Building Interactive Systems

Multimodal Interaction

Professor Bilge Mutlu | Spring 2023

What will we cover today?

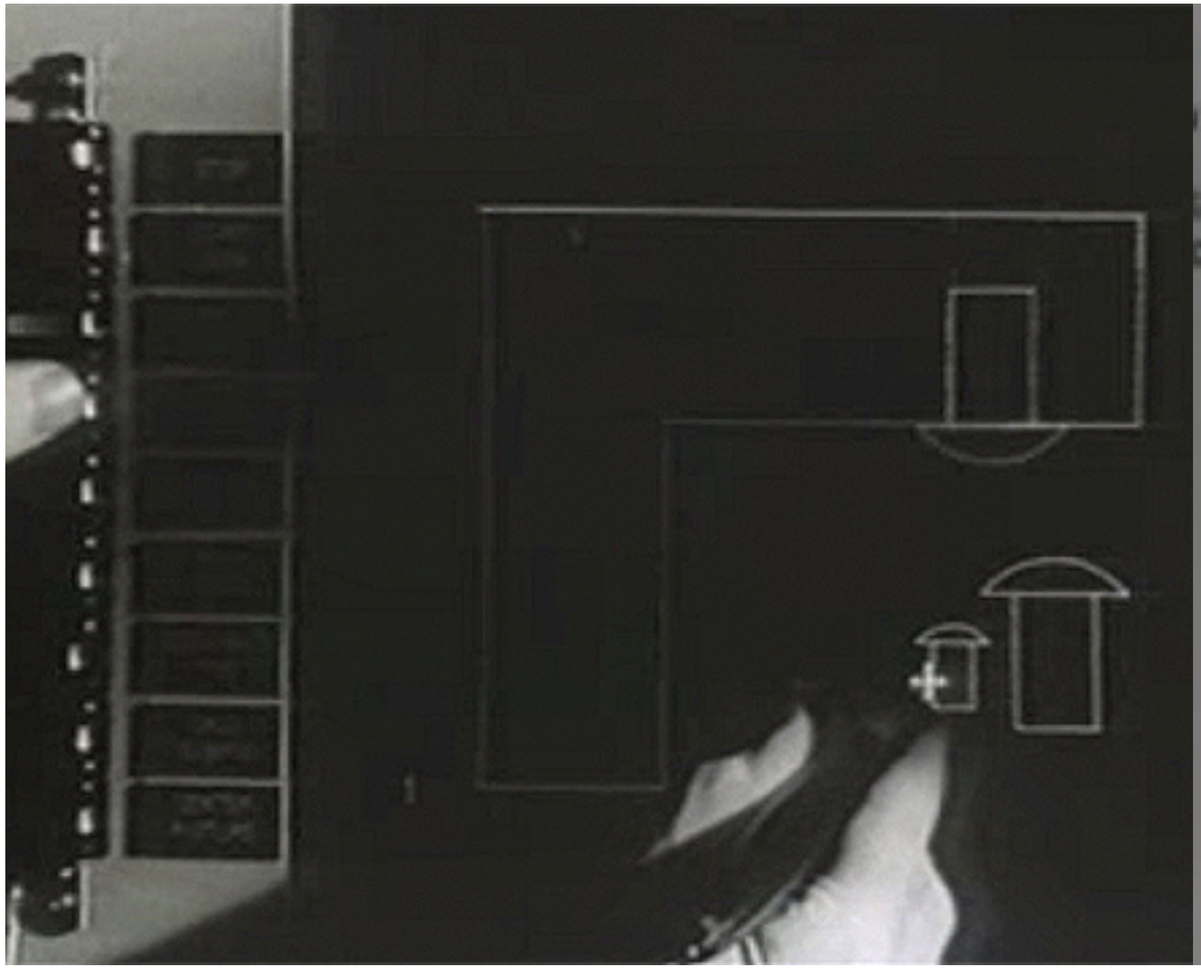
→ **What** is multimodal interaction?

