannot do anything but wait. Events such as the Ctrl-C or ESCAPE-key cannot be used anymore for interrupting the calculation process, unless they are purposely processed inside the calculation loop. This is not a good way of programming, while the application has to take care of all possible events.

In MS-Windows, a slightly different approach is taken. Events are called 'messages', and they are used for all inter-application communications. Because MIAMI does not use MS-Windows, this is not considered further.

Two levels of event-handling can be considered. At a lower level, (GEM, X11) events are generated by the hardware, together with some low-level information. In Motif and Tcl/Tk, a layer is put on top of this, releasing the application from a lot of house-keeping chores.

3.3.1 Within-application event loops: GEM, X11

GEM and X11 applications usually have one loop that cyclic gets an event from the event list and executes this. The event-handler procedure usually consists of a hierarchy of case-statements testing for each possible event, and performing the desired action. These actions must be short, because during the event-handling, no other events can be processed. Lengthy calculations in response to a button click cannot be done in the event handler.

3.3.2 Event-routine binding: Motif, Tcl/Tk

If the application grows, the event-handler will be more and more complex. Usually the application consists of many parts (menus, dialog boxes, etc.), while the event-handler has to process all events meant for all parts.

Both Motif and Tcl/Tk are libraries using X11. Because the event-loop is part of the library, it is now invisible for the application. Tables are kept internally which store a procedure pointer for each possible event, window, class, object, whatever. The application can read and write these tables any time, and therefore influence the behaviour of every event. This solution is called "Event Binding". Because every window keeps its own event binding table, it is easy for an application to serve many windows at the same time.

3.4 Bi- and Multimodal Control

3.4.1 Visual-gestural control

Gestural control is a relevant aspect in multimedia systems. With the current state-of-the-art human movement tracking technology, it is possible to represent most of degrees of freedom of a (part of the) human body in real-time. This allows for a growing number of applications, from advanced human-computer interfaces in multimedia systems to new kinds of interactive multimedia systems (e. g., [175, 247]).

The classification of gestural information patterns in human-machine interaction is an important research topic. A short overview of existing literature is given in Appendix D, mainly referred to robotics and cognitive/psychological studies. The recent research on virtual environments has given a strong acceleration to the research in this field — see for example [247]. From another point of view, it is also important to analyze human activities which involve significant gesture communication. A significant field is related to the study of the languages developed over centuries for dance and music. This field is important also from the point of view of the study of the involved emotional communication [252, 11]. The main areas regard

(i) dance gesture languages and classifications of movement patterns [55, 297, 51, 336, 340];

(ii) languages for music conducting [232];

(iii) the study of the movement involved in music instruments performance [9]; and

(iv) the recent research efforts on "Hyper-Instruments" at MIT and in several other research centers [234, 196, 360] to simulate existing musical instruments and to find out new routes of communicating movement and emotional contents in art performance.

A complete survey on all these four aspects is [234].

Different taxonomies of gesture can be defined with respect to different aspects of gesture communication. The following are some main examples, explored in WP 2.3:

1. Types of movement detected:
   - gross-grain movements (typically full-body);
   - medium-grain movements (arms, legs; e. g., dance patterns);
   - fine-grain movements (e. g., hand gestures).
2. Types of sensoring devices:
   - inertial sensors (inclinometers, accelerometers, gyrometers and gyroscopes);
   - exoskeletons or compliant manipulators, grasped by users or linked to joints of his/her body. Such devices are controlled by a tunable compliance matrix;
   - position, distance sensors (e. g., Costel, MacReflex, V-scope), e. g. based on infrared and ultrasound sensor technology, which should be able to detect in real-time the three-dimensional position of a number of markers, typically worn by the user;
   - devices for gross-movement detection. e.g. a non-linear matrix of active infrared sensors together a sensorized floor, covering a single-user 3D area. Examples of gross movements include rhythmic (part-of-the) body patterns, energy of the movement (which/how many sensors activated in the time unit).[1]
3. Semantics:

   The important aspect is here the meaning associated to different gestures. The use of metaphors and analogies with other modalities is a fundamental aspect to devise evoking associations to gestures.

[1]This last case is adopted in the SoundCage IDMS, a joint project of SoundCage Ltd. and the DIST Computer Music Lab.

3.4.2 Handwriting-visual control

Bimodality in Handwriting-Visual Control can be interpreted in at least two ways:

1. Handwriting as a human output channel (HOC), controlling visual (i. e., graphical) properties of the computer output (COM)
2. Handwriting (HOC), combined with other forms of visible human behavior (HOC)

Ad 1. Originally, in the proposal stages of the MIAMI project, handwriting-visual control was more concerned with handwriting as a human output channel (HOC), controlling graphical properties of the computer output (COM), other than handwriting or drawing ink patterns (WT 2.4, [300, page 21]). Furthermore, the accent lies on discrete, symbolic interaction, as opposed to continuous control, which is dealt with elsewhere (WT 2.3, [300, page 20]).

More specifically, one may consider the control of graphical objects on a CRT screen, by using a pen for discrete selection, i. e., by pointing and tapping (cf. Appendix E.3.1), and the subsequent modification of graphical objects on a screen by the performance of pen gestures (Appendix E.3.3). An interesting aspect of these pen gestures is that they may or may not have iconic properties with respect to the graphical effect which is intended by the user. As an example, the user may draw an 'L'-shape in the vicinity of a rounded corner in order to sharpen the most adjacent vertex of the shown polyhedron, given the current projection on screen (iconic use). Alternatively, a pen gesture may have an arbitrary shape, bound to a command or function, which must be memorized by the user (referential use).

The advantage of pen gestures is that no screen space is being used by widgets such as tool bars and menus. Experienced CAD users already make use of so-called 'Marker' menus, instead of continually referring to possibly remote widget objects on screen. In the area of Personal Digital Assistants (PDAs), experiments are done with the iconic type of pen gestures, in that users are allowed to produce stylized scribbles which iconically depict larger graphical documents. In the user interface, a user-defined scribble then becomes the 'icon' with which the documents may be referred to in later usage. If it is on the screen, selection by tapping on the icon may be performed. However, if the icon is off the screen, the user may produce the (memorized) pen gesture to find the corresponding document. As a more specific example, in a multimedial encyclopedia, a list of pictures of castles may be produced by entering a stylized castle icon with the pen. Artists have shown interest in interfaces in which simple sketched compositions allow for the retrieval of paintings with a similar composition. Although such applications are far from possible with the current state of the art, it is one of the purposes of MIAMI to explore the consequences of these ideas, and uncover basic mechanisms.

Ad 2. Handwriting as a human output channel (HOC), combined with other forms of visible human behavior (HOC) in control and manipulation. Although not originally foreseen, possibilities are present in the area of teleoperation and audio control, if one combines the two-dimensional pen movement in time with another visible signal such as the vertical distance between upper and lower lip (ΔL) as observed by a camera pointed at the user's face. Consider for instance a robotical arm, of which the end-effector position in two dimensions is controlled by the pen on an electronic paper device (see 3.2.3), whereas the effector opening is controlled by the amount of the lips distance ΔL. The third dimension may be controlled by the diameter of the user's face as observed by the camera. For initial experiments, special coloring of the background and the lips may be performed to allow for an easier processing of the camera data.

However, as stated earlier, the control of graphical parameters of objects visualized by the computer, by means of pen gestures, will be the main topic of research in MIAMI. An interesting application refers to the use of a stylized face to represent the state of a handwriting recognition agent. Most recognition systems deliver reliability estimates for the generated word or character class hypotheses. Usually, hypotheses with a subliminal probability are rejected. The facial expression of the miniature face representing the recognizer agent may smile in the case of neat handwriting input, and frown in the case of "Rejects". Research will have to decide whether this information is actually picked up by the user, or is considered as irrelevant marginal graphics of an application.

Although there is a healthy trend in hiding technical details of a system from the user, there may be system components, such as intelligent agents, whose (intermediate) decisions have to be made explicit in meaningful ways. The reason for this is that such decisions may be erroneous. It is extremely demotivating for users of speech and handwriting recognition software not to know *why* recognition sometimes fails. In handwriting-visual control, the use of facial expressions in a miniature live icon may be a good way of externalizing aspects of the internal system state, in a non-invasive manner. Such solutions fall under the "antropomorphic" or "animistic" category, as mentioned in 1.2.2.

3.4.3 Handwriting-speech control

In bimodal handwriting & speech control, the user combines the Human Output Channels (HOCs) of speech and handwriting in a combined way to achieve a specific goal. A distinction must be made between textual input and command input (see Appendix E).

In textual input, the goal is to enter linguistic data into a computer system, either in the form of digitized signals, or as (ASCII)-coded strings. In command input, the user selects a specific command to be executed and adds arguments and qualifiers to it. The term handwriting in the title includes pen-gesture control for the current purpose. Handwriting & speech bimodality in the case of textual input means a potentially increased bandwidth and reliability, provided that the user is able to deal with the combined speech and pen control. Handwriting & speech bimodality in the case of command input allows for a flexible choice [99]. As an example, the user may say */erase/* and circle or tap an object with the pen (↗, i. e. *erase "this"*). Alternatively, the user may draw a deletion gesture and say the name of an object to be deleted.

In the remainder of this section we will consider bimodality in speech and handwriting from two viewpoints: (i) the automatic recognition and artificial synthesis of these HOC data; and (ii), the mere storage and replay of these HOC data. The accent will be on "Control", but we have added some information on computer output media (COM), because of the often encountered confusion with respect to the concepts of recognition vs. synthesis. Furthermore, with speech, we mean the audio signal representing spoken text, with ink, we mean the XY-trajectory representing written text. Both signals are functions of time.

Automatic recognition and artificial synthesis Table 3.4 shows an overview of paradigmatic application examples. They will be described in more detail in the following paragraphs.

	Speech Recognition	Speech Synthesis
Handwriting Recognition	The system improves the throughput in bulk text entry	The user gets voice feedback on handwriting recognition results
Handwriting Synthesis	The user dictates a synthesized handwritten letter	Multimedial text communication by the system

Table 3.4: *Bimodality in handwriting and speech: paradigmatic applications in automatic recognition and artificial synthesis.*

Handwriting Recognition/Speech Recognition: Improved text entry The reliability of recognition systems in the isolated speech or handwriting modality is improved due to the complementary (orthogonal) properties of both human output channels.

HOC (a) hand/arm musculature: position and compliance control
(b) speech musculature, vocal chords, respiratory system: vocal sound production

CIM (a) XY digitizer, handwriting recognition algorithm
(b) Microphone, speech recognition algorithm

COM Text (character fonts) are presented on the CRT

HIC The user gets:
(a) Immediate feedback on speech and handwriting by the intrinsic feedback loop (Figure 1.1)
(b) Mostly visual feedback in the form of text

An excellent overview on on-line handwriting recognition is given in [335]. A comparison between two approaches in on-line cursive script recognition is given in [301].

As regards the algorithmic architectures in integrating handwriting and speech recognition, similar problems as in merging speech recognition with facial speech movements occur. Several forms of merging recognizer output may be considered, of which two will be given:

1. Merging of final output word list on the basis of rank order
2. Merging of an intermediate character search space

Technically, (1) is easiest to implement, but it does not make use of the fact that "the other modality" may fill in ambiguous fragments in the character search space of a given modality. Therefore, merging of hypothesis search spaces as in (2) is the more powerful method. However, since the mapping from phonemes to character is not one-to-one, this is not a trivial task. Within MIAMI, we will try to solve the problem by searching for common singularities in both modalities. As an example, the silence preceding the "/ba/" sound is a relatively easy speech feature to detect, as is the large ascending stroke of the written <b>.

Handwriting Recognition/Speech Synthesis: Handwriting recognizer feedback by synthesized speech

HOC (a) hand/arm musculature: position and compliance control
(b) speech musculature, vocal chords, respiratory system: vocal sound production

CIM (a) XY digitizer, handwriting recognition algorithm
(b) Microphone, speech recognition algorithm

COM Voice-like sound is being produced by the computer, representing recognized text or the state of the recognition agent

HIC The user gets:
(a) Immediate feedback on speech and handwriting by the intrinsic feedback loop (see Figure 1.1)
(b) Auditory feedback

At NICI, a number of experiments have been performed with an experimentation platform called PenBlock running under Unix. In the case of isolated characters (handprint), a spoken spelled letter can be produced, after each character that is entered. If the character is not recognized with sufficient likelihood, an "uh?" sound is produced. The meaning of this is picked up quickly by the subjects, and makes more explicit the fact that an active deciding 'agent' is trying to assess the incoming handwriting products, instead of an infallible machine-like tool. However, the usability of feedback in the form of spoken recognized letters must be studied in more detail. A more neutral "click" sound, after entering each character was considered more acceptable by a small number of informal subjects. Variations on this theme can be designed, giving only speech feedback in case of doubt by the recognition agent.

Other experiments have been performed on a PC, in a small project dubbed "PenTalk" together with Tulip Computers. In this project, the user wrote isolated cursive words (Dutch) which were pronounced after on-line recognition by a speech synthesizer program. Preliminary experiments indicated that the use of speech feedback is limited, especially if the delay is more than about half a second after writing the word. In later versions of the program, speech synthesis synthesised speech feedback was deferred to a drag-and-drop type function. The philosophy was that if it could not be immediate, speech feedback was better placed under active user control. The user decides when and what must be spoken. This has the advantage that the speech synthesis algorithm can pronounce sentences and paragraphs, effectively coding the prosody of speech, which is not possible in isolated words. Based on these preliminary findings it is concluded that speech as an immediate feedback in handwriting recognition is probably most useful as a way of multimodally informing the user on the state of the recognition process or the quality of the handwriting input. Auditory feedback in any case has the advantage that the user does not have to look away from the pen-tip, i.e., where the action takes place. Other forms of (non-vocal) auditory feedback may be considered in MIAMI.

Speech Recognition/Handwriting Synthesis: The user dictates a "written" letter

HOC Speech musculature, vocal chords, respiratory system: vocal sound production

CIM Microphone, speech recognition algorithm

COM Synthesized handwriting in the form of pixel ink, or a handwriting trajectory production animation is presented to the user

HIC The user reads the handwriting from the screen, directly at the time of spoken text dictation, and at a time of later reading, potentially also by another user

In this form of bimodality, spoken words are transformed into synthesized written text. It is unclear whether the underlying algorithm may make use of the fact that handwriting is in fact richer, than an intermediate representation like ASCII is. Handwriting does contain some parameters, like pen pressure, speed, or character size and slant, which could be inferred from the spoken text. Examples: loud in speech → big in handwriting, hasty speech → handwriting more slanted to the right. A writer-dependent handwriting synthesis model costs much time to be developed, although better tools and models are more and more available [303, 271]. Depending on the type of handwriting synthesis model, shape simulation or combined shape and movement simulation can be performed. The first type (shape only) is already available in commercial form.

Handwriting Synthesis/Speech Synthesis: Improved communication through bimodality The combined presentation of linguistic information in the form of speech and handwriting has long been known to be effective by movie directors. Typically, a letter is shown on the screen, and an off-screen voice reads it aloud. The ambiguity in the handwriting style is reduced by the spoken words, and vice versa.

HOC Any form of earlier text entry

CIM Any text-oriented modality

COM (a) Synthesized handwriting in the form of pixel ink, or a handwriting trajectory production animation is presented to the user
(b) synthesized speech, pronouncing the same content as the handwriting is presented to the user

HIC The user reads the handwriting from the screen, and listens to the synthesized speech

Issues here are the optional synchronization of handwriting and speech presentation. Modes are: (i) completely asynchronous, the handwriting can be displayed almost immediately, and is continually present on screen while the speech production takes its time; (ii) synchronous, incremental, where the handwritten words follow the spoken words; and (ii) synchronous, by highlighting the 'current word' in the handwritten ink on screen. These paradigms are relatively well known for the case of speech synthesis and machine-printed fonts on the screen.

Recording and replay of speech and handwriting In this section we will consider the case of stored HOC data. Permutations with the synthesized media of handwriting and speech as above are possible, but will not be considered here. The essential difference with the preceding section is that the media will be stored for later human use, and retrieval or browsing is only possible in a superficial way (i. e., not on the basis of content). Table 3.5 shows an overview of paradigmatic application examples. They will not be described in more detail here, but will be addressed in MIAMI at a later stage. Suffice to state that from the point of view of goal-oriented user behavior, these applications are often perfectly acceptable, and no recognition or synthesis is needed at all.

	Speech Recording	Speech Replay
Handwriting Recording	The system improves the throughput in bulk text entry for later human use	The user annotates voice mail with notes in ink
Handwriting Display	The user summarizes vocally the content of handwritten notes received over E-mail	A multimedial stored speech and ink E-mail message is read by a human recipient

Table 3.5: *Bimodality in handwriting and speech: paradigmatic applications, using plain recording/replaying of HOC data.*

3.4.4 Visual-motoric control

For humans, the most important sensorial system is their vision system[2]. Nearly all actions which are performed are fed back and supervised by *observation*. Therefore, the combination of the two modalities *vision* and *motoric control* is a very natural and intuitive one, leading to a bimodal visual-motoric control. In this section, we do not deal with low-level visual-motoric coupling, like the muscular control of the eye which is necessary to fix an object over time (see 1.2.1), but with the interaction of visual and tactile feedback in motoric control tasks.

With today's standard computer equipment, every human-computer interaction includes some kind of visual-motoric coupling, no matter whether the user types in some text with a keyboard, performs click or drag-and-drop actions with a mouse or trackball, draws a model in a CAD system with a graphic tablet or a 6D input device, or controls a manipulator or mobile robot with a master-slave manipulator or a 6D input device.

In any case, the effects of the actions are — at least — observed on a monitor. But, as far as we know, the influence of visual and tactile feedback to these standard control tasks has not been sufficiently investigated yet. Although several people have performed experiments, usually only small numbers of subjects[3] have been used and only few aspects of device/task combinations have been analyzed and evaluated. Even worse, most researchers did not take into account any tactile feedback, e. g. [102, 58, 12, 13, 68, 235, 198, 197, 100, 152], with the exception of, e. g. [48, 111, 6].

Therefore, we designed several experiments which will be directed towards the

1. Analysis and evaluation of the effect of different input devices for several interaction/manipulation tasks and the
2. Analysis and evaluation of input devices with tactile feedback.

In order to get sufficient sample data, comprehensive tests with a large number of subjects have to be carried out. Otherwise, statistical errors will be introduced and the results obtained might not be transferable. Unfortunately, the number of experiments grows which each additional free variable because of combinatorial explosion. A simple example might illustrate the situation:

> If we limit one of our experiments (a positioning task, see below) to 2D space (#Dimensions Dim = 1), take three different angles (#Angles $\theta_D = 3$), three

[2]This is obviously not true for blind people, which are not considered further within this section.

[3]Usually 5–10, but at least 15 are necessary for each parameter combination in order to compensate for statistical errors (see, e. g., [36].

distances (#Distances D = 3), and use objects with five different sizes (#Sizes S = 5), we will need

$$\begin{aligned} \#Scenes &= Dim * \theta_D * D * S \\ \Rightarrow \#Scenes &= 1 * 3 * 3 * 5 \\ \Leftrightarrow \#Scenes &= 45, \end{aligned}$$

i. e. 45 different scenes (graphical setups). When each of our five devices will be used (see below), and each one is tested with all combinations of feedback modes, we will get a number of 18 different device/feedback mode combinations here (#Modes = 18). Because every test has to be carried out by at least 15 subjects (#Subjects = 15) (see above), the total number of tests will be

$$\begin{aligned} \#Tests &= \#Scenes * \#Modes * \#Subjects \\ \Rightarrow \#Tests &= 45 * 18 * 15 \\ \Leftrightarrow \#Tests &= 12\,150, \end{aligned}$$

which is a rather large number, even under the limitations given above. And it is the number of tests for only one experiment!

The hypothesis which we want to test are, among others, that

- tactile feedback will reduce the execution time and the accuracy in simple 2D positioning tasks, if the same device is used with and without tactile feedback;
- tactile feedback will significantly reduce the execution time and increase the accuracy if the target region is very small;
- the changes in execution time and accuracy will be independent of the angle and distance to the target region;
- the changes described above are more significant if the objects are not highlighted when they are reached by the cursor, i. e. without any visual feedback;
- Fitts' law[4] (see [102]) will hold for input devices with tactile feedback as well.

Various experiments have been designed and implemented to cover basic operations — in which the variable test parameters can be measured exactly — as well as interaction and

[4]Fitts' law states that the *movement time* (MT) of a target-oriented movement to an object with width W and distance D depends linearly on the *index of difficulty* (I_D): $MT = a + b * I_D$, with $a, b = constant$, $I_D = \log_2(2D/W)$.

manipulation tasks which are more oriented towards common applications, like *selection* from a menu or *dragging* an icon. They will be carried out in 2D and 3D space, respectively. Some experiments will be described below:

Positioning: A pointer shall be placed as fast and accurate as possible at a rectangular region (with width W and height H) in an angle θ_D, thereby covering distance D. This will be investigated in 2D as well as in 3D space. The influence of visual and tactile feedback will be determined. The applicability of the well-known Fitts' law [102] will be analyzed under the conditions described above. The results will be relevant for all kinds of graphical interaction tasks.

Positioning and selection: Positioning of an object at a specified target region in a fixed plane which is presented in 2D or 3D space. This is a basic experiment for any drag-and-drop operation.

Selection, positioning, grasping, and displacement: One of several objects shall be grasped, retracted, and put at another position. This includes the features of the experiments described above and extends them with respect to robotic applications like assembly and disassembly.

Positioning and rotation with two-handed input: The first (predominant) hand controls the movement of a mobile robot, whereas the second hand controls the direction of view of a stereo camera system which is mounted on the mobile robot. The task is to find specific objects in the robot's environment. This is a very specialized experiment in which repelling forces of obstacles can be used and in which the possibility to move the camera might be directly related to the velocity of the robot and the potential danger of the situation.[5]

Because most input devices with tactile feedback which are available on the market are either very simple, not available on the market, or really expensive (see 2.2.4), two input devices with tactile and force feedback, respectively, have been designed and built:

Mouse with tactile feedback: Following the idea of Akamatsu and Sato [6], a standard 2-button mouse for an IBM PS/2 personal computer has been equipped with two electromagnets in its base and a pin in the left button. For input, the standard mouse driver is used; for output, the magnets and the pin can be controlled by a bit combination over the parallel printer port by our own software, so that the magnets

[5]This is a very advanced task and it is not clear at the moment whether it may be realized, but the technical equipment (mobile robot PRIAMOS with vision system KASTOR) is available at the University of Karlsruhe, see also 6.2.

will attract the iron mouse pad and the pin will move up and down. Both magnets and the pin can be controlled independently. In order to make the mouse usable with our SGI workstation, a communication between the PC and the workstation is established over the serial communication line. In principle, any standard mouse can be easily equipped with this kind of tactile feedback.

Joystick with force feedback: A standard analog joystick has been equipped with two servo motors and a micro controller board. Communication between the joystick controller and a computer is realized over a serial communication line. The joystick's motors can be controlled in order to impose a force on the stick itself, thus making force reflection possible.

Another device, Logitech's CYBERMAN, has been bought. It is the cheapest device on the market (< 200,- DM) with tactile feedback, although in this case there is only a vibration of the device itself. For the experiments, five devices are available at the moment: the mouse with tactile feedback, the joystick with force feedback, the CYBERMAN, and two 6D input devices, the SPACEMASTER and the SPACEMOUSE. An interesting question is *how* the tactile feedback will be used — considering the hardware as well as the software — in different applications. Some suggestions and comments will be given in the following paragraphs.

Obviously, the devices which are equipped with tactile feedback capabilities realize this feedback in completely different ways. The *mouse with tactile feedback* uses two electromagnets as a kind of "brake", i. e. if a current is applied to them, the movement of the mouse will be more difficult for the user, depending on the current. In addition, a pin in the left mouse button can be raised and lowered frequently, causing a kind of *vibration*. This will motivate the user to press the button. Although in principle the current of the magnets and the frequency of the pin vibration can be controlled continuously, this will usually not be used, therefore we call this kind of feedback *binary*. Logitech's CYBERMAN can also only generate binary feedback: If a special command is sent to the device, it starts to vibrate. Again, the frequency and duration of the vibration can be controlled with parameters, but a continuous feedback is not possible.

The situation changes completely when the *joystick* with force feedback is considered. Here, two servo motors control the position of the joystick, thus allowing a continuous control in the x/y-plane. When the user pushes the stick, but the servo motor tries to move it in the opposite direction, the user gets the impression of force feedback, because the movement becomes more difficult or even impossible.

In order to make the usage of the different devices as easy as possible, a common "meta device driver" (MDD) has been developed for all tools (see figure 3.3). The parameters

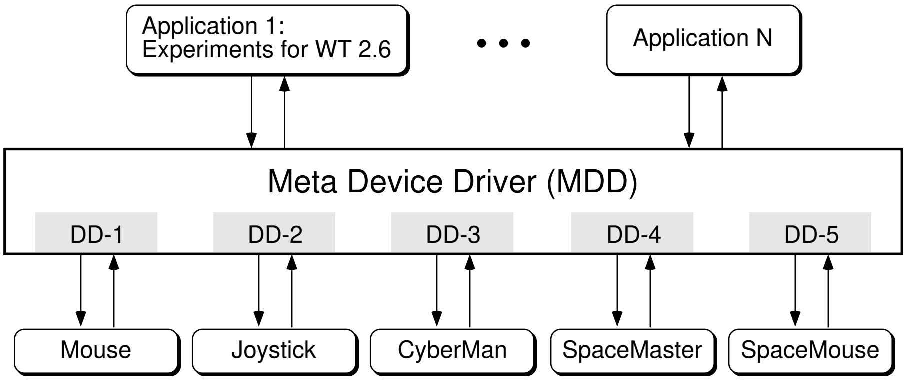

Figure 3.3: *Schematic description of the Meta Device Driver (MDD)*

which are sent to the devices follow the same structure as well as the values received from them. This concept has been developed in order to hide the specific characteristic of a device behind a common interface (cf. figures 3.1 and 3.2). It has been realized as a C++–library and can be linked to any application. If more devices will be available, the MDD can easily be extended.

With respect to the software, several different possibilities exist to give the user a visual and/or tactile feedback. Visual feedback is used by every window manager, e. g. the border of a window is highlighted when it is entered by the mouse cursor. In order to study the effect of tactile feedback, various *feedback schemes* have been developed. Two of them will be described in more detail below:

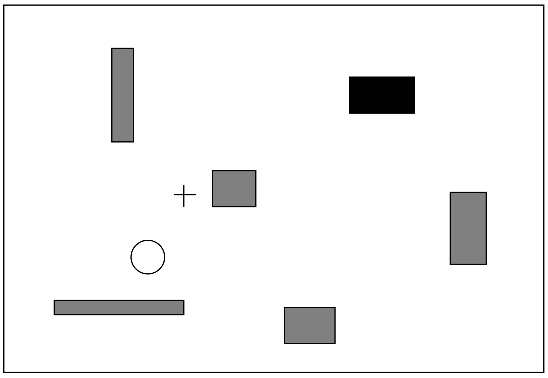

Figure 3.4: *A typical scene which is used for simple 2D positioning tasks with visual and tactile feedback. The circle marks the* start position*, the black object is the* target*, and all grey objects are used as* obstacles*.*

1. The first scheme is used for simple objects in 2D that are divided in *targets* and *obstacles* for the experiments. Figure 3.4 shows a typical scene with five obstacles

and one target. Whenever the cursor enters an obstacle or the target region, the tactile feedback is launched.

For the mouse, the magnets and the pin (or a combination of both) may be used. For the CYBERMAN, the vibration is switched on. For the joystick, things get more complicated. A force function, like the one shown in figure 3.5 needs to be implemented. In this case, the user "feels" some resistance when entering the object, but if the center is approached, the cursor will be dragged into it.

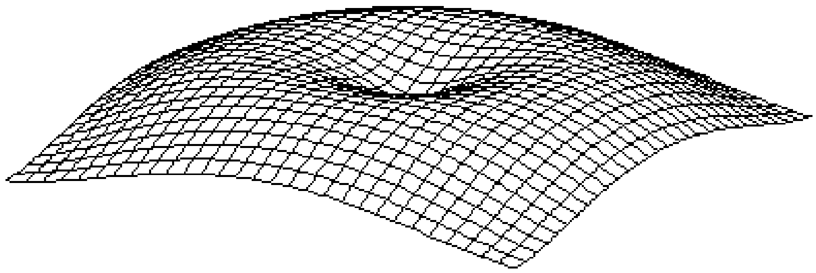

Figure 3.5: *A force function which may be applied to objects in 2D space in order to control the joystick*

2. The second scheme is applied to objects in 3D space which are treated as obstacles, e. g. walls in a mobile robot collision avoidance task. The magnets of the mouse can be used to stop further movement against an obstacle, and the CYBERMAN's vibration can be switched on for the same purpose. Again, the joystick has explicitly to be programmed with a predefined, parametrized function in order to prevent the "mobile robot" from being damaged. Figure 3.6 shows the principle implementation of this function.

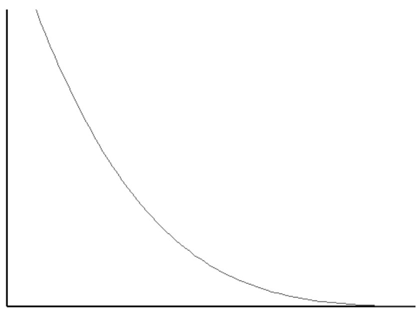

Figure 3.6: *A force function which may be applied to objects in 3D space in order to control the joystick. The x-axis denotes the distance between the cursor and the object, and the y-axis the applied force.*

Chapter 4

Interaction

The two complementary views of human-computer interaction, i. e. *perception* and *control*, have been described in isolation in the two previous chapters. Their combination — with respect to multimodal systems — is the topic of this chapter. We will start with an overview on existing architectures and interaction models. In the second section, these theoretical considerations will be directed towards practical issues of input/output coupling, followed by a section on synchronization. In the last two sections, aspects of Virtual Reality systems and an analysis of interaction will be explained.

4.1 Architectures and Interaction Models for Multimodal Systems

In the literature, many different architectures for and interaction models of multimodal (and other complex) systems can be found. Some of them will be reviewed briefly within this section. The first six paragraphs will mainly deal with different architectures for multimodal systems. The next one presents a layered model for interactions in Virtual Environments which can easily be adapted to a multimodal system. In the following three paragraphs, the focus has been set to the user interface, and in the last paragraph a technique for the development of UIs for multimodal systems will be presented. The basic idea of this section is to give a short overview on existing architectures and models in order to simplify the development of our own architecture which will be developed in MIAMI.

A design space for multimodal systems Following [243], concurrency of processing and the fusion of input/output data are the two main features of a multimodal system[1]. Consequently, Nigay and Coutaz took these features as the basis for their design space and classification method.

The *design space* has been defined along three dimensions[2]: Levels of abstraction, Use of modalities, and Fusion. In order to keep the model as simple as possible, the permitted values of each dimension have been limited. The result is shown in figure 4.1. From a maximum of eight combinations (3 dimensions, 2 values each: $2^3 = 8$), four have been named and are further investigated in the article.

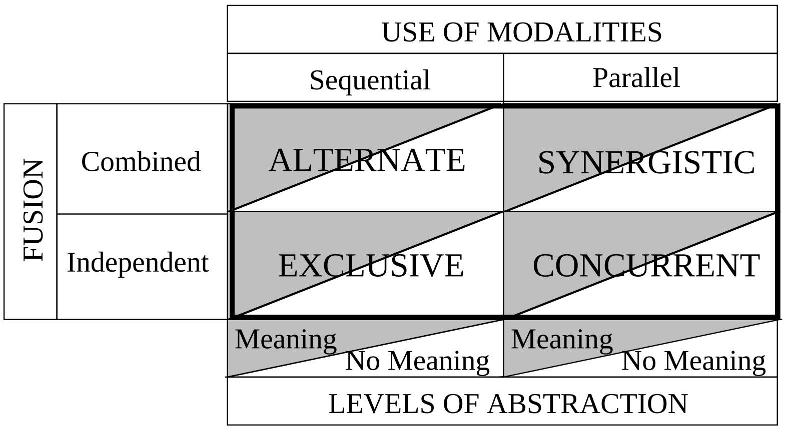

Figure 4.1: *The multi-feature system design space (Taken from [243])*

The design space is then used for a classification scheme for multimodal systems. By introducing weighted features of a system, the location of it within the design space can easily be determined. This is useful in order to test the effect of different features and commands, thus "measuring" the usability of an interface.

The second part of the paper deals with data fusion and concurrent processing. Therefore, a multi agent architecture is introduced which will not be further reviewed here. For data fusion, three levels of fusion have been identified:

Lexical fusion: Binding of hardware primitives to software events; only temporal issues (e. g., data synchronization) are involved

Syntactic fusion: Sequencing of events; combination of data to obtain a complete command

[1]Their definition of *multimodality* can be found in 1.1.4, page 6.

