HCI - MultiModal Systems - Theory 1

Artificial Intelligence

Jacopo Parretti

I Semester 2025-2026

Indice

I Multimodal Interfaces 3

1 Introduction to Multimodal Interfaces 3
1.1 Topics Covered . . . 3

2 Graphical User Interfaces (GUIs) and WIMP 3
2.1 The WIMP Paradigm . . . 3
2.2 Why WIMP Interfaces Are Prevalent . . . 3

3 Post-WIMP Interfaces 4
3.1 Limitations of WIMP . . . 4
3.2 What Are Post-WIMP Interfaces? . . . 4
3.3 Key Characteristics of Post-WIMP Interfaces . . . 4
3.4 Examples of Post-WIMP Interfaces . . . 5
3.5 Virtual Reality, Augmented Reality, and Mixed Reality . . . 5
3.6 Pervasive Computing . . . 6
3.7 Ubiquitous Computing . . . 6
3.8 Disappearing Computing . . . 6
3.9 Ambient Intelligence . . . 7

4 Theoretical Foundations of Post-WIMP Interfaces 7
4.1 Interdisciplinary Grounding . . . 7
4.2 Key Principles Exploited by Post-WIMP Interfaces . . . 7

5 Multimodal Interfaces and Systems 8
5.1 Understanding Multimodal Interfaces . . . 8
5.2 What is Multimodal? . . . 8
5.3 What is Modality? . . . 9
5.4 What is Multimodality? . . . 9
5.5 Multimodality vs. Multimedia . . . 10
5.6 Formal Definition of Multimodal Systems . . . 10
5.7 Input and Output in Multimodal Systems . . . 11
5.8 Examples of Multimodal Systems . . . 12

6 Motivations for Multimodality 13
6.1 Human-Human Communication is Multimodal . . . 14
6.2 Input and Output by the Most Effective Means . . . 14
6.3 Adapting to the Environment . . . 14
6.4 Task Performance and User Preference . . . 15

7 Frameworks for Multimodal Systems 16
7.1 Understanding Frameworks . . . 16
7.2 Types of Frameworks . . . 16
7.3 Example Framework: Layered Architecture . . . 16

8 19

Parte I

Multimodal Interfaces

1 Introduction to Multimodal Interfaces

This section covers the fundamentals of multimodal interfaces, including their history, definitions, and the motivations behind their development. We will explore the differences between multimedia and multimodal systems, and examine various communication paradigms.

1.1 Topics Covered

The main topics we will discuss include:

- Multimodal interfaces and their historical development
- What constitutes multimodal systems
- Understanding modality in human-computer interaction
- Multimedia vs. multimodal systems
- Motivations for multimodality
- Verbal and nonverbal communications
- Social signal processing and affective computing
- Applications and practical examples

2 Graphical User Interfaces (GUIs) and WIMP

2.1 The WIMP Paradigm

The most traditional, consolidated, commonly used, and widespread interfaces are **graphical user interfaces** (GUIs). These interfaces adopt the so-called **WIMP paradigm**, which stands for:

WIMP Paradigm

Windows • **I**cons • **M**enus • **P**ointing devices

Definizione 2.1 (WIMP Paradigm)**.** The WIMP paradigm is a user interface design approach that uses windows to organize content, icons to represent objects and actions, menus to provide options, and pointing devices (such as a mouse) for interaction.

2.2 Why WIMP Interfaces Are Prevalent

WIMP interfaces have become so prevalent because they offer several advantages:

- **Good at abstracting workspaces and documents**: They provide a clear visual metaphor for organizing information
- **Analogous to physical documents**: Windows and icons are analogous to documents as paper sheets or folders, making them intuitive for users
- **Rectangular regions on 2D flat screens**: Their basic representations as rectangular regions on a 2D flat screen make them a good fit for system programmers

- **Suitable for multitasking**: They are well-suited for multitasking work environments, allowing users to work with multiple applications simultaneously

Esempio 2.1. A typical WIMP interface includes:

- Multiple overlapping windows showing different applications
- A menu bar with dropdown options
- Icons representing files, folders, and applications
- A pointer (cursor) controlled by a mouse or trackpad

3 Post-WIMP Interfaces

3.1 Limitations of WIMP

While WIMP interfaces are excellent for many tasks, they are not optimal for all scenarios. Specifically:

- **Not optimal for complex tasks**: WIMP interfaces struggle with complex tasks such as computer-aided design
- **Limited for natural interaction paradigms**: Applications needing more natural interaction paradigms, such as interactive games, require different approaches

3.2 What Are Post-WIMP Interfaces?

Post-WIMP interfaces, as introduced by Van Dam in 1997, aim at overcoming the limitations of traditional WIMP interfaces.