→ **Elements** of multimodal interfaces

→ Multimodal system **architectures**

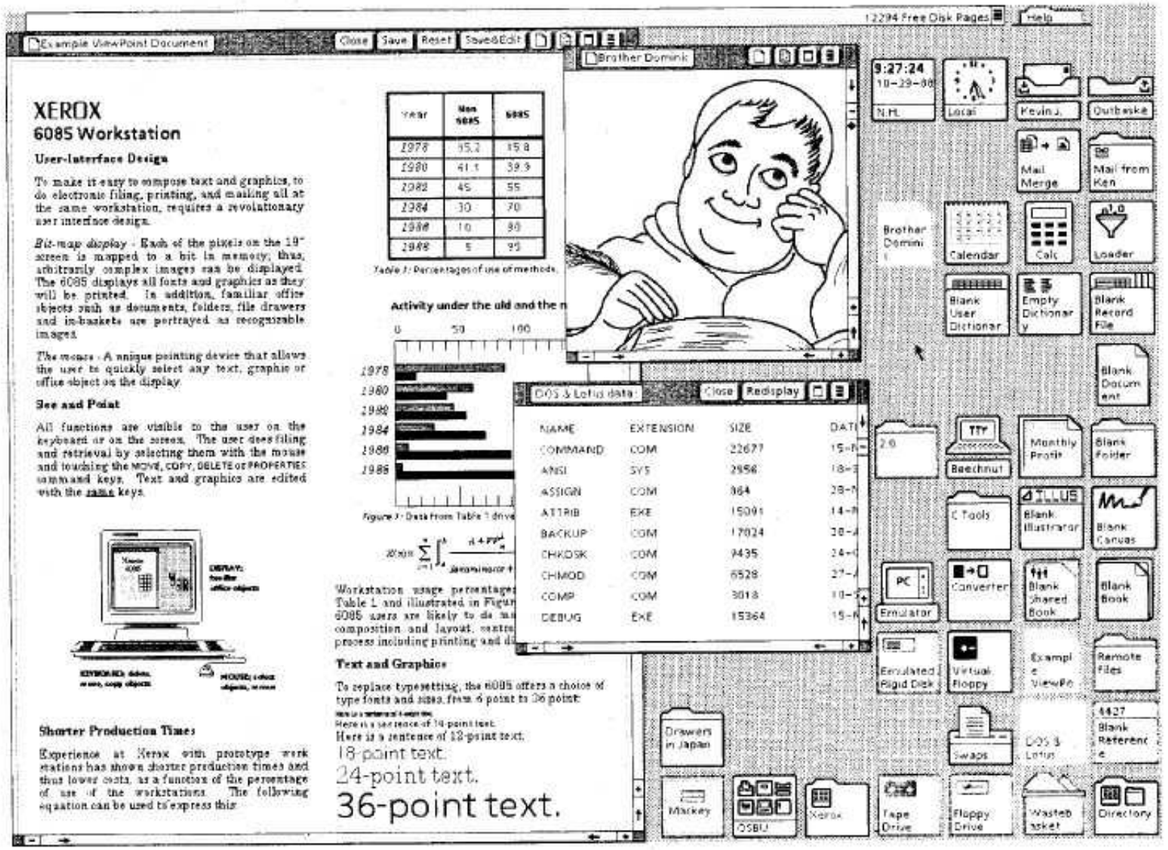
→ **Example** research systems

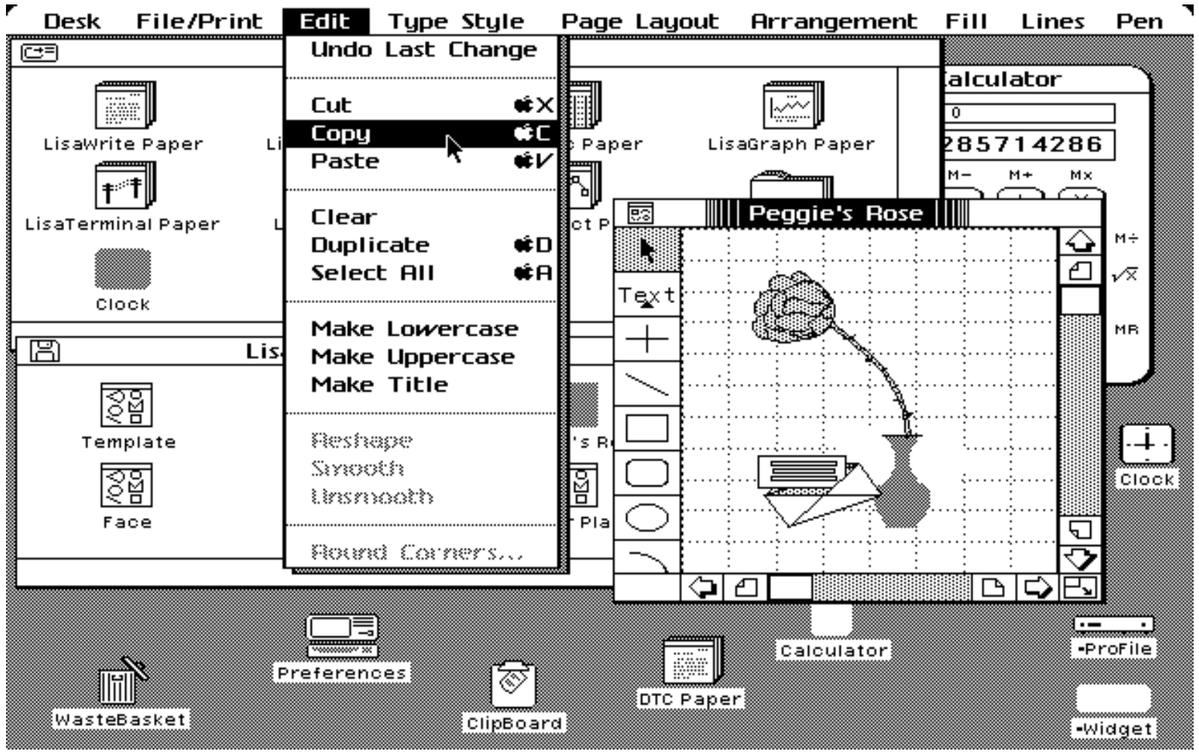
Recap: Direct Manipulation[1]



[1] *Left:* Design World: 50 Years of CAD; *Right:* Forbes: The Mother of All Demos

The WIMP Paradigm[2]





[2] *Left:* Mac history: Apple Lisa; *Right:* Wired: The Xerox Star

What did we see?

WIMP

→ More controls

→ More like a tool that the user has to figure out how to use

→ Screen based

→ Command-based

Post-WIMP

→ Fewer controls (at least not visible)

→ Advanced NLP (especially small talk)

→ More partner than a tool

→ Still screen based

→ Technology might not be there to differentiate between users

→ Dialogue-based

→ More personalized, context-based

Enter Multimodal Interaction

Definition: Multimodal systems process two or more combined user input models—such as speech, pen, touch, manual gestures, gaze, and head/body movements—in a coordinated manner with multimedia system output.[5]

The goal is to capture *naturally occurring forms of human language* (verbal and nonverbal), using *recognition-based technologies*, as input into computer systems.