[2]Compare this to the design space of Frohlich, page 75.

Semantic fusion: Functional combination of commands in order to generate new, more complex commands

The main benefits of the introduced concepts which are also interesting for the prototype systems that are going to be developed within the scope of MIAMI are, according to [243]:

- the introduction of four salient classes of systems which can be used as the extrema of a reference space, and
- the chance to characterize and reason about the I/O properties of interactive systems by using their reference space.

A framework for the design space of interfaces In [109], Frohlich presents a framework for describing the design space of human computer interfaces. His framework is based on four columns: the *mode*, the *channel*, the *medium*, and the *style*. Inside these, the following items have been identified:

- Modes:
  language — action
- Channel:
  audio — visual — haptic/kinesthetic
  All of these can be combined with both modes.
- Medium:
  - Media for the language mode are:
    speech — text — gesture
  - Media for the action mode are:
    sound — graphics — motion
- Style:
  - Styles in the language mode are:
    programming language — command language — natural language — field filling — menu selection
  - Styles in the action mode are:
    window — iconic — pictorial

Frohlich states that the information in interfaces based on the language mode is used *symbolically*, whereas in interfaces based on action mode it is used more *literally*. The application of each style to each medium within a given modality suggests that styles are modality specific rather than medium specific, although they are not completely medium independent. The two main advantages of this framework, according to Frohlich, are

1. that they "help to clarify the terminology used to describe interfaces and various aspects of their design." and
2. that they can be used in "classifying past and present interface designs."

In the end, he argues that his framework may be extended in several ways, naming three potential candidates for new dimensions: *tasks*, *techniques*, and *devices*. For a multimodal human-computer interface, these extensions definitely have to be applied.

Architecture of a multimodal dialogue interface In the former ESPRIT II project 2474, "MMI2: A Multi Mode Interface for Man Machine Interaction with knowledge based systems", the architecture shown in figure 4.2 has been developed for a multimodal UI [22]. The interesting thing to notice about this approach is that it deals not only with informations but also with the *meaning* of it.[3] In figure 4.2, this is represented by the arrows marked with *CMR*, the *Common Meaning Representation*. With respect to multimodality, the MMI2 system supports a graphic and a gesture mode as well as several different language modes (command, Spanish, French, English). One of the main benefits is the integration of all these modes which are managed by the system itself.

One of the basic assumptions that lead to this architecture is that

> "mode integration should mainly be achieved by an integrated management of a single, generalized, discourse context."

The second basic architectural principle is that

> "there is a meaning representation formalism, common to all modes, which is used as a vehicle for internal communication of the semantic content of interactions inside the interface itself and also used as a support for semantic and pragmatic reasoning."

Although this architecture has been especially designed with the application of knowledge based systems in mind, it can easily be adapted for the purposes of MIAMI. The tasks of

[3]Which is exactly what we are going to do in MIAMI.

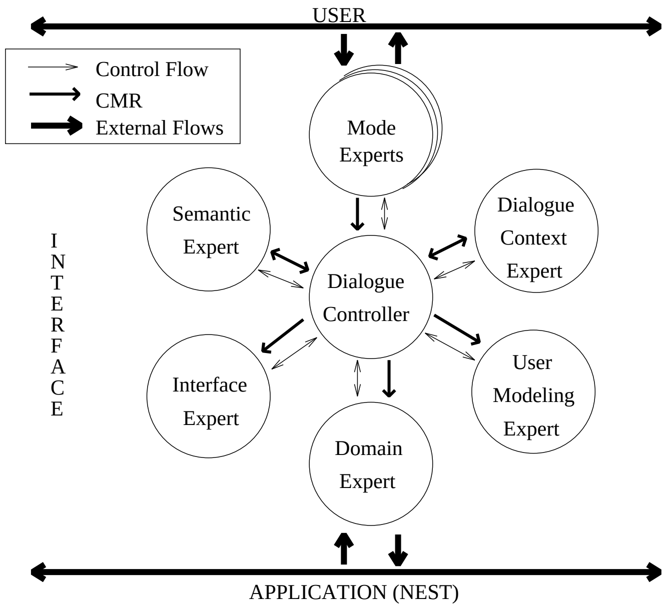

Figure 4.2: *General architecture of the MMI2 system [CMR: Common Meaning Representation] (Taken from [22])*

the 'experts' shown in figure 4.2 will not be further described here with two exceptions: The *user modeling expert* which maintains and exploits the user mode' and the *semantic expert* which has all the knowledge about the general properties of meaning concepts. These two are of special interest to MIAMI because both a user discourse model and a semantical representation might be necessary in order to strive for meaning or to find the user's intention.

Software structure of UIMS The authors of [50] have identified some requirements of systems for the development of user interfaces (UIs): consideration of standards, openness to all interaction styles, and provision of comfortable design tools. They argue that most of the existing systems fulfill only some of those requirements. Therefore, they present a layered model for the interface between an application's functionality and its UI which will especially be useful for the design of UIs for multi-tasking, multi-windowing systems that support a free choice of input and output media. Two pictures shown in figure 4.3 are best suited to show their approach.

By introducing these layers, the authors want to achieve "a clear separation of application functionality from dialog functionality, and a clear responsibility for the different actions that are involved in a dialog.", i. e. different actions are located in different layers. The responsibilities (and functionalities, resp.) have been defined as follows:

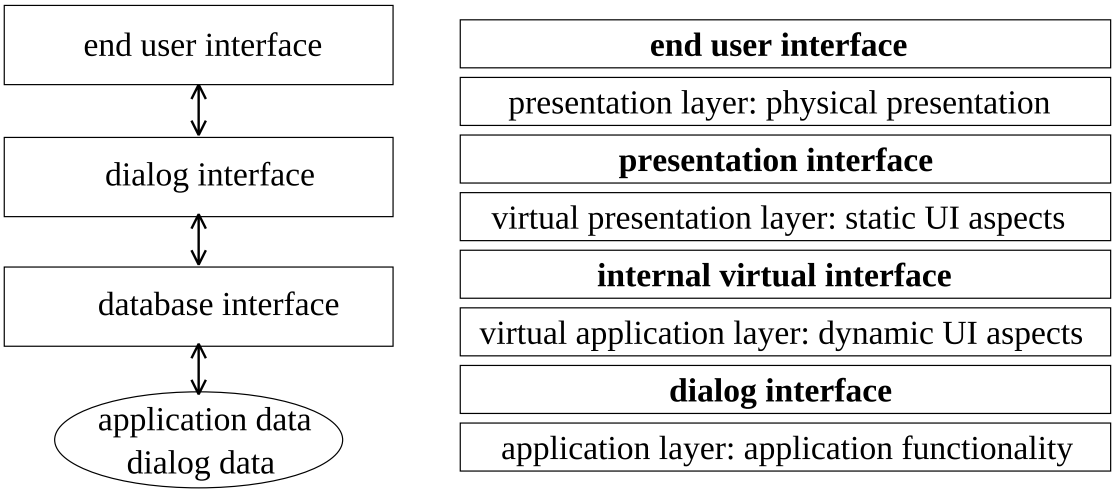

Figure 4.3: *The basic software architecture (left picture) and the corresponding layered model for human-computer interaction (right picture). (Adapted from [50])*

presentation layer: output to the screen; handling of inputs from the user; toolkit functionalities

virtual presentation layer: separation of input and output [...] from the dialog; definition of logical devices and virtual terminals; machine- and device-independent presentation; handles all static aspects of the UI

virtual application layer: contains the semantics of the respective application; dialog between the user and the application; handles all dynamic aspects of the UI

application layer: application's functionality

Another aspect of this implementation is the object-oriented approach which allows the layers to communicate via *messages*. Such a well-defined, layered model seems to be a good approach for an independent, thus flexible implementation of a human-computer interface which can also be used for a multimodal application.

The cognitive coprocessor architecture In 1989, Robertson et al. developed an architecture for interactive user interfaces in order to solve both, the *Multiple Agent Problem* and the *Animation Problem* [293]. Their approach is based on the assumption (called the *three agent model*) that

> "The behavior of an interactive system can be described as the product of the interactions of (at least) three agents: a *user*, a *user discourse machine*, and a task machine or *application*."

First, the two problems mentioned above shall be described:

- The *Multiple Agent Problem* arises from the fact that in a human-machine dialogue the agent's[4] capabilities vary significantly.
- Another difficulty will emerge when providing smooth interactive animation *and* solving the Multiple Agent Problem simultaneously. The problem of balancing the restricted computational power between these two is called the *Animation Problem*.

Although the architecture has not been designed with multimodal systems in mind, the idea of an interface with 3D and animation support is very appealing for MIAMI, too. The name *cognitive coprocessor* has been derived from the cognitive assistance it provides the user and other agents, as well as the coprocessing of multiple interacting agents that it supports. In addition to the three agent model, the architecture contains intelligent agents and smooth interactive animation.

Architectural qualities and principles In [141], Hill and his coauthors have defined five qualities for multimodal and multimedia interfaces:

1. *Blended modalities*: The user should be able to blend modes at any time.
2. *Inclusion of ambiguity*: Input modes that yield ambiguous or probabilistic input are desirable when appropriate.
3. *Protocol of cooperation*: Intervention on input and output modules should be available at any time.
4. *Full access to the interaction history*: The interpreted levels of interaction history must be accessible on-line as well as after the finishing of the interaction session.
5. *Evolution*: The interfaces have to be open to improvements without the need for a complete reimplementation.

By investigating traditional interface architectures, they found that semantics and pragmatics are usually not shared across modalities. Another drawback is the missing of "felicitous interruption, information volunteering, and intelligent displays", which are considered to be important interface building techniques. The different modalities of traditional architectures are therefore "noncommunicating components", as shown in figure 4.4.

[4] In this model the user is an agent as well as several "intelligent" processes running on the computer.

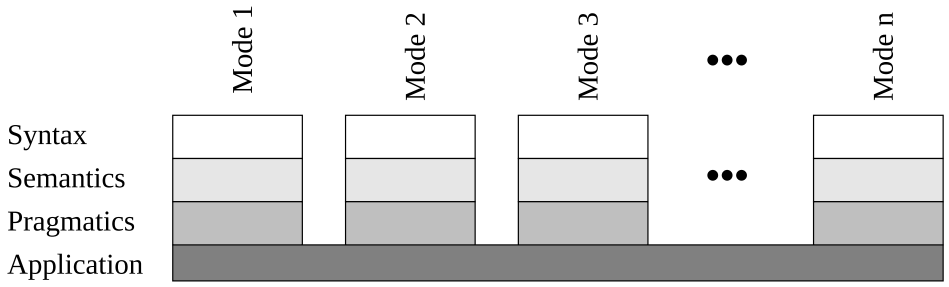

Figure 4.4: *The traditional structure of modality cooperation (Taken from [141])*

In the new structure proposed by Hill et al., some of the processing levels have been unified across input modalities, whereas others have been opened. Thus, "applications [...] interact with a semantic representation of user activities rather than a syntactic one." The principle architecture according to this approach is shown in figure 4.5. It is based on the *principle of uniform access*, which guarantees for a separation of interface aspects from the application; and the *compromise of almost homogeneous representation*, which tries to balance between performance (mainly: speed) achieved by specialized device drivers and access functions on the one hand, and homogeneity that allows the blending of modalities on the other hand.

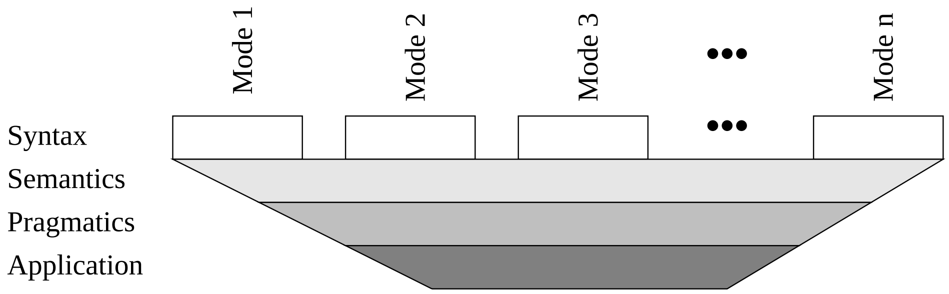

Figure 4.5: *A better structure for modality blending (Taken from [141])*

Interactions in Virtual Environments P. Astheimer et al. [7] are presenting *interaction diagrams* for interactions in 2D (figure 4.6) as well as in 3D (figure 4.7). The different layers of their model are identical, but the entities inside the layer may differ.

Although these tables have been especially designed for interactions in Virtual Environments (VE), they can also serve as an underlying model for interactions in a multimodal system. In this case, the application level has to be changed significantly, but interaction techniques like *navigation*, *object manipulation*, *gesture recognition*, *natural speech recognition*, or *eye tracing* will stay the same.

Applications	Office Process Control Solid Modeling ...
UI-Toolkits	Open LOOK OSF/Motif Windows ...
Metaphors	Desktop Metaphor Direct Manipulation Windows ...
Interaction Tasks	Valuator Pick Selection ...
Interaction Techniques	Menus Drag Click Icon Buttons ...
Events	Button up/down Motion Key Press x/y Pos. ...
Input Devices	Mouse Tablet Touch-Screen Joystick Keyboard Trackball ...

Figure 4.6: Layered model for 2D interactions

Applications	Virtual Reality Scientific Modeling Simulation Architecture CAD ...
UI-Toolkits	GIVEN Vis-a-Vis WorldToolKit ...
Metaphors	Virtual Reality Direct Manipulation Pyramid Metaphor ...
Interaction Tasks	Navigation Move Identification Select Rotation Scale Modification ...
Interaction Techniques	Grabbing Release Pointing Gesture Language Virtual Sphere 3D-Menu ...
Events	Gestures 6D-Motion Button Click Force 2D-Motion Torque ...
Input Devices	DataGlove SpaceBall 3D-Mouse 6D-Tracker Eye-Tracker Joystick ...

Figure 4.7: Layered model for 3D interactions

Man-machine communication In his book on man-machine communication [120], Geiser gives a good overview on the topic. The chapters comprise theoretical aspects of information processing and ergonomical requirements as well as technical aspects of the computer system. The basic model of a man-machine system, shown in figure 4.8, is very similar to our model (compare with figure 1.1).

Geiser has reviewed several different models of human-computer interaction. A very short description of them follows:

The layered model [246]: Norman has introduced a simple conceptual model with seven different layers of user action: On top, the user formulates a *goal*. In order to execute it, three layers are needed (top-down): *planning*, *action specification*, and *action execution*. To perceive feedback, another three layers are involved (bottom-up): *perception*, *interpretation*, and *valuation* of the system's state.

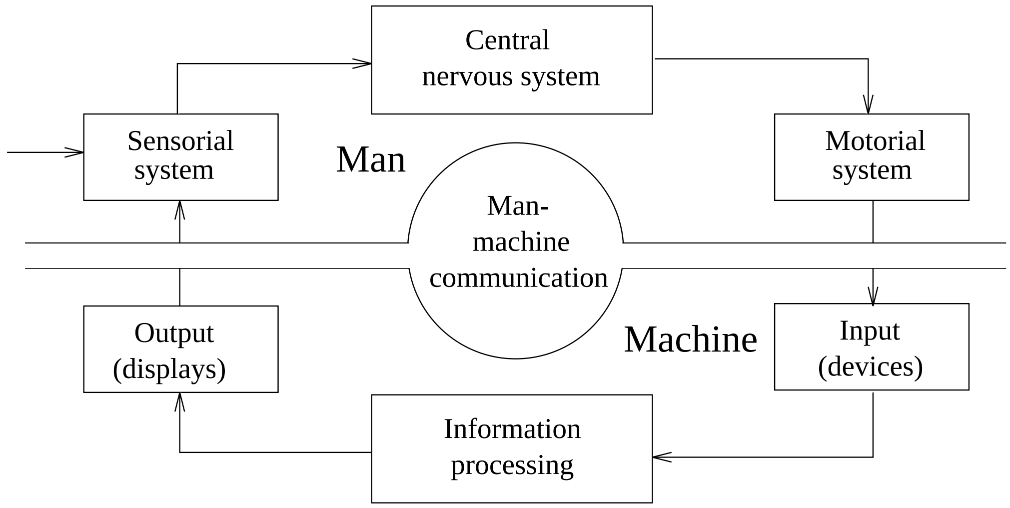

Figure 4.8: *Structure of a man-machine system (Taken and translated from [120])*

The 3-level model [283]: Rasmussen's model[5] is a conceptual model of the user, comprising three hierarchical levels of behavior: skills, rules, and knowledge (SRK). The layers represent different levels of abstraction in the human (cognitive) processing system. A more detailed description of this model follows in the next paragraph.

The GOMS model [60]: The GOMS model describes the user's task representation as a conceptual model in terms of **G**oals, **O**perations, **M**ethods, and **S**election rules. It is especially useful for the prediction of the user's behavior, but it has a major drawback: Abnormal behavior of the user is not considered.

The keystroke-level model (KLM) [59]: This model only considers *one* aspect of man-machine communication: How long does it take for an *experienced user* to perform a standard task *without any error*? The model's benefit is the quantitative measurement of observable parameters in man-machine interactions. Drawbacks of the KLM are that cognitive aspects are difficult to include and that in many cases the measurements performed are not very accurate.

The theory of cognitive complexity [156]: Based on the GOMS model, Kieras and Polson have developed a theory in order to allow a quantitative analysis of the complexity of the man-machine dialogue. The theory's purpose is to support the design process of dialogues with respect to the expense and the transfer of learning, the execution time, and the user friendliness of the system. The *cognitive complexity* is defined as the content, the structure, and the knowledge that the user needs in order to use a device. Additionally, two different knowledge representation schemes

[5] Rasmussen is also the coauthor of the article reviewed in the following paragraph.

have been defined: One for the knowledge of the user concerning the task, the other one for the device itself. Unfortunately, the predicted results could not be proved in practice, and the model has been refined several times by other researchers.

Ecological interface design A more theoretical approach has been taken by Vicente and Rasmussen. In [346], they are introducing a framework called EID (*Ecological Interface Design*), which is based on the skills, rules, knowledge taxonomy of cognitive control (see the *3-level model* in the last paragraph. Their method has especially been developed for complex human-machine systems with direct manipulation interfaces. Therefore, it could be useful for multimodal systems, too.

At first, they distinguish between three different types of events with which the user has to deal in complex domains:

1. *Familiar events*: Due to experience, the operator's skills are sufficient to deal with this type of event.

2. *Unfamiliar, but anticipated events*: Because this type of event has been anticipated by the interface designer, the operator is sufficiently supported when it occurs.

3. *Unfamiliar and unanticipated events*: Improvision by the operator is necessary and might lead to wrong reactions.

The EID has been developed in order to deal with all three kinds of events and to offer the operator the most appropriate support in any situation. Therefore, an abstraction hierarchy with multiple levels of cognitive control, based on the SRK taxonomy, has been used. SRK stands for the three different levels of cognitive control:

- *SBB* – *Skill*-based behavior: Automated behavioral patterns are used at this level. SBB's principle is as follows:

  "To support interaction via time-space signals, the operator should be able to act directly on the display, and the structure of the displayed information should be isomorphic to the part-whole structure of movements."

- *RBB* – *Rule*-based behavior: Cognitive control on this level depends on a set of cue-action mappings with the following principle:

  "Provide a consistent one-to-one mapping between the work domain constraints and the cues or signs provided by the interface."

- *KBB – Knowledge*-based behavior: At this level of cognition, problem solving operations on a symbolic representation come into play. Hence, KBB's principle can be formulated like that:

  "Represent the work domain in the form of an abstraction hierarchy to serve as an externalized mental model that will support knowledge-based problem solving."

The first two behaviors are concerned with perception and action (which "is fast, effortless, and proceeds in parallel"), whereas KBB is used in analytical problem solving on a symbolic representation (which "is slow, laborious, and proceeds in a serial fashion"). The latter one will most often be used when unfamiliar situations arise. When designing an interface, the user's processing level should be kept as low as possible, because operation times will be reduced and the process will be less error-prone.

Intelligent user interfaces

> "The great thing about the title *Intelligent User Interfaces* is that it is ambiguous: is the book about interfaces for intelligent users, or intelligent interfaces for any users? I think the answer isn't simple."
>
> W. Mark in the foreword of [329].

The book *Intelligent User Interfaces*, published in 1991, contains many ideas, concepts, and realized systems for sophisticated UIs [329]. Although it is impossible to review all chapters within the scope of this report, a few comments have to be made. The book is of some significance for MIAMI for its first two parts:

The first part, *Multimodal Communication*, though written by experts of the field, doesn't hold what its name promises. The term multimodality is used within a very narrow sense, mainly meaning a combination of natural speech and deictic gesture input (following the "put-that-there" paradigm). In our opinion, this is not enough to justify the word multimodality. The more interesting thing to notice about this part is the description of user and discourse models, which may be useful in MIAMI, too.

The latter also holds true for the second part of the book, *Models, Plans, and Goals*. Modeling aspects of the user, providing the UI with adaptive interaction capabilities, and using intelligent agents for the user's access to the system's functions are the main topics of this part. At the moment, it seems to be clear that without any coverage of the higher levels of a UI, the ambitious goal of modeling cognitive aspects can not be reached (see section 5).

Wizard of Oz technique for multimodal systems In [294], Salber and Coutaz describe the application of the Wizard of Oz technique (WOz) to multimodal systems. The basic idea of a WOz system is the modeling of a system or system behavior which is not yet or only partly available by a human (the hidden "wizard") and to hide this fact from the user. By analyzing the performed operations, the user's needs can be identified in advance which may lead to a better design of the final system.

This idea is especially interesting for the design of multimodal systems because at the moment the understanding of how to design such a system is very limited. Therefore, the authors argue that the WOz technique is an appropriate approach to the identification of sound design solutions. The whole article can not be reviewed here in depth, but the requirements from the wizard's perspective as well as those from the system perspective shall be mentioned:

- Requirements from the wizard's perspective:
  - *Task complexity*: In a multimodal system, the wizard must simulate more complex functions such as the synergistic combination of modalities than in a traditional system.
  - *Information bandwidth*: Due to the chance of using more than one modality for user input, the input processing bandwidth of a multimodal system will exceed that of a traditional system.
  - *Multiwizard configuration*: One way to overcome the two problems mentioned above may be to use more than one wizard.
- Requirements from the system perspective:
  - *Performance*: The system has to provide the effective foundations for efficiency.
  - *Configuration flexibility*: In a multiwizard, multimodal system, one has to deal with 'a variable number of wizards as well as the variety of input and output devices.
  - *Communication protocol flexibility*: This is a critical aspect because i) multiple wizards may be engaged and ii) information exchange over the communication channels may be on any level of abstraction.

In addition, powerful tools for the analysis and the evaluation of the observed data are necessary in order to successfully apply the Wizard of Oz technique to a multimodal system. This has been identified to be one of the major weak points during the tests performed by Salber and Coutaz. Their second result is that "the organization of the wizards' work requires a lot of testing and experiments; [...]". Nevertheless, the WOz technique could be an appropriate solution in MIAMI when the integration of software packages won't be possible due to missing pieces.

4.2 Input/Output Coupling

In fact, the complete dissociation between Input and Output, or sensory and motor channels is meaningless, because of the tight coupling between input and output in the human information processing system. First, of all, perception of an external world is meaningless in a system which cannot perform actions on that world. Second, it can be argued that none of the sensory modalities is completely independent of motor output. Vision relies to such an extent on the motor output of the oculomotor muscles that a stationary image will disappear within a second after paralysis of these muscles (due to the absence of optic flow). Hearing depends in two ways on motor output: the head orientation helps in sound localization, and the dynamic range of sound perception is regulated by tiny muscles stretching the tympanic membrane. Propriocepsis is controlled by the gamma-efferents, where the central nervous system directly controls the sensitivity of the muscle spindle receptors. Indeed, for each sensory modality or sub-modality it is possible to identify one or more reflexes, which are (more or less automatic) control systems, either open-loop (as the vestibulo-ocular reflex) or closed-loop (as in the stretch reflex). These reflexes (i. e. the appropriate activation/stimulation of them by an artificial system) are important for virtual reality but not necessarily for advanced (multimodal) man-machine interaction.

In any case, the strongest form of integrated perception & action occurs in somesthesia, kinesthesia, sense of touch, haptic perception. Many people (Aristotle as well as his followers, probably including Muller) take for granted that the sense of touch, as the perceptual channel which operates while we touch/manipulate objects, only has to do with the skin, forgetting the crucial role of the muscle/joint receptors. The pure stimulation of the skin without a concurrent stimulation of joint/muscle receptors, which is being attempted in some "advanced" data gloves, is highly non-physiological and virtually pointless from the man-machine point of view. The concept of haptic perception captures the essence of this problem, although it still has no nationality in neurobiology; Kandel & Schwartz, for example, speak of the "Somatic Sensory System" to include in an integrated way pain, thermal sensation, touch-pressure, position sense and kinesthesia. The main point is that in the tactile exploration of an object it is the "conjunction" of information from skin, muscle, joint receptors, not their "disjunction", which is essential. Slight misalignments of the different components can destroy the internal coherence (the Gestalt) of the haptic-perceptual process. Kandel & Schwartz, differently from Shepherd, do not define "sensory modalities" per se but subdivide the analysis of sensory-perceptual processes into three main "Systems": the somatic, visual, and auditory sensory systems, respectively. They consider the "sense of balance" as a component of the motor control system and forget smell and taste, mainly due to their limited cortical representations. We suggest to drop the term "sense of touch" because it is misleading: It can be interpreted as all-inclusive or in an extremely selective way. It is better to distinguish , along with Shepherd, between "somesthesia" and "kinesthesia" and to define "haptic perception" as the conjunction of the two: not a mere sum but some kind of integration process which builds a common internal representation of space and objects. From this point of view, haptic perception is multimodal by itself.

4.3 Synchronization

4.3.1 Object synchronization

As stated in 2.3.1, audiovisual integration depends on the acoustical and visual signals interaction and complexity. The following taxonomy captures the essential aspects of the interaction and complexity. We are not providing here numerical data characterizing the integration processes as they are highly dependent on the particular signals used. Fundamental to the underlying integration and cross-modal effects are the following properties:

Synchronized objects This category refers to audiovisual objects whose all relevant spatial and/or temporal properties as measured by single senses, are overlapping. Synchrony of representations may refer to many different aspects: spatial location, temporal changes, motion. The synchronized representation leads to a fast, strong and 'natural' enhancement of the representation which might be perhaps best described that single-sensory object descriptions give rise to the creation of an 'enhanced' multisensory object. This enhancement lead in turn to many specific cross-modal effects, one of them is for example enhancement of intelligibility of audio-visual speech perception [208].

Nonsynchronized objects This term refers to the situation when at least some of the spatial and/or temporal properties of objects, as measured by the single senses, are in conflict. Here one can consider as a reasonable hypothesis, that because of its goal of building of a consistent representation the sensory integration system will try to put a lot of effort to integrate the conflicting data, even at a cost of sacrificing the precision from a single sense. Only if this process of building of an integrated representation fails, the data will be interpreted as coming from many objects. One of the most prominent effects of this kind is ventriloquism [321], where by a clever visual stimulation, ventriloquist uses its own voice but creates strong illusion of a talking puppet. This is achieved by perfect integration of visual and acoustical stimuli. However, it seems that the integration effect in the spatial domain is much more pronounced than in the temporal domain.

4.3.2 Complexity of information

Apart of building the representation of the environment, senses play important communication roles. For the communication purposes, raw information signals are grouped into units which convey complex content. These groupings can be created, retrieved and manipulated in an endless way, reflecting unlimited complexity of information which can be created.

We can differentiate at least among several levels of information complexity which the sensory processing system must be dealing with:

Elementary signals These are signals which can not be broken into simpler ones, as regards their spatial, temporal, or other specific properties.

Simple syntactical units This is a sequence of elementary signals producing an elementary component which will be a base for complex information representation.

Semantic sequences A sequence of syntactical units enables representation of complex information.

All these types of stimulations can appear as inputs to the visual and acoustical systems. The question arises how integrated audio-visual processing of them is organized. The information complexity can be paired with the object synchronization properties to produce a picture of audio-visual integration. The following rules can be formulated in this respect:

1. For synchronized audio-visual objects, the higher the complexity of information conveyed, the more enhanced their integrated representation becomes. This results in higher system performance in extracting the information from bimodal perception than from single senses.

2. For nonsynchronized objects, the processes involved are more complicated. In the case of lack of temporal synchronization, the building of an integrated representation is highly disturbed. However, in the case of spatial desynchronization, the building of

an integrated representation is preferred and the conflicting data coming from single senses are attenuated. Usually the visual system is dominant over the acoustical system, that is its data has more weight for building of the integrated representation.

4.4 Virtual Reality?

> **VRR**: What do you think is the greatest obstacle facing the VR industry and why?
> **Latta**: Lack of basic research. The issues of having the most intimate form of human computer interface in Virtual Reality necessitate a thorough understanding of human perceptual, muscle and psychological systems. Yet that research foundation does not exist.
> Dr John Latta interviewed in *Virtual Reality Report*, **2** (7) p 4.
> Taken from: [195, page 43]

In real life, humans use machines, e. g. a car or a machine tool, and interact with them in a multimodal way, i. e. by means of perception (vision/hearing/haptic perception), and action (knob/wheel/lever/button control). In the case of virtual reality, the purpose is simply to substitute the sensory-motor flow elicited by a physical ambient environment with a computer generated sensory-motor flow which imitates/emulates (a) physical reality to a sufficient degree. Flight simulators are a good example. The purpose is not to improve or optimize man-machine interaction but to reproduce the standard one, with the goal of training the user or simply entertain him. Imitation/emulation is a possible paradigm of man-machine interaction, but it is not the only one. Due to the decoupling of the sensory-motor flow from the physical world, new environments can be generated based on environments which are physical in itself, but of a completely different scale (molecular surface landscapes vs galactic environments based on astronomical data), and which would normally never elicit an ambient sensory-motor flow.

More generally, one can define paradigms in which multimodality is used for improving the interaction between the user and reality via a computer:

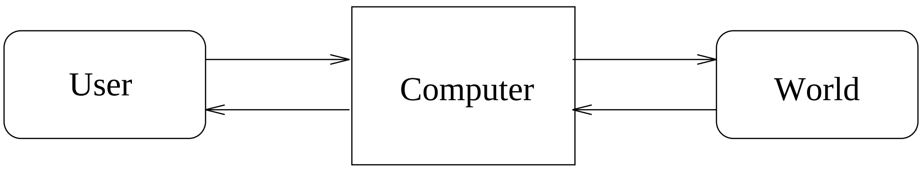

Figure 4.9: *The computer as a mediator (or* agent*, see 5.2) between the user and the world*

In [123], the following applications of VR are described in detail:

- Operations in hazardous or remote environments
- Scientific visualization
- Architectural visualization
- Design
- Education and training
- Computer supported cooperative work
- Space exploration
- Entertainment

It can be seen from this list that VR will have a great impact in the future in many different domains. The techniques and devices which are needed in order to let the user immerse in the virtual world created by a computer are currently under development. In this sense, and because of the money that will be spent for VR research, it may be called one of the driving forces of multimodal technology.

Ellis (in [89]) and Null and Jenkins(in [247]) from NASA labs on Virtual Reality (VR) argue that Virtual Environments (they use this name instead of VR), presented via head-mounted computer-driven displays, provide a new medium for man-machine interaction. Like other media, they have both physical and abstract components. Paper, for example, as a medium for communication, is itself one possible physical embodiment of the abstraction of a two-dimensional surface onto which marks may be made. The corresponding abstraction for head-coupled, virtual image, stereoscopic displays that synthesize a coordinated sensory experience is an environment. It is important to collocate VR (or VEs, following NASA) with respect to man-machine interaction.

In summary, we can understand multimodal interaction as a process characterized as follows: (i) the computer is able to capture the largest possible part of the human motor outflow, (ii) human movements are as little constrained as possible, (iii) the human receives from the computer a perceptual inflow which maximally utilizes the different available channels, (iv) the inflow and the outflow are optimally tuned, in relation with the specific task.

4.5 Analysis of Interaction

A number of formalisms have been developed to describe the user behavior in human-computer interaction. Two basic approaches can be identified: (1) formal modeling, (2) statistical modeling. In formal modeling, the ideal user behavior is described in the form of a grammar. Several formalisms have been proposed, such as the well-known Backus-Naur Form (BNF), Task Action Grammar (TAG) [260], Goals/Operators/Selectors/Methods (GOMS) [60], and others. A more modern (executable) example is SOAR [170]. These formalisms are either conceptual or operational simulation languages. The models can be developed theoretically without experiments when the task goals and the application constraints are sufficiently known. A weak point however, of the formal methods is their inability to handle or describe human error, and the differences in user styles (e. g., perceptual/cognitive/motorical styles). The second group of models is more of a statistical nature, using the post-hoc identification of State Transition Networks, Petri nets, or probabilistic grammar inference. Both method categories (formal vs statistical) are much more suited for symbolical and discrete-action human-computer (HOC) interaction than for describing details of continuous control in HOC interaction, such as the dragging of an object on screen.

