UiO : University of Oslo

Multimodal Aspects of Stochastic Interaction Management

Pierre Lison
Language Technology Group

Trial Lecture
21st February 2014

Outline of the lecture

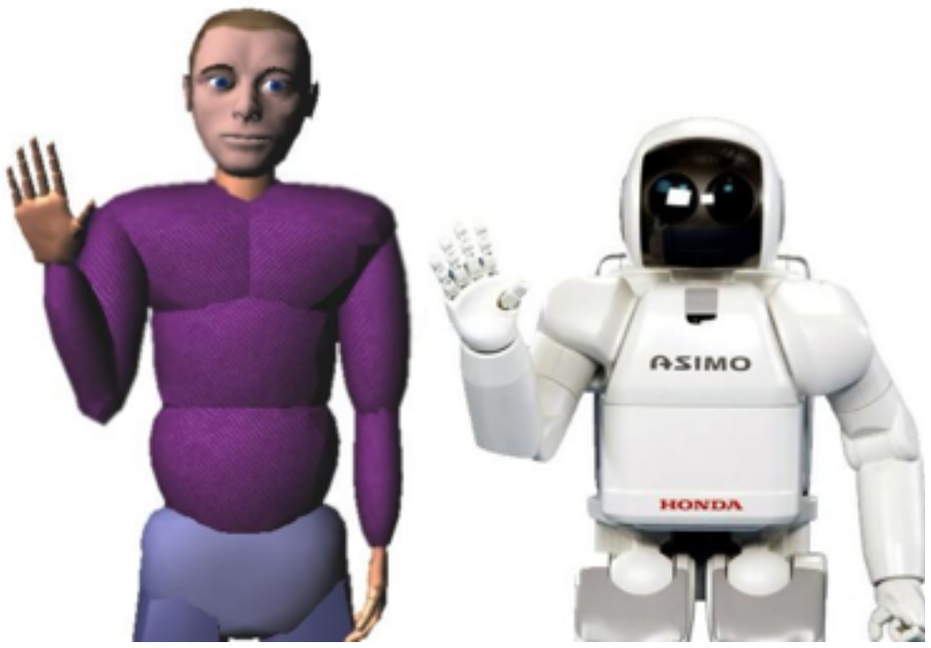
1. What is a multimodal system?
2. Multimodal architectures
3. Interaction management
4. Conclusion

Outline of the lecture

1. **What is a multimodal system?**
2. Multimodal architectures
3. Interaction management
4. Conclusion

Multimodal interfaces

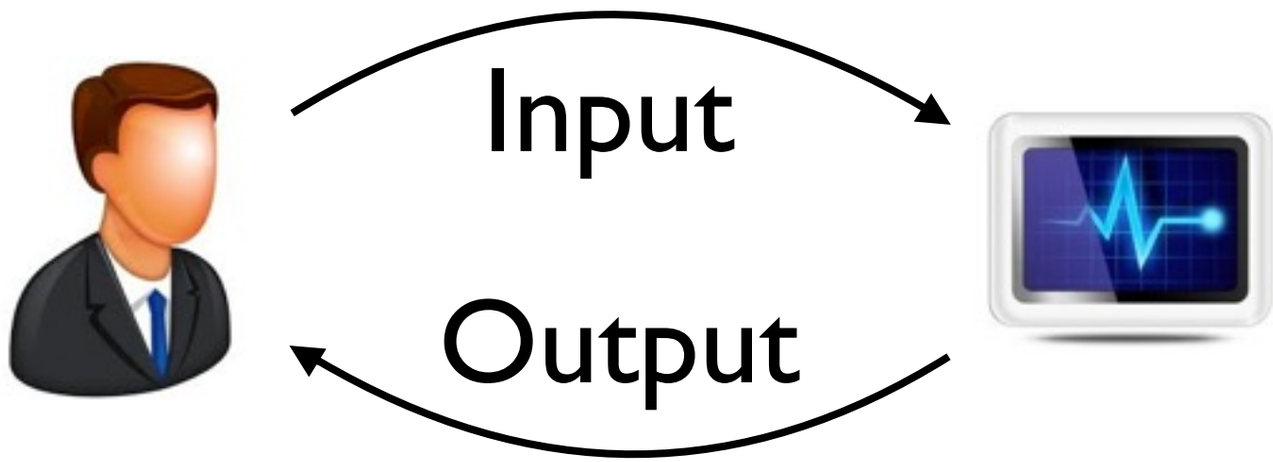
A **multimodal interface** is a computer interface that provides the user with more than one "path of communication"