[Naturally occurring language] + [recognition-based technologies]

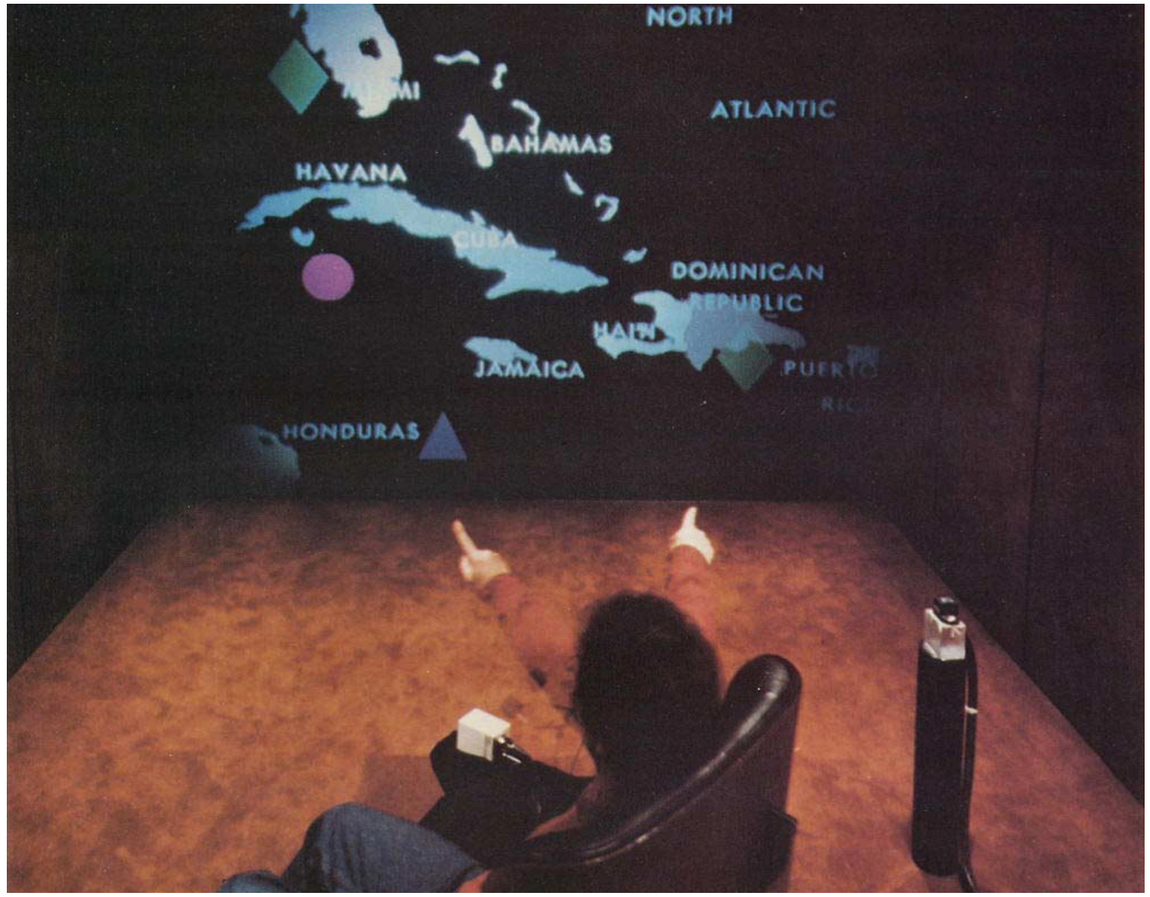
[5] Oviatt (2003). Multimodal Interfaces.*The human-computer interaction handbook*

The Birth of Multimodal Interfaces

The Media Room[6]

Move [that] to the right of the green square.
Put [that] [there].
Make [that] like [that].
Call [that] ... the calendar.

Referential communication; deixis



[6] Bolt (1980), Put-that-there: voice and gesture at the graphics interface. *Computer Graphics*.

Elements of Multimodal Interfaces

1. Natural forms of multimodal language
2. Recognition-based technologies
3. Multimodal fusion
4. Multimodal fission

Element 1: Language Forms[8]

Challenge: How can we identify naturally occurring modalities that effectively convey user intent?

Modality	Example
Visual	Face location, gaze, facial expression, lipreading, face-based identity, gesture, sign language
Auditory	Speech input, non-speech audio
Touch	Pressure, location/selection, gesture
Other sensors	Sensor-based motion capture

[8] Blattner & Glinert (1996). Multimodal integration. *IEEE multimedia*.

CARE Model[9]

- → **Complementarity:** Multiple complementary modalities are necessary to understand intent (e.g., speech + pointing gesture in "Put that there").
- → **Assignment:** Only one modality communicates user intent (e.g., steering wheel in a car).
- → **Redundancy:** Multiple modalities, each of which are sufficient, can communicate intent.
- → **Equivalence:** Multiple modalities that can interchangeably used (e.g., speech and keyboard can both be used to write text).

[9] Coutaz et al. (1995). Four easy pieces for assessing the usability of multimodal interaction: the CARE properties. *Interact'95*.

CARE Model: Complementarity

Modalities complement each other to convey meaning (each modality is insufficient to convey the same meaning).