Chapter 5

Cognition

As explained in the introduction, *cognition* is one of the most important topics for multimodal systems. Without it, data can not be interpreted sufficiently by a computer, and the meaning and intention of the user's actions will not be recognized by the machine. Therefore, cognition is the key issue to separate multimodal research from the more technically oriented multimedia and VR research. We start this chapter with a short overview on cognition in humans, followed by a part on learning in humans as well as in computers. In the last section, we will review the concept of intelligent agents which is today's most interesting approach to realize 'intelligent', autonomous, and intentional behavior.

5.1 Cognition in Humans

It is essential to be aware of the fact that the "raw data" coming from the senses are not at the epistemological level of interest in MIAMI. Taking for granted the existing and well-known "unimodal" psychophysical constraints at this level, it is what happens at the higher levels of "information integration in bimodality" which needs our focused attention. The whole issue of sensory-motor channel capacities which originates in the field of psychophysics (Weber's law, Fechner's law, Stevens's law from which the channel capacity data are obtained) is virtually useless within MIAMI, because it considers only a single channel at a time, in an isolated manner which tends to greatly underestimate the global or Gestalt-like features of the human perceptual-motor system. The main rationale of MIAMI is to exploit the (hidden) synergies between different channels in a "natural" environment (see the concept of "ecological perception" advocated by J. J. Gibson [122]).

Modality, in the neurobiological literature, implies the adjective "sensory". So it is not possible to speak of motor modalities. As mentioned earlier (Table 1.1), the classification among different modalities is mainly in terms of different forms of energies, which are transduced and detected. A similar classification can be applied to motor control aspects, which produce different kinds of work (mechanical work, acoustic work, chemical energy, etc.). However, these distinctions are physical and do not capture the informational aspect, which is tied to task and context, i. e., the meaning of an event or action. For example, a sound (emitted and/or detected) can be an utterance (only relevant for its timing), a musical tune (with its pitch/duration/intensity/timbre), or a spoken word. It can be man-generated or machine-generated and, dually, machine-perceived or man-perceived. Similarly, the term "gesture" is too vague and strongly task-dependent. On the other hand, the notion of channel is too restrictive and inappropriate when dealing with multimodal interaction, because the whole business of exploring multimodality is that in biology the ensemble is much more than the pure sum of the parts: emergent properties and functionalities can emerge if the parts are carefully matched.

As a consequence, the attempt to classify input/output devices is an exercise in futility if it is not grounded in a specific context, i. e. is not made task-dependent. Therefore, the taxonomy document should terminate with a preliminary sketch of the different application paradigms.

In order to structure the concepts in this area, the following representation levels are proposed [183], in increasing order of abstraction:

Signals A signal refers to the N-dimensional waveform representation of a modality. It is characterized by spectral content, and a required sampling frequency and resolution can be identified. The signal directly corresponds to a physical entity in a quantitative fashion. In the sound modality, signals refer to the acoustical or waveform representation of a sound. In computer models, signals are digitally representated by an array of numbers. In audio, for CD-quality, a sampling rate of 44100 sa/sec and 16 bit resolution is often used. In music research, it is sometimes necessary to perform classical digital signal processing operations on musical signals, such as fourier transform or wavelet transform (see for example [80]).

Perceptual Mappings A perceptual mapping represents transformed relevant aspects of a Signal in a condensed representation. In this framework, a Mapping is assumed to be a state or snapshot of the neural activity in a brain region during a defined time interval. It is modelled as an ordered array of numbers (a vector). For example, the most complete auditory mapping (i. e., the one closest to the Signal) is assumed to occur at the level of the auditory nerve. From this mapping all other mappings can be derived. For example, at the cochlear nucleus auditory, processing becomes differentiated and more specialized and some neurons perform onset detection. According to [182], Schemata and Mental Representations should be taken into account for a classification of auditory mappings (indeed, this classification scheme can be generalized also to other modalities).

Schemata A schema is a categorical information structure which reflects the learned functional organization of neurons as response structure. As a control structure it performs activity to adapt itself and guide perception. Basic schemata features are presented in [57].

Schemata are multifunctional. In the present framework, adaptation to the environment is seen as a long-term process taking several years. It is data-driven because no schema is needed in adapting other schemata to the environment. Long-term data-driven adaptation is distinguished from short-term schema-driven control. The latter is a short-term activity (e. g. 3 to 5 sec) responsible for recognition and interpretation. It relies on the schema and is therefore called schema-driven.

Schema responses to signal-based auditory maps are also considered maps. In that sense, the response of a schema to an map is also a map. But the schema has an underlying response and control structure which is more persistent than maps. The structure contained in a schema is long-term, while the information contained in a map is short-term. The latter is just a snapshot of an information flow.

As observed by Leman [182], neurophysiological studies provide evidence for the existence of different kinds of schemata (e. g., [328]). An example of schemata in auditory modeling is Leman's two-dimensional array of artificial neurons in a Kohonen-type network [182]. This schema has been trained with short pieces of music and is able to classify tone centers and chords in input signals.

Mental Representations Mental representations are knowledge structures that refer to a "mental" world. They are used in solving specific tasks. Techniques of multidimensional scaling depict the data of the tests as mental spaces — with both analogical and topological properties. Schemata are computer implementations of mental representations. The latter can serve as metaphors for the schemata. One of the aims of modelling perception and cognition is to show that the representational categories are causally related to each other. Signals are transformed into maps, and maps organize into schemata and are controlled by these schemata. By looking for correlations between the model and psychological data, one may try to relate the cognitive brain maps to mental representations. As such, the notion of mental representation can be incorporated in the knowledge-base. It is important to mention that this approach offers, contrary to the static world of mental representations, a dynamic point of view. This dynamics is introduced at two levels: (i) how the cognitive map comes into existence, (ii) how it is used in a perception task. The categories have been useful in a study on tone semantics [183, 57] and are part of the present framework of a hybrid AI-based signal manipulation system.

The following of this section describes in more detail the basic requirements of representation and reasoning systems able to model these high-level aspects of cognition. This viewpoint reflects the fact that the taxonomy should not reduce only to the I/O channels.

5.1.1 Symbolic, subsymbolic, and analogical

To avoid confusion and misunderstandings, we adopt the terms *symbolic*, *subsymbolic*, and *analogical* with the following meaning. "Symbolic" stands for a representation system in which the *atomic* constituents of representations are, in their turn, representations. Such a representation system has a compositional syntax and semantics. The typical case of symbolic system is an interpreted logical theory.

We call a representation "subsymbolic" if it is made by constituent entities that are not representations in their turn, e. g., pixels, sound images as perceived by the ear, signal samples; subsymbolic units in neural networks can be considered particular cases of this category [317].

The term "analogical" refers to a representation in which the constituents and their relations are one-to-one with the represented reality. In this category we include mental models as defined by Johnson-Laird [151], mental imagery and diagrammatic representations. Analogical representations play a crucial role in multimedia knowledge representation. Note that, in this sense, analogical is not opposite to digital: in the hybrid model we are developing, analogical components are implemented by computer programs.

5.1.2 High-level representations: Basic issues and requirements

We need formalisms able to manage the different structure and levels of abstraction of multimedia objects, from the symbolic, abstract representations to the subsymbolic representations (e. g., the signal perceived by the human ear).

Moreover, different views of the same objects are often necessary: according to the reasoning perspective and the goal to fulfill, multimedia object representations can vary from an atom in a symbolic high-level representation to a stream of low-level signals in the deepest view of the same material.

Metaphors are a crucial issue regarding representation and reasoning capabilities in multimedia systems. Metaphors are widely used in reasoning by humans and are at the basis of languages for the integration of different multimedia knowledge. For example, in the problem of a robot navigating in a three-dimensional space (e. g., a "museal robot" or a "robotic actor" on a theater stage), a bipolar force field can be a useful metaphor: in the mental model, a moving robot can correspond to an electric charge, a target to be reached corresponds to a charge of opposite sign, and obstacles correspond to charges of the same sign. Music languages are rich of metaphors derived from the real world dynamics — see for example [55]. In general, the terms and descriptions in one modality can be used to express intuitively "similar" concepts in other modalities. We deem that metaphors are the basic "glue" for integrating different modalities, e. g., sound/music and movement/dance representations. The issue of reasoning based on metaphors has been widely studied from different points of view in AI, psychology and philosophy. Steps toward an approach to model metaphors can be found for example in [115]: his theory analyzes metaphors in terms of similarities of topological structures between dimensions in a conceptual space.

Furthermore, formalisms able to support users should provide mechanisms for reasoning on actions and plans, for analyzing alternatives, strategies, starting from user requirements and goals. They should provide both formal and informal analysis capabilities for inspecting the objects represented.

Another point is learning, i. e., how to automatically update the system knowledge (new analysis data, new planning strategies), for example by means of generalization processes starting from examples presented by the user. The solutions proposed in the AI literature, such as the purely symbolic approaches, and the learning systems based on neural networks, are preliminary attempts in this direction.

Lastly, an emerging aspect regards the modeling and communication of emotions in multimedia systems. Preliminary work is available in literature [252, 288] and is currently experimented at Carnegie Mellon in the framework of interactive agent architectures [11].

5.1.3 Human learning and adaptation

A good review of the different topics in human learning can be found in [192], particularly in chapter 13, where they list a number of "laws of learning":

1 – Law of effect An action that leads to a desirable outcome is likely to be repeated in simular circumstances.

2 – Law of causal relationship For an organism to learn the relationship between a specific action and an outcome, there must be an apparent causal relation between them.

3 – Law of causal learning (for desirable outcomes) The organism attempts to repeat those particular actions that have an apparent causal relation to the desired outcome.

4 – Law of causal learning (for undesirable outcomes) The organism attempts to avoid particular actions that have an apparent causal relation to the undesirable outcome.

5 – The law of information feedback The outcome of an event serves as an information about that event.

As regards the content of the cognitive processes and their acquisition, the fundamental work was done by Piaget [268] who distinguished four stages in the intellectual development of children:

1. Sensorimotor development (0–2 years)
2. Preoperational thought (2–7 years)
3. Concrete operations (7–11 years)
4. Formal operations (11–15 years)

5.1.4 Hybrid interactive systems

Hybrid systems integrate different, complementary representations, each dealing with particular aspects of the domain. Combining different, complementary representations, as discussed in [78, pages 29–30], is therefore a key issue (a multi-paradigm approach, from the point of view of the computational mechanism).

For example, taxonomic representations in terms of semantic networks are appropriate for inheritance and classification inference mechanisms [39]; production rules are appropriate for representing logical implications; reasoning on actions and plans requires still further mechanisms, such as temporal reasoning; finally, reasoning on the domain by means of "metaphors" is crucial for integrated multimedia representations.

In this framework, integrated agent architectures are promising paradigms for the development of hybrid systems for multimodal integration and multimedia representation [288, 101, 201]. Intelligent agencies are shortly reviewed later on in this chapter, see 5.2.

5.2 (Intelligent) Agents and Multimedia

Intelligent agents are an emergent research area of particular interest in multimedia systems. Recent proposals — e. g., [101, 288] — regard the development of agent architectures for the distributed control of complex tasks, in the real-world and in real-time. Operating in the real world means to cope with unexpected events at several levels of abstraction both in time and space. This is one of the main requirements of intelligent agent architectures, and is also the typical scenario for a multimodal interaction in multimedia systems. It is therefore of crucial importance to develop and experiment integrated agent architectures in the framework of multimedia systems.

For an overview on intelligent agents, we suggest to read the paper by Wooldridge and Jennings [368]. It covers theoretical aspects of agents as well as agent architectures and programming languages which have been specifically designed to be used with agents. Another interesting survey of existing agent architectures is [101], and a first attempt of an agent architecture developed in the MIAMI project is described in [56]. As stated in [101], "the intention is to provide an implementation of a software control architecture which is competent, functionally rich, behaviourally diverse, and which encourages and readily facilitates extensive experimental evaluation." MIAMI experiments involve a set of autonomous agents, each delegated to a precise skill (e. g., input modality), characterized by flexible and modular integration in the different experimental scenarios.

For a multimodal system, several aspects of intelligent agents are especially interesting, e. g. intentional behavior, believe, autonomy, negotiating capabilities, etc. All of them will significantly contribute to an intelligent user interface which is able to interpret the user's actions, decide for itself the best way to present information to the user, and cope with one of several input modalities selected by the user without the need to explicitly tell the computer.

Relations between metaphors and diagrammatic or pictorial representations are also other interesting recent developments [239, 124], which can be useful in the design of multimedia systems.

In cognitive musicology, mental models and analogical representations based on metaphors are widespread. They are mostly related to the problem of music imagery. For example, Todd [219] argues that musical phrasing has its origin in the kinematics and the self-stimulation of (virtual) self-movement. This is grounded on the psycho-physical structure of the human auditory system. Another example comes from robot navigation in 3D space. In this task domain, a bipolar force field is a useful metaphor: the moving robot corresponds to an electric charge, and a target to be reached corresponds to a charge of opposite sign. Obstacles correspond to charges of the same sign. Metaphorical reasoning implies the use of multiple representational levels and environment-based representations.

5.2.1 Application Scenarios

As a general assumption, in MIAMI we follow a bottom-up approach to AI, grounding representation and reasoning models on the psycho-physical and perceptual aspects. Examples of bottom-up approaches starting from the signal, perceptual level of representation of musical signals can be found in [80, 8, 183].

According to our bottom-up approach to AI and multimedia, we are mainly interested in the study and development of "autonomous" multimedia systems, that is, autonomous agents characterized by multimodal interaction with user(s) in an unstructured and evolving environment [201, 320]. In particular, examples of real world scenarios that are considered in MIAMI are the following:

- a theatrical machine: a system delegated to manage and integrate sound, music, and either three-dimensional computer animation of humanoid figures (e.g., dance movements) or the movement of real autonomous agents in a theater stage (e.g., a real vehicle on wheels, equipped with on-board sensors, a computer for the low-level processing of sensorial data, etc.). In both cases we have an agent which should be able to move, navigate, react to events happening on stage (e.g., actions performed by the actors), to acquire sounds from the environment, and possibly execute musical tasks;

- a museal machine, based on an autonomous robot, very similar in its architecture to the theatrical machine, operates in real time in a museum exhibition area, and is able to welcome, entertain, guide, and instruct visitors. See [319] for a different AI-based model developed for a similar museal domain. In both cases, a musical competence is part of the requirements for the systems.

- interactive systems for the acquisition of human movement (at various levels of abstraction) and their use in various applications: e.g., use of this information to control computer tasks going from advanced user interfaces to entertainment tools.

The main keypoints considered in the design of experimental system architectures in these application scenarios are: (i) high-level integration of different modalities and skills; (ii) high level, multimodal interaction with the real world in real time: for example, such a system builds up representations and reasons on a realistic model of what is happening on stage, e. g., for deciding how to interpret or generate a music object, far more than a simple "triggering" mechanism: interaction therefore means a deeper mechanism than a simple temporal synchronization of chunks of music or graphics/animation data. Given the complexity of the problem domain, we deem that a single knowledge representation and reasoning formalism is not sufficient for all the aspects of such complex domains.

Chapter 6

Scenarios & Dreams

All former chapters have more or less introduced the 'state of the art' of their particular topic. In this last chapter, we will dare to look in the future and describe some possible scenarios, or 'dreams', which might be implemented within the scope of MIAMI. The first section deals with the application of multimodal techniques in the domain of music and art. In the second part, a multimodal control architecture for a mobile robot with an active stereo vision system will be described.

6.1 The Multimodal Orchestra

A possible application scenario which can be relevant for entertainment, education, and clinical evaluation, is one in which the user is immersed in a multimodal experience but not a conventional virtual reality that he alone can perceive. Rather, we can think of an audio-visual environment which can be communicated to other humans, either other actors participating in the same event or external spectators of the action. For example, we can think of a sound/music/light/image/speech synthesis system which is driven/tuned by the movements of the actors using specific metaphors for reaching, grasping, turning, pushing, navigating, etc. Specific scenarios could regard:

- multimodal musical instruments
- multimodal orchestras
- multimodal choreography
- multimodal kinetic painting
- multimodal teaching by showing

Multimodality technology requires an instrumented environment, instrumented dresses, and instrumented tools which allow to capture natural human movements. Learning problems have two sides and at least some of them can be translated to a dimensionality-reduction paradigm:

1. on the human side, like when learning to drive a car, there is the need to perceive the complex but smooth and continuous associations between movements and actions, avoiding cognitive bottlenecks which may arise with discontinuous and/or symbolic feedback;

2. on the computer side, there is the dimensionality-reduction problem of extracting "principal components" from a high-dimensional space of redundant degrees of freedom and this may or may not imply the recognition of gestures.

In any case, we need learning because in multimodal systems there is not, in general, a simple one-to-one translation of signals and events as in the VR systems.

6.2 Multimodal Mobile Robot Control

One of the most interesting research areas in robotics is the development of fully autonomous systems. In order to achieve autonomy, the mobile systems have to be equipped with sophisticated sensory systems, massive computing power, and what might be called an "intelligent controller". Because this is not possible with today's technology, many researchers try to incorporate the human operator in the sense and control loop, thus replacing parts of the system where appropriate. For instance, the recognition and decision capabilities of humans are much better than that of a mobile robot at the present time. Although a mobile system might be equipped with a vision system, the image processing and object recognition still takes to much time to drive autonomously with reasonable speed. On the other hand, the sensory systems of the robot which act locally can perceive and process data which is usually not detectable by humans, like the reflection of ultrasound or gas molecules which might be "smelt" by a special detector. Thus, a teleoperation system where parts of the controller will run autonomously whereas others will need the operator's interaction seems to be a very promising approach.

The control of a mobile vehicle (or robot) with advanced multimodal methods and sophisticated I/O devices is both, very attractive and useful. Take, e. g., the mobile platform PRIAMOS (shown in figure 6.1) which is currently under development at the University of Karlsruhe [188]. It is equipped with a multisensor system which includes the following:

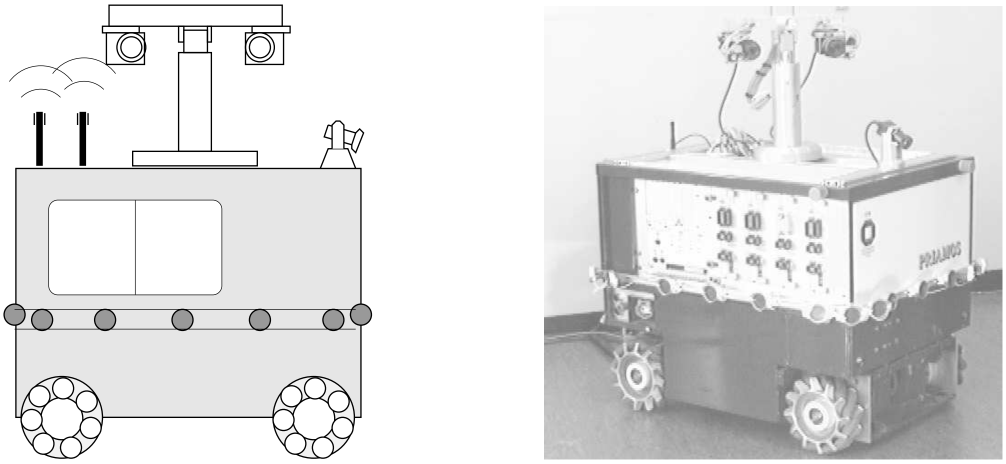

Figure 6.1: *A sketch and a real picture of the mobile robot system* Priamos

Ultrasonic sensors 24 sensors are placed as a ring around the robot. The system offers several modes and each sensor can be addressed independently.

Active vision system The active stereo vision system Kastor is mounted on top of the platform, allowing the operator to get a panoramic view of the remote world. 18 degrees of freedom can be controlled independently [357].

Structured light Two laser diodes emit structured light which is especially easy to "see" for a third camera due to the filters applied. The intersection of the two laser lines allows an easy detection of obstacles on the floor.

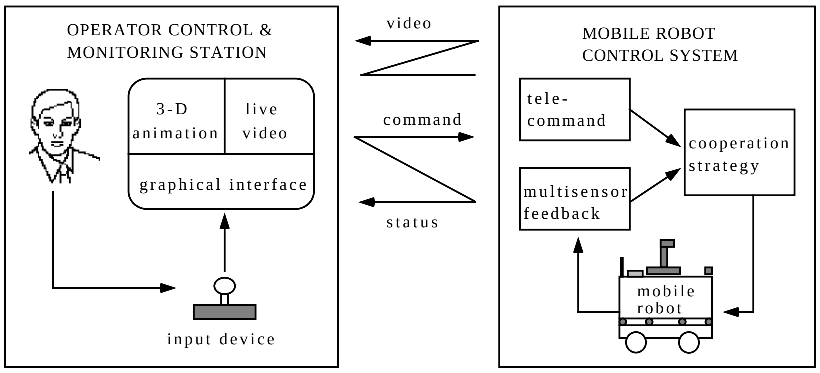

Figure 6.2: *The controller configuration of* Priamos *(taken from [188])*

The control concept of Priamos which is shown in figure 6.2 follows the "shared autonomy" approach (see [142]), i. e. the robot receives its programs and data from a supervisory station and carries out its tasks autonomously. If an unforeseen event occurs, the robot asks the human operator for assistance. In addition, the operator can take over control at any time. Supervision of the robot's tasks is performed with the sensory data sent to the operator and a simulation model which runs in parallel[1] to the real execution.

The software and hardware architecture of PRIAMOS is suitable to be extended with multimodal functionalities. One of several possible system configurations will be described in the following scenario (which is only fictitious at the moment):

> "The mobile robot's platform operates in either one of two different modes. In *Autonomous Mode*, the robot uses an internal map and its sensory system to navigate without any help from the operator who only supervises the movements. In the *Supervisory Mode*, the operator remotely controls the robot. For this task he uses a joystick which not only controls all degrees of freedom (x- and y-translation, z-rotation) but also reflects several characteristics of the robot's environment.
> Therefore, the joystick has been equipped with force feedback capabilities. Whenever the robot approaches an obstacle which is sensed by the ultrasonic sensors, the operator is able to feel it when the distance falls short of a certain threshold. Thus, collisions can be avoided without charging the operator's "vision system" with additional information.
> This is especially necessary because the operator is already burdened with the images of the active stereo vision system KASTOR. Both cameras send their images independently to one of the two displays in the Head-Mounted Display (HMD) that the operator is wearing. Thus, he gets a full 3D impression of the remote world. The camera head can also operate in autonomous or supervisory mode. In the latter, the operator uses his second hand to move the cameras in the desired direction. By fully controlling all 18 degrees of freedom with a kind of master-slave manipulator or a 6D mouse, the operator is able to focus any object in the view of the cameras.
> When using the helmet, the operator is no longer able to use a keyboard for additional command input. This is no drawback, because all commands can simply be spoken as the control system is equipped with a natural language processing (NLP) unit. Although only keywords and values can be recognized by it, it is comfortable and easy to use because it is a speaker independent system which doesn't have to be trained. Therefore, simple commands like "stop" or "faster" can be executed in a very intuitive way.
> Instead of the force feedback described above, another way to inform the operator about nearby obstacles is to convert the signals of the ultrasonic sensors

[1] Or even some time ahead as a kind of *prediction* mechanism.

> into acoustical information. For instance, the reflection of a wall detected by the sensors can be artificially created by a sound generator, thus giving the operator the impression of "walking" through the remote environment. Of course, alarm signals are transmitted to him by acoustical stimulus, too."

The scenario described above has not been realized yet, but it is based exclusively on techniques which are either already existing or under current development. The most difficult problem will be the complexity of this system, which also complicates the integration of all the separate system components. We are convinced that the techniques and methods which are going to be developed in MIAMI will contribute significantly to this integration process.

Appendix A

An Introduction to Binaural Technology

Humans, like most vertebrates, have two ears that are positioned at about equal height at the two sides of the head. Physically, the two ears and the head form an antenna system, mounted on a mobile base. This antenna system receives elastomechanical (acoustic) waves of the medium in which it is immersed, usually air. The two waves received and transmitted by the two ears are the physiologically adequate input to a specific sensory system, the auditory system.

The peripheral parts of the auditory system transform each of the two waves into neural spike trains, after having performed a running spectral decomposition into multiple frequency channels, among other preprocessing. The multi-channel neural spike trains from each of the two ears are then combined in a sophisticated way to generate a running "binaural-activity pattern" somewhere in the auditory system. This binaural-activity pattern, most probably in combination with monaural-activity patterns rendered individually by each ear's auditory channels, forms the auditory input to the cortex, which represents a powerful biologic multi-purpose parallel computer with a huge memory and various interfaces and in- and output ports. As an output, the cortex delivers an individual perceptual world and, eventually, neural commands to trigger and control specific motoric expressions.

It goes without saying that a number of constraints must to hold for this story to be true. For example, the acoustic waves must be in the range of audibility with respect to frequency range and intensity, the auditory system must be operative, and the cortex must be in a conscious mode, ready to accept and interpret auditory information. Further, it makes sense to assume that multiple sources of feedback are involved in the processes of reception, processing and interpretation of acoustic signals. Feedback clearly occurs between the modules of the subcortical auditory system, and between this system and the cortex. Obvious feedback from higher centers of the central nervous system to the motoric positioning system of the ears-and-head array can also be observed whenever position-finding movements of the head are induced.

Although humans can hear with one ear only - so called monaural hearing - hearing with two functioning ears is clearly superior. This fact can best be appreciated by considering the biological role of hearing. Specifically, it is the biological role of hearing to gather information about the environment, particularly about the spatial positions and trajectories of sound sources and about their state of activity. Further, it should be recalled in this context that interindividual communication is predominantly performed acoustically, with brains deciphering meanings as encoded into acoustic signals by other brains.

In regard of this generic role of hearing, the advantage of binaural as compared to monaural hearing stands out clearly in terms of performance, particularly in the following areas [27]:

1. localization of single or multiple sound sources and, consequently, formation of an auditory perspective and/or an auditory room impression;
2. separation of signals coming from multiple incoherent sound sources spread out spatially or, with some restrictions, coherent ones;
3. enhancement of the signals from a chosen source with respect to further signals from incoherent sources, as well as enhancement of the direct (unreflected) signals from sources in a reverberant environment.

It is evident that the performance features of binaural hearing form a challenge for engineers in terms of technological application. In this context a so-called *Binaural Technology* has evolved during the past three decades, which can operationally be defined as follows.

Binaural Technology is a body of methods that involve the acoustic input signals to both ears of the listener for achieving practical purposes, e. g., by recording, analyzing, synthesizing, processing, presenting and evaluating such signals.

Binaural Technology has recently gained in economic momentum, both on its own and as an enabling technology for more complex applications. A specialized industry for Binaural Technology is rapidly developing. It is the purpose of this chapter to take a brief look at this exciting process and to reflect on the bases on which this technology rests, i. e., on its experimental and theoretical foundations. As has been discussed above, there are basically three “modules” engaged in the reception, perception and interpretation of acoustical signals: the ears-and-head array, the subcortical auditory system, and the cortex. Binaural

Technology makes use of knowledge of the functional principles of each. In the following three sections, particular functions of these three modules are reviewed in the light of their specific application in Binaural Technology.

A.1 The Ears-and-Head Array: Physics of Binaural Hearing

The ears-and-head array is an antenna system with complex and specific transmission characteristics. Since it is a physical structure and sound propagation is a linear process, the array can be considered to be a linear system. By taking an incoming sound wave as the input and the sound pressure signals at the two eardrums as the output, it is correct to describe the system as a set of two self-adjusting filters connected to the same input. Self-adjusting, in the sense used here, means that the filters automatically provide transfer functions that are specific with regard to the geometrical orientation of the wavefront relative to the ears-and-head array.

Physically, this behavior is explained by resonances in the open cavity formed from pinna, ear canal and eardrum, and by diffraction and reflection by head and torso. These various phenomena are excited differently when a sound wave impinges from different directions and/or with different curvatures of the wavefront. The resulting transfer functions are generally different for the two filters, thus causing “interaural” differences of the sound-pressure signals at the two eardrums. Since the linear distortions superimposed upon the sound wave by the two “ear filters” are very specific with respect to the geometric parameters of the sound wave, it is not far from the mark to say that the ears-and-head system encodes information about the position of sound sources in space, relative to this antenna system, into temporal and spectral attributes of the signals at the eardrums and into their interaural differences. All manipulations applied to the sound signals by the ears-and-head array are purely physical and linear. It is obvious, therefore, that they can be simulated. As a matter of fact, there is one important branch of Binaural Technology that attempts to do just this.

It makes sense at this point to begin the technological discussion with the earliest, and still a very important application, of Binaural Technology, namely, authentic auditory reproduction. Authentic auditory reproduction has been achieved when listeners hear exactly the same in a reproduction situation what they would hear in an original sound field, the latter existing at a different time and/or location. As a working hypothesis, Binaural Technology begins with the assumption that listeners hear the same in a reproduction situation as in an original sound field when the signals at the two ear-drums are exactly the same during reproduction as in the original field. Technologically, this goal is achieved by means of so-called artificial heads which are replicas of natural heads in terms of acoustics, i. e. they realize two self-adjusting ear filters like natural heads.

Artificial heads, in combination with adequate playback equipment, are a basic instrumentation for a number of economically appealing applications. The playback equipment, needed for this application, is usually based on headphones. Yet, under specific, restricted conditions, loudspeakers can also be used. A first category of application in this context is subsumed under the following section.

A.1.1 Binaural recording and authentic reproduction

These applications exploit the capability of Binaural Technology to archive the sound field in a perceptually authentic way, and to make it available for listening at will, e. g., in entertainment, education, instruction, scientific research, documentation, surveillance, and telemonitoring. It should be noted here that binaural recordings can be compared in direct sequence (e. g., by A/B comparison), which is often impossible for the original sound situations. Since the sound-pressure signals at the two ear-drums are the physiologically adequate input to the auditory system, they are considered the basis for auditory-adequate measurement and evaluation, both in a physical and/or auditory way [31]. Consequently, we have the further application category discussed below.

A.1.2 Binaural measurement and evaluation

In physical binaural measurement, physically based procedures are used, whereas in the auditory case, human listener serve as measuring and evaluating instruments. Current applications of binaural measurement and evaluation can be found in areas such as noise control, acoustic-environment design, sound-quality assessment (for example, in speech-technology, architectural acoustics and product-sound design), and in specific measurements on telephone systems, headphones, personal hearing protectors, and hearing aids [30, 305]. For some applications, scaled-up or scaled-down artificial heads are in use, for instance, for the evaluation of architectural scale models [91, 92, 371, 373].

Since artificial heads, basically, are just a specific way of implementing a set of linear filters, one may think of other ways of realizing such filters, e. g., electronically. For many applications this adds additional degrees of freedom, as electronic filters can be controlled at will over a wide range. This idea leads to yet another category of application, as follows.

A.1.3 Binaural simulation and displays

There are many current applications in binaural simulation and displays, with the potential of an ever-increasing number. The following list provides examples: binaural mixing [277] binaural room simulation [180, 181] advanced sound effects (for example, for computer games), provision of auditory spatial-orientation cues (e. g., in the cockpit or for the blind), auditory display of complex data, and auditory representation in teleconference, telepresence and teleoperator systems.

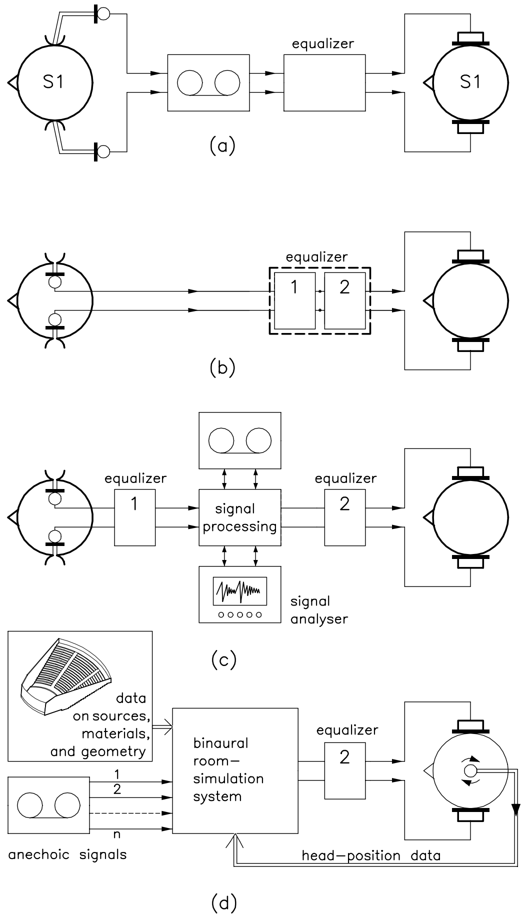

Figure A.1: *Binaural-Technology equipment of different complexity: (a) probe-microphone system on a real head, (b) artificial-head system, (c) artificial-head system with signal-processing and signal-analysis capabilities, (d) binaural room-simulation system with head-position tracker for virtual-reality applications.*

Figure A.1, by showing Binaural-Technology equipment in an order of increasing complexity, is meant to illustrate some of the ideas discussed above. The most basic equipment is obviously the one shown in panel (a). The signals at the two ears of a subject are picked up by (probe) microphones in a subject's ear canal, then recorded, and later played back to the same subject after appropriate equalization. Equalization is necessary to correct linear distortions, induced by the microphones, the recorder, and the headphones, so that the signals in the subject's ear canals during the playback correspond exactly to those in the pick-up situation. Equipment of this kind is adequate for personalized binaural recordings. Since a subject's own ears are used for the recording, maximum authenticity can be achieved.