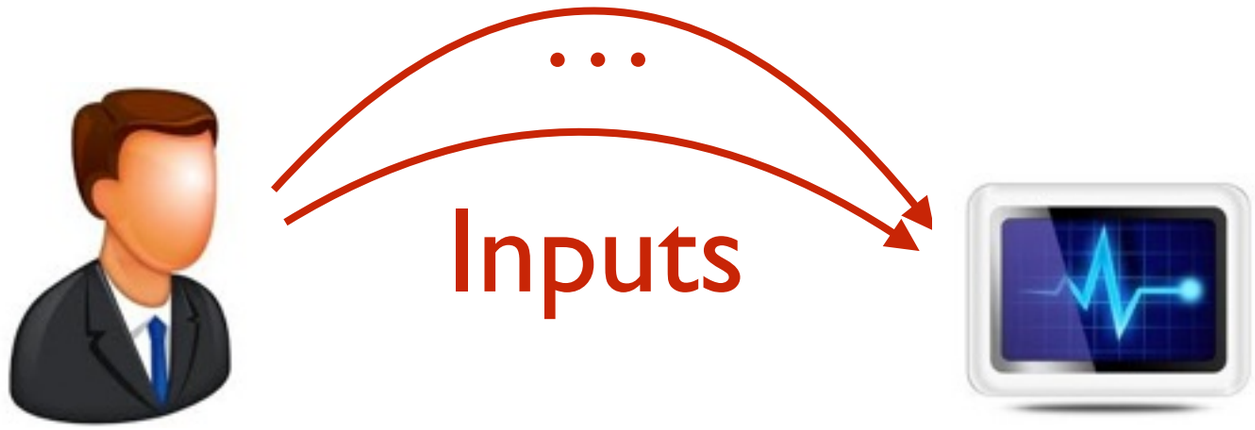
What is a modality?

In human-computer interaction, a "modality" is a *channel of communication* between the user and the machine

- Relation to *human senses*: vision, audition, touch, etc.
- Includes both the system *inputs* and *outputs*



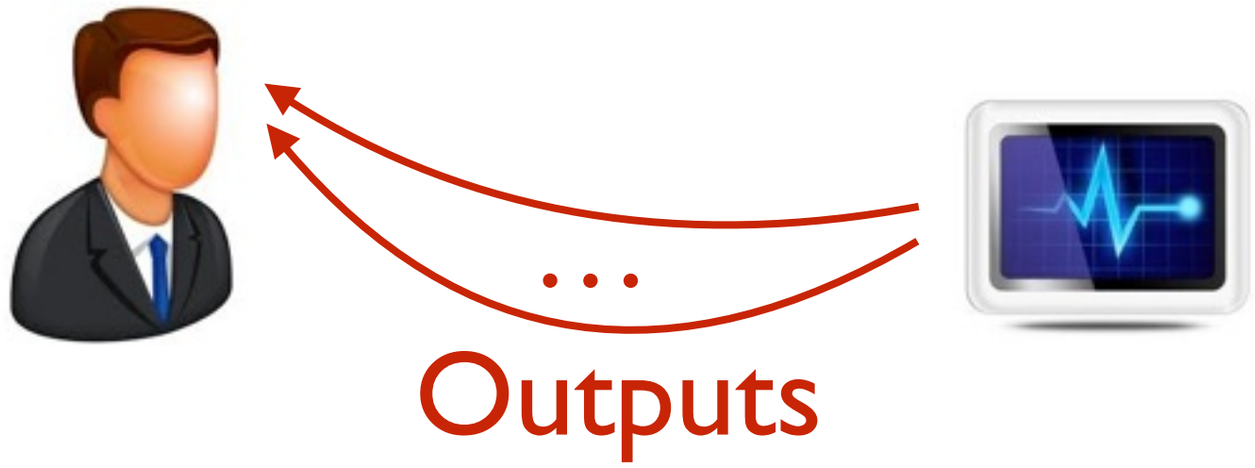
Multimodal inputs



Why use multiple *input* modalities?