Some natural, complementary combinations:

→ **Speech + gaze direction**: The system infers that the user is speaking to it.

→ **Speech + gestures**: Gestures disambiguate referential speech.

CARE Model: Assignment

Modalities are assigned to specific functions.

Examples:

→ **Speech:** Use for dictation.

→ **Gesture:** Scrolling, panning, zooming.

→ **Pointing & Clicking:** Selection, direct manipulation.

CARE Model: Redundancy

Multiple modalities that trigger the same function are used simultaneously in a redundant fashion.

Examples (very few real-world examples):

→ **Pointing + verbal disambiguation**: The tall, red bottle [pointing toward the bottle].

CARE Model: Equivalence

Multiple modalities trigger the same function.

Examples:

→ Keyboard arrows / trackpad gestures → scrolling

→ Keyboard shortcuts / menu items / trackpad gestures → navigation, actions (flag, archive, snooze)

Element 2: Recognition-Based Technologies[10]

	GUI (e.g., WIMP)	MUI
User input	Single	Multiple
Interpretation	Atomic, deterministic	Continuous, probabilistic
Processing	Sequential	Parallel
Architecture	Centralized	Distributed & time-sensitive

[10] Dumas et al. (2009). Multimodal interfaces: A survey of principles, models and frameworks. *Human machine interaction: Research results of the mmi program.*

Why do we have to recognize?

Unimodal input

→ *User perspective:* explicit

→ *Communication perspective:* pass-through

→ *System perspective:* simple triggers

Multimodal input

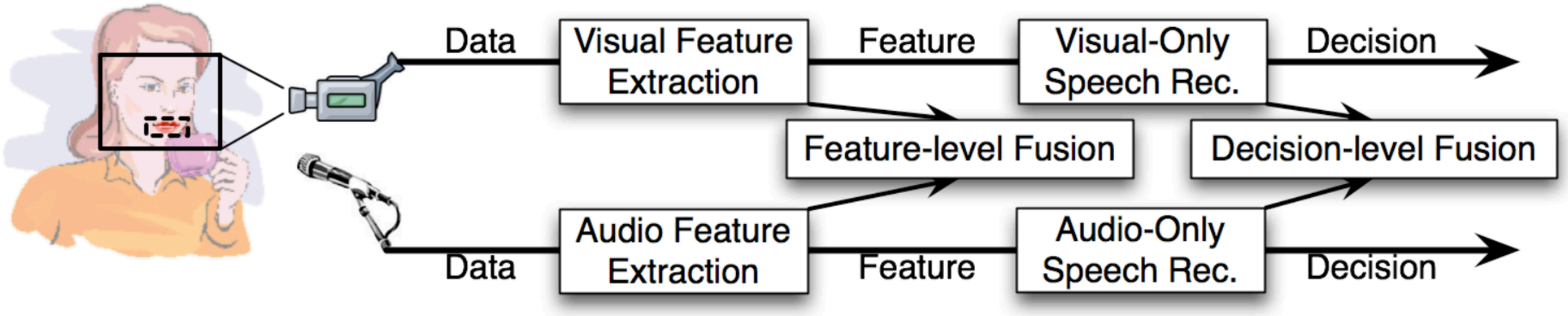
→ *User perspective:* implicit

→ *Communication perspective:* fusion of multiple low-level signals into high-level inference

→ *System perspective:* complex states

Element 3: Multimodal Fusion[10]

Challenge: How do systems infer user intent from multimodal input?



[10] Dumas et al. (2009). Multimodal interfaces: A survey of principles, models and frameworks. *Human machine interaction: Research results of the mmi program.*

	Data-level	Feature-level	Decision-level[10]
Input type	Raw data of same type	Closely coupled modalities	Loosely coupled modalities
Level of information	High detail	Moderate detail	Mutual disambiguation by combining modalities
Noise/failure sensitivity	Highly susceptible	Less sensitive	Highly resistant
Usage	Not commonly used	Used to combine particular modalities	Most widely used
Application examples	Fusion of two video streams	Speech recognition from voice and lip movement	Pen/speech interaction

Pros and cons of *early* vs. *mid-level* vs. *late* integration models[12]

[10] Dumas et al. (2009). Multimodal interfaces: A survey of principles, models and frameworks. *Human machine interaction: Research results of the mmi program.*

[12] Turk (2014). Multimodal interaction: A review. *Pattern recognition letters.*

Feature-fusion (FF), *decision-fusion (DF)*, and *hybrid* fusion strategies:[13]