Artificial heads (panel b) have practical advantages over real heads for most applications; for one thing, they allow for auditory real-time monitoring of a different location. One has to realize, however, that artificial heads are usually cast or designed from a typical or representative subject. Their directional characteristics will thus, in general, deviate from those of an individual listener. This fact can lead to a significant decrease in perceptual authenticity. For example, errors such as sound coloration or front-back confusion may appear. Individual adjustment is only partly possible, namely, by equalizing the headphones specifically for each subject. To this end, the equalizer may be split into two components, a head equalizer (1) and a headphone equalizer (2). The interface between the two allows some freedom of choice. Typically, it is defined in such a way that the artificial head features a flat frequency response either for frontal sound incidence (free-field correction) or in a diffuse sound field (diffuse-field correction). The headphones must be equalized accordingly. It is clear that individual adjustment of the complete system, beyond a specific direction of sound incidence, is impossible in principle, unless the directional characteristics of the artificial head and the listener's head happen to be identical.

Panel (c) depicts the set-up for applications were the signals to the two ears of the listener are to be measured, evaluated and/or manipulated. Signal-processing devices are provided to work on the recorded signals. Although real-time processing is not necessary for many applications, real-time play back is mandatory. The modified and/or unmodified signals can be monitored either by a signal analyzer or by binaural listening.

The most complex equipment in this context is represented by panel (d). Here the input signals no longer stem from a listener's ears or from an artificial head, but have been recorded or even generated without the participation of ears or ear replicas. For instance, anechoic recordings via conventional studio microphones may be used. The linear distortions which human ears superimpose on the impinging sound waves, depending on their direction of incidence and wave-front curvature, are generated electronically via a so-called ear-filter bank (electronic head). To be able to assign the adequate head-transfer function to each incoming signal component, the system needs data of the geometry of the sound field. In a typical application, e. g. architectural-acoustics planning, the system contains a sound-field simulation based on data of the room geometry, the absorption features of the materials implied, and the positions of the sound sources and their directional characteristics. The output of the sound-field modeling is fed into the electronic head, thus producing so-called binaural impulse responses. Subsequent convolution of these impulse responses with anechoic signals generates binaural signals as a subject would observe in a corresponding real room. The complete method is often referred to as binaural room simulation.

To give subjects the impression of being immersed in a sound field, it is important that perceptual room constancy is provided. In other words, when the subjects move their heads around, the perceived auditory world should nevertheless maintain its spatial position. To this end, the simulation system needs to know the head position in order to be able to control the binaural impulse responses adequately. Head position sensors have therefore to be provided. The impression of being immersed is of particular relevance for applications in the context of virtual reality.

All of the applications discussed in this section are based on the provision of two sound-pressure signals to the ear-drums of human beings, or on the use of such signals for measurement and application. They are built on our knowledge of what the ears-and-head array does, i. e., on our understanding of the physics of the binaural transmission chain in front of the eardrum. We shall now proceed to the next section, which deals with the signal processing behind the eardrum and its possible technical applications.

A.2 The Subcortical Auditory System: Psychophysics of Binaural Hearing

As mentioned above, the subcortical auditory system converts incoming sound waves into neural spike trains which are then processed in a very sophisticated way. Among the things that we know from physiological experiments are the following. The signals are decomposed into spectral bands that are maintained throughout the system. Autocorrelation of the signals from each of the ears, as well as cross-correlation of the signals from both ears, are performed. Specific inhibition and excitation effects are extensively present.

Models of the function of the subcortical auditory system take our knowledge of its physiology into account, but are usually oriented primarily towards the modeling of psychoacoustic findings. Most models have a signal-driven, bottom-up architecture. As an output, a (running) binaural-activity pattern is rendered that displays features corresponding to psychoacoustic evidence and/or allows for the explanation of binaural performance features. Since psychoacoustics, at least in the classical sense, attempts to design listening experiments in a "quasi-objective" way, psychoacoustic observations are, as a rule, predominantly associated with processes in the subcortical auditory system.

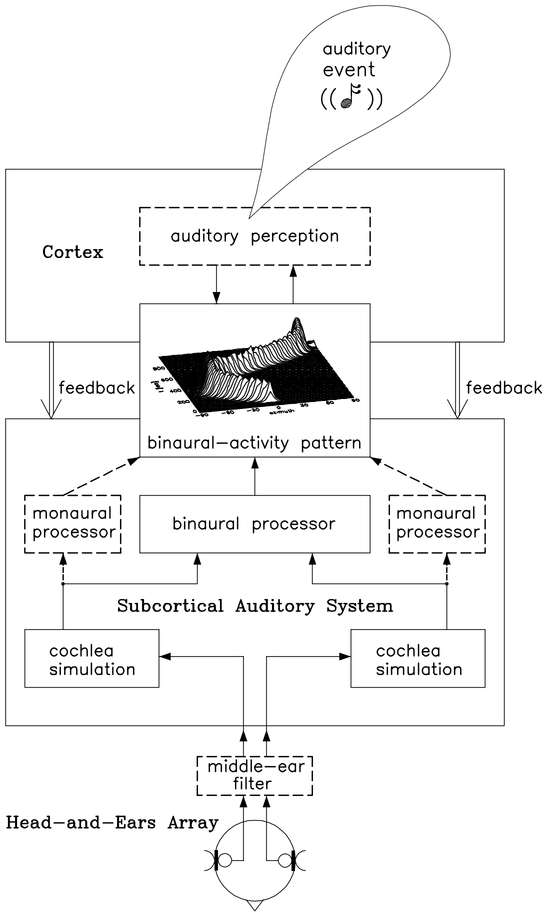

Figure A.2: *Architecture for an application oriented model of binaural hearing: Binaural signals as delivered by the ear-and-head array (or its electronic simulation) are fed into a model of the subcortical auditory system, implying simulation of the function of the cochleae and of binaural interaction as essential modules. The interface between the subcortical auditory model and the evaluation stages on top of it is provided by a running binaural-activity pattern.*

There seems to be the following consensus among model builders. A model of the subcortical auditory system must at least incorporate three functional blocks to simulate binaural performance in the areas as listed above (figure A.2):

1. a simulation of the functions of the external ear, including head (skull), torso, pinnae, ear canal, and eardrum; plus, eventually, the middle ear;

2. a simulation of the inner ears, i.e. the cochleae, including receptors and first neurons; plus a set of binaural processors to identify interaurally correlated contents of the signals from the two cochleae and to measure interaural arrival-time and level differences; along with, eventually, additional monaural processors.

3. a set of algorithms for final evaluation of the information rendered by the preceding blocks with respect to the specific auditory task to be simulated.

The first block corresponds to the head-and-ears array as discussed in the preceding section, with the exception of the middle ear. As a matter of fact, detailed modeling of the middle ear is deemed unnecessary in current Binaural Technology. The middle ear is approximated by a linear time-invariant bandpass, thus neglecting features such as the middle-ear reflex. Nevertheless, more elaborate models of the middle ear were readily available from literature, if needed, [146, 145, 147, 32].

The second block includes two essential modules, cochlea simulation and simulation of subcortical binaural interaction. They will now be discussed in this order. The cochlea model simulates two primary functions, namely, a running spectral analysis of the incoming signals, and a transformation of the (continuous) mechanical vibrations of the basilar membrane into a (discrete) nerve-firing pattern: physiological analog-to-digital conversion. In doing so, it has to be considered that both spectral selectivity and A/D conversion depend on the signal amplitude, i. e., behave nonlinearly. The simplest approximation for the spectral selectivity to be simulated is by means of a bank of adjacent band-pass filters, each, for example, of critical bandwidth. This realization is often used when computing speed is more relevant than precision. More detailed modeling is achieved by including the spectrally-selective excitation at each point of the basilar membrane. The amplitude dependence of excitation and selectivity can optionally be included into the model by simulating active processes, which are supposed to be part of the functioning of the inner ear.

A more precise simulation of the physiological A/D conversion requires a stochastic receptor-neuron model to convert movement of the basilar membrane into neural-spike series. Such models have indeed been implemented for simulations of some delicate binaural effects. However, for practical applications, it is often not feasible to process individual neural impulses. Instead, one can generate deterministic signals that represent the time function of the firing probability of a bundle of nerve fibers. For further simplification, a linear dependence of the firing probability on the receptor potential is often assumed. The receptor potential is sufficiently well described for many applications by the time function of the movement of the basilar membrane, half-wave rectified and fed through of first order low-pass with a 800 Hz cut-off frequency. This accounts for the fact that, among other things, in the frequency region above about 1.5 kHz, binaural interaction works on the envelopes rather than on the fine structure of the incoming signals.

With regard to the binaural processors, the following description results from work performed in the author's lab at Bochum (e. g., [190, 191, 113].) First, a modified, interaural running-cross-correlation function is computed, based on signals originating at corresponding points of the basilar membranes of the two cochlea simulators , i. e., points which represent the same critical frequency. The relevance of cross-correlation to binaural processing has been assumed more than once and is, moreover, physiologically evident. A Bochum modification of cross-correlation consists in the employment of a binaural contralateral inhibition algorithm. Monaural pathways are further included in the binaural processors to allow for the explanation of monaural-hearing effects.

Some details of the binaural processors are given in the following. The first stage of the processor is based on the well known coincidence-detector hypothesis. A way to illustrate this is by assuming two complementary tapped delay lines - one coming from each ear - whose taps are connected to coincidence cells which fire on receiving simultaneous excitation from both side's delay lines. It can be shown that this stage renders a family of running interaural cross-correlation functions as output. Thus we arrive at a three-dimensional pattern (interaural arrival-time difference, critical-band frequency, cross-correlation amplitude) which varies with time and can be regarded as a running binaural-activity pattern. The generation of the running cross-correlation pattern is followed by application of a mechanism of contralateral inhibition based on the following idea. Once a wavefront has entered the binaural system through the two ears, it will consequently give rise to an activity peak in the binaural pattern. Consequently, inhibition will be applied to all other possible positions of activity in each band where excitation has taken place. In each band where signals are received, the first incoming wavefront will thus gain precedence over possible activity being created by later sounds which are spectrally similar to the first incoming wavefront, such as reflections. The actual amount of inhibition is determined by specific weights which vary as a function of position and time, such as to fit psychoacoustical data. Inhibition may, for example, continue for a couple of milliseconds and then gradually die away until it is triggered again. Using this concept as well as specific algorithm of contralateral inhibition, in combination with the inclusion of monaural pathways into the processor, the processing of interaural level differences by the binaural system is properly modeled at the same time. For certain combinations of interaural arrival-time and interaural level differences, e. g. "unnatural" ones, the model will produce multiple peaks in the inhibited binaural activity pattern, thus predicting multiple auditory events - very much in accordance with the psychoacoustical data [114].

To deal with the problem of natural interaural level differences being much higher at high frequencies than at low ones, the binaural processors must be adapted to the external-ear transfer functions used in the model. To this end, additional inhibitory weighting is implemented on the delay lines of the coincidence networks in such a way that the binaural processors are always excited within their "natural" range of operation. This additional weighting is distributed along the delay lines. The complete set of binaural processors can, thus, be conceptualized as an artificial neural network, more specifically, as a particular kind of time-delay neural network. The adaptation of this network to the particular set of external-ear transfer functions used is accomplished by means of a supervised learning procedure.

The output of the binaural processor, a running binaural-activity pattern, is assumed to be interfacing to higher nervous centers for evaluation. The evaluation procedures must be defined with respect to the actual, specific task required. Within the scope of our current modeling, the evaluation process is thought of in terms of pattern recognition. This concept can be applied when the desired output of the model system is a set of sound-field parameters, such as the number and the positions of the sound source, the amount of auditory spaciousness, reverberance, coloration etc. Also, if the desired output of the model system is processed signals, such as a monophonic signal which has been improved with respect to its S/N ratio, the final evaluative stage may produce a set of parameters for controlling further signal processing.

Pattern-recognition procedures have so far been projected for various tasks in the field of sound localization and spatial hearing, such as lateralization, multiple image phenomena, summing localization, auditory spaciousness, binaural signal enhancement, and parts of the precedence effect (see [29] for cognitive components of the precedence effect). Further, effects such as binaural pitch, dereverberation and/or decoloration are within the scope of the model.

We shall now consider the question of whether the physiological and psychoacoustic knowledge of the subcortical auditory system, as manifested in models of the kind described above, can be applied for Binaural Technology. Since we think of the subcortical auditory system as a specific front-end to the cortex that extracts and enhances certain attributes from the acoustic waves for further evaluation, signal-processing algorithms as observed in the subcortical auditory system may certainly be applied in technical systems to simulate performance features of binaural hearing. Progress in signal-processor technology makes it feasible to implement some of them on microprocessor hardware for real-time operation. Consequently, a number of interesting technical applications have come into reach of today's technology. A first category is concerned with spatial hearing, as described below.

A.2.1 Spatial hearing

Auditory-like algorithms may decode information from the input signals to the ear that allows assessment of the spatial position of sound sources. They may further be used for predictions of how humans form the positions and spatial extents of their auditory events, how they establish an auditory perspective, and how they suppress echoes and reverberance. Typical applications are: source-position finders, tools for the evaluation of architectural acoustics and sound systems (such as spaciousness meters, echo detectors, and precedence indicators,) tools for the evaluation of auditory virtual environments and for psychoacoustic research. There are further perceptual features of auditory events, besides position and spatial extent, which are based on binaural rather than monaural information. Following a usage in the field of product-sound design, they may be called binaural psychoacoustic descriptors, as discussed in the next section.

A.2.2 Binaural psychoacoustic descriptors

Binaural psychoacoustic descriptors include binaural loudness, binaural pitch, binaural timbre and binaural sensory-consonance. Algorithms taken from binaural auditory models may be used to generate estimates of these descriptors. There is an increasing demand for such tools, e. g., in the area of sound-quality evaluation. The most tempting field of application for binaural auditory models, however, concerns the ability of binaural hearing to process signals from different sources selectively, and to enhance one of them with regard to the others. A key word for this area of application could be binaural signal enhancement.

A.2.3 Binaural signal enhancement

A well-known term in the context of binaural signal enhancement is the so-called "cocktail-party effect," denoting that, with the aid of binaural hearing, humans can concentrate on one talker in the presence of competing ones. It has further been established that with binaural hearing a desired signal and noise can be separated more effectively than with monaural hearing. Binaural auditory models may help to simulate these capabilities by providing front-ends that allow for better separation of a mix of sound sources. In a specific Bochum version of a so-called "cocktail-party" processor, i. e., a processor to enhance speech in a "cocktail-party" situation, the binaural processor of the auditory model is used to control a Wiener filter [35, 34]. This is accomplished by first identifying the position of a desired talker in space, and then estimating its S/N ratio with respect to competing talkers and other noise signals. The system performs its computation within critical-bands. In the case of two competing talkers, the desired signal can be recovered to reasonable intelligibility, even when its level is 15 dB lower than that of the competing one. Application possibilities for this kind of systems are numerous, such as tools for editing binaural recordings, front ends for signal-processing hearing aids, speech-recognizers and hands-free telephones. In general, binaural signal enhancement may be used to build better "microphones" for acoustically adverse conditions. As stated above, the cues provided by models of the subcortical auditory system, and contained in binaural-activity patterns, must consequently be evaluated in adequate ways. The next section deals with this problem.

A.3 The Cortex: Psychology of Binaural Hearing

Most models of the subcortical auditory system assume a bottom-up, signal-driven process up to their output, the running binaural-activity pattern. The cortex, consequently, takes this pattern as an input. The evaluation of the binaural-activity pattern can be conceived as a top-down, hypothesis-driven process. According to this line of thinking, cortical centers set up hypotheses, e. g., in terms of expected patterns, and then try to confirm these hypotheses with appropriate means, e. g., with task-specific pattern-recognition procedures. When setting up hypotheses, the cortex reflects on cognition, namely, on knowledge and awareness of the current situation and the world in general. Further, it takes into account input from other senses, such as visual or tactile information. After forming hypotheses, higher nervous stages may feed back to more peripheral modules to prompt and control optimum hypothesis testing. They may, for example, induce movement of the ears-and-head array or influence the spectral decomposition process in the subcortical auditory system.

The following two examples help to illustrate the structure of problems that arise at this point from a technological point of view. First, in a "cocktail-party" situation a human listener can follow one talker and then, immediately, switch his attention to another. A signal-processing hearing aid should be able to do the same thing, deliberately controlled by its user. Second, a measuring instrument to evaluate the acoustic quality of concert halls will certainly take into account psychoacoustic descriptors like auditory spaciousness, reverberance, auditory transparency, etc. However, the general impression of space and quality that a listener develops in a room, may be coØdetermined by visual cues, by the specific kind of performance, by the listener's attitude, and by factors like fashion or taste, among other things.

There is no doubt that the involvement of the cortex in the evaluation process adds a considerable amount of "subjectivity" to binaural hearing, which poses serious problems to Binaural Technology. Engineers, as most scientists, are trained to deal with the object as being independent of the observer (assumption of "objectivity") and prefer to neglect phenomena that cannot be measured or assessed in a strictly "objective" way. They further tend to believe that any problem can be understood by splitting it up into parts, and analyzing these parts separately. At the cortical level, however, we deal with percepts, i. e. objects that do not exist as separate entities, but as part of a subject-object (perceiver-percept) relationship. It should also be noted that listeners normally listen in a "gestalt" mode, i. e., they perceive globally rather than segmentally. An analysis of the common engineering type may thus completely miss relevant features.

Perceiver and percept interact and may both vary considerably during the process of perception. For example, the auditory events may change when listeners focus on specific components such as the sound of a particular instrument in an orchestra. Further, the attitude of perceivers towards their percepts many vary in the course of an experimental series, thus leading to response modification.

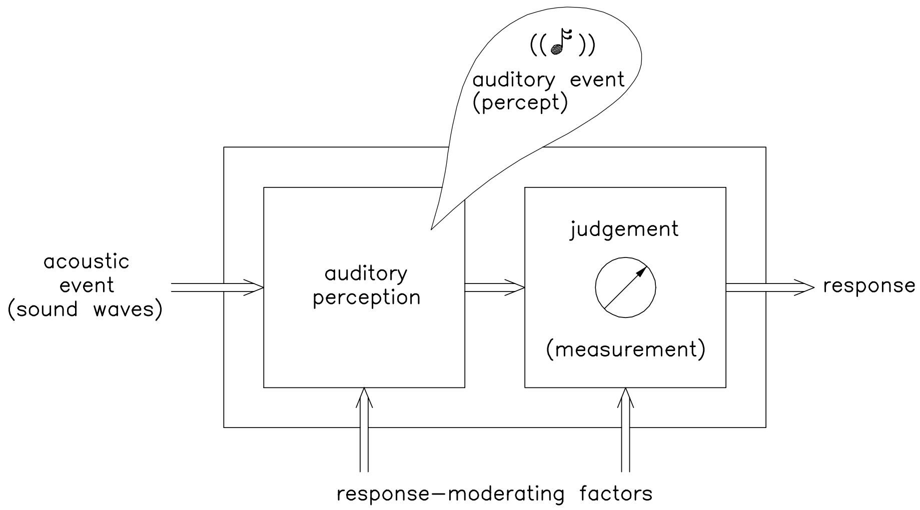

Figure A.3: *Schematic of a subject in a listening experiment. Perception as well as judgement are variant, as modeled by the assumption of response-moderating factors.*

A simple psychological model of the auditory perception and judgment process, shown in figure A.3, will now be used to elaborate on the variance of listeners' auditory events in a given acoustic setting and the variance of their respective responses. The schematic symbolizes a subject in a listening experiment. Sound waves impinge upon the two ears, are preprocessed and guided to higher centers of the central nervous system, where they give rise to the formation of an auditory event in the subject's perceptual space. The auditory event is a percept of the listener being tested, i. e., only he/she has direct access to it. The rest of the world is only informed about the occurrence of the said percept, if the subject responds in such a way as to allow conclusion to be made from the response to the percept (indirect access). In formal experiments the subject will usually be instructed to respond in a specified way, for example by formal judgement on specific attributes of the auditory event. If the response is a quantitative descriptor of perceptual attributes, we may speak of measurement. Consequently, in listening experiments, subjects can serve as an instrument for the measurements of their own perception, i. e., as both the object of measurement and the "meter". The schematic in Fig.3 features a second input into both the auditory-perception and the judgement blocks where "response-moderating factors" are fed in to introduce variance to the perception and judgement processes.

Following this line of thinking an important task of auditory psychology can be to identify such response-moderating factors and to clarify their role in binaural listening. Many of these factors represent conventional knowledge or experience from related fields of perceptual acoustics, e. g., noise- and speech-quality evaluation. It is well known that the judgements of listeners on auditory events may depend on the cognitive "image" which the listeners have with respect to the sound sources involved (source-related factors). It may happen, for instance, that the auditory events evoked by sources that are considered aggressive (e. g., trucks), are judged louder than those from other sources (e. g., passenger cars) - given the same acoustical signals. The "image" of the source in the listeners' minds may be based, among other things, on cues from other senses (e. g., visual) and/or on prior knowledge. Situative factors are a further determinant in this context, i. e., subjects judge an auditory event bearing the complete (multi-modal) situation in mind in which they occur. Another set of factors is given by the individual characteristics of each listener (personal factors), for example his/her subjective attitude towards a specific sound phenomenon, an attitude that may even change in the course of an experiment. Response-moderating factors that draw upon cognition tend to be especially effective when the sounds listened to transmit specific information, i. e., act as carriers of meaning. This is obvious in the case of speech sounds, but also in other cases. The sound of a running automobile engine, for instance, may signal to the driver that the engine is operating normally.

The fact that response moderating factors do not only act on judgements but also on the process of perception itself, may seem to be less obvious at a first glance, but is, nevertheless, also conventional wisdom. We all know that people in a complex sound situation have a tendency to miss what they do not pay attention to and/or do not expect to hear. There is psychoacoustical evidence that, e. g., the spectral selectivity of the cochlea is influenced by attention. At this point, the ability to switch at will between a global and an analytic mode of listening, should also be noted. It is commonly accepted amongst psychologists that percepts are the result of both the actual sensory input at a given time and of expectation.

If we want to build sophisticated Binaural-Technology equipment for complex tasks, there is no doubt that psychological effects have to be taken into account. Let us consider, as an example, a binaural-surveillance system for acoustic monitoring of a factory floor. Such a system must know the relevance and meaning of many classes of signals and must pay selective attention to very specific ones, when an abnormal situation has been detected. A system for the evaluation of acoustic qualities of spaces for musical performances must detect and consider a range of different shades of binaural signals, depending on the kind and purpose of the performances. It might even have to take into account the taste of the local audience or that of the most influential local music reviewer. An intelligent binaural hearing aid should know to a certain extent, which components of the incoming acoustic signals are relevant to its user, e. g., track a talker who has just uttered the users name.

As a consequence, we shall see in the future of Binaural Technology that psychological models will be exploited and implemented technologically, though, may be not, for a while, in the form of massively parallel biologic computing as in the cortex. There are already discussions about and early examples of combinations of expert systems and other knowledge-based systems with artificial heads, auditory displays and auditory-system models. When we think of applications like complex human/machine interfaces, multi-media systems, interactive virtual environments, and teleoperation systems, it becomes obvious that conventional Binaural Technology must be combined with, or integrated into, systems that are able to make decisions and control actions in an intelligent way. With this view in mind it is clear that Binaural Technology is still in an early stage of development. There are many relevant technological challenges and business opportunities ahead.

Acknowledgement

Many of the ideas discussed in this chapter have evolved from work at the author's lab. The author is especially indebted to his doctoral students over the years who have helped to provide a constant atmosphere of stimulating discussion. Several completed doctoral dissertations which are correlated to this chapter, along with references to tagged publications in English, are given in the following list of citations: M. Bodden [33, 35, 34, 28], J.-P. Col [72, 29], H. Els [90, 91, 92], W. Gaik [112, 114, 113], H. Hudde [144, 146, 145], H. Lehnert [179, 180, 181, 28], U. Letens [184, 32], W. Lindemann [189, 190, 191], W. Pompetzki [275], Ch. Pösselt [276, 306], D. Schlichthärle [298], J. Schröter [304, 30, 305], H. Slatky [314], S. Wolf [365], N. Xiang [372, 371, 373].

Appendix B

Audio-Visual Speech Synthesis

The bimodal information associated to speech is conveyed through audio and visual cues which are coherently produced at phonation time and are transmitted by the auditory and visual channel, respectively. Speech visual cues carry basic information on visible articulatory places and are associated to the shape of the lips, the position of the tongue, and the visibility of the teeth. This kind of information is usually complementary to that conveyed by the acoustic cues of speech which are mostly correlated to the voicing/unvoicing character.

The synthesis of speech visual cues represents a problem with a large variety of possible solutions ranging from continuous approaches were a small set of reference mouth images are variously "distorted" to obtain approximations of any possible mouth shape, to parametric approaches were a synthetic mouth icon is "animated" by simply changing its parameters.

Despite the specific methodology which is employed, the refresh frequency of the mouth shape must be high enough for guaranteeing the visual representation of rapid transients of speech, very important for comprehension. The temporal refresh frequency is usually higher than 25 Hz implying a resolution of 40 ms of speech. The basic articulators which must be reproduced are the lips, teeth and tongue. Secondary indicators like cheeks, chin and nose are usually very useful. A minimum spatial resolution of 100x100 pixel is needed.

B.1 Visual Speech Synthesis from Acoustics

Speech production is based on the basic mechanisms of phonation, related to the vibration of the vocal cords, and of vocal articulation, related to the time-varying geometry of the vocal tract responsible of the phonemic structure of speech. Forcing the diaphragm up, air is pushed out from the lungs into the trachea, glottis (the gap between the two vocal cords) and larynx before reaching the upper part of the vocal tube, called vocal tract, formed by pharynx, nasal and oral concavities. The periodical closure of the glottis interrupts the airflow generating a periodic variation of the air pressure whose frequency can be raised to the acoustic range. The harmonic components of this acoustic wave, multiples of the fundamental (pitch frequency), are then modified as long as the air flows through the vocal tract depending on its geometry. The vocal tract, in fact, can be shaped variously by moving the jaw, tongue, lips and velum: in this way the vocal tract implements a time-varying system capable to filter the incoming acoustic wave, reshaping its spectrum and modifying the produced sound.

Speech is the concatenation of elementary units, phones, generally classified as vowels if they correspond to stable configurations of the vocal tract or, alternatively, as consonants if they correspond to transient articulatory movements. Each phone is then characterized by means of a few attributes (open/closed, front/back, oral/nasal, rounded/unrounded) which qualify the articulation manner (fricative like /f/,/s/, plosive like /b/, /p/, nasal like /n/, /m/, ...) and articulation place (labial, dental, alveolar, palatal, glottal).

Some phones, like vowels and a subset of consonants, are accompanied by vocal cords vibration and are called "voiced" while other phones, like plosive consonants, are totally independent of cords vibration and are called "unvoiced". In correspondence of voiced phones the speech spectrum is shaped, as previously described, in accordance to the geometry of the vocal tract with characteristic energy concentrations around three main peaks called "formants", located at increasing frequencies F1, F2 and F3.

An observer skilled in lipreading is able to estimate the likely locations of formant peaks by computing the transfer function from the configuration of the visible articulators. This computation is performed through the estimation of four basic parameters: (i) the length of the vocal tract L, (ii) the distance d between the glottis and the place of maximum constriction; (iii) the radius r of the constriction; (iv) the ratio $\frac{A}{L}$ between the area A of the constriction and L. While the length L can be estimated a priori taking into account the sex and age of the speaker, the other parameters can be inferred, roughly, from the visible configuration. If the maximum constriction is located in correspondence of the mouth, thus involving lips, tongue and teeth as it happens for labial and dental phones, this estimate is usually reliable. In contrast, when the maximum constriction is non visible like in velar phones (/k/, /g/) the estimate is usually very poor.

Lipreading represents the highest synthesis of human expertise in converting visual inputs into words and then into meanings. It consists of a personal database of knowledge and skills constructed and refined by training, capable to associate virtual sounds to specific mouth shapes, generally called "viseme", and therefore infer the underlying acoustic message. The lipreader attention is basically focused on the mouth, including all its components like lips, teeth and tongue, but significant help in comprehension comes also from the entire facial expression.

In lipreading a significant amount of processing is performed by the lipreader himself who is skilled in post-filtering the converted message to recover from errors and from communication lags. Through linguistic and semantic reasoning it is possible to exploit the message redundancy and understand by context: this kind of knowledge-based interpretation is performed by the lipreader in real-time.

Audio-visual speech perception and lipreading rely on two perceptual systems working in cooperation so that, in case of hearing impairments, the visual modality can efficiently integrate or even substitute the auditory modality. It has been demonstrated experimentally that the exploitation of the visual information associated to the movements of the talker lips improves the comprehension of speech: the Signal to Noise Ratio (SNR) is incremented up to 15 dB and the auditory failure is most of times transformed into near-perfect visual comprehension. The visual analysis of the talker face provides different levels of information to the observer improving the discrimination of signal from noise. The opening/closing of the lips is in fact strongly correlated to the signal power and provides useful indications on how the speech stream is segmented. While vowels, on one hand, can be recognized rather easily both through hearing and vision, consonants are conversely very sensitive to noise and the visual analysis often represents the only way for comprehension success. The acoustic cues associated to consonants are usually characterized by low intensity, a very short duration and fine spectral patterning.

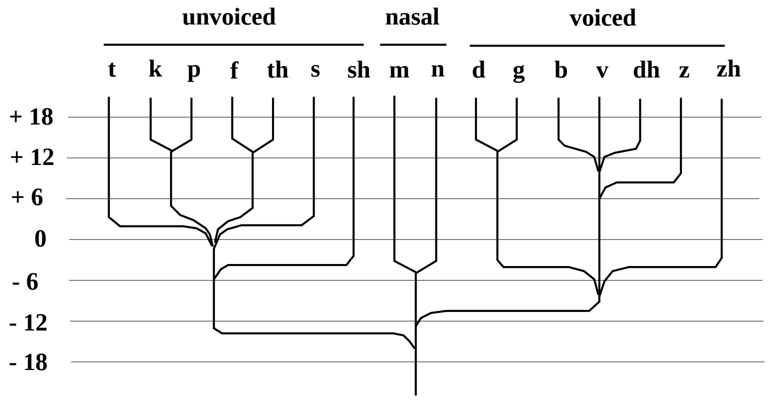

Figure B.1: *Auditory confusion of consonant transitions CV in white noise with decreasing Signal to Noise Ratio expressed in dB. (From B. Dodd, R. Campbell, "Hearing by eye: the psychology of lipreading", Lawrence Erlbaum Ass. Publ.)*

The auditory confusion graph reported in figure B.1 shows that cues of nasality and voicing are efficiently discriminated through acoustic analysis, differently from place cues which are easily distorted by noise. The opposite situation happens in the visual domain, as shown in figure B.2, where place is recognized far more easily than voicing and nasality.