- Increased *usability* and *accessibility*
- More meaningful and reliable interpretations
- Better grasp of the user's current state (i.e. intention, attention, affect)

Multimodal outputs



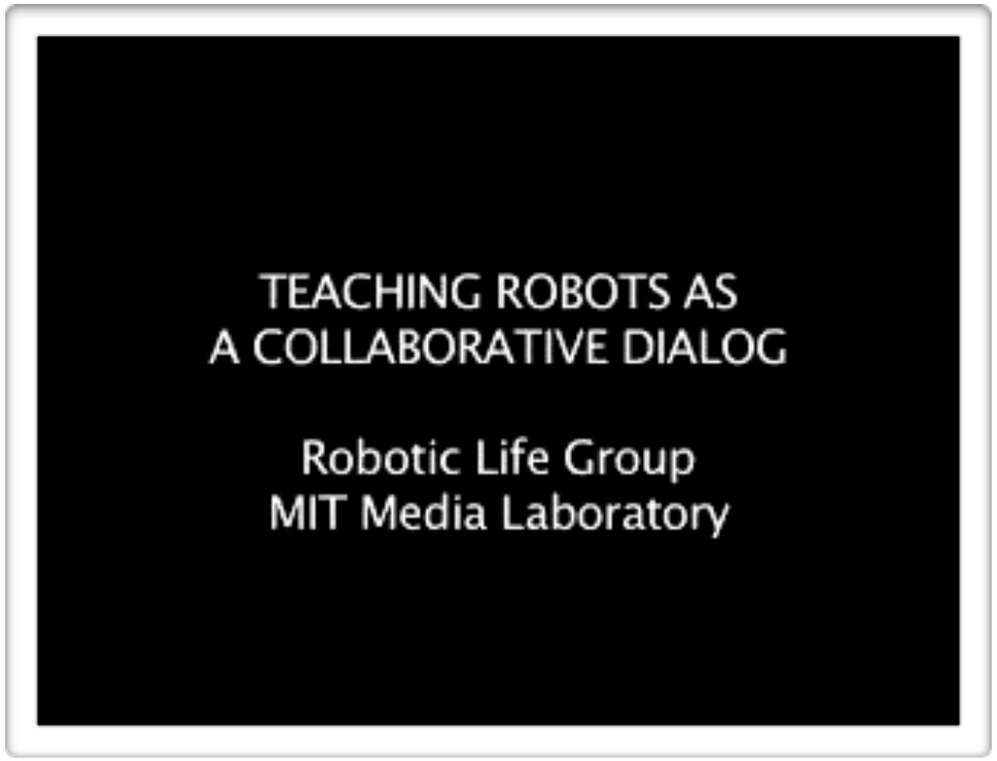
Why use multiple *output* modalities?

- *Tailor* the system outputs to the situation
- Enrich generated content
- Increase user *engagement*

Types of inputs/outputs

Major modalities:

- Vision
- Audition
- Touch



Modalities in human communication

Human face-to-face communication is fundamentally multimodal

- Speech, gaze, gestures, body pose
- Continuous, bidirectional exchange of information
- Interactive alignment of behaviour