a. Analysis unit

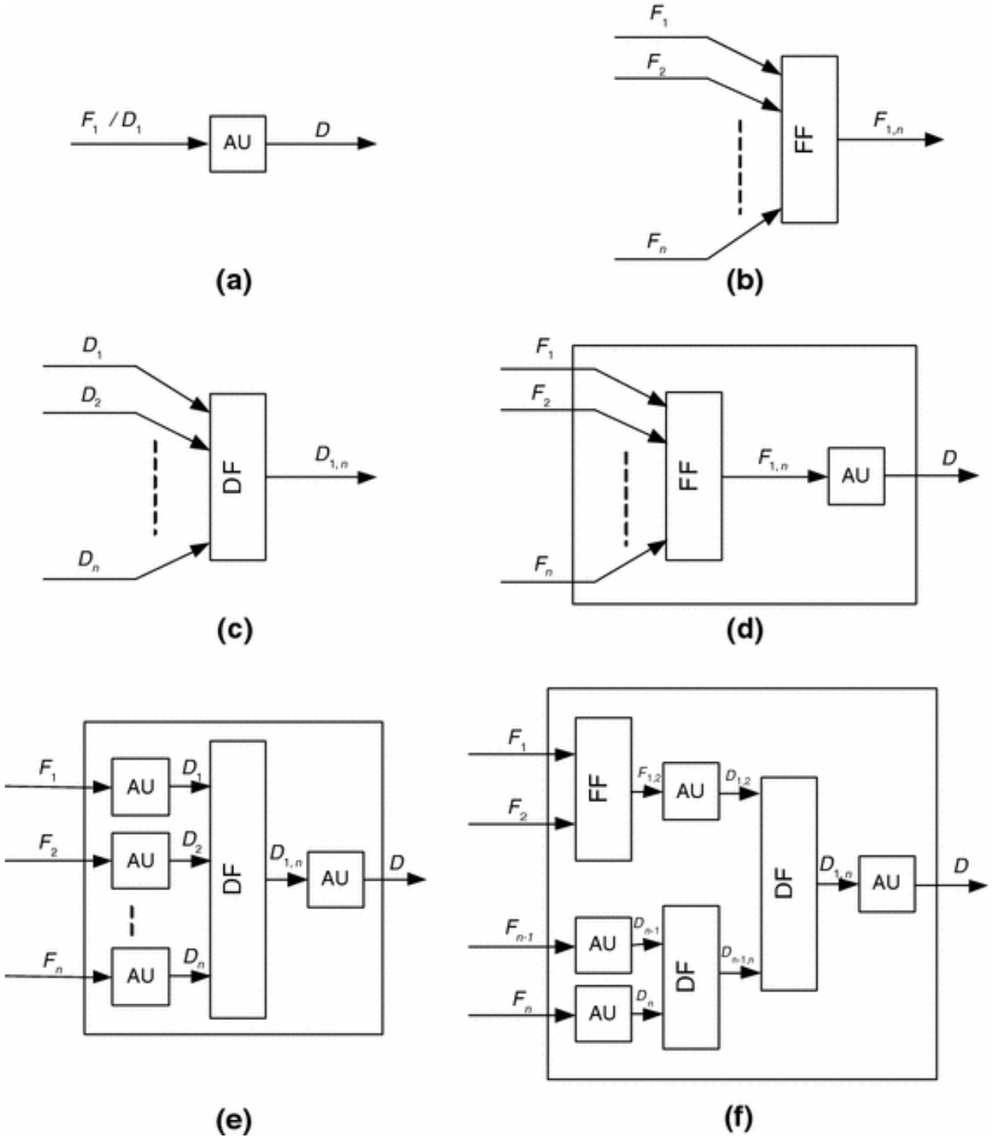
b. Feature fusion unit

c. Decision fusion unit

d. Feature level multimodal analysis

e. Decision level multimodal analysis

f. Hybrid multimodal analysis



[13] Atrey et al. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*.

Multimodal Fusion Methods[13]

1. **Rule-based methods:** *linear weighted fusion*, majority voting rule, custom-defined rule

   $$I = \sum_{i=1}^{n} w_i \times I_i \text{ or } I = \prod_{i=1}^{n} {I_i}^{w_i}$$

2. **Classification-based methods:** SVM, *Bayesian inference*, Dampster-Shafer theory, dynamic Bayesian networks, neural networks, maximum entropy model

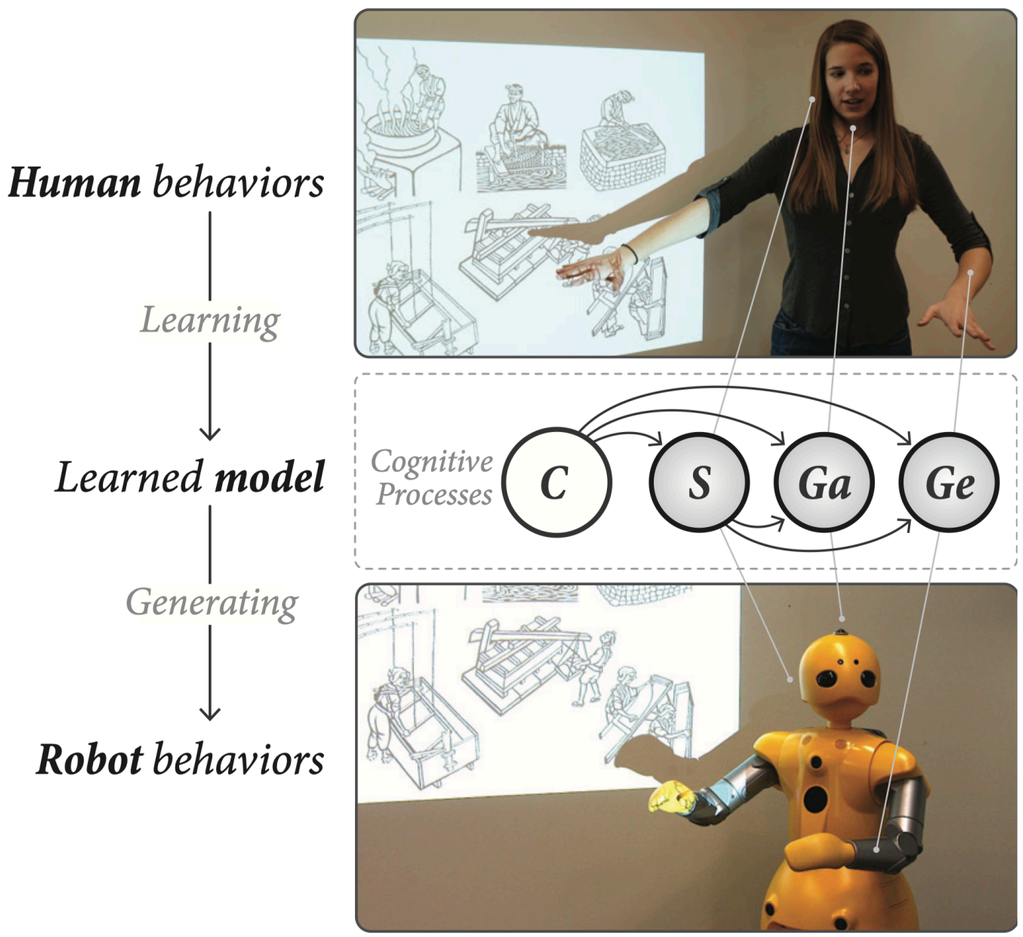
   $$p(H|I_1, I_2, \ldots, I_n) = \frac{1}{N} \prod_{k=1}^{n} p(I_k|H)^{w_k} \text{ where } \hat{H} = \mathbf{argmax}_{H \in E}\, p(H|I_1, I_2, \ldots, I_n).$$