Definizione 3.1 (Post-WIMP Interfaces)**.** Post-WIMP interfaces consist of **widgetless user interfaces**, including:

- **Virtual reality systems**
- User interfaces based on **gestures**
- User interfaces based on **speech**
- User interfaces based on **physical controls**

3.3 Key Characteristics of Post-WIMP Interfaces

Post-WIMP interfaces have two fundamental characteristics:

Key Characteristics of Post-WIMP

1. Multiple sensory channels (input)
2. Multimedia output (engaging multiple senses)

1. **Multiple sensory channels**: They integrate input from **several sensory channels**, allowing for richer interaction
2. **Multimedia output**: They produce **multimedia output**, engaging multiple senses simultaneously

Osservazione. The shift from WIMP to Post-WIMP represents a move towards more natural and intuitive human-computer interaction, leveraging the full range of human sensory and motor capabilities.

3.4 Examples of Post-WIMP Interfaces

Post-WIMP interfaces are increasingly common in modern technology:

- **Wall-size displays**: Large interactive displays used in collaborative environments
- **Multi-touch displays**: Touchscreens that can detect multiple simultaneous touch points, enabling gestures like pinch-to-zoom
- **Vehicle applications**: In-car infotainment systems with touchscreens and voice control
- **Public kiosks**: Interactive information terminals in public spaces

Esempio 3.1. Modern smartphones exemplify Post-WIMP interfaces by combining:

- Multi-touch gestures (swipe, pinch, tap)
- Voice assistants (Siri, Google Assistant)
- Accelerometers and gyroscopes for motion-based interaction
- Haptic feedback for tactile responses

Nota. The evolution from WIMP to Post-WIMP interfaces reflects the changing needs of users and the expanding capabilities of technology. As computing devices become more diverse (smartphones, tablets, wearables, AR/VR headsets), the rigid WIMP paradigm becomes less suitable, and more flexible, multimodal approaches become necessary.

3.5 Virtual Reality, Augmented Reality, and Mixed Reality

Post-WIMP interfaces encompass various reality technologies that blend the physical and digital worlds in different ways:

Reality Technologies

VR: Fully artificial environment
AR: Virtual objects overlaid on real world
MR: Virtual + real world combined

Definizione 3.2 (Virtual Reality (VR))**. Virtual Reality** provides a fully artificial environment where users are completely immersed in a virtual world. Users experience full immersion in a virtual environment, typically through head-mounted displays (HMDs) and controllers.

Definizione 3.3 (Augmented Reality (AR))**. Augmented Reality** overlays virtual objects on the real-world environment. The real world is enhanced with digital objects, allowing users to see both simultaneously.

Definizione 3.4 (Mixed Reality (MR))**. Mixed Reality** combines virtual environments with the real world, allowing users to interact with both the real world and the virtual environment simultaneously. Virtual and physical objects coexist and interact in real-time.

Technology	**Characteristics**
Virtual Reality (VR)	Fully artificial environment; Full immersion in virtual environment
Augmented Reality (AR)	Virtual objects overlaid on real-world environment; The real world enhanced with digital objects
Mixed Reality (MR)	Virtual environment combined with real world; Interact with both the real world and the virtual environment

Tabella 1: Comparison of VR, AR, and MR technologies

3.6 Pervasive Computing

Definizione 3.5 (Pervasive Computing)**.** Pervasive Computing involves integrating computation into everyday objects and activities, making computing accessible anytime and anywhere.

Key aspects of pervasive computing include:

- **Computers exiting everywhere**: Embedded into fridges, washing machines, door locks, cars, furniture, and even people
- **Mobile** portable computing devices
- **Wireless** communication
- Examples include wearable health trackers, smart home devices, and IoT sensors

3.7 Ubiquitous Computing

Definizione 3.6 (Ubiquitous Computing)**.** Ubiquitous Computing is a vision where computing devices are seamlessly integrated into the environment and used naturally without conscious thought.