Hand gestures

Symbolic

Iconic

Metaphoric

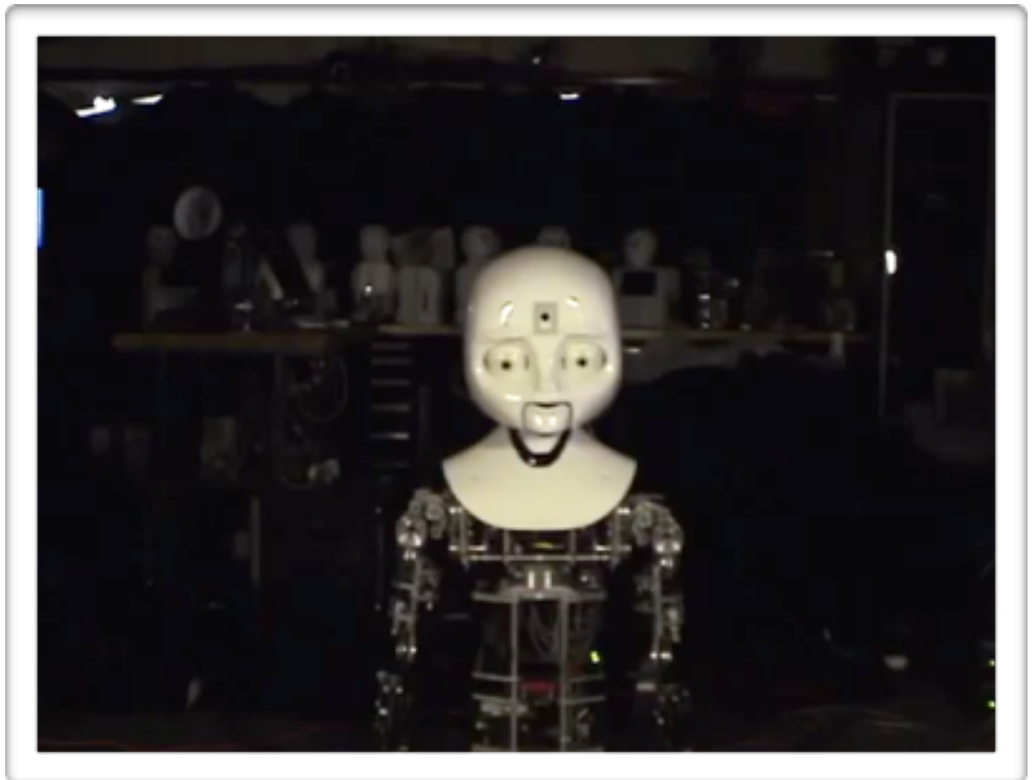
Deictic

Beat

[D. McNeill (2008), "*Gesture and Thought*", University of Chicago Press]

Non-verbal signals

- More than gestures!
- Gaze, facial expressions & body posture also convey important signals



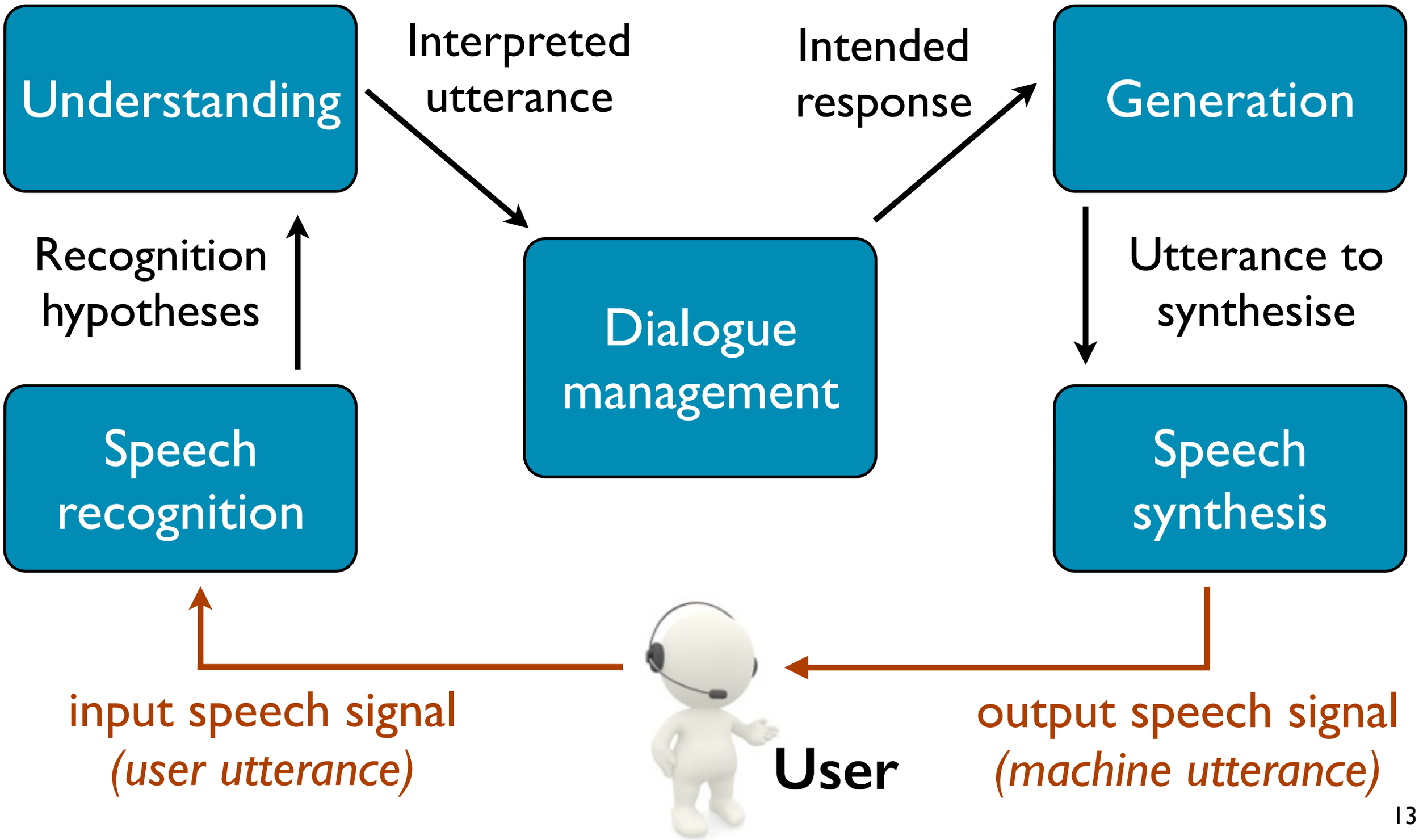
- Used to control *turn-taking*, *attention*, *grounding*, and *affect*