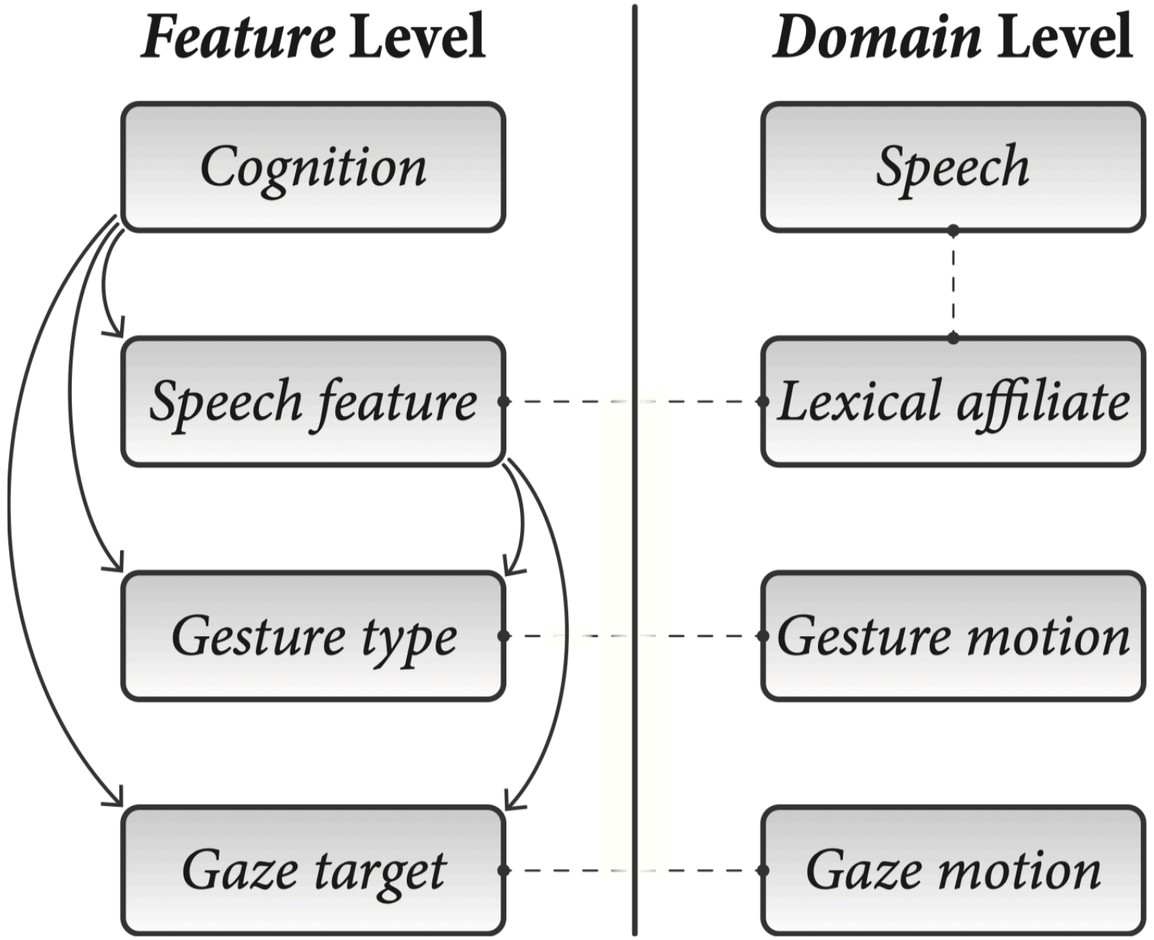
3. **Estimation-based methods:** *Kalman filter*, extended Kalman filter, particle filter

   $$x(t) = A(t)x(t-1) + B(t)I(t) + w(t) \text{ and } y(t) = H(t)x(t) + v(t)$$

[13] Atrey et al. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*.