Characteristics of ubiquitous computing:

- **More user and application-driven** compared to Pervasive computing
- Both keywords are often used interchangeably
- Examples: Smart home devices, like lights and thermostats, that automatically adjust based on your preferences without needing direct control

Osservazione. While pervasive computing focuses on the physical distribution of computing devices, ubiquitous computing emphasizes the seamless and invisible integration of these devices into our daily lives.

3.8 Disappearing Computing

Definizione 3.7 (Disappearing Computing)**.** Disappearing Computing is the idea that technology becomes so embedded and integrated into the environment that it effectively "disappears" from the user's perception.

Examples of disappearing computing include:

- **Smart fabrics and wearables**: Clothing with embedded sensors and computing capabilities
- **Voice-activated assistants**: Such as Alexa, which respond to natural language commands
- **Smart buildings**: Environments that adapt automatically to occupants' needs
- **Augmented reality contact lenses**: Technology so integrated it becomes part of the user's vision

Nota. Disappearing computing represents the ultimate goal of ubiquitous computing: technology that is so well-integrated into our environment that we interact with it naturally and unconsciously, without being aware of the underlying computational systems.

3.9 Ambient Intelligence

Definizione 3.8 (Ambient Intelligence)**.** Ambient Intelligence refers to environments that are sensitive and responsive to the presence of people, utilizing embedded systems and artificial intelligence to enhance user experience.

Key features of ambient intelligence:

- **Sensitive environments**: Detect and respond to human presence and behavior
- **Embedded systems**: Computing integrated throughout the environment
- **Artificial intelligence**: Systems that learn and adapt to user preferences
- **Enhanced user experience**: Seamless interaction without explicit commands

Esempio 3.2. Examples of ambient intelligence include:

- Automated lighting that adjusts based on occupancy and time of day
- Climate control systems that learn user preferences
- Traffic lights that adapt in real-time to traffic conditions
- Cars that communicate with each other to improve safety and traffic flow
- Smart homes that anticipate user needs based on patterns and context

4 Theoretical Foundations of Post-WIMP Interfaces

4.1 Interdisciplinary Grounding

Post-WIMP interfaces are grounded on models and theories from multiple disciplines:

- **Psychology**: Understanding human perception, cognition, and behavior
- **Physiology**: How the human body processes sensory information
- **Biomechanics**: Human movement and physical interaction
- **Neurosciences**: Brain function and neural processing
- **Cognitive sciences**: Mental processes and information processing
- **Social sciences**: Human social behavior and communication

4.2 Key Principles Exploited by Post-WIMP Interfaces

Post-WIMP interfaces leverage several fundamental human capabilities:

7 Key Principles

Human sensory system • Multiple sensory channels • Sensory interconnections
Non-verbal communication • Affect & emotion • Social signals • Art & human sciences

1. **Exploit the human sensory system**: Utilizing vision, hearing, touch, and other senses
2. **Exploit multiple sensory channels**: Engaging several senses simultaneously for richer interaction
3. **Exploit the deep interconnections between sensory channels**: Leveraging how different senses work together (e.g., audio-visual integration)

4. **Exploit non-verbal communication**: Using gestures, facial expressions, body language
5. **Exploit affect and emotion**: Recognizing and responding to emotional states
6. **Exploit social signals**: Understanding social cues and interpersonal dynamics
7. **Exploit theories from art and human sciences**: Drawing from music, choreography, theatre, and cinema to create engaging experiences

Osservazione. The effectiveness of Post-WIMP interfaces stems from their ability to align with natural human capabilities and behaviors, rather than forcing users to adapt to rigid computational paradigms. This human-centered approach makes interactions more intuitive, efficient, and satisfying.

Nota. By integrating knowledge from diverse fields, Post-WIMP interfaces create more natural and holistic interaction experiences. This interdisciplinary approach is essential for developing systems that truly understand and respond to human needs, emotions, and social contexts.

5 Multimodal Interfaces and Systems

5.1 Understanding Multimodal Interfaces

Post-WIMP interfaces are often **multimodal interfaces**, meaning they exploit multiple **sensory modalities**.

Multimodal = Multiple Sensory Channels

Integration of information from several different sensory modalities

Definizione 5.1 (Sensory Modality)**.** A **sensory modality** is the sensory channel through which information is perceived. It refers to the type of communication channel used for transferring or acquiring information.