[K. Jokinen, H. Furukawa, M. Nishida and S. Yamamoto (2013), "Gaze and turn-taking behavior in casual conversational interactions", *ACM TiiS*.]

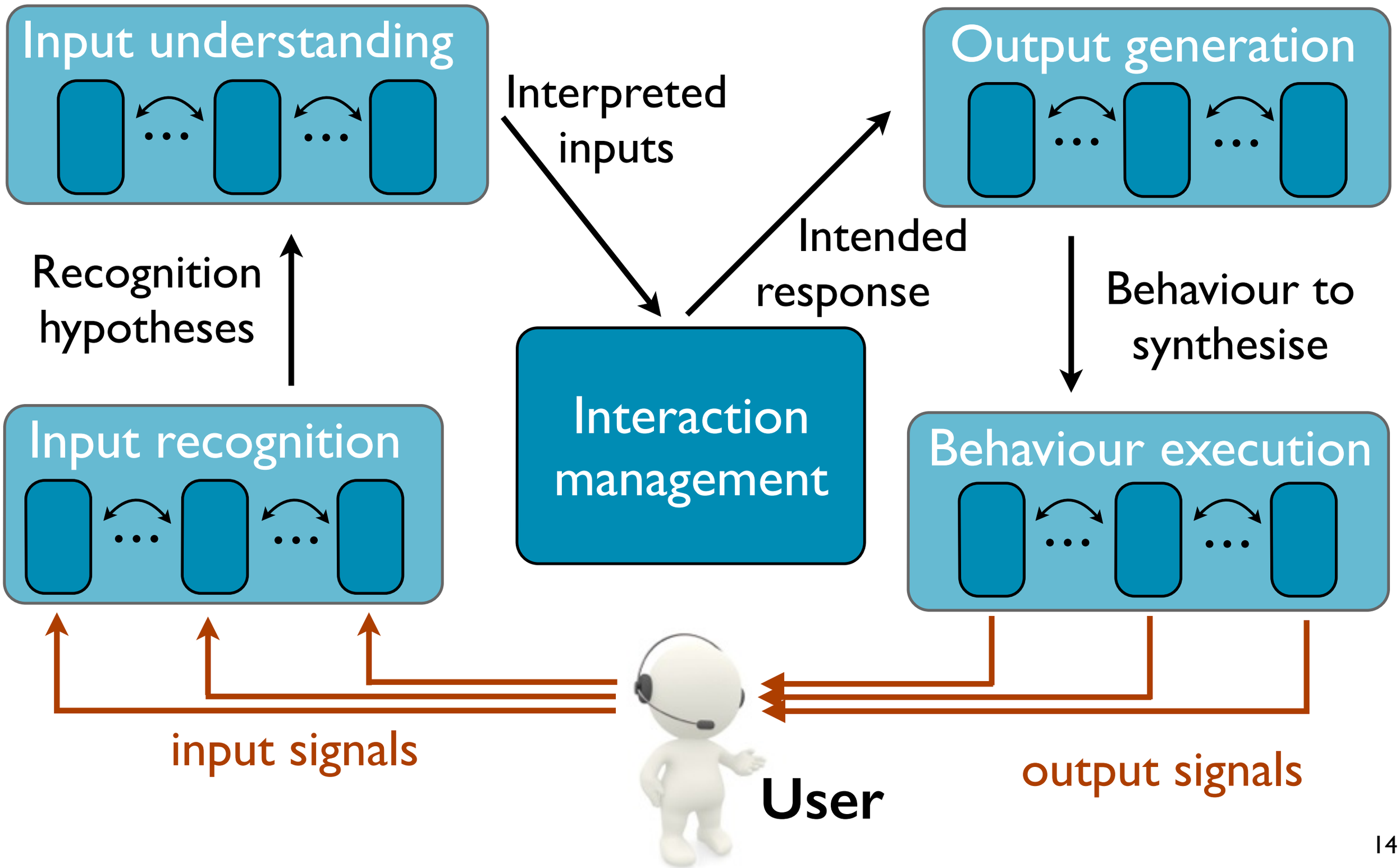
Outline of the lecture