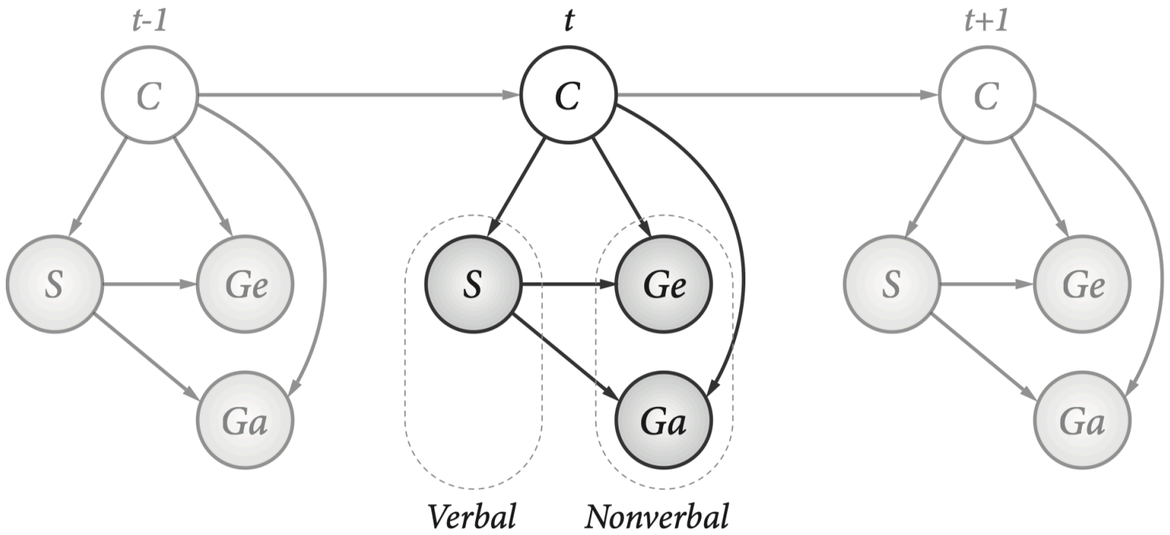
Example Fusion Using DBNs[14]





[14] Huang & Mutlu (2014). Learning-based modeling of multimodal behaviors for humanlike robots. *HRI 2014*.

Gesture Type	Speech Features		
Deictic gestures	Concrete reference *"a big pot"*	Abstract reference *"the first step"*	Pronoun *"this person"*
Iconic gestures	Concrete object *"two boards"*	Descriptive verb *"peel it off"*	Non-descriptive action *"make it"*
Metaphoric gestures	Abstract concept *"for six hours"*	Abstract process *"how paper is made"*	Abstract object *"the water soluble elements"*
Beat gestures	Important information *"at least ten times of water"*	New information *"for example"*	Connector *"so that"*



Element 4: Multimodal Fission[10]

Definition: Generating the system's response to the user in the most appropriate modality/modalities, choosing from or integrating text-to-speech synthesis, audio cues, visual cues, haptic feedback or animated agents.

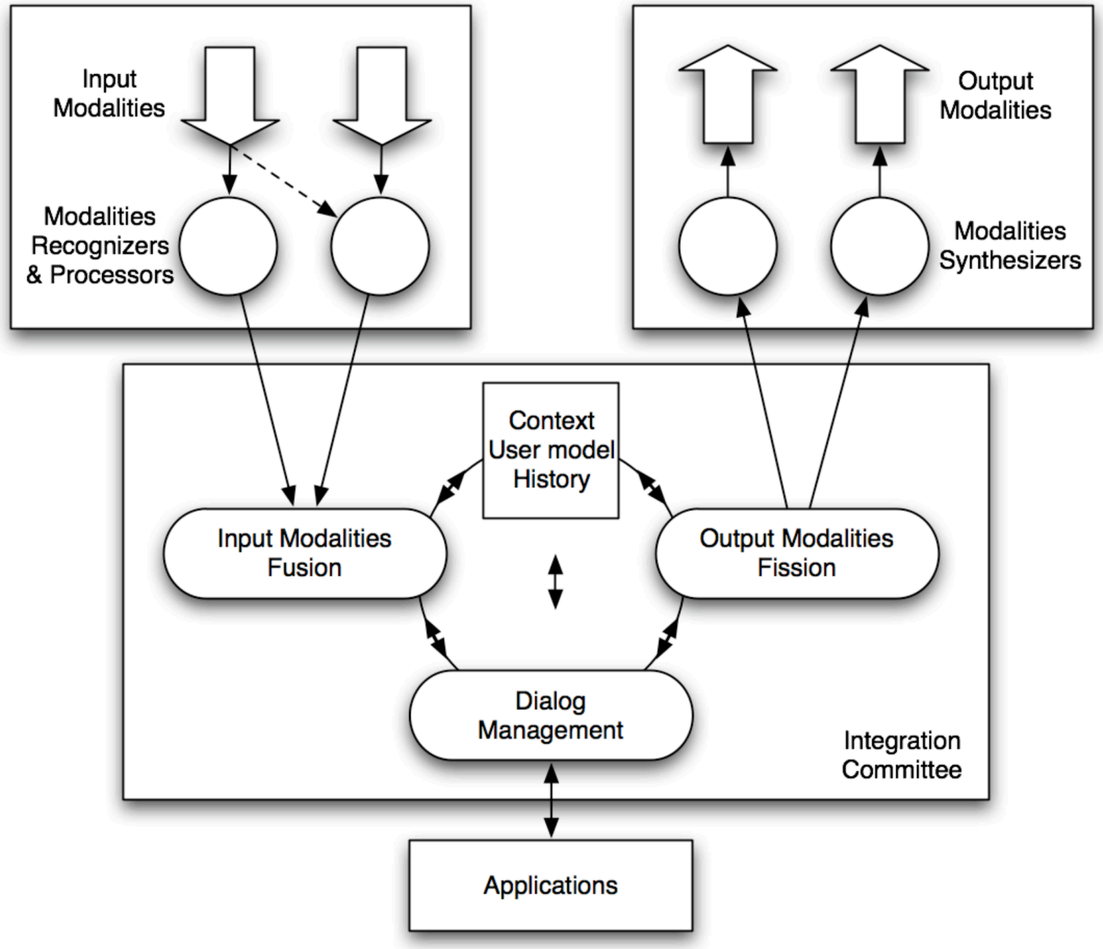
Three key tasks:

1. **Message construction**, usually through schema- or plan-based approaches
2. **Output channel selection**, based on context, user profile, etc.
3. **Message syncronization** by coordinating outputs in different modalities

[10] Dumas et al. (2009). Multimodal interfaces: A survey of principles, models and frameworks. *Human machine interaction: Research results of the mmi program.*

Components:[10]

1. Dialogue management
2. Consideration of user context
3. Output modality selection
4. Modality synthesis



[10] Dumas et al. (2009). Multimodal interfaces: A survey of principles, models and frameworks. *Human machine interaction: Research results of the mmi program.*

Adaptive Multimodal Fission

Modality selection: CARE model: Complementarity, Assignment, Redundancy, Equivalence[9]

Output coordination: Physical layout, temporal coordination, referring expressions

Example systems:

1. GUIDE “Gentle User Interface for Elderly people”[15]
2. Proximity Toolkit[16]

[9] Coutaz et al. (1995). Four easy pieces for assessing the usability of multimodal interaction: the CARE properties. *Interact’95*.

[15] Costa & Duarte (2011). Adapting multimodal fission to user’s abilities. UAHCI 2011.

[16] Greenberg et al. (2011). Proxemic interactions: the new ubicomp? *interactions*.

Multimedia System Architectures[5]

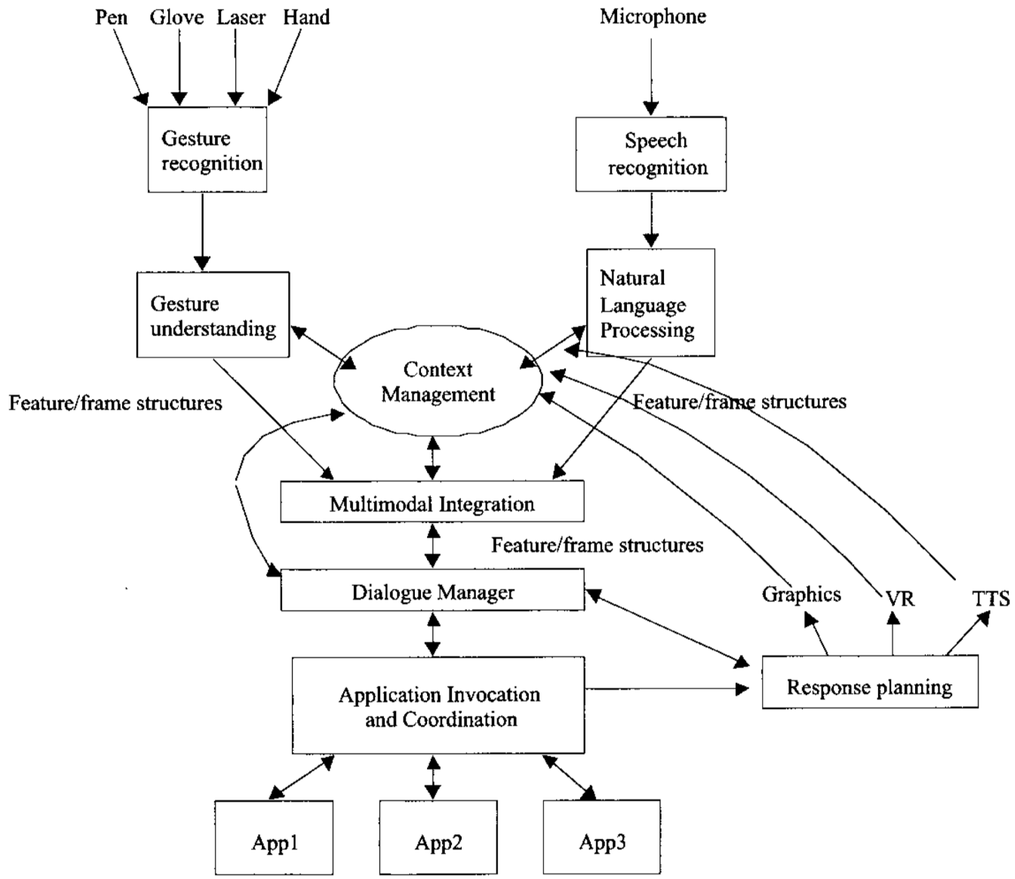
The four elements:

→ Natural forms of multimodal language

→ Recognition-based technologies

→ Multimodal fusion

→ Multimodal fission



[5] Oviatt (2003). Multimodal Interfaces.*The human-computer interaction handbook*

Example Multimodal Systems

Figaro[17]

Music: Summer from Bensound.com

[17] Porfirio et al. (2021). Figaro: A tabletop authoring environment for human-robot interaction. *CHI 2021.*

Tabula[18]

[18] Porfirio et al.(2023). Sketching Robot Programs On the Fly. *HRI 2023*.

To Learn More

→ ACM International Conference on Multimodal Interaction

→ ICMI Proceedings