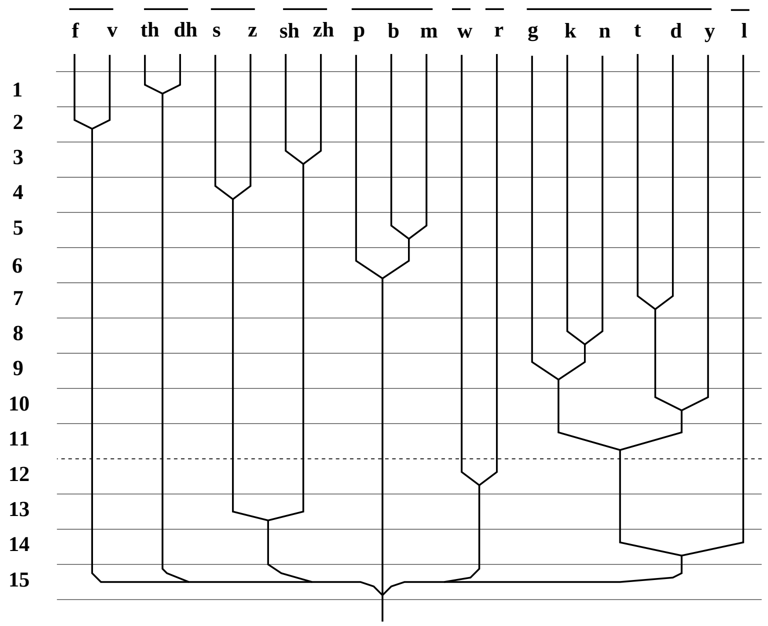

Figure B.2: *Visual confusion of consonant transitions CV in white noise among adult hearing impaired persons. By decreasing the Signal to Noise Ratio consonants, which are initially discriminated, are progressively confused and clustered. When the 11-th cluster is formed (dashed line), the resulting 9 groups of consonants can be considered distinct visemes. (From B. Dodd, R. Campbell, :'"Hearing by eye: the psychology of lipreading", Lawrence Erlbaum Ass. Publ.)*

Place cues are associated, in fact, to mid-high frequencies (above 1 kHz) which are usually scarcely discriminated in most of hearing disorders, on the contrary of nasality and voicing which reside in the lower part of the frequency spectrum. Cues of place, moreover, are characterized by short-time fine spectral structure requiring high frequency and temporal resolution, differently from voicing and nasality cues which are mostly associated to unstructured power distribution over several tens of milliseconds.

In any case, seeing the face of the speaker is evidently of great advantage to speech comprehension and almost necessary in presence of noise or hearing impairments: vision directs the auditor attention, adds redundancy to the signal and provides evidence of those cues which would be irreversibly masked by noise.

In normal verbal communication, the analysis and comprehension of the various articulation movements relies on a bimodal perceptive mechanism for the continuous integration of coherent visual and acoustic stimuli [331] In case of impairments in the acoustic channel, due to distance, noisy environments, transparent barriers like a pane of glass, or to pathologies, the prevalent perceptive task is consequently performed through the visual modality. In this case, only the movements and the expressions of the visible articulatory organs are exploited for comprehension: vertical and horizontal lips opening, vertical jaw displacement, teeth visibility, tongue position and other minor indicators like cheeks inflation and nose contractions.

Results from experimental phonetics show that hearing impaired people behave differently from normal hearing people in lipreading [94, 254]. In particular, visemes like bilabial /b, p, m/, fricative /f, v/ and occlusive consonants /t, d/ are recognized by each of them, while other visemes like /k, g/ are recognized only by hearing impaired people. The occurrence of correct recognition for each viseme is also different between normal and hearing impaired people: as an example, hearing impaired people recognize much more successfully nasal consonants /m, n/ than normal hearing people. These two specific differences in phoneme recognition can be hardly explained since velum, which is the primary articulator involved in phonemes like /k, g/ or /m, n/, is not visible and its movements cannot be perceived in lipreading. A possible explanation, stemming from recent results in experimental phonetics, relies on the exploitation of secondary articulation indicators commonly unnoticed by the normal observer.

If a lipreadable visual synthetic output must be provided through the automatic analysis of continuous speech, much attention must be paid to the definition of suitable indicators, capable to describe the visually relevant articulation places (labial, dental and alveolar) with the least residual of ambiguity [205, 204, 206]. This methodological consideration has been taken into account in the proposed technique by extending the analysis-synthesis region of interest also to the region around the lips, including cheeks and nose.

B.1.1 Articulatory description

Besides the main purpose of describing and quantizing the characteristics of the articulatory modes and places within the production of each single phone in different languages [149, 69, 167] and providing fruitful comparison with conventional articulatory classification [168, 169] the use of modern and sophisticated experimental techniques allows the definition of a far more accurate taxonomy of the articulatory parameters. Among the many techniques those which are providing the most relevant hints to experimental phonetics are X-ray kinematography, X-ray kymography, electro-palategraphy, electrokymography, labiography.

When articulatory movements are correlated with their corresponding acoustic output, the task of associating each phonetic segment to a specific articulatory segment becomes a critical problem. Differently from a pure spectral analysis of speech where phonetic units exhibit an intellegible structure and can be consequently segmented, the articulatory analysis does not provide, by its own, any univoque indication on how to perform such segmentation.

A few fundamental aspects of speech bimodality have inspired, since a fairly long time, interdisciplinary studies in neurology [173, 49, 225], physiology [307], psychology [361, 106], and linguistics [248].

Experimental phonetics has demonstrated that, besides speed and precision in reaching the phonetic target (that is the articulatory configuration corresponding to a phoneme), speech exhibits high variability due to multiple factors such as:

- psychological factors (emotions, attitudes);
- linguistic factors (style, speed, emphasis);
- articulatory compensation;
- intra-segmental factors;
- inter-segmental factors;
- intra-articulatory factors;
- inter-articulatory factors;
- coarticulatory factors.

To give a idea of the interaction complexity among the many speech components, it must be noticed that emotions with high psychological activity increase automatically the speed of speech production, that high speed usually determins articulatory reduction (Hipo-speech) and that a clear emphasized articulation is produced (Hyper-speech) in case of particular communication needs.

Articulatory compensation takes effect when a phono-articulatory organ is working under unusual constraints: as an example when someone speaks while he is eating or with the sigarette between his lips [200, 104, 1].

Intra-segmental variability indicates the variety of articulatory configurations which correspond to the production of the same phonetic segment, in the same context and by the same speaker. Inter-segmental variability, on the other hand, indicates the interaction between adjacent phonetic segments [193, 119] and can be expressed in "space", like a variation of the articulatory place, or in "time", meaning the extension of the characteristics of a phone to the following ones.

Intra-articulatory effects take effect when the same articulator is involved in the production of all the segments within the phonetic sequence. Inter-articulatory effects indicate the interdependencies between independent articulators involved in the production of adjacent segments within the same phonetic sequence.

Coarticulatory effects indicate the variation, in direction and extension, of the articulators movements during a phonetic transition [19, 134, 155]. Forward coarticulation takes effect when the articulatory characteristics of a segment to follow are anticipated to previous segments, while backward coarticulation happens when the articulatory characteristics of a segment are maintained and extended to following segments. Coarticulation is considered "strong" when two adjacent segments correspond to a visible articulatory discontinuity [110] or "smooth" when the articulatory activity proceeds smoothly between two adjacent segments.

The coarticulation phenomenon represents a major obstacle in lipreading as well as in artificial articulatory synthesis when the movements of the lips must be reconstructed from the acoustic analysis of speech, since there is no strict correspondence between phonemes and visemes [15]. The basic characteristic of these phenomena is the non-linearity between the semantics of the pronounced speech (despite the particular acoustic unit taken as reference), and the geometry of the vocal tract (representative of the status of each articulatory organ). It is apparent that speech segmentation cannot be performed by means of the only articulatory analysis. Articulators start and complete their trajectories asynchronously exhibiting both forward and backward coarticulation with respect to the speech wave.

Many theories exist and many cognitive models have been proposed on speech production and perception making reference to the different concepts of spatial acoustic or orosensorial target. Spatial target refers to the theoretical approach to speech production based on the hypothesis that speech is produced by adaptating the articulators toward a final target configuration, through an embedded spatial reference system. Acoustic target, as opposed to spatial target, implies that speech is produced by the adaptating the articulators in order to reach a final acoustic target [245]. Orosensorial target, in contrast with the previous two theories, explains speech production in terms of tactile feedbacks coming from receptors located at the mouth premises, used to control the articulatory adaptation [324, 265, 266].

Three basic different models are currently adopted, namely the closed loop model, the *open loop model* and the *coordinative model*. The closed loop model stems from Cybernetics (Cartesio machine) and considers the sensorial feedback (tactile and/or acoustic) to be the intrinsec control of the articulatory activities during speech production. The articulatory trajectories are planned on-line depending on the measured feed-backs until the target (spatial or acoustic) is reached. The open loop model stems from Computer Science and Artificial Intelligence (Turing machine) and is based on the metaphor "brain like a computer". It assumes that each articulatory activity is controlled through a sequence of instructions [106]. The articulatory trajectories are pre-programmed (off-line) point by point and are executed deterministically. The coordinative model [307, 106, 163], is based on the neurophysiology metaphor. It is in contrast to cybernetic and algorithmical models. It is based on coordinative structures, that is functional groups of muscles behaving like coordinated units according to rules learnt by training. The articulatory trajectories are the result of the interactive work of coordinative structures aimed at reaching a "qualitative" target in a flexible and adaptive way.

Independently from the particular approach, a very large set of signal observations is necessary to faithfully estimate the articulatory parameters. Classic methods generally aim at an accurate phoneme recognition and at a consequent synthesis by rule to associate the corresponding viseme. In [176, 270], two estimation algorithms have been proposed for the statistical analysis of the cepstrum space for detecting phonetic clusters, representative of stable configurations of the articulatory organs, and for modeling the transition paths linking clusters, representative of coarticulation.

B.1.2 Articulatory synthesis

In face-to-face verbal communication, since the analysis and comprehension of the various articulatory movements relies on a bimodal perceptive mechanism for the continuous integration of coherent visual and acoustic stimula [331], the most natural way for visualizing the cues of speech on a a screen is that of synthesizing speech on a "talking human face". In this way a message from the computer (output) can be represented according to a bimodal model and visualized through a multimedia interface.

Worldwide interest is recently increasing around the possibility of employing anatomic models of a human face, suitable to synthetic animation through speech, for applications like the realization of multimedia man-machine interfaces, new algorithms for very low bitrate model-based videophony, advanced applications in the fields of education, culture and entertainment.

Different approaches have been investigated and proposed to define the basic set of parameters, necessary to model human mouth expressions including lips, teeth and tongue movements. In [38] 14 elementary visual units, responsible of the mouth shape, have been defined and associated to a corresponding acoustic class. Through suitable phonemic classification, a consistent sequence of visemes is obtained to visualize stable acoustic frames, while coarticulation transients are approximated by interpolation. In [139] a phonemic classifier is employed to determine a basic set of mouth parameters through sinthesis by rule, while in [231] the intermediate stage of phonemic classification is bypassed and input acoustic parameters are straightforward associated to suitable output visual parameters. All these approches, however, are unsuited to specific applications where a very faithful reproduction of the labial expressions is required such as those oriented to synthetic lipreading. The main problems stem from the lack of a suitable mouth anatomical model and from the coarse representation of the coarticulation phenomena.

A recent approach [61] based on Time-Delay Neural Networks (TDNN) employs a novel fast learning algorithm capable to provide very accurate and realistic control of the mouth parameters. As proven by the experimental results, visible coarticulation parameters are correctly estimated by the network and the corresponding visualization is smooth enough to be comprehended by lipreading. In this work human faces have been modeled through a flexible wire-frame mask suitably adapted to their somatic characteristics with increasing resolution in correspondence to high detail features like eyes and mouth. This kind of wire-frame is generally referred to as $2D\frac{1}{2}$, meaning that it isn't neither a 2D nor a true 3D structure, but something in-between. A comparison can be done with a carnival mask reproducing the characteristics of a human face on a thin film of clay: the 3D information is incomplete since only one half of the head is modeled by the mask. The wire-frame structure is organized as a set of subregions, each of them affected by predefined deformation rules producing time-varying variations to simulate muscle contraction. Animation procedures have been designed on the basis of Ekman and Friese [88] psychological studies, formalizing the relationship between the subjective information content of the speaker's audio-visual communication and his corresponding facial expression. In particular, the set of basic animation units (AUs) defined in [5] for facial mimics description and coding has been implemented with reference the Parke model [257].

B.2 Audio-Visual Speech Synthesis from Text

B.2.1 Animation of synthetic faces

In the last two decades, a variety of synthetic faces have been designed all over the world with the objective of their animation. The quality of facial models goes from a simple electronic curve on an oscilloscope, through a wide range of pre-stored human face shapes and more or less caricatural 2D vector-driven models of the most salient human face contours, to a very natural rendering of a 3D model (by mapping of real photographs on it, for instance). I think that, as in the case of acoustic speech synthesis, one must differentiate the final model from its animation technique. I suggest the reader refer to [46] for a turorial presentation of the techniques and methods used in facial animation. To sum up, Figure B.3 gives a classification of the most noticeable publications of designed systems along these two criteria. For simplification purposes, the table does not consider the work done by investigators to develop or apply the models cited, nor to synchronise them with speech synthesizers. Of course, the control parameters of a rule-driven model may be given by a code-book, after analysis of the acoustic wave-form, so the Y-axis could have been presented differently. However, the table aims at showing the basic characteristics of the various approaches by assigning the most suitable animation technique to a given model.

Facial Model : / **Animation technique :**	**Lissajou-like** (Electronics)	**2D vectors or shapes**	**3D wire or raster-graphics**
from acoustics	Boston, 1973 Erber, 1978	Simons, 1990 *(through stochastic networks)*	Morishima, 1990 Lavagetto, 1994 *(through stochastic networks)*
by rules (control parameters)		Brooke, 1979 Guiard-Marigny, 1992	Parke, 1974 Platt, 1981 Waters, 1987 Magnenat-Thalmann,1988 Cohen, 1993 Guiard-Marigny, 1994
by code-books (quantization) **or key-frames** (interpolation)		Montgomery, 1982 Matsuoka, 1988 Woodward, 1991 Mohamadi, 1993	Parke, 1972 Bergeron, 1985 Aizawa, 1987 Naha, 1988

Figure B.3: *Classification of best-known face synthesizers based on the kind of facial model developed (X-axis), and on its underlying method of animation (Y-axis). Only the first author and date of publication are quoted.*

Whatever the facial model, it may be animated for speech by three main methods:

1. A direct mapping from acoustics to geometry creates a correspondence between the energy in a filtered bandwith of the speech signal and the voltage input of an oscilloscope, and therefore with the horizontal or vertical size of an elliptical image [37, 96]. An indirect mapping may also be achieved through a stochastic network which outputs a given image from a parametrization of the speech signal, after training of the network in order to make it match any possible input with its corresponding output as accurately as possible. Simons and Cox [313] used a Hidden Markov network, whereas Morishima et al. [230] preferred a connectionist network, for instance. In both, the inputs of the networks were LPC parameters calculated from the speech signal.

2. When the facial model has previously been designed so that it can be animated through control parameters, it is possible to elaborate rules which simulate the gestural movements of the face. These commands can be based on a geometrical parametrization, such as jaw rotation or mouth width for instance, in the case of the 2D models by Brooke [45] or Guiard-Marigny [132], or the 3D models by Parke [258], Cohen and Massaro [71], or Guiard-Marigny et al. [133]. The commands can also be based on an anatomical description, such as the muscle actions simulated in the 3D models by Platt and Badler [273], Waters [355], Magnenat-Thalmann et al. [202], or Viaud [345].

3. Facial models may only mimic a closed set of human expressions, whatever the tool used to create them:

   - a set of 2D vectors [227, 366],
   - simplified photos [16]; (Mohamadi, 1993),
   - hand-drawn mouth-shapes [215]),
   - 3D reconstructions of 2D images [256]),
   - 3D digitizing of a mannequin using a laser scanner [203, 238, 160, 166], or even
   - direct computer-assisted sculpting [255].

   If no control parameters can be applied to the structure obtained in order to deform it and so generate different expressions, and if digitizing of multiple expressions is impossible, hand-modification by an expert is necessary in order to create a set of relevant key-frame images. The pre-stored images can then be concatenated as in cartoons, so that a skilled animator may achieve coherent animation. Such a technique has been widely employed, since it only requires a superficial model of the face, and as little physiological knowledge of speech production is needed to modify the external texture so that natural images can be duplicated (rotoscoping technique).

I also want to mention two other techniques that directly rely on human gestures:

4. Expression slaving consists of the automatic measurement of geometrical characteristics (reference points or anatomical measurements) of a speakers face, which are mapped onto a facial model for local deformation [18, 337, 363, 259, 177]. In the computed-generated animation *The Audition* [224], the facial expressions of a synthetic dog are mapped onto its facial model from natural deformations automatically extracted from a human speakers face [259].

5. Attempts have also been made to map expressions to a facial model from control commands driven in real-time by the hand of a skilled puppeteer [79]. In a famous computer-generated French TV series, *Canaille Peluche*[1], the body and face gestures of Mat, the synthetic ghost, are so created. Finally, a deformation that is rather based on the texture than on the shape of the model can be achieved by a mapping of various natural photos on the model [166].

The basic principles of these various animation techniques are represented in figure B.4.

B.2.2 Audio-visual speech synthesis

Experiments on natural speech (see ICP-MIAMI report 94-1) allow us to anticipate that similar effects will be obtained with a TtAVS synthesizer: Even if the current quality of (most) TtS systems is not as bad as highly degraded speech, it is obvious that under very quiet conditions, synthesizers are much less intelligible than humans. Moreover, it is realistic to predict that in the near future, the spread of speech synthesizers will lead to wide use in noisy backgrounds, such as in railway stations. Such adverse conditions will necessitate a synchronized presentation of the information from another modality, for instance, the orthographic display of the text, or the animation of a synthetic face (especially for foreigners and illitirates). There are hence several reasons for the study and use of Audio-Visual Speech Synthesis.

- Audio-Visual Speech Synthesis allows investigators to accurately control stimuli for perceptual tests on bimodality:

[1] The series is produced by Canal Plus and Videosystem

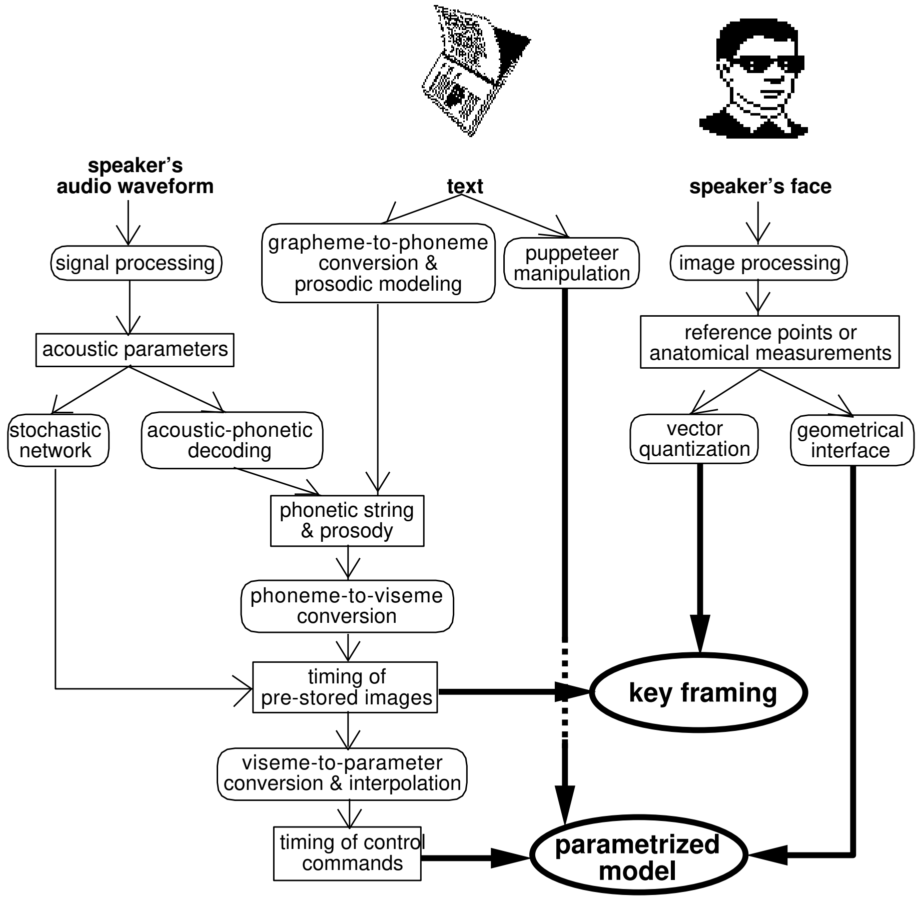

Figure B.4: *General architecture of Facial Animation showing the possible techniques.*

Massaro and Cohen [212] studied how speech perception is influenced by information presented to ear and eye by dubbing acoustic tokens generated by a speech synthesizer (in [158]) onto a sequence of images generated by a facial model (see [258]), as modified by Pearce et al. [261]). Highly controlled synthetic stimuli thus allowed them to investigate into details the McGurk effect.

- Audio-Visual Speech Synthesis is also a tool for basic research on speech production:

  Pelachaud studied the relationship between intonation and facial expression by means of natural speech [262] and the facial model developed by Platt [272].

- Thanks to the increasing capacities of computer graphics, highly natural (or hyper-realistic) rendering of 3D synthetic faces now allows movie producers to create synthetic actors whose facial gestures have to be coherent with their acoustic production, due to their human-like quality and the demands of the audience. Short computer-generated movies clearly show this new trend:

Tony de Peltrie [18]), *Rendez-vous Montral* [203]); *Sextone for President* [159]); *Tin Toy* [285]); *Bureaucrat* [356]); *Hi Fi Mike* [370]); and *Don Quichotte* [117]), among others.

- It is hence necessary for computer-assisted artists to be equiped with software facilities so that the facial gestures and expressions of their characters are easily, quickly, and automatically generated in a coherent manner. Several attempts to synchronize synthetic faces with acoustic (natural or synthetic) speech may be found in literature [186, 140, 177, 215, 238, 70, 230, 263, 367, 226], among others, for (British & American) English, Japanese, and French.
- Few audio-visual synthesizers have ever been integrated into a text-to-speech system [367, 226, 71, 138].

Unfortunately, most of the authors only reported informal impressions from colleagues about the quality of their system, but as far as I am aware none of them has ever quantified the improvement in intelligibility given by adding visual synthesis to the acoustic waveflow. I strongly support the idea that assessment methodologies should be standardized so that the various approaches can be compared to one another. Next report will present results of intelligibility tests run at the ICP with various visual (natural and synthetic) displays of the lips, the jaw and the face under different condition of background noise added to the original acoustic signal.

Appendix C

Audio-Visual Speech Recognition

Automatic speech recognition (ASR) promises to be of great importance in human-machine interfaces, but despite extensive effort over decades, acoustic-based recognition systems remain too inaccurate for the vast majority of conceivable applications, especially those in noisy environments (automobiles, factory floors, crowded offices, etc.). While incremental advances may be expected along the current ASR paradigm, additional, novel approaches — in particular those utilizing visual information as well — deserve serious study. Such hybrid (acoustic and visual) ASR systems have already been shown to have superior recognition accuracy, especially in noisy conditions, just as humans recognize speech better when also given visual information (the "cocktail party effect").

C.1 Integration Models of Audio-Visual Speech by Humans

The objective of this section is to present existing models of integration specifically proposed for speech perception. It is introduced by a discussion on how fusion of sensory captors relies basically on two disciplines: Cognitive Psychology and Engineering Sciences. This report finally proposes a taxonomy of the integration models that can be found in the literature. Later on, we will make another report on the quantitative implementation of such models. Here again, those techniques come from two scientific domains, namely *Cognitive Psychology* and *Automatic Speech Recognition*.

C.1.1 General principles for integration

Multimodal perception is not specific to speech, nor to audition/vision. Information transfer and intermodal integration are classical problems widely studied in psychology and neuropsychology [136, 137, 358, 135, 321]. Therefore, interactions between various modalities have been studied in literature:

- vision and audition [222, 358, 137]
- vision and tactile [284, 136, 137, 358];
- vision and grasping [137]
- vision and proprioception [358, 137]
- vision and infrared perception [135]
- audition and grasping [137]
- audition and proprioception [358]

From literature (see [137], for a review), three main classes of intersensory fusion models can be considered:

1. use of language (or of symbolic attributes) as a basis for fusion,
2. recoding from a modality to an other (considered as dominant for a given perceptual task), and
3. use of amodal continuous representations (which are characteristic of the physical properties of the distal source, independently of the proximal sensations, see e. g., the theory of direct realistic perception by Gibson [121].

We below detail those three cases.

In the first case, information is categorized in symbols which provide the interaction with a support. Classification thus precedes fusion: Such models can be called as late integration [348]. In order for the transfer between modalities to be done at a symbolic level, the subject is supposed to have learned the symbols. Since infants and animals are capable of visual and tactile interaction, the hypothesis of intermodal transfer through language is obsolete for visual-tactile perceptions [136].

The second category assumes that one modality dominates the other for a given task. For instance, it is considered that vision is dominant for spatial localization [281] and that audition is dominant for judgments dealing with the temporal organization of stimuli [137]. The non-dominant (or submissive?) modality is then recoded into the dominant one. It is only after this recoding process that the interaction (transfer, fusion) take place. According to Hatwell [136, 137], transfer from vision to the tactile modality must be explained by a model which falls into this category:

> "Spatial perceptions coming from the dominant modality (vision) remain in their own code, whereas those coming from the non-dominant modality (tactile) are, as best as possible, recoded into the dominant modality and adjusted to it".
>
> [136, page 340]

In the third case, it is considered that the physical signal is used by the sensory system in order to retrieve information about the source responsible for its generation. Sensations are thus transcoded into an amodal space before to be integrated. This amodal space might be isomorphic to the physical properties of the source, as assumed in the *Theory of Direct Perception* (cf. Fowler [108]). It might also be made of internal dimensions that are independent from the modality, such as intensity, duration, etc. It is hence assumed that there is no direct transfer between modalities, but that each sensation is projected upon a same amodal space where interactions then take place.

C.1.2 Five models of audio-visual integration in speech perception

In the area of speech perception, several models have been proposed to account for speech recognition in noise or for the perception of paradoxical stimuli [52, 208, 209, 53, 292, 348]. However, Summerfield addressed the main theoretical issues and discussed them in the light of experimental data and of general theories of speech perception [332]. He described five different architectural structures that may solve the problem of audio-visual fusion in the perception of speech. We here revisit those five models in order to better fit them into the general principles presented in the above section. The proposed structures are presented in the following sections. They are finally put into relation with the general principles.

Direct Identification Model (DI) The Direct Identification model of audio-visual configurations is based upon the *Lexical Access From Spectra* by Klatt [157]. It has been turned into a *Lexical Access from Spectra and Face Parameters*. The input signals are here directly transmitted to the bimodal classifier. This classifier may be, for instance, a bimodal (or bivectorial) lexicon in which the prototype the closest to the input is selected (see figure C.1).

Basics of the model: There is no common representation level over the two modalities between the signal and the percept.

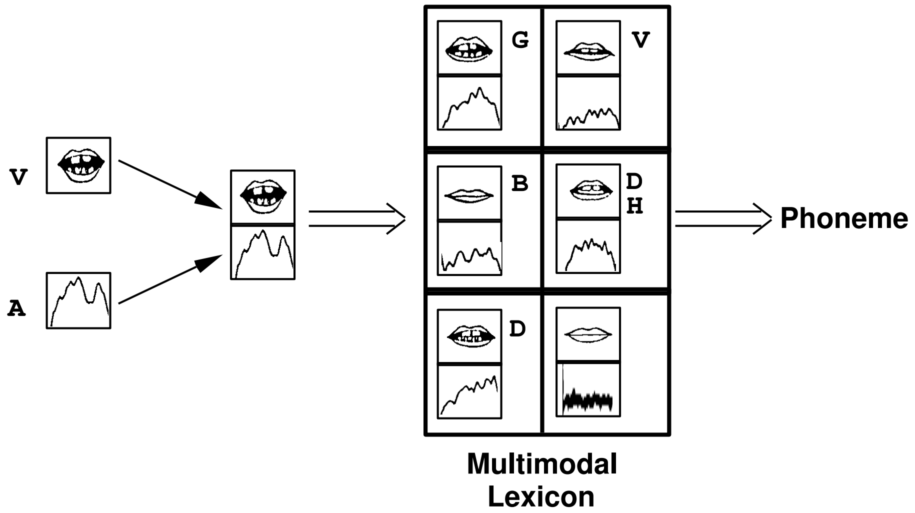

Figure C.1: *Schematic of the Direct Identification (DI) model*

Separated Identification Model (SI) In the Separated Identification model, the visual and the auditory input are separately identified through two parallel identification processes. Fusion of the phonemes or phonetic features across each modality occurs after this identification. Fusion can be processed with logical data, such as in the VPAM model (Vision Place Auditory Mode) where each modality deals with a set of phonetic features [220, 332]. Thus, the place of articulation of the output is that of the visual input and the mode of articulation (voiced, nasal, etc.) is that of the auditory input (see figure C.2).

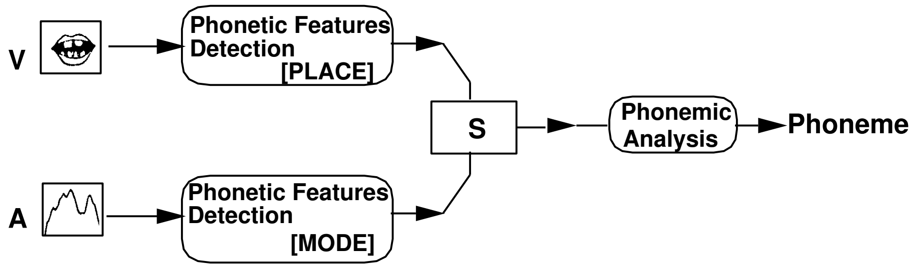

Figure C.2: *Schematic of the VPAM model (example of an SI structure) (after [332])*

Fusion can also be processed with probabilistic data (fuzzy logic). Each input can be matched against (unimodal) prototypes in order to get two scores for each category.

Then, each pair of data corresponding to the same category is fused through probabilistic computation. The FLMP (Fuzzy Logical Model of Perception) falls into this class of models (see [208, 52]).

Basics of the model: The inputs are matched against prototypes (or even classified) before being integrated.

Dominant Modality Recoding Model (RD) The model of Recoding into the Dominant Modality sees the auditory modality as a dominant modality in the perception of speech since it is better suited to it. The visual input is thus recoded into a representation of this dominant modality. The selected representation of the auditory modality is the transfer function of the vocal tract. This transfer function is separately evaluated from the auditory input (e. g., through cepstral processing) and from the visual input (through association). Both evaluations are then fused. The source characteristics (voiced, nasal, etc.) are evaluated only from the auditory information. The so evaluated pair source / vocal tract is then sent to the phonemic classifier (see figure C.3).

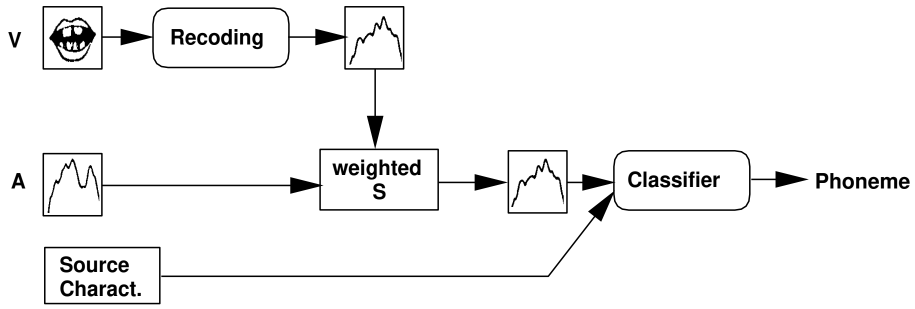

Figure C.3: *Schematic of the Dominant Modality Recoding (RD) model*

Basics of the model: The visual modality is recoded into an auditory metric. Both representations are then integrated into the auditory space where the categorization takes place.

Motor Space Recoding Model (MR) In the Motor Space Recoding Model, the two inputs are projected onto an amodal space (which is neither auditory nor visual). They are then fused within this common space. This amodal space is the articulatory space of vocal tract configurations. The visual modality is projected only on the vocal tract space dimensions where to information can be carried. For instance, the visual modality may bring information on lip rounding, not on the velum position. The final representation depends on two possibilities. when there is a single projection on a dimension (from the auditory input), this projection makes the final decision.

When information from the two modalities is simultaneously projected upon one dimension (e. g. jaw height), the final value comes from a weighted sum of both inputs. The final representation thus obtained is given by the phonemic classifier. This architecture fits well the *Motor Theory of Speech Perception* [187]) and the *Realistic Direct Theory of Speech Perception* [108].

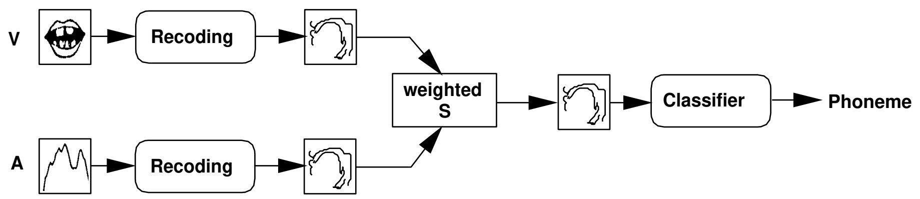

Figure C.4: *Schematic of the Motor Space Recoding (MR) model*

Basics of the model: The two modalities are projected upon a common motor space where they are integrated before final categorization.