Definizione 5.2 (Multimodal)**.** In our context, **multimodal** refers to the integration of information from several different **sensory channels**.

Osservazione. Multimodal interfaces leverage the human ability to process information through multiple senses simultaneously, creating richer and more natural interaction experiences.

5.2 What is Multimodal?

The term "multimodal" can be understood in different ways depending on the context:

5.2.1 Multimodal Distribution

In a statistical or mathematical context, **multimodal distribution** refers to:

- Multiple modes (i.e., peaks) in a probability density function
- A distribution with several local maxima

Esempio 5.1. A histogram showing multiple peaks represents a multimodal distribution, where different groups or clusters in the data create distinct modes.

5.2.2 Multimodal in HCI Context

In the context of human-computer interaction, "multimodal" specifically refers to the use of multiple sensory channels for communication and interaction.

5.3 What is Modality?

Definizione 5.3 (Modality)**. Modality** refers to the way in which something is expressed or perceived.

Modalities can be understood at different levels of abstraction:

5.3.1 Raw Modalities vs. Abstract Modalities

- **Raw Modalities** (closest from sensor):
  - Direct sensory input from physical sensors
  - Examples: Speech signal, Image data
- **Intermediate Modalities**:
  - Processed information from raw data
  - Examples: Language (from speech), Detected objects (from images)
- **Abstract Modalities** (farthest from sensor):
  - High-level interpretations and semantic information
  - Examples: Sentiment intensity, Object categories

Esempio 5.2. Consider processing visual information:

1. **Raw modality**: Image pixels captured by a camera
2. **Intermediate modality**: Detected objects in the image (e.g., "person", "car")
3. **Abstract modality**: Object categories and their relationships (e.g., "pedestrian crossing street")

5.4 What is Multimodality?

Definizione 5.4 (Multimodality)**. Multimodality** is related to sensory modalities and encompasses:

- Touch
- Speech
- What you see (visual attention)
- Hearing
- Taste
- Smell

The Five Human Senses

Vision • Hearing • Touch • Taste • Smell

The five primary human senses form the basis of multimodal interaction:

- **Vision**: Visual perception and attention
- **Hearing**: Auditory information processing
- **Touch**: Tactile and haptic feedback
- **Taste**: Gustatory perception (less common in HCI)

- **Smell**: Olfactory perception (emerging in HCI research)

Osservazione. While vision, hearing, and touch are the most commonly exploited modalities in current HCI systems, emerging technologies are beginning to incorporate taste and smell for more immersive experiences.

5.5 Multimodality vs. Multimedia

It is important to distinguish between **multimodality** and **multimedia**, as these terms are often confused:

Definizione 5.5 (Multimodality vs. Multimedia). • **Multimodality**:

  - *Modality* refers to a certain type of information and/or the representation of the information
  - Focuses on **sensory modalities** (vision, hearing, touch, etc.)
  - Concerns how information is perceived and processed by humans

- **Multimedia**:
  - *Medium* is the instrument for storing or communicating information
  - Focuses on the **technical medium** used to convey information
  - Concerns the format and delivery mechanism of content

Esempio 5.3. A TV show is a **medium** that uses **auditory** and **visual modalities**:

- The **medium** is television (the instrument for communicating)
- The **modalities** are auditory (sound) and visual (images)

Aspect	**Multimodality**	**Multimedia**
Focus	Sensory channels	Communication instruments
Definition	Type of information/representation	Medium for storing/communicating
Examples	Vision, hearing, touch, smell, taste	TV, radio, internet, print
In HCI	How users perceive information	How systems deliver information

Tabella 2: Comparison between Multimodality and Multimedia

Nota. Understanding the distinction between multimodality and multimedia is crucial in HCI:

- **Multimedia** refers to the technical infrastructure (e.g., video, audio, text files)
- **Multimodality** refers to the human sensory experience (e.g., seeing, hearing, touching)

A multimodal interface uses multiple sensory channels for interaction, while a multimedia system uses multiple media formats for content delivery. Modern systems often combine both approaches.

5.6 Formal Definition of Multimodal Systems

Definizione 5.6 (Multimodal Systems - W3C Definition). Multimodal systems are "systems that support a user communicating with an application by using **different modalities** such as **voice** (in a human language), **gesture**, **handwriting**, **typing**, **audio-visual speech**, etc."

