

Human-Machine Interaction

Interaction Modalities &
Technologies

Multimodality

Dr. Patrick Chan
patrickchan@scut.edu.cn

South China University of Technology

1

Multimodality

Human Sensory Integration

- Humans **naturally integrate multiple senses**
- Multimodal interaction **aligns with natural perception processes**
- **Well-designed** multimodal systems **feel intuitive and robust**



2

Why Not Single Modality?

- **Single-modality is fragile**
 - Bright sunlight reduces screen visibility
 - Loud environments reduce speech clarity
 - Physical movement limits visual attention
- **Multimodal systems adapt to environmental constraints**



Why Not Single Modality?



Multimodal Fusion

- **Combining** information from multiple channels to support perception, decision, and action
- There are two types:
 - **Complementary**
 - Each modality provides **different** information
 - Increases richness
 - Example: **visual** path + **auditory** warning
 - **Redundant**
 - **Same** critical information delivered via multiple channels
 - Increases safety
 - Example: **flashing red** + **alarm** sound

5

Challenges

- **Data Fusion**
 - **Synchronize** different sensor **rates**
Example: 30 fps camera + 100 Hz IMU + 16 kHz audio
- **Conflict Resolution**
 - Handle **contradictory** inputs
Example: wave (gesture) vs “Stop” (voice)
- **Power Consumption**
 - Running vision, speech, and gaze AI together drains mobile **battery**

6

Complementarity

- **Combining multiple modalities** allows one channel to **compensate** for the **limitations** of another
- Example



- **Vision**

- **Pro:** Spatial information
Persistent display
High data bandwidth
- **Con:** Overload
Occlusion
Limited field of view

- **Audio**

- **Pro:** Temporal alerts
Attention capture
Directional cues
- **Con:** Noise
Desensitization
Limited capacity

7

Redundancy

- **Critical information** can be presented through **different channels**
 - If **one** channel **fails**, **another** remains **active**
- Redundancy **increases:**
 - **Detection** probability
 - **Reliability**
 - Reaction **speed**



8

Design Principle

• Temporal Synchronization

- Timing alignment is critical
- E.g. **Confusion** when
 - Visual alert appears **before** sound
 - Sound triggers **without** visual confirmation

• Conflict Resolution

- Multimodality disagrees
- System must:
 - Detect inconsistencies
 - Prioritize more reliable modality
 - Provide clear clarification



9

Design Principle

• Modality Dominance

- Dominance should match urgency
- Maintain **balanced** awareness in normal situations

• Reliability-Based Weighting

- Not all modalities are equally reliable
- Emphasize more reliable channel
- Adaptive reliability improves robustness



10

Design Principle

• Attention Management

- Multi-signal **competes** for **limited attention**
- Should:
 - **Avoid** simultaneous **overload**
 - Sequence **signals logically**
 - **Escalate gradually**
 - **Prevent** **startle effects**

• Adaptive Modality Switching

- **Adapt modality** based on:
 - **Task phase**
 - **User workload**
 - **Environmental condition**
 - **Risk level**
- **Example:**
 - **Low risk:** Visual only
 - **High risk:** Visual + audio

11

Failure Handling

• Failure-aware fusion increases safety

- When one modality **fails:**
 - **Detect** the **failure**
 - **Notify** the user
 - **Increase reliance** on **remaining modalities**

• **Never** leave users **unaware** of **degraded capability**

12

Vision + Audio

- Combining vision and audio is common in HMI
 - **Vision** usually provides the main information continuously
 - **Audio** is often used for alerts, or urgent notifications
- E.g. Driving Interface
 - **Vision**: speed and map shown on the screen
 - **Audio**: warning sound when risk is detected
 - Audio supports vision by drawing attention to urgent events



13

Vision + Speech

- **Vision** shows the user what is available or happening
- **Speech** lets the user interact naturally based on what they see
- E.g. In-Car Navigation System
 - **Vision**: the screen shows the map and nearby gas stations
 - **Speech**: the driver gives the command by voice
 - Links the spoken request to the visual information on the screen and selects the correct destination



14

Gesture + Vision

- **Gaze** → **Target** selection (Which object?)
- **Gesture** → **Action** execution (What to do?)
- **Example**
Look at a smart lamp → swipe up → brighten light
- **Benefit**
 - System only **listens to gestures aimed** at **selected** object
 - Greatly **reduces false positives**



15

Gesture + Speech

- **Semantic Fusion**
 - Combine **gesture + speech** meaning
Example: “Put that” + pointing → object
“there” + pointing → location
- **Error Correction**
 - **Voice confirms gesture** action
Example: “Did you mean to delete? Say Yes.”
- **Contextual Clues**
 - **Speech sets interaction** mode
Example: saying “Design Mode” → hand becomes 3D brush



16

Gesture + Haptics

- **Passive Haptics**

- Use **real surfaces** for tactile **cues**
Example:
Gesture ends on desk → feel “stop”

- **Ultrasound Mid-Air Haptics**

- **Acoustic pressure** creates **virtual touch** in air
- **Simulates pressing** a floating button

- **Wearable Haptics**

- Devices provide **vibration feedback**
Example: ring vibrates when gesture recognized