Model of Motor Dynamics This last model is the one least developed by Summerfield. Each modality gives an estimation of the motor dynamics (mass and force) of the vocal tract. Such an estimation might be based upon kinematics data (position, velocity and acceleration). The two pieces of estimation (from each modality) of motor dynamics are fused and then categorized, describing the identity of phonemes in terms of time-varying functions rather than articulatory states traversed [332, page 44]. The important difference between this model and the previous one relies in the nature of the motor representation, as seen as the projection and integration of the sensory inputs. In fact, the previous model could be given the name AR (Articulatory Recoding) whereas the current model could be entitled MR in itself.

However, introducing such a distinction which is relevant to deal with theories of speech perception is somewhat disputable when dealing with architectures of audio-visual fusion. We could thus think of other distinctions that make use of the various propositions found in literature on the nature of the preprocessing at the input or of the level of the linguistic units at the output. This is why we prefer to take only in consideration the architecture called MR, as described in figure C.4, since it can regroup "static" as well as "dynamic" representations of the motor configurations.

C.1.3 Conclusion

In connection with the general theories of multimodal integration (see section C.1.1), the DI and SI models fall into the first category of models in which fusion occurs at the symbolic level. The RD model falls obviously into the second category since the visual modality is recoded into the auditory modality (which is assumed to be dominant). Finally, the MR model falls into the third category of neo-Gibsonian models.

In connection with the principles of sensors integration proposed by Crowley and Demazeau [75], it can be noticed that the first three principles (common lexicon and coordinates system) cannot be avoided. However, the fourth and the fifth principle (uncertainty and confidence) are seldom considered, even though they are somehow in the essence of Massaro's FLMP. The four models are schematically represented in figure C.5.

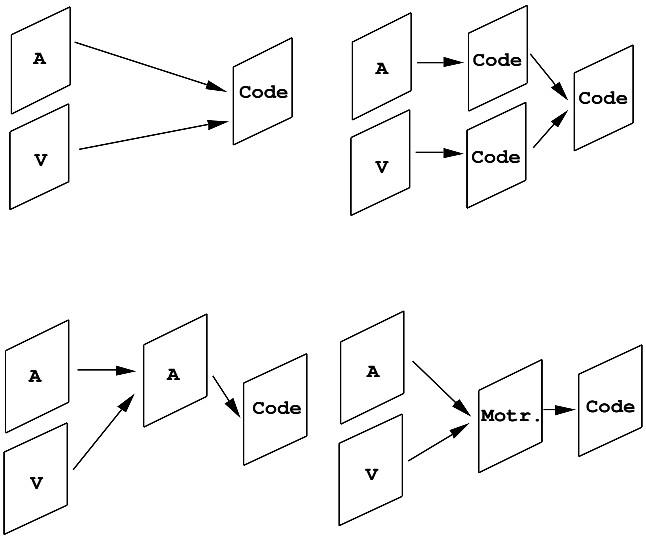

Figure C.5: *Schematic of the four integration models*

Several variations of these models could be developed. Continuous links from a model to the other could even be elaborated, as it is the case in the modeling area, so that differences would be smoothed at the same time as the structures would become more complex. On the opposite, and in a spirit of methodological control, we will later on fix the architectures in their "canonical" form, so that the different options can be clearly compared.

C.1.4 Taxonomy of the integration models

We will here make an attempt to classify the above presented architectures under a synthetic form and to rely them to the general models of cognitive psychology. To do so, we will organize the architectures through three questions:

1. *Does the interaction between modalities imply or not a common intermediate representation?*

The answer allows a first opposition between the DI model (for which it is NO) an the other architectures.

2. *If there is an intermediate representation, does it rely on the existence of prototypes or of early classification processes, i. e., at a symbolic level? In other words, is it a late or an early integration (see [348, page25])?*

   We will say that integration is late when it occurs after the decoding processes (SI case), even though these processes give continuous data (as with the FLMP). Otherwise, integration is early when it applies to continuous representations, common to the two modalities, and which are obtained through low-level mechanisms which do not rely on any decoding process: It is the case with the RD and MR models.

3. *Is there at last any dominant modality which can give a common intermediate representation in an early integration model (RD)? Or, is this common representation amodal (such as in the MR model)?*

   Given this, it leads to the taxonomy presented in figure C.6.

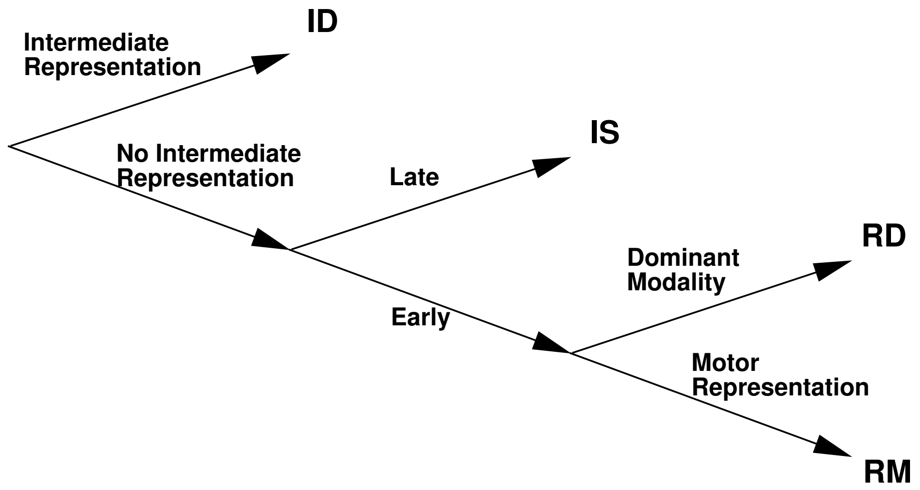

Figure C.6: *Taxonomy of the integration models*

C.2 Audio-Visual Speech Recognition by Machines

In this section, we describe some existing automatic speech recognition systems based mainly on optical and acoustical information flows. We will finally present some results obtained with the automatic audio-visual speech recognizer implemented at ICP.

C.2.1 Audio-visual speech perception by humans

Research with human subjects has shown that visible information of the talker's face provides extensive benefit to speech recognition in difficult listening conditions [330, 21]. Studies have shown that visual information from a speaker's face is integrated with auditory information during phonetic perception. The McGurk effect demonstrates that an auditory /ba/ presented with a video /ga/ produces the perception of /da/ [220]. It indicates that the perceived place of articulation can be influenced by the visual cues. Other researchers have shown that bilabials and dentals are more easily perceived visually than alveolar and palatals [240]. All these experiments have demonstrated that speech perception is bimodal for all normal hearers, not just for profoundly deafs, as formally noticed by Cotton in 1935 [73]. For more an extended review of the intelligibility of audio-visual speech by humans, see the ICP-MIAMI 94-1 report by Benoit. For detailed information on the integration of such auditory and visual information by humans, see the ICP-MIAMI 94-2 report by Robert-Ribes.

C.2.2 Automatic visual speech recognition

The visible facial articulations cues can be exploited to attempt automatic visual speech recognition, which could enhance the performance of conventional, acoustical speech recognizers, especially when the acoustic signal is degraded by noise. In this section, we present some prototypical systems, processing the optical flow as a video input.

The optical automatic speech recognition system developed by Brooke and Petajan in 1986 [47] was based on the magnitude of the radial vector changes measured from the midpoint of a line between the inner lip corners to points on the inner line of the lips at equal angular intervals, with the zero angle at the left corner. The objective of this research was to discover an *Optical Automatic Speech Recognition* algorithm using this radial vector measure. The recognition experiment consisted of distinguishing between three vowels /a/, /i/, /u/ in consonant- vowel-consonant utterances using a similarity metric of the radial vector measure with two speakers. The results show that, if the threshold of similarity is held sufficiently high and consonant coarticulation effects are non-existent or neglected, a perfect vowel recognition is achieved on the three vowels considered.

Pentland has developed an optical speech recognition system for isolated digit recognition [264]. He used optical flow techniques to capture the movement of muscles in four regions of the lower facial area, as evidenced by the changes in muscle shadows under normal illumination. The recognition rate was 70% for ten digits. Smith used optical information from the derivatives of the area and the height of oral-cavity forms to distinguish among four words that an acoustic automatic speech recognizer had confused [316]. Using the two derivatives, he managed to distinguish perfectly among the four acoustically confused words. Goldschen built an optical speech recognizer for continuous speech [126]. The system used a probabilistic approach based on Hidden Markov Models (HMMs). The database consisted of sequences of binary images of contours or edges depicting the nostrils. The recognition rate was 25% of complete sentences.

C.2.3 Automatic audio-visual speech recognition

Although acoustically-based automatic speech recognition system have witnessed enormous improvements over the past ten years, they still experience difficulties in several areas, including operation in noisy environments. However using visual cues can enhance significantly the recognition rate. In this section, we provide an overall survey on the existing audio-visual speech recognizers. The first audio-visual speech recognition system using oral-cavity region was developed by Petajan [267].

The primary focus of his research was demonstrating that an existing acoustic automatic speech recognizer could achieve a greater recognition percentage when augmented with information from the oral-cavity region. This system used a single speaker to perform isolated-word recognition. Petajan used a commercial acoustic automatic speech recognizer and built the audio-visual recognition system and utilized four (static) features from the oral-cavity region of each image frame in decreasing order by weight: height, area, width, and perimeter. The commercial acoustic automatic speech recognizer alone had a recognition rate of 65%, nearly typical of the technology commercially available at that time, however the audio-visual automatic speech recognizer achieved a recognition rate of 78%.

He further expanded his research building an improved version of the recognizer. In attempting to approximate real-time performance, the optical processor capturing the images utilized fast contour and region coding algorithms based on an extended variation of predictive differential quantization [143]. The optical automatic speech recognizer utilized vector quantization to build a codebook of oral-cavity shadows. This new system had a faster performance than the earlier described while using only the area features of the oral-cavity region. The entire system was tested for speaker-dependent experiments on isolated words. On two experiments, optical identification is 100% on 16 samples of each digit for one speaker and 97.5% on eight samples of each digit for another.

Nishida, while working at the MIT Media Laboratory, used optical information from the oral-cavity to find word boundaries for an acoustic automatic speech recognizer [244]. His works focused on the derivative of dark areas as obtained from the difference between two consecutive binary images. Yuhas, in his doctoral research in 1989, used a neural network and both optical and acoustic information to recognize nine phonemes [374]. The goal of his research was to accurately recognize the phonemes from optical information, independently of the acoustic noise level. As the acoustic recognition degraded with noise, the optical system maintained its performance.

Stork et al. used a time delay neural network (TDNN) to achieve a speaker-independent recognition of ten phoneme utterances [327]. The optical information consisted of four distances: (i) the distance between the nose and chin (lowering of the jaw), (ii) the height at the center of the upper and lower lips, (iii) the height at the corner of the lips, and (iv) the width corner of the lips. The TDNN achieved an optical recognition accuracy of 51% and a combined acoustic and optical recognition of 91%. Bregler used the TDNN as a preprocessor to the dynamic time warping algorithm to recognize 26 German letters in connected speech [41]. The recognition rate for the combined acoustical and optical information was 97.2% which was slightly greater than the separate acoustic (97.2%) and optical (46.9%) recognition rates, but improved more significantly in noisy environments.

C.2.4 Current results obtained at ICP

Experimental paradigm

The main purpose of our research is to build an automatic audio-visual speech recognizer. It is important to note that the corpus used is limited compared to other databases usually used to test speech recognizers. As we will explain further in this report, the system has to identify an item out of only 18 candidates, so that the test conditions are rather simple when compared to a recognition problem. Nevertheless, our goal is to study the benefit of an automatic lipreading on automatic speech recognition. In that sense, any training/test paradigm will be as informative as another. This is why we have first chosen a simple procedure.

Technique used

A probabilistic approach, i. e. the HMM is used as the main classifier. The system has first been developed using an off-line analysis of the lips [171]. In order to extract the necessary visual information from the talker's face, the speaker's lips are made up in blue. A chroma-key turns it on-line into a saturated black color. Therefore the vermilion area of the lips is easily detected by thresholding the luminance of the video signal. Finally, these images are used as the input of an image processing software which measures the parameters used for the visual recognition system. Figure C.7 gives an example of the input of our image processing tool (left panel), and the measured parameters (right panel).

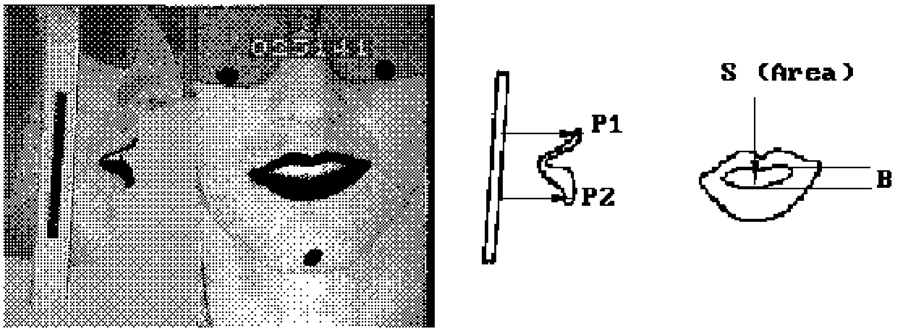

Figure C.7: *An example of an image frame and the visual parameters*

The parameters are extracted from each video field (50 fields in a one second sequence). They are stored in a file used as the main "optical data" input of our visual speech recognizer. The audio input is a vector composed of a total of 26 elements such as Mel-frequency cepstral coefficients, energy, delta coefficients. In order to have a complete synchronization between audio and visual information cues, a 20 msec audio analysis window has been used. Therefore, visual and audio vectors are in synchrony during all the process. As a first step towards building an audio-visual speech recognizer, we followed the architecture of the Direct Identification Model (see MIAMI report 94-2 by Robert-Ribes). The video parameters are appended to each audio vector. Several tests have been run in order to choose the best combination of visual parameters. Finally, only four parameters have been selected as the input of the speech recognition system. They are presented in figure C.7. The corpus is made of 18 non-sense words of the form VCVCV. Each word was uttered nine times so that the audio database consists of 162 words. In all the following results, we divided the corpus into a training set made of 18x7 utterances and a test set made of the remaining 18x2 utterances. The HMM is composed of 9 states. For each state, the transition and the output probabilities and other parameters are computed using forward-backward algorithm and Baum-Welch re-estimation.

Preliminary results

Two different types of training have been performed in order to check the efficiency of the audio only and the audio-visual speech recognizer. We first trained the HMMs using clear acoustic data, whatever the S/N in the test condition . Second, we trained the HMMs onto corrupted acoustic data by adding a white noise whose intensity N corresponded to that of each S/N condition during the test.

Results with clear acoustic training In this case, training process is run with clear audio data and the system is tested with degraded audio data. First, the acoustic recognizer is tested and second, visual parameters are added to the input. Figure C.8 summarizes these results. The horizontal line represents the score obtained using only the four visual parameters (V). The bottom curve represents the scores obtained with auditory only data during the training and the test session (A). The top curve represents the scores obtained with audio-visual data during the training and the test session (AV). The audio-visual recognition rate (AV) is lower than the recognition rate using visual parameters only (A) in the left area of the vertical dashed line. This might be due to the unbalanced weight of the audio and visual data in our HMM vectors: 26 A + 4 V parameters. In the right area only, a gain in the audio-visual recognition rate (AV) has been achieved compared to the visual alone (V). In all cases, AV scores are higher than A scores. These first results are unsatisfactory, since visual information is not processed at its best by the HMMs.

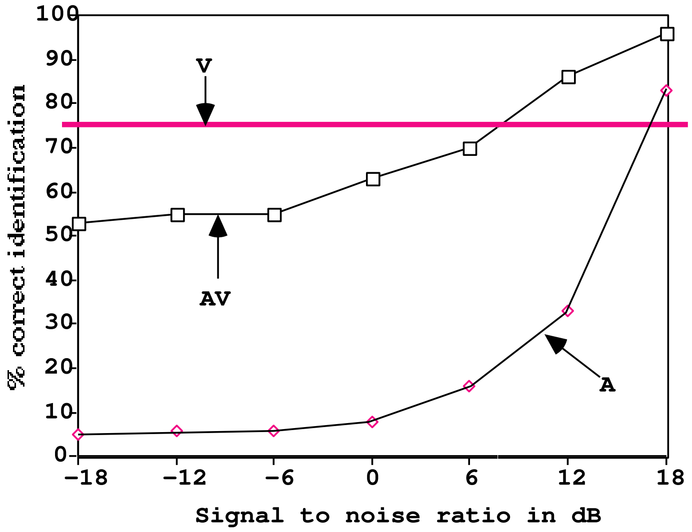

Figure C.8: *Test scores of the audio (A) and audio-visual (AV) recognizers after training the HMMs in clear acoustic conditions*

Results with degraded acoustic training

In this case, for each S/N condition, the HMMs are trained using auditory data degraded with an additive white noise of the same intensity as that used during the test session. Figure C.9 shows the results of this experiment. Scores are presented as in figure C.8. The AV score in the left area (1) is highly influenced by the noisy data from the auditory vector so that the resulting recognition rate is lower than the V score. In the center area (2), AV scores are higher than both A and V scores. In the right area (3), AV scores are similar or even lower than A scores. These last results confirm a well known fact that HMMs have to be trained with (even artificially) degraded data to improve the recognition scores when the test conditions are degraded.

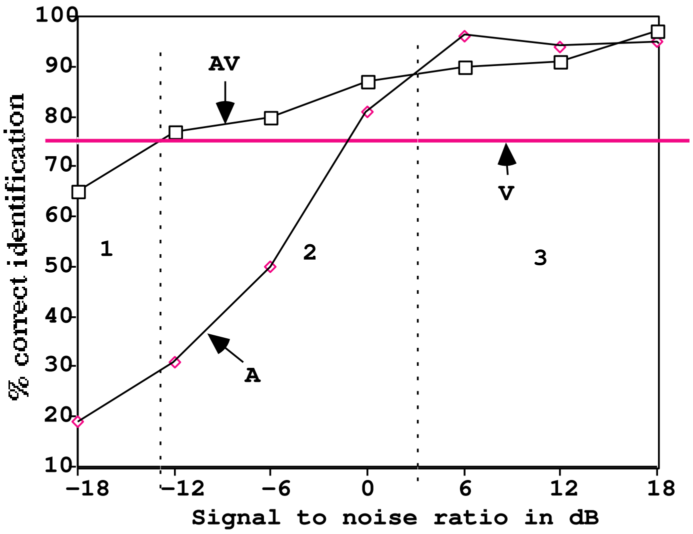

Figure C.9: *Test results obtained from the audio-visual (AV) and audio recognizer (A). System has been trained for each test condition.*

C.2.5 Forecast for future works

The first results seem to be a great step towards designing an audio-visual speech recognition system but many problems still have to be solved. First of all, in our model, weighting factors of visual and auditory parameters have been neglected; so that the system still relies on unbalanced information. Second, other models for audio-visual integration have to be implemented in order to select the best way of combining audio and visual information in our case. For a complete on-line audio-visual speech recognition, we have planned to design a video analysis board capable of extracting the lip contours and computing the four necessary visual parameters in a short delay (< 40 msec). Then it will be possible to collect a much larger amount of audio-visual data within a decent period of time. It should allow the audio-visual speech recognizer to run in real-time. This would then be an ideal interface for bimodal speech dialogue on a multimedia platform.

Appendix D

Gesture Taxonomies

[289] characterize several classes of gesture:

Symbolic: conventional, context-independent, and typically unambiguous expressions (e. g., “OK” or peace signs)

Deictic: gestures that point to entities, analogous to natural language deixis (e. g. “this is not that” or “put that there”)

Iconic: gestures that are used to display objects, spatial relations, and actions (e. g., illustrating the orientation of two robots at a collision scene)

Pantomimic: gestures that display an invisible object or tool (e. g., making a fist and moving to indicating a hammer)

D.1 Hand Gestures Taxonomy

[161] define a system able to parse 3D and time-varying gestures. Their system captures gesture features such as postures of the hand (straight, relaxed, closed), its motion (moving, stopped), and its orientation (up, down, left, right, forward, backward — derived from normal and longitudinal vectors from the palm). Over time, a stream is of gesture is them abstracted into more general *gestlets* (e. g., Pointing ⇒ attack, sweep, end reference). Similarly, low level eye tracking input was classified into classes of events (fixations, saccades, and blinks). They integrate these multimodal features in a hybrid representation architecture.

It should be noted that the term and concept of 'gesture' is also used in pen computing. Although here, the recorded action takes place in the 2D plane, similar phenomena play a role as in the case of 3D hand gesturing, but with a much easier signal processing involved. Other interesting references are [233, 287, 364, 341].

Appendix E

Two-dimensional Movement in Time: Handwriting, Drawing, and Pen Gestures

Handwriting and Drawing are two different means of human information storage and communication, produced by the same single two-dimensional output system: a pointed writing implement, usually driven by the arm, which leaves a visible trace on a flat surface. Handwriting conveys symbolical data, whereas Drawing conveys iconic data. The third data type, Pen Gestures, consists of unique symbols, as used by book editors and in pen computers, requiring a function to be executed. Contrary to speech, handwriting is not an innate neural function, and must be trained over several years. During the training process, handwriting evolves from a slow feedback process involving active attention and eye-hand coordination to a fast automatic and ballistic process. The atomic movement unit in handwriting is a stroke, which is a movement trajectory bounded by two points of high curvature and a corresponding dip in the tangential movement velocity. The typical modal stroke duration is of the order of 100 ms, and varies less for increased movement amplitudes than one would expect: For a large range, writers exert an increased force in order to maintain a preferred rhythm of movement (10 strokes/s, 5 Hz). Once "fired" cortically, such a stroke cannot be corrected by visual feedback. The handwriting process evolves in a continuous process of concurrent planning and execution, the planning process being 2–3 characters in advance of the execution process. Writing errors convey the fact that during writing several processes take place, including phonemic-to-graphemic conversion (spelling), and graphemic-to-allographic conversion ("font choice").

Table E.1 gives an overview of parameters that may be controlled with a pen.

x, y	position (velocity, acceleration...)
p	pen force ("pressure")
	binary pen-up/pen-down switch analog axial force transducer
z	height
ϕ_x, ϕ_y	angles
Switching	by thresholding of pen force p (above) or with additional button(s) on the pen

Table E.1: *Parameters controlled by a pen*

E.1 The Pen-based CIM/HOC as a Means of Human-Computer Interaction

In what follows, the accent will lie on the Human Output Channel (HOC) aspects of pen-based interaction. The CIM aspects will occasionally be touched upon. For now it suffices to mention the fact that in pen-based computing the concept of "electronic ink" as a typical fundamental data type, differing substantially from other basic data types such as "pixels", "vectors" and "sound". As such the ink object has a number of typical properties. It refers to a pen-tip trajectory, possibly including pen angle and axial pen force, and it has a translation along the x and y axis, a scale, and a slant. Further rendering attributes are color, line thickness and brush. In a limited number of areas, also the concept of "invisible ink", e. g., for pen-up movements, is needed. Aspects of electronic ink processing will be dealt with elsewhere. As regards computer output modalities (COM), there are no fundamental problems in using state-of-the art computer graphics in rendering electronic ink. Modeling user interfaces in pan-based COM is sometimes referred to as Pen User Interface (PUI) design, similar to the well-known term Graphical User Interface (GUI) design.

For the time being, however, we will return to the HOC aspects of pen-based user interaction.

The mouse is basically a selection instrument for selection objects, and menu items. Some continuous control like the dragging of screen objects can be performed by the user after some training but continuous control is not a strong point of the mouse. Also drawing and sketching are very difficult. With a pen, however, the same user-actions can be performed as with a single-button mouse, but there are additional functions which are typical for the pen. More elaborate *data entry* is possible, in the form of linguistic data (text) or graphical data (ink). Also, the user has a more direct control over objects on the screen: there is a higher degree of *Direct Manipulation*, similar as in finger touch screens, but with a much higher spatial resolution in case of the pen.

A basic distinction in pen input is between (i) Textual Data Input, (ii) Command Input, (iii) Graphical Input. Signature Verification will not be considered in this context.

E.2 Textual Data Input

Involved are letters, digits and punctuations. There are a number of generic styles:

- block capitals
- isolated handprint
- run-on handprint/mixed-cursive
- fully connected cursive
- combinations of these.

However, there are large differences between writers, especially of differing nationality. Recently a new artificial style has been proposed by Goldberg, which has the advantage of being recognized with almost 100% accuracy [125]. However, the shape of these unistroke characters has nothing in common with the known alphabet. The shapes are immediately replaced by their machine font variant.

E.2.1 Conversion to ASCII (handwriting recognition)

In this case the recognized data can be processed and used as regular computer text, allowing for all storage and retrieval methods, rendering in machine fonts, database searching, E-mail. The following three methods are sorted in order of increasing reliability.

Free Text Entry

Hierarchically, handwriting consists of the following components:

stroke → character → word → sentence → paragraph → page

Larger components have not been approached in current technology. It should be noted, that apart from the already difficult classification of shape (character, word), also the segmentation preceding it is erroneous. Writers will often write words in a sentence with a horizontal spacing which is of the same magnitude within and between words. The histograms of horizontal distances ΔX for within and between-word horizontal spacing will generally have a high overlap.

Difficult issues in current recognition and user interfacing technology are the multitude of styles over writers, the variability of shape within a writer, the processing of delayed pen actions, such as the dotting of i's and j's at a much later stage in the input process, and all kinds of delayed graphical corrections to letters (crossing, adding ink, adding drawings).

The following three categories of *Free Text Entry* may be defined, depending on the number and nature of existing constraints on the handwriting process.

1. Fully unconstrained (size, orientation, styles)
   This type of input, like on the normal office and household notes or PostIts, cannot be recognized well at this point in time.

2. Lineated form, no prompting, free order of actions
   Here there are weak geometrical constraints, leading to a much more regular input quality. There are no temporal constraints.

3. Prompted
   In order to reduce the error introduced by segmentation into sentences or words, the user is required to indicate the boundaries of an input chunk. In the QWERTY keyboard, a similar segmenting functionality is provided by the <Return> or <Enter> key.

   The following methods are being used:

   - "OK" dialog box chunk ending. A "Cancel" button allows for rewriting misspelled input. This method works reliable but interrupts the writing process.
   - Time-out. Characters and Words can be segmented by a time out. If the writer lifts the pen and does not produce any ink within T time, a segmentation event is assumed to have occurred. A typical time out value for words is 800 ms, but this varies over users. Long time out values cause merging of ink chunks, short time out values lead to premature processing of unfinished chunks.
   - Gesture (see below). In this case, a simple gesture, like tapping with the pen to the right of the last word will indicate the segmentation event.

Boxed Forms

By constraining the position, orientation and size of the handwritten input, the system can force a reduced variability of input, and can make rigid use of context, e.g. by only considering digits in the recognition process if an "amounts" field must be filled in. Often even handwriting style is prescribed. Mostly, but no exclusively, isolated handprint characters and block print capitals are allowed with current recognition technology.

Virtual keyboard

The size of the keyboard limits the miniaturization of handhold computers (Personal Digital Assistants, PDAs). A QWERTY keyboard can be presented on the LCD/digitizer and individual virtual keys can be tapped with the pen. In fact reasonable speeds can be reached with the proper layout [342]. In LCD/digitizer telephones with a touch sensitive surface it was observed that users prefer tapping with the pen over pressing with their finger on the glass screen [82].

E.2.2 Graphical text storage

In this type of textual input, there is no need for computer recognition since the end user of the produced text is a human again. Applications are "Ink E-mail", faxing, notekeeping. Attaching context information to notes (date, time, operator etc.), already allows for a limited form of automatic processing and retrieval.

E.3 Command Entry

E.3.1 Widget selection (discrete selection)

Here the same user actions are considered as the "clicking" of a mouse. The terminology in pen computing is "tapping". A special problem of the pen on integrated LCD/digitizers is that the hand may cover parts of the screen. A distinction must be made between left-handed and right-handed users.

A problem in selecting text printed on the screen is the font size in E.2 type applications (Textual Data Input). The font size must be of about the same size as the user's handwriting or larger. Otherwise, serious text selection problems occur.

E.3.2 Drag-and-drop operations (continuous control)

As with the mouse, this type of operation is rather difficult and must be learned. The typical error is the premature Drop by lifting the pen or pressing not hard enough.

E.3.3 Pen gestures

Gestures are a limited size and duration pen-tip trajectory of unique shape, mostly (but not exclusively) different from known character shapes. A problem is the discrimination between character shapes from gesture shapes, both by recognition algorithms on the CIM side, and by the human user (Cognition and HOC side), who must remember the pattern shapes and their meaning. Ideally, a pen-based user interface is "mode-less", but often a gesture mode must be invoked to address the gesture recognizer explicitly. Also, the gestures can be produced in dedicated screen areas. Two different basic pen gestures may be defined:

Position-independent gestures

System functions are activated, regardless of the pen position, such as starting an application program.

Position-dependent context gestures

Here the X/Y-position of the pen is used to relate the gesture to objects on the screen. An example is the crossing of words.

E.3.4 Continuous control

The fact that the pen delivers continuous signals may be used in analog control. For instance, in some applications the pen force is used to encode ink trace thickness.

E.4 Handwriting and Pen Gestures Computer Output Media (COM)

The abundance of handwriting synthesis models in literature allows for the production of handwriting computer output, which is "faithful" both as regards the static image quality, and the temporal movement characteristics. Not all models are equally suitable in practical applications, and not all are good as a model for the "real" neural handwriting system. For completeness, however, it is good to realize that much as in speech, both recognition and synthesis can be realized in the current state of the art. Apart from modeling the pen-tip trajectory only, recent models have begun to address the whole movement of forearm, hand and finger joints, as well as the pen tip. However, in the context of MIAMI, it may suffice to use standard inverse kinematics approaches to produce handwriting with a mechanical or graphical arm where applicable. Storage and graphical rendering of already produced human handwriting is a technical, rather than a scientific problem.

E.5 Graphical Pattern Input

With Graphical Pattern Input, we mean those geometrical patterns which do not directly refer to linguistic information. The following categories may be discerned, varying in the degree to which symbolic, iconic and structural information is involved.

E.5.1 Free-style drawings

This category of pen-based input mainly consists of sequentially entered iconic representations. The pen as the input device offers a tremendous advantage over any other form of computer input device. However, there are also some disadvantages as compared to the normal graphical tools like pencil or brush on real paper. The opaque tablet with an inking ballpoint allows for a more or less natural drawing behavior on paper, but the result on the screen can only be seen by looking up to the CRT monitor. The electronic paper solution immediately (see above) reveals the graphical result in screen resolution and color, but here, the use of the pen is not as easy as in the pencil & paper situation. The general phenomenon which is missing is the mechanical effect of producing a trace, i. e., proprioceptively "feeling the friction" of the pen-to-surface contact. Two basic principles are worth mentioning: (i) free-style drawing as a process in which pixel ink is deposited in a bitmap, and (ii), free-style drawing as a process in which a position vector is produced in time. The latter type of recording can be played back in time easily. This also allows for a stroke-based "undo" operation during the interaction, which is more difficult to achieve in the case of bitmap-oriented drawing.

E.5.2 Flow charts and schematics

Here, the type of graphical data entered is of a mixed iconic and structural representation. Mostly, the use of the drawing itself is rather useless without textual annotation and clarification. The algorithms involved here try to detect basic shapes from a library (i. e., lines, triangles, circles, rectangles, squares, etc.) and replace the sloppy input with the nearest neat library shape of the same size. Provisions are needed for closing holes and interactively improving the drawing. The systems are mostly object oriented in the sense that the objects have “widgets”: anchor points to tap on for resizing and moving. Basic rules for ”neatness” are usually applied, like making connections to a side of a rectangle only on anchor points at $1/2, 1/3, \ldots, 1/6$ of the length of that side. This property is sometimes called “magnetic joining”. Research in this area dates from the late sixties, were combinations of flow-chart entry and programming in Fortran were studied.

E.5.3 Miscellaneous symbolic input

In this category, the entry of mathematical formulas or musical scores is considered. In both these types of graphical behavior there is an aspect of symbolical and structural information. This means that apart from a symbol identity like “Greek alpha”, its geometrical relation to the other graphical objects plays an essential role. Interestingly, the entering of formulas is sometimes used as an example of the ease of *Direct Manipulation*. It has become clear however, that the correct entry of formulas with the pen is a tedious process, if rules for graphical style are not applied within the interface algorithm. As a worst case example, the user must enter four similar formulas with a tantalizing accurate positioning of subscripts and superscripts etc. four times, with the risk of very annoying spacing differences between these formulas. A combination of rule-based approaches, as in LaTeX, and the actual use of interactive pen input is the best solution to this problem. Finally, both in the automatic recognition of mathematical formulas and musical scores, context information is needed and is in practice actually used to solve shape classification ambiguities.