Source: W3C Multimodal Interaction Working Group, Multimodal Interaction Requirements, W3C NOTE 8 January 2003

Definizione 5.7 (Multimodal HCI System)**.** A multimodal HCI system is simply one that responds to inputs in more than **one modality** or **communication channel** (e.g., speech, gesture, writing, and others).

Osservazione. The key characteristic of multimodal systems is their ability to accept and process input through multiple channels, providing users with flexibility in how they interact with the system.

5.7 Input and Output in Multimodal Systems

5.7.1 User Input and Output

In multimodal systems, interaction is bidirectional:

- **User provides input** in one or more modalities
- **User receives output** in one or more modalities

This bidirectional communication allows for rich, natural interaction patterns that mirror human-to-human communication.

5.7.2 Classification of Input Modalities

Input in multimodal systems may be classified into three categories:

Three Types of Input

Sequential • Simultaneous • Composite

Definizione 5.8 (Sequential Input)**. Sequential** input is received on a single modality, though that modality can change over time.

Esempio 5.4. A user first speaks a command ("Open the file"), then switches to typing the filename. The modalities are used one at a time, in sequence.

Definizione 5.9 (Simultaneous Input)**. Simultaneous** input is received on multiple modalities, and treated separately.

Esempio 5.5. A user speaks while also using hand gestures, but the system processes speech and gestures as independent input streams without combining them.

Definizione 5.10 (Composite Input)**. Composite** input is received on multiple modalities at the same time and treated as a single, integrated "composite" input.

Esempio 5.6. A user points at a map (gesture) while saying "zoom in here" (speech). The system combines both inputs to understand that the user wants to zoom in at the specific location being pointed to. The meaning emerges from the integration of both modalities.

Input Type	**Characteristics**
Sequential	Single modality at a time; modality can change over time
Simultaneous	Multiple modalities used at once; processed separately
Composite	Multiple modalities used at once; integrated into single input

Tabella 3: Classification of input modalities in multimodal systems

5.7.3 Multimodal Output

The output generated by a multimodal system can take various forms:

- **Audio**: Spoken feedback, sounds, music
- **Visual**: Graphics, text, animations, video
- **Haptic feedback**: Vibrations, force feedback, tactile sensations
- **Lighting**: Visual cues through ambient or directed lighting
- And other forms of sensory output

Osservazione. Multimodal output allows systems to communicate with users through the most appropriate channel(s) for the context, enhancing comprehension and user experience.

5.8 Examples of Multimodal Systems

Multimodal systems perform several key functions to enable natural interaction:

5.8.1 Observing and Analyzing Users

Multimodal systems excel at:

- **Gathering** information from several modalities
- **Analyzing** the collected multimodal data
- **Integrating** information from different sources

5.8.2 Building Internal Representations

Multimodal systems build internal representations of the user(s), for example, in terms of:

- **Cognitive states**: Goals, beliefs, intentions, attention
- **Emotional states**: Mood, emotion, stress levels, engagement

Esempio 5.7 (Virtual Therapist System)**.** A virtual therapist system tracks a user's:

- Facial expressions
- Tone of voice
- Body language

The system uses these multiple modalities to assess the user's emotional state. It builds an internal model to understand their stress levels and mental health, tailoring responses and suggestions accordingly.

5.8.3 Generating Real-Time Multimedia Feedback

Multimodal systems generate real-time multimedia feedback for the user(s), based on:

- Analysis of the input
- Internal models of the user
- The tasks at hand

Esempio 5.8 (Smart Fitness Coach App)**.** A smart fitness coach app analyzes a user's exercise performance through:

- Video (analyzing form and movement)
- Sensor data (heart rate, speed, repetitions)

It provides real-time feedback through:

- **Audio prompts**: "Straighten your back"
- **Visual cues**: Highlighting correct form on screen
- **Haptic vibrations**: Alerting to incorrect posture

This multimodal feedback helps correct form and enhance performance in real-time.

5.8.4 Providing Natural Interfaces for Complex Tasks

Multimodal systems provide users with a multimodal interface to the machine for the execution of complex tasks needing natural interfaces.