E.6 Known Bimodal Experiments in Handwriting

E.6.1 Speech command recognition and pen input

A number of application areas are possible [74]:

1. ink and speech annotation of existing documents
2. pointing by pen, data or modifier input by voice
3. the combined pen/voice typewriter: text input

The interesting functionality is derived from the complementarity of the two human output channels. Pointing to objects by speech is dull, slow and error-prone. Thus for pointing, the pen may be used. Similarly, for symbolic data entry, speech is a fast and natural channel. In text input, the combined recognition of speech and handwriting may considerably improve recognition rates.

E.6.2 Handwriting recognition and speech synthesis

At NICI, experience has been done with combining cursive handwriting recognition and (Dutch) speech synthesis. The speech synthesis module was produced by Boves et al. Preliminary results indicate that speech and sound may be a good way of giving recognizer feedback to the user, provided that the response is very fast. Problems arise in the prosody coding, which operates at the sentence level, and not at the level of single words. Single character speech feedback is currently only possible with isolated handprint styles.

Bibliography

[1] J. H. Abbs and G. L. Eilenberg. Pheripheral mechanism of speech motor control. In N. J. Lass, editor, *Contemporary Issue in Experimental Phonetics*, pages 139–166. 1976.

[2] C. Abry, L. J. Boe, and J. L. Schwartz. Plateaus, Catastrophes and the Structuring of Vowel Systems. *Journal of Phonetics*, 17:47–54, 1989.

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

[4] M. L. Agronin. The Design and Software Formulation of a 9-String 6-Degree-of-Freedom Joystick for Telemanipulation. Master's thesis, University of Texas at Austin, 1986.

[5] K. Aizawa, H. Harashima, and T. Saito. Model-based analysis-synthesis image coding (mbasic) system for person's face. *Image Communication*, 1:139–152, 1989.

[6] M. Akamatsu and S. Sato. A multi-modal mouse with tactile and force feedback. *Int. Journ. of Human-Computer Studies*, 40:443–453, 1994.

[7] P. Astheimer et al. Die Virtuelle Umgebung – Eine neue Epoche in der Mensch-Maschine-Kommunikation. *Informatik-Spektrum*, 17(6), December 1994.

[8] D. Baggi, editor. *Readings in Computer-Generated Music*. IEEE CS Press, 1992.

[9] J. Baily. Music structure and human movement. In P. Howell, I. Cross, and R. West, editors, *Musical Structure and Cognition*, pages 237–258. Academic Press, 1985.

[10] R. Balakrishnan, C. Ware, and T. Smith. Virtual hand tool with force feedback. In C. Plaison, editor, *Proc. of the Conf. on Human Factors in Computing Systems, CHI'94*, Boston, 1994. ACM/SIGCHI.

[11] J. Bates. The role of emotions in believable agents. *Comm. of the ACM*, 37(7):122–125, 1994.

[12] R. J. Beaton et al. An Evaluation of Input Devices for 3-D Computer Display Workstations. In *Proc. of the SPIE Vol. 761*, pages 94–101, 1987.

[13] R. J. Beaton and N. Weiman. User Evaluation of Cursor-Positioning Devices for 3-D Display Workstations. In *Proc. of the SPIE Vol. 902*, pages 53–58, 1988.

[14] A. P. Benguerel and H. A. Cowan. Coarticulation of upper lip protrusion in French. *Phonetica*, 30:41–55, 1974.

[15] A. P. Benguerel and M. K. Pichora-Fuller. Coarticulation effects in lipreading. *Journal of Speech and Hearing Research*, 25:600–607, 1982.

[16] C. Benoît, M. T. Lallouache, T. Mohamadi, and C. Abry. A set of french visemes for visual speech synthesis. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Designs*, pages 485–504. Elsevier Science Publishers B. V., North-Holland, Amsterdam, 1992.

[17] C. Benoît, T. Mohamadi, and S. Kandel. Effects of phonetic context on audio-visual intelligibility of French speech in noise. *Journal of Speech & Hearing Research*, (in press), 1994.

[18] P. Bergeron and P. Lachapelle. Controlling facial expressions and body movements in the computer generated animated short 'Tony de Peltrie'. In *SigGraph '85 Tutorial Notes, Advanced Computer Animation Course*. 1985.

[19] N. Bernstein. *The Coarticulation and Regulation of Movements*. London Pergamon Press, 1967.

[20] P. Bertelson and M. Radeau. Cross-modal bias and perceptual fusion with auditory visual spatial discordance. *Perception and Psychophysics*, 29:578–584, 1981.

[21] C. A. Binnie, A. A. Montgomery, and P. L. Jackson. Auditory and visual contributions to the perception of consonants. *Journal of Speech & Hearing Research*, 17:619–630, 1974.

[22] J.-L. Binot et al. Architecture of a Multimodal Dialogue Interface for Knowledge-Based Sytems. ESPRIT II, Project No. 2474.

[23] E. Bizzi. Central and peripheral mechanisms in motor control. In G. Stelmach and J. Requin, editors, *Advances in psychology 1: Tutorials in motor behavior*, pages 131–143. Amsterdam: North Holland, 1980.

[24] E. Bizzi, A. Polit, and P. Morasso. Mechanisms underlying achievement of final head position. *Journal of Neurophysiology*, 39:435–444, 1976.

[25] M. M. Blattner and R. Dannenberg, editors. *Multimedia Interface Design (Readings)*. ACM Press/Addison Wesley, 1992.

[26] M. M. Blattner et al. Sonic enhancement of two-dimensional graphic display. In G. Kramer, editor, *Auditory Display*, pages 447–470, Reading, Massachusetts, 1994. Santa Fe Institute, Addison Wesley.

[27] J. Blauert. *Hearing - Psychological Bases and Psychophysics*, chapter Psychoacoustic binaural phenomena. Springer, Berlin New York, 1983.

[28] J. Blauert, M. Bodden, and H. Lehnert. Binaural Signal Processing and Room Acoustics. *IEICE Transact. Fundamentals (Japan)*, E75:1454–1458, 1993.

[29] J. Blauert and J.-P. Col. *Auditory Psychology and Perception*, chapter A study of temproal effects in spatial hearing, pages 531–538. Pergamon Press, Oxford, 1992.

[30] J. Blauert, H. Els, and J. Schröter. A Review of the Progress in External Ear Physics Regarding the Objective Performance Evaluation of Personal Ear Protectors. In *Proc. Inter-Noise '80*, pages 643–658, USA New York, 1980. Noise-Controil Found., Noise-Control Found.

[31] J. Blauert and K. Genuit. Sound-Environment Evaluation by Binaural Technology: Some Basic Consideration. *Journ. Acoust. Soc. Japan*, 14:139–145, 1993.

[32] J. Blauert, H. Hudde, and U. Letens. Eardrum-Impedance and Middle-Ear Modeling. In *Proc. Symp. Fed. Europ. Acoust. Soc., FASE, 125-128, PLisboa*, 1987.

[33] M. Bodden. *Binaurale Signalverarbeitung: Modellierung der Richtungserkennung und des Cocktail-Party-Effektes (Binaural signal processing (Modeling of direction finding and of the cocktail-party effect)*. PhD thesis, Ruhr-Universität Bochum, 1992.

[34] M. Bodden. Modeling Human Sound Source Localization and the Cocktail-Party-Effect. *Acta Acustica 1*, 1:43–55, 1993.

[35] M. Bodden and J. Blauert. Separation of Concurrent Speech Signals: A Cocktail-Party Processor for Speech Enhancement. In *Proc. ESCA Workshop on: Speech Processing in Adverse Conditions (Cannes Mandleieu)*, pages 147–150. ESCA, 1992.

[36] J. Bortz. *Lehrbuch der Statistik*. 1977.

[37] D. W. Boston. Synthetic facial animation. *British Journal of Audiology*, 7:373–378, 1973.

[38] H. H. Bothe, G. Lindner, and F. Rieger. The development of a computer animation program for the teaching of lipreading. In *Proc. of the 1st TIDE Conference, Bruxelles*, pages 45–49. 1993.

[39] R. J. Brachman and J. G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9:171–216, 1985.

[40] L. D. Braida. Crossmodal integration in the identification of consonant segments. *Quaterly Journal of Experimental Psychology*, 43:647–678, 1991.

[41] C. Bregler, H. Hild, S. Manke, and W. A. Improving connected letter recognition by lipreading. In *International Joint Conference of Speech and Signal Processing*, volume 1, pages 557–560, Mineapolis, MN, 1993.

[42] S. A. Brewster, P. C. Wright, and A. D. N. Edwards. A detailed investigation into the effectiveness of earcons. In G. Kramer, editor, *Auditory Display*, pages 471–498, Reading, Massachusetts, 1994. Santa Fe Institute, Addison Wesley.

[43] D. E. Broadbent. The magic number seven after fifteen years. In A. Kennedy and A. Wilkes, editors, *Studies in long term memory*, pages 3–18. London: Wiley, 1975.

[44] E. R. Brocklehurst. The NPL Electronic Paper Project. Technical Report DITC 133/88, National Physical Laboratory (UK), 1994.

[45] N. M. Brooke. Development of a video speech synthesizer. In *Proceedings of the British Institute of Acoustics, Autumn Conference*, pages 41–44, 1979.

[46] N. M. Brooke. Computer graphics synthesis of talking faces. In G. Bailly and C. Benoît, editors, *Talking Machines: Theories, Models and Designs*, pages 505–522. Elsevier Science Publishers B. V., North-Holland, Amsterdam, 1992.

[47] N. M. Brooke and E. D. Petajan. Seeing speech: Investigation into the synthesis and recognition of visible speech movement using automatic image processing and computer graphics. In *Proceedings of the International Conference on Speech Input and Output*, pages 104–109, 1986.

[48] F. P. Brooks, Jr. et al. Project GROPE - Haptic Displays for Scientific Visualization. *ACM Computer Graphics*, 24(4):177–185, Aug. 1990.

[49] H. W. Buckingham and H. Hollien. A neural model for language and speech. *Journal of Phonetics*, 6:283, 1993.

[50] J. Burgstaller, J. Grollmann, and F. Kapsner. On the Software Structure of User Interface Management Systems. In W. Hansmann et al., editors, *EUROGRAPHICS '89*, pages 75–86. Elsevier Science Publishers B. V., 1989.

[51] T. W. Calvert et al. The evolution of an interface for coreographers. In *Proc. INTERCHI*, 1993.

[52] H. W. Campbell. *Phoneme recognition by ear and by eye: a distinctive feature analysis.* PhD thesis, Katholieke Universiteit te Nijmegen, 1974.

[53] R. Campbell. Tracing lip movements: making speech visible. *Visible Language*, 8(1):33–57, 1988.

[54] R. Campbell and B. Dodd. Hearing by eye. *Quaterly Journal of Experimental Psychology*, 32:509–515, 1980.

[55] A. Camurri et al. Dance and Movement Notation. In P. Morasso and V. Tagliasco, editors, *Human Movement Understanding.* North-Holland, 1986.

[56] A. Camurri et al. Music and Multimedia Knowledge Representation and Reasoning: The HARP System. *Computer Music Journal (to appear)*, 1995.

[57] A. Camurri and M. Leman. AI-based Music Signal Applications. Techn. rep., IPEM - Univ. of Gent and DIST - Univ. Of Genova, 1994.

[58] S. Card, W. K. English, and B. J. Burr. Evaluations of Mouse, Rate-controlled Isometric Joystick, Step Keys, and Text Keys for Text Selection on a CRT. *Ergonomics*, 21(8):601–613, Aug. 1978.

[59] S. K. Card, T. P. Moran, and A. Newell. The Keystroke-Level Model for User Performance Time with Interactive Systems. *Communications of the ACM*, 23(7):396–410, 1980.

[60] S. K. Card, T. P. Moran, and A. Newell. *The Psychology of Human-Computer Interaction.* Lawrence Erlbaum Ass., Publishers, 1983.

[61] E. Casella, F. Lavagetto, and R. Miani. A time-delay neural network for speech to lips movements conversion. In *Proc. Int.Conf. on Artificial Neural Networks, Sorrento, Italy*, pages 26–27. 1994.

[62] M. A. Cathiard. Identification visuelle des voyelles et des consonnes dans le jeu de la protrusion-retraction des levres en francais. Technical report, Memoire de Maitrise, Departement de Psychologie, Grenoble, France, 1988.

[63] M. A. Cathiard. La perception visuelle de la parole : apercu des connaissances. *Bulletin de l'Institut de Phonetique de Grenoble*, 17/18:109–193, 1988/1989.

[64] M. A. Cathiard, G. Tiberghien, A. Tseva, M. T. Lallouache, and P. Escudier. Visual perception of anticipatory rounding during pauses: A cross-language study. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, volume 4, pages 50–53, Aix-en-Provence, France, 1991.

[65] M. Chafcouloff and A. Di Cristo. Les indices acoustiques et perceptuels des consonnes constrictives du francais, application la synthese. In *Actes des 9emes Journes d'Etude sur la Parole, Groupe Communication Parlee du GALF*, pages 69–81, Lannion, France, 1978.

[66] A. Chapanis and R. M. Halsey. Absolute Judgements of Spectrum Colors. *Journ. of Psychology*, pages 99–103, 1956.

[67] H. J. Charwat. *Lexikon der Mensch-Maschine-Kommunikation.* Oldenbourg, 1992.

[68] M. Chen, S. J. Mountford, and A. Sellen. A Study in Interactive 3-D Rotation Using 2-D Control Devices. *ACM Computer Graphics*, 22(4):121–129, Aug. 1988.

[69] N. Chomsky and M. Halle. *Sound Pattern of English.* Harper and Row, New-York, 1968.

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

[71] M. M. Cohen and D. W. Massaro. Modelling coarticulation in synthetic visual speech. In M.-T. . Thalmann, editor, *Proceedings of Computer Animation '93*, Geneva, Switzerland, 1993.

[72] J.-P. Col. *Localisation auditiv d'un signal et aspects temporels de l'audition spatiale (Auditory localization of a signal and temporal aspects of spatial hearing).* PhD thesis, Marseille, 1990.

[73] J. Cotton. Normal 'visual-hearing'. *Science*, 82:582–593, 1935.

[74] H. D. Crane and D. Rtischev. Pen and voice unite: adding pen and voice input to today's user interfaces opens the door for more natural communication with your computer. *Byte*, 18:98–102, Oct. 1993.

[75] J. L. Crowley and Y. Demazeau. Principles and techniques for sensor data fusion. *Signal Processing*, 32:5–27, 1993.

[76] S. D. Visual-neural correlate of speechreading ability in normal-hearing adults: reliability. *Journal of Speech and Hearing Research*, 25:521–527, 1982.

[77] R. B. Dannenberg and A. Camurri. Computer-Generated Music and Multimedia Computing. In *IEEE ICMCS Intl. Conf. on Multimedia Computing and Systems, Proc. ICMCS 94*, pages 86–88. IEEE Computer Society Press, 1994.

[78] R. Davis, H. Shrobe, and P. Szolovits. What is a Knowledge Representation? *AI Magazine*, 14(1), 1993.

[79] B. de Graf. Performance facial animation notes. In *Course Notes on State of the Art in Facial Animation*, volume 26, pages 10–20, Dallas, 1990. SigGraph '90.

[80] G. De Poli, A. Piccialli, and C. Roads, editors. *Representations of Musical Signals.* MIT Press, 1991.

[81] W. N. Dember and J. S. Warn. *Psychology of Perception - 2nd Edition.* Holt, Rinehart & Winston, New York, 1979.

[82] L. Dikmans. Future intelligent telephone terminals: A method for user interface evaluation early in the design process. Technical report, *IPO/Philips rapport '94*, Eindhoven: Institute for Perception Research, 1994.

[83] N. F. Dixon and L. Spitz. The detection of audiovisual desynchrony. *Perception*, (9):719–721, 1980.

[84] B. Dodd and R. Campbell, editors. *Hearing by Eye: The Psychology of Lip-reading*, Hillsdale, New Jersey, 1987. Lawrence Erlbaum Associates.

[85] E. H. Dooijes. *Analysis of Handwriting Movements.* PhD thesis, University of Amsterdam, 1983.

[86] R. A. Earnshaw, M. A. Gigante, and H. Jones, editors. *Virtual Reality Systems.* Academic Press, 1993.

[87] B. Eberman and B. An. EXOS Research on Force Reflecting Controllers. *SPIE Telemanipulator Technology*, 1833:9–19, 1992.

[88] P. Ekman and W. V. Friesen. *Facial Action Coding System.* Consulting Psychologists Press, Stanford University, Palo Alto, 1977.

[89] S. R. Ellis. Nature and Origins of Virtual Environments: A Bibliographical Essay. *Computing Systems in Engineering*, 2(4):321–347, 1991.

[90] H. Els. *Ein Meßsystem für die akustische Modelltechnik (A measuring system for the technique of acoustic modeling).* PhD thesis, Ruhr-Universität Bochum, 1986.

[91] H. Els and J. Blauert. Measuring Techniques for Acoustic Models - Upgraded. In *Proc. Internoise'85, Schriftenr. Bundesanst. Arbeitsschutz, Vol. Ib 39/II*, pages 1359–1362. Bundesanst. Arbeitsschutz, 1985.

[92] H. Els and J. Blauert. A Measuring System for Acoustic Scale Models. In *12th Int. Congr. Acoust., Proc. of the Vancouver Symp. Acoustics & Theatre Planning for the Performing Arts*, pages 65–70, 1986.

[93] N. P. Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech & Hearing Research*, 12:423–425, 1969.

[94] N. P. Erber. Auditory, visual and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research*, 15:413–422, 1972.

[95] N. P. Erber. Auditory-visual perception of speech. *Journal of Speech & Hearing Disorders*, 40:481–492, 1975.

[96] N. P. Erber and C. L. de Filippo. Voice/mouth synthesis and tactual/visual perception of /pa, ba, ma/. *Journal of the Acoustical Society of America*, 64:1015–1019, 1978.

[97] C. W. Eriksen and H. W. Hake. Multidimensional Stimulus Differences and Accuracy of Discrimination. *Psychological Review*, 67:279–300, 1955.

[98] P. Escudier, C. Benoît, and M. T. Lallouache. Identification visuelle de stimuli associes a l'opposition /i/ - /y/ : etude statique. In *Proceedings of the First French Conference on Acoustics*, pages 541–544, Lyon, France, 1990.

[99] C. Faure. Pen and voice interface for incremental design of graphics documents. In *Proceedings of the IEE Colloquium on Handwriting and Pen-based input, Digest Number 1994/065*, pages 9/1–9/3. London: The Institution of Electrical Engineers, March 1994.

[100] W. Felger. How interactive visualization can benefit from multidimensional input devices. *SPIE*, 1668:15–24, 1992.

[101] I. A. Ferguson. *TouringMachines: An Architecture for Dynamic, Rational, Mobile Agents.* Phd thesis, University of Cambridge, 1992.

[102] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6):381–391, June 1954.

[103] J. D. Foley. Neuartige Schnittstellen zwischen Mensch und Computer. In *Spektrum der Wissenschaft*, number 12, pages 98–106. Dec. 1987.

[104] J. W. Folkins and J. H. Abbs. Lip and jaw motor control during speech: motor reorganization response to external interference. *J. S. H. R.*, 18:207–220, 1975.

[105] C. Fowler, P. Rubin, R. Remez, and M. E. Turvey. Implications for speech production of a general theory of action. In *Language Production, Speech and Talk*, volume 1, pages 373–420. Academic Press, London, 1980.

[106] C. A. Fowler. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 1980.

[107] C. A. Fowler. Current perspective on language and speech production: A critical overview. In *Speech Science*, pages 193–278. Taylor and Francis, London, 1985.

[108] C. A. Fowler. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14:3–28, 1986.

[109] D. M. Frohlich. The Design Space of Interfaces. Technical report, Hewlett-Packard Company, 1991.

[110] O. Fujimura. Elementary gestures and temporal organization. what does an articulatory constraint means? In *The cognitive representation of speech*, pages 101–110. North Holland Amsterdam, 1981.

[111] Y. Fukui and M. Shimojo. Edge Tracing of Virtual Shape Using Input Device with Force Feedback. *Systems and Computers in Japan*, 23(5):94–104, 1992.

[112] W. Gaik. *Untersuchungen zur binauralen Verarbeitung kopfbezogener Signale (Investigations into binaural signal processing of head-related signals)*. PhD thesis, Ruhr-Universität Bochum, 1990.

[113] W. Gaik. Combined Evaluation of Interaural Time and Intensity Differences: Psychoacoustic Results and Computer Modeling. *Journ. Acoust. Soc. Am.*, 94:98–110, 1993.

[114] W. Gaik and S. Wolf. Multiple Images: Psychological Data and Model Predictions. In J. Duifhuis, J. W. Horst, and H. P. Wit, editors, *Basic Issues of Hearing*, pages 386–393, London, 1988. Academic Press.

[115] P. Gårdenfors. Semantics, Conceptual Spaces and the Dimensions of Music. In Rantala, Rowell, and Tarasti, editors, *Essays on the Philosophy of Music*, volume 43 of *Acta Philosophica Fennica*, pages 9–27, Helsinki, 1988.

[116] W. R. Garner. An Informational Analysis of Absolute Judgments of Loudness. *Journ. of Experimental Psychology*, 46:373–380, 1953.

[117] F. Garnier. Don Quichotte. Computer-generated movie, 2:40, 1991.

[118] W. W. Gaver. Using and creating auditory icons. In G. Kramer, editor, *Auditory Display*, pages 417–446, Reading, Massachusetts, 1994. Santa Fe Institute, Addison Wesley.

[119] T. H. Gay. Temporal and spatial properties of articulatory movements: evidence for minimum spreading across and maximum effects within syllable boundaries. In *The cognitive representation of speech*, pages 133–138. North Holland Amsterdam, 1981.

[120] G. Geiser. Mensch-Maschine Kommunikation. Oldenbourg, 1990.

[121] J. J. Gibson. *The senses considered as perceptual systems.* Houghton Mifflin, Boston, 1966.

[122] J. J. Gibson. *The ecological approach to visual perception.* Hougton Miffin, Boston, 1979.

[123] M. A. Gigante. Virtual Reality: Definitions, History and Applications. In R. A. Earnshaw, M. A. Gigante, and H. Jones, editors, *Virtual Reality Systems*, chapter 1. Academic Press, 1993.

[124] J. Glasgow and D. Papadias. Computational Imagery. *Cognitive Science*, 16:355–394, 1992.

[125] D. Goldberg and C. Richardson. Touch-Typing with a Stylus. In *InterCHI '93 Conference Proceedings*, pages 80–87. Amsterdam, 1993.

[126] A. J. Goldschen. *Continuous Automatic Speech Recognition by Lip Reading.* PhD thesis, School of Engineering and Applied Science of the George Washington University, 1993.

[127] M. Good. Participatory Design of A Portable Torque-Feedback Device. In P. Bauersfeld, J. Bennett, and G. Lynch, editors, *Proc. of the Conf. on Human Factors in Computing Systems, CHI'92*, pages 439–446. ACM/SIGCHI, 1992.

[128] K. W. Grant and L. D. Braida. Evaluating the articulation index for auditory-visual input. *Journal of the Acoustical Society of America*, 89:2952–2960, 1991.

[129] K. P. Green, E. B. Stevens, P. K. Kuhl, and A. M. Meltzoff. Exploring the basis of the McGurk effect: Can perceivers combine information from a female face and a male voice? *Journal of the Acoustical Society of America*, 87:125, 1990.

[130] L. Grimby, J. Hannerz, and B. Hedman. Contraction time and voluntary discharge properties of individual short toe extensors in man. *Journal of Physiology*, 289:191–201, 1979.

[131] R. Gruber. *Handsteuersystem für die Bewegungsführung.* PhD thesis, Universität Karlsruhe, 1992.

[132] T. Guiard-Marigny. *Animation en temps reel d'un modele parametrise de levres.* PhD thesis, Institut National Polytechnique de Grenoble, 1992.

[133] T. Guiard-Marigny, A. Adjoudani, and C. Benoît. A 3-D model of the lips for visual speech synthesis. In *Proceedings of the 2nd ESCA-IEEE Workshop on Speech Synthesis*, pages 49–52, New Paltz, NY, 1994.

[134] R. Hammarberg. The metaphysics of coarticulation. *Journal of Phonetics*, 4:353–363, 1976.

[135] P. H. Hartline. Multisensory convergence. In G. Adelman, editor, *Encyclopedia of Neuroscience*, volume 2, pages 706–709. Birkhauser, 1987.

[136] Y. Hatwell. Toucher l'espace. La main et la perception tactile de l'espace. Technical report, Universitaires de Lille, 1986.

[137] Y. Hatwell. Transferts intermodaux et integration intermodale. 1993.

[138] C. Henton and P. Litwinowicz. Saying and seeing it with feeling: techniques for synthesizing visible, emotional speech. In *Proceedings of the 2nd ESCA-IEEE Workshop on Speech Synthesis*, pages 73–76, New Paltz, NY, 1994.

[139] D. R. Hill, A. Pearce, and B. Wyvill. Animating speech: an automated approach using speech synthesised by rules. *The Visual Computer*, 3:176–186, 1988.

[140] D. R. Hill, A. Pearce, and B. Wyvill. Animating speech: an automated approach using speech synthesised by rules. *The Visual Computer*, 3:277–289, 1989.

[141] W. Hill et al. *Architectural Qualities and Principles for Multimodal and Multimedia Interfaces*, chapter 17, pages 311–318. ACM Press, 1992.

[142] G. Hirzinger. Multisensory Shared Autonomy and Tele-Sensor-Programming – Key Issues in Space Robotics. *Journ. on Robotics and Autonmous Systems*, (11):141–162, 1993.

[143] X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[144] H. Hudde. *Messung der Trommelfellimpedanz des menschlichen Ohres bis 19kHz (Measurement of eardum impedance up to 19 kHz).* PhD thesis, Ruhr-Universität Bochum, 1980.

[145] H. Hudde. Estimation of the Area Function of Human Ear Canals by Sound-Pressure Measurements. *Journ. Acoust. Soc. Am.*, 73:24–31, 1983.

[146] H. Hudde. Measurement of Eardrum Impedance of Human Ears. *Journ. Acoust. Soc. Am.*, 73:242–247, 1983.

[147] H. Hudde. Measurement-Related Modeling of Normal and Reconstructed Middle Ears. *Acta Acustica, submitted*, 1994.

[148] H. Iwata. Artificial Reality with Force-feedback: Development of Desktop Virtual Space with Compact Master Manipulator. *Computer Graphics*, 24(4):165–170, 1990.

[149] R. Jakobson, G. Fant, and M. Halle. *Preliminaries to Speech Analysis. The distinctive features and their correlates.* MIT Press, Cambridge MA, 1951.

[150] B. M. Jau. Anthropomorphic Exoskeleton dual arm/hand telerobot controller. pages 715–718, 1988.

[151] P. N. Johnson-Laird. *Mental models.* Cambridge University Press, Cambridge, 1983.

[152] P. Kabbash, W. Buxton, and A. Sellen. Two-Handed Input in a Compound Task. *Human Factors in Computing Systems*, pages 417–423, 1994.

[153] E. R. Kandel and J. R. Schwartz. *Principles of neural sciences.* Elsevier North Holland, 1993. (1981 is the first edition; there is a new one with some update on neurotransmitters and molecular biology aspects, probably dated 1993).

[154] J. A. S. Kelso, D. Southard, and D. Goodman. On the nature of human interlimb coordination. *Science*, 203:1029–1031, 1979.

[155] R. D. Kent and F. D. Minifie. Coarticulation in recent speech production models. *Journal of Phonetics*, 5:115–133, 1977.

[156] D. Kieras and P. G. Polson. An Approach to the Formal Analysis of User Complexity. *Int. Journ. of Man-Machine Studies*, 22:365–394, 1985.

[157] D. H. Klatt. Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7:279–312, 1979.

[158] D. H. Klatt. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:971–995, 1980.

[159] J. Kleiser. Sextone for president. Computer-generated movie, 0:28.

[160] J. Kleiser. A fast, efficient, accurate way to represent the human face. *Course Notes on State of the Art in Facial Animation, SigGraph '89*, 22:35–40, 1989.

[161] D. B. Koons, C. J. Sparrel, and K. R. Thorisson. Integrating Simultaneous Output from Speech, Gaze, and Hand Gestures. In M. Maybury, editor, *Intelligent Multimedia Interfaces*, pages 243–261. Menlo Park: AAAI/MIT Press, 1993.

[162] G. Kramer. In introduction to auditory display. In G. Kramer, editor, *Auditory Display*, pages 1–77, Reading, Massachusetts, 1994. Santa Fe Institute, Addison Wesley.

[163] N. Kugler, J. A. S. Kelso, and M. T. Turvey. On the concept of coordinative structures as dissipative structures: I. theoretical line. In *Tutorials in motor behaviour*, pages 32–47. North Holland Amsterdam, 1980.

[164] P. N. Kugler, J. A. S. Kelso, and M. T. Turvey. On control and coordination of naturally developing systems. In J. A. S. Kelso and J. E. Clark, editors, *The development of movement control and coordination*, pages 5–78. New York: Wiley, 1982.

[165] P. K. Kuhl and A. N. Meltzoff. The bimodal perception of speech in infancy. *Science*, 218:1138–1141, 1982.

[166] T. Kurihara and K. Arai. A transformation method for modeling and animation of the human face from photographs. In N. M.-T. . D. Thalmann, editor, *Computer Animation '91*, pages 45–58. Springer-Verlag, 1991.

[167] P. Ladefoged. Phonetics prerequisites for a dinstictive feature theory. In *Papers in linguistics and phonetics to the memory of Pierre Delattre, Mouton, The Heague*, pages 273–285. 1972.

[168] P. Ladefoged. *A course in phonetics.* Hardcourt Brace Jovanovich Inc. NY, 1975.

[169] P. Ladefoged. What are linguistic sounds made of? *Language*, 56:485–502, 1980.

[170] J. Laird, A. Newell, and P. Rosenbloom. Soar: an architecture for general intelligence. *Artificial Intelligence*, 33:1–64, 1987.

[171] M. T. Lallouache. *Un poste "visage-parole" couleur. Acquisition et traitement automatique des contours des levres.* PhD thesis, Institut National Polytechnique de Grenoble, 1991.

[172] D. R. J. Laming. *Information theory of choice-reaction times.* London: Academic Press, 1968.

[173] K. S. Lashley. The problem of serial order in behaviour. In L. A. Jeffres, editor, *Cerebral Mechanisms in Behaviour*, pages 112–136. 1951.

[174] H. G. Lauffs. *Bediengeräte zur 3-D-Bewegungsführung.* PhD thesis, RWTH Aachen, 1991.

[175] B. Laurel, R. Strickland, and T. Tow. Placeholder: Landscape and Narrative in Virtual Environments. *Computer Graphics*, 28(2):118–126, 1994.

[176] D. Lavagetto, M. Arzarello, and M. Caranzano. Lipreadable frame animation driven by speech parameters. In *IEEE Int. Symposium on Speech, Image Processing and neural Networks*, pages 14–16, Hong Kong, April 1994.

[177] B. Le Goff, T. Guiard-Marigny, M. Cohen, and C. Benoît. Real-time analysis-synthesis and intelligibility of talking faces. In *Proceedings of the 2nd ESCA-IEEE Workshop on Speech Synthesis*, pages 53–56, New Paltz, NY, 1994.

[178] M. Lee, A. Freed, and D. Wessel. Real time neural network processing of gestural and acoustic signals. In *Proc. Intl. Computer Music Conference*, Montreal, Canada, 1991.

[179] H. Lehnert. *Binaurale Raumsimulation: Ein Computermodell zur Erzeugung virtueller auditiver Umgebungen (Binaural room simulation: A computer model for generation of virtual auditory environments).* PhD thesis, Ruhr-Universität Bochum, 1992.

[180] H. Lehnert and J. Blauert. A Concept for Binaural Room Simulation. In *Proc. IEEE-ASSP Workshop on Application of Signal Processing to Audio & Acoustics, USA-New Paltz NY*, 1989.

[181] H. Lehnert and J. Blauert. Principles of Binaural Room Simulation. *Journ. Appl. Acoust.*, 36:259–291, 1992.

[182] M. Leman. Introduction to auditory models in music research. *Journal of New Music Research*, 23(1), 1994.

[183] M. Leman. Schema-Based Tone Center Recognition of Musical Signals. *Journ. of New Music Research*, 23(2):169–203, 1994.