Esempio 5.9 (Smart Home Assistant)**.** A smart home assistant allows users to control their home environment using:

- **Voice commands**: "Set the temperature to 22 degrees"
- **Touchscreens**: Tapping controls on a display
- **Gestures**: Waving to turn on lights

Users can adjust lighting, set thermostats, and control entertainment systems through a natural, integrated interface. The system accepts input through whichever modality is most convenient for the user at that moment.

Nota. These examples demonstrate how multimodal systems enhance user experience by:

1. Providing multiple ways to interact, increasing accessibility and convenience
2. Understanding users more deeply through multiple information channels
3. Responding through the most appropriate output modalities for the context
4. Enabling more natural, human-like interaction patterns

The power of multimodal systems lies in their ability to combine information from multiple sources to create a richer understanding of user intent and context, leading to more intelligent and responsive interactions.

6 Motivations for Multimodality

Understanding why multimodal interfaces are important helps us design better human-computer interaction systems. There are several key motivations for adopting multimodality in interface design.

Four Key Motivations

1. Human-human communication is multimodal
2. Input/output by the most effective means
3. Adapting to the environment
4. Task performance and user preference

6.1 Human-Human Communication is Multimodal

Osservazione. Human-human communication is inherently multimodal. Unimodal communication is an artifact of communication technology, not natural human behavior.

When people communicate with each other, they naturally use:

- Speech and language
- Facial expressions
- Hand gestures
- Body language
- Eye contact and gaze
- Tone of voice and prosody

Nota. The restriction to single modalities (e.g., text-only chat, voice-only phone calls) is a limitation imposed by technology, not a natural preference. As technology advances, interfaces should move toward supporting the multimodal nature of human communication.

6.2 Input and Output by the Most Effective Means

Different types of information are best expressed through different modalities. Multimodal interfaces allow users to choose the most effective means for each type of communication.

6.2.1 Non-Verbal Modalities for Spatial Information

Certain kinds of content are most easily expressed in specific **non-verbal modalities**.

Esempio 6.1 (Spatial Information)**.** Consider giving directions or describing a region:

- **Visual/gestural**: Drawing the borders of a region on a map is intuitive and precise
- **Verbal**: Describing "the left bank of the river" verbally can be ambiguous or cumbersome

Example: Drawing the borders of a region on a map is much more natural than trying to describe the boundaries verbally.

6.2.2 Verbal Communication for Abstract Concepts

Some information is better suited to **verbal communication**.

Esempio 6.2 (Abstract Descriptions)**.** Describing location with phrases like "the left bank of the river" is more natural verbally than trying to convey the same information through gestures alone.

Osservazione. The principle of using the most effective modality for each type of information leads to more efficient and natural interactions. Users can seamlessly switch between modalities based on what they need to communicate.

6.3 Adapting to the Environment

Multimodal interfaces enable rapid adaptation to changing environments, by switching to the most suitable modality or by complementing different modalities.

6.3.1 Adaptation to Physical Environment

Environmental conditions can make certain modalities more or less effective:

- **Ambient noise**: In noisy environments, visual or haptic feedback may be more effective than audio
- **Darkness/brightness**: Lighting conditions affect the usability of visual interfaces
- **Hands-free situations**: When hands are occupied (e.g., driving, cooking), voice commands become essential

Esempio 6.3. A smartphone can adapt to a noisy environment by switching from audio alerts to vibration (haptic feedback) or visual notifications.

6.3.2 Adaptation to Social Environment

Social context also influences modality preferences:

- **Single user vs. multiple users**: Interfaces may need to support individual or collaborative interaction
- **Social interaction**: In public spaces, users may prefer silent input methods (typing, gestures) over voice commands
- **Collaborative applications**: Interfaces for group work need to support multiple simultaneous users

Esempio 6.4. In a quiet library, a user might prefer typing or using touch gestures rather than speaking to a voice assistant. In contrast, while driving alone, voice commands are the safest option.

Nota. The ability to adapt to environmental changes is crucial for creating robust, usable systems. Multimodal interfaces provide flexibility that unimodal systems cannot match, ensuring that users can interact effectively regardless of their context.

6.4 Task Performance and User Preference

6.4.1 Empirical Evidence

Many empirical studies have demonstrated that multimodal interfaces improve task performance and are preferred by users over unimodal interfaces.