[184] U. Letens. *Über die Interpretation von Impedanzmessungen im Gehörgang anhand von Mittelohr-Modellen (Interpretation of impedance measurements in the ear canal in terms of middle-ear models).* PhD thesis, Ruhr-Universität Bochum, 1988.

[185] J. S. Lew. Optimal Accelerometer Layouts for Data Recovery in Signature Verification. *IBM Journal of Research & Development*, 24(4):496–511, 1980.

[186] J. P. Lewis and F. I. Parke. Automated lip-synch and speech synthesis for character animation. In *Proceedings of CHI '87 and Graphics Interface '87*, pages 143–147, Toronto, Canada, 1987.

[187] A. Liberman and I. Mattingly. The motor theory of speech perception revisited. *Cognition*, 21:1–36, 1985.

[188] I.-S. Lin, F. Wallner, and R. Dillmann. An Advanced Telerobotic Control System for a Mobile Robot with Multisensor Feedback. In *Proc. of the 4th Intl. Conf. on Intelligent Autonomous Systems (to appear)*, 1995.

[189] W. Lindemann. *Die Erweiterung eines Kreuzkorrelationsmodells der binauralen Signalverarbeitung durch kontralaterale Inhibitionsmechanismen (Extension of a cross-correlation model of binaural signal processing by means of contralateral inhibition mechanisms).* PhD thesis, Ruhr-Universität Bochum, 1985.

[190] W. Lindemann. Extension of a Binaural Cross-Correlation Model by Means of Contralateral Inhibition. I. Simulation of Lateralization of Stationary Signals. *Journ. Acoust. Soc. Am.*, 80:1608–1622, 1986.

[191] W. Lindemann. Extension of a Binaural Cross-Correlation Model by Means of Contralateral Inhibition. II. The Law of the First Wave Front. *Journ. Acoust. Soc. Am.*, 80:1623–1630, 1986.

[192] P. H. Lindsay and D. A. Norman. *Human Information Processing.* Academic Press, New York, 1977.

[193] J. F. Lubker. Representation and context sensitivity. In *The Cognitive Representation of Speech*, pages 127–131. North Holland Amsterdam, 1981.

[194] F. J. Maarse, H. J. J. Janssen, and F. Dexel. A Special Pen for an XY Tablet. In W. S. F.J. Maarse, L.J.M. Mulder and A. Akkerman, editors, *Computers in Psychology: Methods, Instrumentation, and Psychodiagnostics*, pages 133–139. Amsterdam: Swets and Zeitlinger, 1988.

[195] L. MacDonald and J. Vince, editors. *Interacting with Virtual Environments.* Wiley Professional Computing, 1994.

[196] T. Machover and J. Chung. Hyperinstruments: Musically intelligent and interactive performance and creativity systems. In *Proc. Intl. Computer Music Conference*, Columbus, Ohio, USA, 1989.

[197] I. S. MacKenzie and W. Buxton. Extending Fitts' Law to Two-Dimensional Tasks. In P. Bauersfeld, J. Bennett, and G. Lynch, editors, *Human Factors in Computing Systems, CHI'92 Conf. Proc.*, pages 219–226. ACM/SIGCHI, ACM Press, May 1992.

[198] I. S. MacKenzie, A. Sellen, and W. Buxton. A comparison of input devices in elemental pointing and dragging tasks. In S. P. Robertson, O. G.M., and O. J.S., editors, *Proc. of the ACM CHI'91 Conf. on Human Factors*, pages 161–166. ACM-Press, 1991.

[199] A. MacLeod and Q. Summerfield. Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, 21:131–141, 1987.

[200] P. F. MacNeilage. Motor control of serial ordering of speech. *Psychol. Review*, 77:182–196, 1970.

[201] P. Maes, editor. *Designing autonomous agents: Theory and practice from biology to engineering and back*, Cambridge, MA, 1990. The MIT Press/Bradford Books.

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

[203] N. Magnenat-Thalmann and D. Thalmann. The direction of synthetic actors in the film Rendez-vous Montral. *IEEE Computer Graphics & Applications*, 7(12):9–19, 1987.

[204] E. Magno-Caldognetto et al. Automatic analysis of lips and jaw kinematics in vcv sequences. In *Proc. Eurospeech '92*, pages 453–456. 1992.

[205] E. Magno-Caldognetto et al. Liprounding coarticulation in italian. In *Proc. Eurospeech '92*, pages 61–64. 1992.

[206] E. Magno-Caldognetto et al. Articulatory dynamics of lips in italian /'vpv/ and /'vbv/ sequences. In *Proc. Eurospeech '93*. 1993.

[207] C. Marsden, P. Merton, and H. Morton. Latency measurements compatible with a cortical pathway for the stretch reflex in man. *Journal of Physiology*, 230:58–59, 1973.

[208] D. W. Massaro. Categorical partition: A fuzzy-logical model of categorization behaviour. 1987.

[209] D. W. Massaro. Multiple book review of 'speech perception by ear and eye'. *Behavioral and Brain Sciences*, 12:741–794, 1989.

[210] D. W. Massaro. Connexionist models of speech perception. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, volume 2, pages 94–97, Aix-en-Provence, France, 1991.

[211] D. W. Massaro and M. M. Cohen. Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, 9:753–771, 1983.

[212] D. W. Massaro and M. M. Cohen. Perception of synthesized audible and visible speech. *Psychological Science*, 1:55–63, 1990.

[213] D. W. Massaro and D. Friedman. Models of integration given multiple sources of information. *Psychological Review*, 97:225–252, 1990.

[214] T. H. Massie and J. K. Salisbury. The PHANToM Haptic Interface: a Device for Probing Virtual Objects. In *Proc. of the ASME Winter Annual Meeting, Symp. on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, Chicago, 1994.

[215] K. Matsuoka, K. Masuda, and K. Kurosu. Speechreading trainer for hearing-impaired children. In J. Patrick and K. Duncan, editors, *Training, Human Decision Making and Control*. Elsevier Science, 1988.

[216] M. Maybury, editor. *Intelligent Multimedia Interfaces*. Menlo Park: AAAI/MIT Press, 1993.

[217] N. Mayer. XWebster: Webster's 7th Collegiate Dictionary, Copyright ©1963 by Merriam-Webster, Inc. On-line access via Internet, 1963.

[218] S. McAdams and E. Bigand, editors. *Thinking in Sound - The Cognitive Psychology of Human Audition*. Clarendon Press, Oxford, 1993.

[219] N. P. McAngus Todd. The Auditory "Primal Sketch": A Multiscale Model of Rhytmic Grouping. *Journ. of New Music Research*, 23(1), 1994.

[220] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[221] M. L. Meeks and T. T. Kuklinski. Measurement of Dynamic Digitizer Performance. In R. Plamondon and G. G. Leedham, editors, *Computer Processing of Handwriting*, pages 89–110. Singapore: World Scientific, 1990.

[222] M. A. Meredith and B. E. Stein. Interactions among converging sensory inputs in the superior colliculus. *Science*, 221:389–391, 1983.

[223] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity to process information. *Psychological Review*, 63:81–97, 1956.

[224] G. S. P. Miller. The audition. Computer-generated movie, 3:10, 1993.

[225] A. G. Mlcoch and D. J. Noll. Speech production models as related to the concept of apraxia of speech. In *Speech and Language. Advances in basic research and practise*, volume 4, pages 201–238. Academic Press NY, 1980.

[226] T. Mohamadi. *Synthese partir du texte de visages parlants: ralisation d'un prototype et mesures d'intelligibilit bimodale*. PhD thesis, Institut National Polytechnique, Grenoble, France, 1993.

[227] A. A. Montgomery and G. Soo Hoo. ANIMAT: A set of programs to generate, edit and display sequences of vector-based images. *Behavioral Research Methods and Instrumentation*, 14:39–40, 1982.

[228] P. Morasso and V. Sanguineti. Self-Organizing Topographic Maps and Motor Learning. In Cliff et al., editors, *From Anamals to Animats 3*, pages 214–220. MIT Press, 1994.

[229] S. Morishima, K. Aizawa, and H. Harashima. Model-based facial image coding controlled by the speech parameter. In *Proc. PCS-88, Turin*, number 4. 1988.

[230] S. Morishima, K. Aizawa, and H. Harashima. A real-time facial action action image synthesis driven by speech and text. In *Visual Communication and Image processing '90, the Society of Photo optical Instrumentation Engineers*, volume 1360, pages 1151–1158, 1990.

[231] S. Morishima and H. Harashima. A media conversion from speech to facial image for intelligent man-machine interface. *IEEE Journal on Sel. Areas in Comm.*, 9(4):594–600, 1991.

[232] H. Morita, S. Hashimoto, and S. Otheru. A computer music system that follows a human conductor. *IEEE COMPUTER*, 24(7):44–53, 1991.

[233] P. Morrel-Samuels. Clarifying the distinction between lexical and gestural commands. *Intl. Journ. of Man-Machine Studies*, 32:581–590, 1990.

[234] A. Mulder. Virtual Musical Instruments: Accessing the Sound Synthesis Universeas a Performer. In *Proc. First Brazilian Symposium on Computer Music, 14th Annual Congress of the Brazilian Computer Society*, Caxambu, Minas Geiras, Brazil, 1994.

[235] A. Murata. An Experimental Evaluation of Mouse, Joystick, Joycard, Lightpen, Trackball and Touchscreen for Pointing - Basic Study on Human Interface Design. In H.-J. Bullinger, editor, *Human Aspects in Computing: Design and Use of Interactive Systems and Work with Terminals*, 1991.

[236] L. E. Murphy. Absolute Judgements of Duration. *Journ. of Experimental Psychology*, 71:260–263, 1966.

[237] E. D. Mynatt. Auditory representations of graphical user interfaces. In G. Kramer, editor, *Auditory Display*, pages 533–553, Reading, Massachusetts, 1994. Santa Fe Institute, Addison Wesley.

[238] M. Nahas, H. Huitric, and M. Saintourens. Animation of a B-Spline figure. *The Visual Computer*, (3):272–276, 1988.

[239] N. H. Narayanan, editor. *Special issue on Computational Imagery*, volume 9 of *Computational Intelligence*. Blackwell Publ., 1993.

[240] N. P. Nataraja and K. C. Ravishankar. Visual recognition of sounds in Kannada. *Hearing Aid Journal*, pages 13–16, 1983.

[241] K. K. Neely. Effect of visual factors on the intelligibility of speech. *Journal of the Acoustical Society of America*, 28:1275–1277, 1956.

[242] N. Negroponte. From Bezel to Proscenium. In *Proceedings of SigGraph '89*, 1989.

[243] L. Nigay and J. Coutaz. A design space for multimodal systems - concurrent processing and data fusion. In *INTERCHI '93 - Conference on Human Factors in Computing Systems, Amsterdam*, pages 172–178. Addison Wesley, 1993.

[244] Nishida. Speech recognition enhancemant by lip information. *ACM SIGCHI bulletin*, 17:198–204, 1986.

[245] S. G. Nooteboom. The target theory of speech production. In *IPO Annual Progress Report*, volume 5, pages 51–55. 1970.

[246] D. A. Norman. Cognitive Engineering. In D. A. Norman and S. W. Draper, editors, *User Centered System Design*, pages 31–61. Lawrence Erlbaum Association, 1986.

[247] C. H. Null and J. P. Jenkins, editors. *NASA Virtual Environment Research, Applications, and Technology. A White Paper*, 1993.

[248] S. Oehman. Numerical models of coarticulation. *J.A.S.A.*, 41:310–320, 1967.

[249] A. O'Leary and G. Rhodes. Cross Modal Effects on Visual and Auditory Object Perception. *Perception and Psychophysics*, 35:565–569, 1984.

[250] J. R. Olson and G. Olson. The growth of cognitive modeling in human-computer interaction since goms. *Human-Computer Interaction*, 5:221–265, 1990.

[251] P. L. Olson and M. Sivak. Perception-response time to unexpected roadway hazard. *Human Factors*, 26:91–96, 1986.

[252] P. O'Rorke and A. Ortony. Explaining Emotions. *Cognitive Science*, 18(2):283–323, 1994.

[253] O. Ostberg, B. Lindstrom, and P. O. Renhall. Contribution to speech intelligibility by different sizes of videophone displays. In *Proc. of the Workshop on Videophone Terminal Design*, Torino, Italy, 1988. CSELT.

[254] E. Owens and B. Blazek. Visems observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28:381–393, 1985.

[255] A. Paouri, N. Magnenat-Thalmann, and D. Thalmann. Creating realistic three-dimensional human shape characters for computer-generated films. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation'91*, pages 89–99. Springer-Verlag, 1991.

[256] F. I. Parke. Computer-generated animation of faces. In *Proceedings of ACM National Conference*, volume 1, pages 451–457, 1972.

[257] F. I. Parke. Parameterized models for facial animation. *IEEE Computer Graphics and Applications*, 2:61–68, 1981.

[258] F. I. Parke. *Facial animation by spatial mapping.* PhD thesis, University of Utah, Department of Computer Sciences, 1991.

[259] E. C. Patterson, P. Litwinowicz, and N. Greene. Facial animation by spatial mapping. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation'91*, pages 31–44. Springer-Verlag, 1991.

[260] S. J. Payne. Task action grammar. In B. Shackel, editor, *Proc. Interact '84*, pages 139–144. Amsterdam: North-Holland, 1984.

[261] A. Pearce, B. Wyvill, G. Wyvill, and D. Hill. Speech and expression: A computer solution to face animation. In *Graphics Interface '86*, pages 136–140, 1986.

[262] C. Pelachaud. *Communication and coarticulation in facial animation.* PhD thesis, University of Pennsylvania, USA, 1991.

[263] C. Pelachaud, N. Badler, and M. Steedman. Linguistics issues in facial animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation'91*, pages 15–30. Springer-Verlag, 1991.

[264] A. Pentland and K. Masi. Lip reading: Automatic visual recognition of spoken words. Technical Report 117, MIT Media Lab Vision Science Technical Report, 1989.

[265] J. S. Perkell. Phonetic features and the physiology of speech production. In *Language Production*, pages 337–372. Academic Press NY, 1980.

[266] J. S. Perkell. On the use of feedback in speech production. In *The Cognitive Representation of Speech*, pages 45–52. North Holland Amsterdam, 1981.

[267] E. Petajan. *Automatic Lipreading to Enhance Speech Recognition.* PhD thesis, University of Illinois at Urbana-Champain, 1984.

[268] J. Piaget. *The Origins of Intelligence in Children.* International University Press, New York, 1952.

[269] K. Pimentel and K. Teixeira. *Virtual Reality: through the new looking glass.* Windcrest Books, 1993.

[270] B. Pinkowski. LPC Spectral Moments for Clustering Acoustic Transients. *IEEE Trans. on Speech and Audio Processing*, 1(3):362–368, 1993.

[271] R. Plamondon and F. J. Maarse. An evaluation of motor models of handwriting. *IEEE Transactions on Systems, Man and Cybernetics*, 19:1060–1072, 1989.

[272] S. M. Platt. *A structural model of the human face.* PhD thesis, University of Pennsylvania, USA, 1985.

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

[274] I. Pollack. The Information of Elementary Auditory Displays. *Journ. of the Acoustical Society of America*, 25:765–769, 1953.

[275] W. Pompetzki. *Psychoakustische Verifikation von Computermodellen zur binauralen Raumsimulation (Psychoacoustical verification of computer-models for binaural room simulation).* PhD thesis, Ruhr-Universität Bochum, 1993.

[276] C. Pösselt. *Einfluss von Knochenschall auf die Schalldämmung von Gehörschützern (Influence of bone conduction on the attenuation of personal hearing protectors).* PhD thesis, Ruhr-Universität Bochum, 1986.

[277] C. Pösselt et al. Generation of Binaural Signals for Research and Home Entertainment. In *Proc. 12th Int. Congr. Acoust. Vol. I, B1-6, CND-Toronto*, 1986.

[278] W. K. Pratt. *Digital Image Processing.* Wiley, New York, 1991.

[279] J. Psotka, S. A. Davison, and S. A. Lewis. Exploring immersion in virtual space. *Virtual Reality Systems*, 1(2):70–92, 1993.

[280] M. Radeau. Cognitive Impenetrability in Audio-Visual Interaction. In J. Alegria et al., editors, *Analytical Approaches to Human Cognition*, pages 183–198. North-Holland, Amsterdam, 1992.

[281] M. Radeau. Auditory-visual spatial interaction and modularity. *Cahiers de Psychologie Cognitive*, 13(1):3–51, 1994.

[282] M. Radeau and P. Bertelson. Auditory-Visual Interaction and the Timing of Inputs. *Psychological Research*, 49:17–22, 1987.

[283] J. Rasmussen. *Information Processing and Human-Machine Interaction. An Approach to Cognitive Engineering.* North-Holland, 1986.

[284] C. M. Reed, W. M. Rabinowitz, N. I. Durlach, and L. D. Braida. Research on the tadoma method of speech comunication. *Journal of the Acoustical Society of America*, 77(1):247–257, 1985.

[285] W. T. Reeves. Simple and complex facial animation: Case studies. In *Course Notes on State of the Art in Facial Animation*, volume 26. SigGraph '90, 1990.

[286] D. Reisberg, J. McLean, and G. A. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd and R. Campbell, editors, *Hearing by eye: The psychology of lip-reading*, pages 97–114. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[287] J. Rhyne. Dialogue Management for Gestural Interfaces. *Computer Graphics*, 21(2):137–142, 1987.

[288] D. Riecken, editor. *Special Issue on Intelligent Agents*, volume 37 of *Communications of the ACM*, 1994.

[289] B. Rime and L. Schiaratura. Gesture and speech. In F. R. S. and R. B., editors, *Fundamentals of Nonverbal Behaviour*, pages 239–281. New York: Press Syndacate of the University of Cambridge, 1991.

[290] A. Risberg and J. L. Lubker. Prosody and speechreading. Quaterly Progress & Status Report 4, Speech Transmission Laboratory, KTH, Stockholm, Sweden, 1978.

[291] J. Robert. Integration audition-vision par reseaux de neurones: une etude comparative des modeles d'integration appliques la perception des voyelles. Technical report, Rapport de DEA Signal-Image-Parole, ENSER, Grenoble, France, 1991.

[292] J. Robert-Ribes, P. Escudier, and J. L. Schwartz. Modeles d'integration audition-vision: une etude neuromimetique. Technical report, ICP, 1991. Rapport Interne.

[293] G. G. Robertson, S. K. Card, and J. D. Mackinlay. The Cognitive Coprocessor Architecture for Interactive User Interfaces. *ACM*, pages 10–18, 1989.

[294] D. Salber and J. Coutaz. Applying the Wizard of Oz Technique to the Study of Multimodal Systems, 1993.

[295] V. J. Samar and D. C. Sims. Visual evoked responses components related to speechreading and spatial skills in hearing and hearing-impaired adults. *Journal of Speech & Hearing Research*, 27:162–172, 1984.

[296] B. Scharf. *Loudness*, chapter 6, pages 187–242. Academic Press, New York, 1978.

[297] T. Schiphorst et al. Tools for Interaction with the Creative Process of Composition. In *Proc. of the CHI '90*, pages 167–174, 1990.

[298] D. Schlichthärle. *Modelle des Hörens - mit Anwendungen auf die Hörbarkeit von Laufzeitverzerrungen (Models of hearing - applied to the audibility of arrival-time distortions)*. PhD thesis, Ruhr-Universität Bochum, 1980.

[299] E. M. Schmidt and J. S. McIntosh. Excitation and inhibition of forearm muscles explored with microstimulation ofprimate motor cortex during a trained task. In *Abstracts of the 9th Annual Meeting of the Society for Neuroscience*, volume 5, page 386, 1979.

[300] L. Schomaker et al. MIAMI — Multimodal Integration for Advanced Multimedia Interfaces. Annex i: Technical annex, Commission of the European Communities, December 1993.

[301] L. R. B. Schomaker. Using Stroke- or Character-based Self-organizing Maps in the Recognition of On-line, Connected Cursive Script. *Pattern Recognition*, 26(3):443–450, 1993.

[302] L. R. B. Schomaker and R. Plamondon. The Relation between Pen Force and Pen-Point Kinematics in Handwriting. *Biological Cybernetics*, 63:277–289, 1990.

[303] L. R. B. Schomaker, A. J. W. M. Thomassen, and H.-L. Teulings. A computational model of cursive handwriting. In R. Plamondon and M. L. Suen, C. Y. andSimner, editors, *Computer Recognition and Human Production of Handwriting*, pages 153–177. Singapore: World Scientific, 1989.

[304] J. Schröter. *Messung der Schalldämmung von Gehörschützern mit einem physikalischen Verfahren-Kunstkopfmethode (Measurement of the attenuation of personal hearing protectors by means of a physical technique - dummy-head method).* PhD thesis, Ruhr-Universität Bochum, 1983.

[305] J. Schröter. The Use of Acoustical Test Fixures for the Measurement of Hearing-Protector Attenuation, Part I: Review of Previous Work and the Design of an Improved Test Fixure. *Journ. Acoust. Soc. Am.*, 79:1065–1081, 1986.

[306] J. Schröter and C. Pösselt. The Use of Acoustical Test Fixures for the Measurement of Hearing-Protector Attenuation, Part II: Modeling the External Ear, Simulating Bone Conduction, and Comparing Test Fixure and Real-Ear Data. *Journ. Acoust. Soc. Am.*, 80:505–527, 1986.

[307] J. A. Scott Kelso. The process approach to understanding human motor behaviour: an introduction. In J. A. Scott Kelso, editor, *Human motor behaviour: an introduction*, pages 3–19. Lawrence Erlbaum Ass. Pub., Hillsdale NJ, 1982.

[308] G. M. Shepherd. *Neurobiology.* Oxford Univ. Press, 2nd edition edition, 1988.

[309] K. B. Shimoga. A Survey of Perceptual Feedback Issues in Dexterous Telemanipulation: Part II. Finger Touch Feedback. In *Proc. of the IEEE Virtual Reality Annual International Symposium.* Piscataway, NJ : IEEE Service Center, 1993.

[310] B. Shneiderman. *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* New York: Addison-Wesley, 1992.

[311] D. Silbernagel. Taschenatlas der Physiologie. Thieme, 1979.

[312] R. Simon. Pen Computing Futures: A Crystal Ball Gazing Exercise. In *Proc. of the IEEE Colloquium on Handwriting and Pen-based Input, Digest Number 1994/065*, page 4. London: The Institution of Electrical Engineers, March 1994.

[313] A. D. Simons and S. J. Cox. Generation of mouthshapes for a synthetic talking head. In *Proceedings of the Institute of Acoustics*, volume 12, pages 475–482, Great Britain, 1990.

[314] H. Slatky. *Algorithmen zur richtungsselektiven Verarbeitung von Schallsignalen eines binauralen Cocktail-Party-Prozessors (Algorithms for direction-selective processing of sound signals by means of a binaural cocktail-party processor).* PhD thesis, Ruhr-Universität Bochum, 1993.

[315] P. M. T. Smeele and A. C. Sittig. The contribution of vision to speech perception. In *Proceedings of 13th International Symposium on Human Factors in Telecommunications*, page 525, Torino, 1990.

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

[317] P. Smolensky. A proper treatment of connectionism. *Behavioural and Brain Sciences*, 11:1–74, 1988.

[318] H. E. Staal and D. C. Donderi. The Effect of Sound on Visual Apparent Movement. *American Journal of Psychology*, 96:95–105, 1983.

[319] L. Steels. Emergent frame recognition and its use in artificial creatures. In *Proc. of the Intl. Joint Conf. on Artificial Intelligence IJCAI-91*, pages 1219–1224, 1991.

[320] L. Steels. The artificial life roots of artificial intelligence. *Artificial Life*, 1(1-2):75–110, 1994.

[321] B. E. Stein and M. A. Meredith. *Merging of the Senses.* MIT Press, Cambridge, London, 1993.

[322] R. Steinmetz. *Multimedia-Technologie.* Springer-Verlag, 1993.

[323] K. Stevens. The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. D. Jr and P. B. Denes, editors, *Human communication: A unified view*, pages 51–66. McGraw-Hill, New-York, 1972.

[324] K. N. Stevens and J. S. Perkell. Speech physiology and phonetic features. In *Dynamic aspects of speech production*, pages 323–341. University of Tokyo Press, Tokyo, 1977.

[325] S. S. Stevens. On the Psychophysical Law. *Psychological Review*, 64:153–181, 1957.

[326] R. J. Stone. Virtual Reality & Telepresence – A UK Iniative. In *Virtual Reality 91 – Impacts and Applications. Proc. of the 1st Annual Conf. on Virtual Reality*, pages 40–45, London, 1991. Meckler Ltd.

[327] D. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. In *International Joint Conference of Neural Networks*, Baltimore, 1992.

[328] N. Suga. Auditory neuroethology and speech processing: complex-sound processing by combination-sensitive neurons. In G. M. Edelmann, W. Gall, and W. Cowan, editors, *Auditory Function: Neurobiological Bases of Hearing.* John Wiley and Sons, New York, 1988.

[329] J. W. Sullivan and S. W. Tyler, editors. *Intelligent User Interfaces.* ACM Press, Addison-Wesley, 1991.

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

[331] A. Q. Summerfield. Use of visual information for phonetic perception. *Phonetica*, 36:314–331, 1979.

[332] Q. Summerfield. Comprehensive account of audio-visual speech perception. In B. D. . R. Campbell, editor, *Hearing by eye: The psychology of lip-reading.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[333] Q. Summerfield. Visual perception of phonetic gestures. In G. Mattingly and M. Studdert-Kennedy, editors, *Modularity and the Motor Theory of Speech Perception.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1991.

[334] I. E. Sutherland. The ultimate display. In *Information Processing 1965, Proc. IFIP Congress*, pages 506–508, 1965.

[335] C. C. Tappert, C. Y. Suen, and T. Wakahara. The State of the Art in On-line Handwriting Recognition. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 12:787–808, 1990.

[336] L. Tarabella. Special issue on Man-Machine Interaction in Live Performance. In *Interface*, volume 22. Swets & Zeitlinger, Lisse, The Netherlands, 1993.

[337] D. Terzopoulos and K. Waters. Techniques for realistic facial modeling and animation. In N. Magnenat-Thalmann and D. Thalmann, editors, *Computer Animation'91*, pages 59–74. Springer-Verlag, 1991.

[338] H. L. Teulings and F. J. Maarse. Digital Recording and Processing of Handwriting Movements. *Human Movement Science*, 3:193–217, 1984.

[339] M. T. Turvey. Preliminaries to a theory of action with reference to vision. In R. Shaw and J. Bransford, editors, *Perceiving, Acting and Knowing: Toward an ecological psychology*, pages 211–265. Hillsdale, NJ: Erlbaum, 1977.

[340] T. Ungvary, S. Waters, and P. Rajka. NUNTIUS: A computer system for the interactive composition and analysis of music and dance. *Leonardo*, 25(1):55–68, 1992.

[341] Väänänen, K. and Böhm, K. *Gesture Driven Interaction as a Human Factor in Virtual Environments – An Approach with Neural Networks*, chapter 7, pages 93–106. Academic Press Ltd., 1993.

[342] T. van Gelderen, A. Jameson, and A. L. Duwaer. Text recognition in pen-based computers: An empirical comparison of methods. In *InterCHI '93 Conference Proceedings*, pages 87–88, Amsterdam, 1993.

[343] D. Varner. Olfaction and VR. In *Proceedings of the 1993 Conference on Intelligent Computer-Aided Training and Virtual Environment Technology, Houston, TX*, 1993.

[344] B. Verplank. Tutorial notes. In *Human Factors in Computing Systems, CHI'89*. New York: ACM Press, 1989.

[345] M.-L. Viaud. *Animation faciale avec rides d'expression, vieillissement et parole*. PhD thesis, Paris XI-Orsay University, 1993.

[346] K. J. Vicente and J. Rasmussen. Ecological Interface Design: Theoretical Foundations. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(4):589–606, Aug. 1992.

[347] P. Viviani and N. Stucchi. Motor-perceptual interactions. In J. Requin and G. Stelmach, editors, *Tutorials in Motor Behavior*, volume 2. Elsevier Science Publishers B. V., North-Holland, Amsterdam, 1991.

[348] J. H. M. Vroomen. *Hearing voices and seeing lips: Investigations in the psychology of lipreading.* PhD thesis, Katolieke Univ. Brabant, Sep. 1992.

[349] W. J. Wadman. *Control mechanisms of fast goal-directed arm movements.* PhD thesis, Utrecht University, The Netherlands, 1979. Doctoral dissertation.

[350] W. J. Wadman, W. Boerhout, and J. J. Denier van der Gon. Responses of the arm movement control system to force impulses. *Journal of Human Movement Studies*, 6:280–302, 1980.

[351] E. A. Wan. Temporal Back-propagation for FIR Neural Networks. In *Proc. Int. Joint Conf. on Neural Networks*, volume 1, pages 575–580, San Diego CA, 1990.

[352] J. R. Ward and M. J. Phillips. Digitizere Technology: Performance and the Effects on the User Interface. *IEEE Computer Graphics and Applications*, (April '87):31–44, 1987.

[353] D. H. Warren. Spatial Localization Under Conflicting Conditions: Is There a Single Explanation? *Perception and Psychophysics*, (8):323–337, 1979.

[354] D. H. Warren, R. B. Welch, and T. J. McCarthy. The role of visual-auditory compellingness in the ventriloquism effect: implications for transitivity among the spatial senses. *Perception and Psychophysics*, 30:557–564, 1981.

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

[356] K. Waters. Bureaucrat. Computer-generated movie, 1990.

[357] P. Weckesser and F. Wallner. Calibrating the Active Vision System KASTOR for Real-Time Robot Navigation. In J. F. Fryer, editor, *Close Range Techniques and Machine Vision*, pages 430–436. ISPRS Commision V, 1994.

[358] R. B. Welch and D. H. . Warren. *Handbook of perception and human performanceHandbook of perception and human performance*, chapter Chapter 25: Intersensory interactions. 1986.

[359] E. M. Wenzel. Spatial sound and sonification. In G. Kramer, editor, *Auditory Display*, pages 127–150, Reading, Massachusetts, 1994. Santa Fe Institute, Addison Wesley.

[360] D. Wessel. Improvisation with high highly interactive real-time performance systems. In *Proc. Intl. Computer Music Conference*, Montreal, Canada, 1991.

[361] W. A. Wickelgren. Context-sensitive coding, associative memory and serial order in speech behaviour. *Psycho. Rev.*, 76:1–15, 1969.

[362] N. Wiener. *Cybernetics: or control and communication in the animal and the machine.* New York: Wiley, 1948.

[363] L. Williams. Performance driven facial animation. *Computer Graphics*, 24(3):235–242, 1990.

[364] C. G. Wolf and P. Morrel-Samuels. The use of hand-drawn gestures for text editing. *Intl. Journ. on Man-Machine Studies*, 27:91–102, 1987.

[365] S. Wolf. *Lokalisation von Schallquellen in geschlossenen Räumen (Localisation of sound sources in enclosed spaces).* PhD thesis, Ruhr-Universität Bochum, 1991.

[366] P. Woodward. *Le speaker de synthese.* PhD thesis, ENSERG, Institut National Polytechnique de Grenoble, France, 1991.

[367] P. Woodward, T. Mohamadi, C. Benoît, and G. Bailly. Synthese partir du texte d'un visage parlant francais. In *Actes des 19emes Journees d'Etude sur la Parole*, Bruxelles, 1992. Groupe Communication Parlee de la SFA.

[368] M. Wooldridge and N. R. Jennings. Intelligent Agents: Theory and Practice. *(submitted to:) Knowledge Engineering Review*, 1995.

[369] R. H. Wurtz and C. W. Mohler. Organization of monkey superior colliculus enhanced visual response of superficial layer cells. *Journal of Neurophysiology*, 39:745–765, 1976.

[370] B. L. M. Wyvill and D. R. Hill. Expression control using synthetic speech. In *SigGraph '90 Tutorial Notes*, volume 26, pages 186–212, 1990.

[371] N. Xiang, , and J. Blauert. A Miniature Dummy Head for Binaural Evaluation of Tenth-Scale Acoustic Models. *Journ. Appl. Acoust.*, 33:123–140, 1991.

[372] N. Xiang. *Mobile Universal Measuring System for the Binaural Room-Acoustic-Model Technique.* PhD thesis, Ruhr-Universität Bochum, 1991.

[373] N. Xiang and J. Blauert. Binaural Scale Modelling for Auralization and Prediction of Acoustics in Auditoria. *Journ. Appl. Acoust.*, 38:267–290, 1993.

[374] B. P. Yuhas, M. H. Goldstein Jr., and T. J. Sejnowski. Integration of acoustic and visual speech signal using neural networks. *IEEE Communications Magazine*, pages 65–71, 1989.