Osservazione. Research consistently shows that multimodal interaction leads to:

- Faster task completion
- Higher accuracy
- Greater user satisfaction
- Reduced cognitive load

6.4.2 Research Findings

Several landmark studies have demonstrated the advantages of multimodal interfaces:

- **Oviatt (1996)**: Clear advantages over unimodal speech for map-based tasks
- **Cohen et al. (1998)**: Multimodal interfaces were faster than GUI for map-based tasks

- **Nishimoto et al. (1995)**: Multimodal interfaces were faster than GUI for drawing applications
- **Hauptmann (1989)**: User preference for speech and gesture in object manipulation tasks

Nota. These studies provide strong empirical support for multimodal design. The benefits are not merely theoretical—they translate into measurable improvements in real-world tasks. This evidence should guide interface designers toward multimodal solutions when appropriate.

7 Frameworks for Multimodal Systems

7.1 Understanding Frameworks

Frameworks and conceptual models have been proposed for multimodal systems to help organize and understand their structure.

Definizione 7.1 (Frameworks vs. Architectures)**.** Frameworks are **not architectures**. Rather, they are a level of abstraction above an architecture.

- Frameworks do not indicate how components are allocated to hardware devices and the communication among devices
- They provide conceptual organization and design principles
- They help understand the flow of information and processing stages

7.2 Types of Frameworks

Frameworks have been developed for different types of multimodal systems:

- **Verbal communication frameworks**: Focusing on multimodal systems that emphasize speech and language
- **Non-verbal communication frameworks**: Focusing on multimodal systems that emphasize gestures, expressions, and other non-verbal modalities

Osservazione. In this course, we will be focusing on **nonverbal communication** frameworks, as they are particularly relevant for understanding modern multimodal interfaces that leverage gestures, facial expressions, and body language.

7.3 Example Framework: Layered Architecture

A typical framework for multimodal systems uses a layered architecture that processes information at different levels of abstraction:

Four-Layer Framework

Layer 1: Physical Signals & Virtual Sensors
Layer 2: Low-Level Features
Layer 3: Mid-Level Features
Layer 4: Qualities Communication (Emotions & Social Signals)

7.3.1 Layer 1: Physical Signals and Virtual Sensors

The lowest layer deals with raw sensory input:

- **Physical sensors**: Hardware devices that capture raw data
- **Virtual sensors**: Software abstractions of sensor data

Esempio 7.1 (RGB-D Sensor)**.** Layer 1 might include an RGB-D sensor such as Kinect, which provides:

- 3D trajectories of specific body parts
- The silhouette of the tracked bodies
- Captured depth image
- RGB color image

This raw sensor data forms the foundation for higher-level processing.

7.3.2 Layer 2: Low-Level Features

The second layer extracts basic features from the raw sensor data:

- **Sub-layered processing**: From physical space to model spaces
- Feature extraction and preprocessing
- Noise reduction and normalization

Esempio 7.2. From the Kinect depth data, Layer 2 might extract:

- Joint positions and angles
- Body pose estimation
- Hand positions and orientations

7.3.3 Layer 3: Mid-Level Features

The third layer processes features at an intermediate level of abstraction:

- **Points or trajectories in multidimensional spaces** (amodal)
- Temporal patterns and sequences
- Spatial relationships

Esempio 7.3. Layer 3 might track:

- Hand movement trajectories over time
- Velocity and acceleration of gestures
- Relative positions of body parts

7.3.4 Layer 4: Qualities Communication

The highest layer interprets the meaning of the processed data:

- **Nonverbal emotions and social signals**
- High-level semantic interpretation
- User intent and affective states

Esempio 7.4. Layer 4 might recognize:

- Specific gestures (e.g., "wave", "point", "swipe")
- Emotional states (e.g., "frustrated", "engaged", "confused")

- Social signals (e.g., "greeting", "agreement", "attention")

Nota. This layered framework illustrates how multimodal systems progressively transform raw sensor data into meaningful interpretations. Each layer builds upon the previous one, moving from low-level physical signals to high-level semantic understanding. This abstraction is crucial for:

1. **Modularity**: Each layer can be developed and tested independently
2. **Reusability**: Lower layers can support multiple higher-level interpretations
3. **Clarity**: The framework makes the processing pipeline explicit and understandable

Understanding these frameworks helps designers and developers create more effective multimodal systems by providing a clear conceptual structure for organizing complex processing pipelines.

8

Aggiungere appunti fino alla fine del primo pdf 1 intro.pdf