1. What is a multimodal system?
2. **Multimodal architectures**
3. Interaction management
4. Conclusion

Classical dialogue architecture

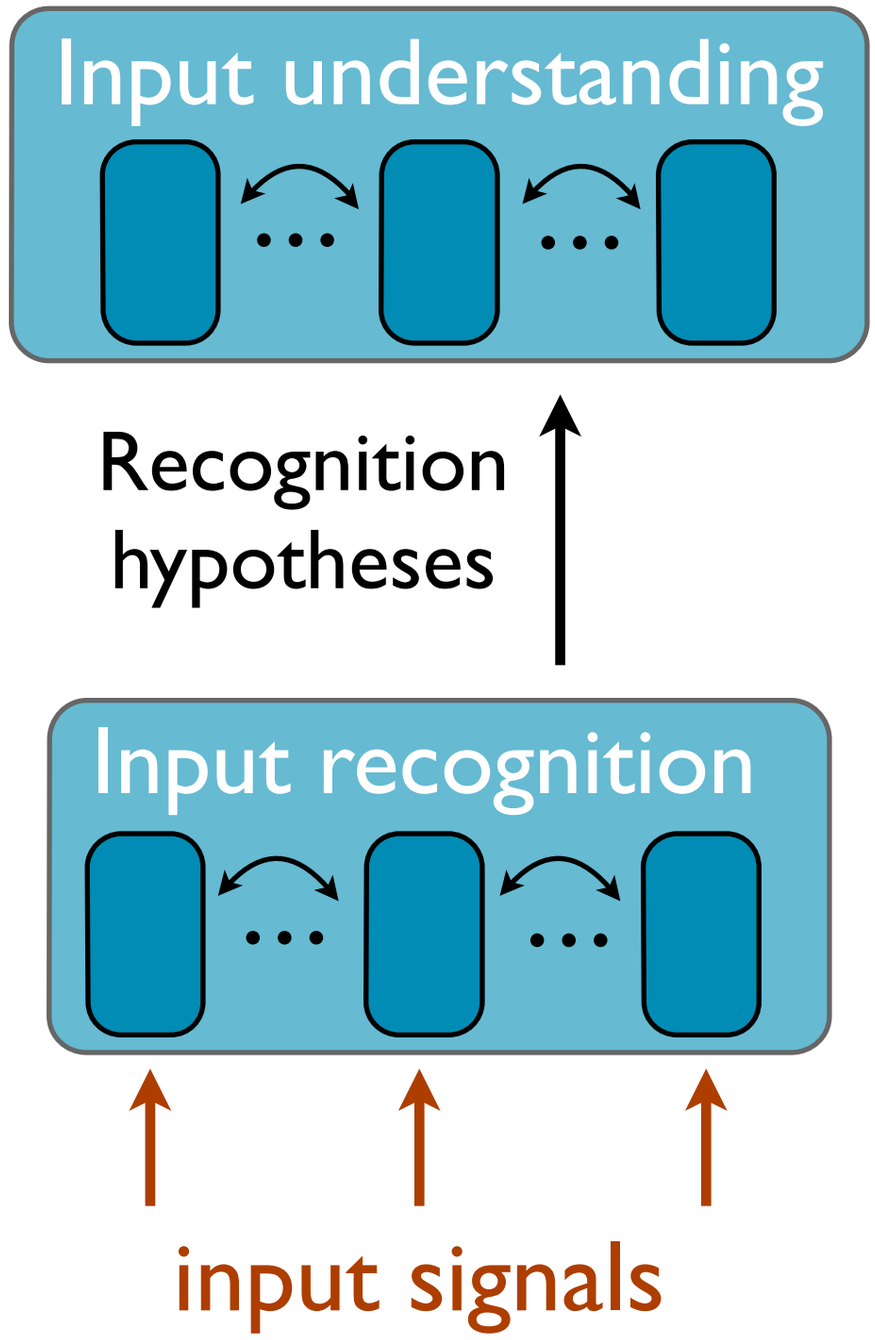


Multimodal architecture



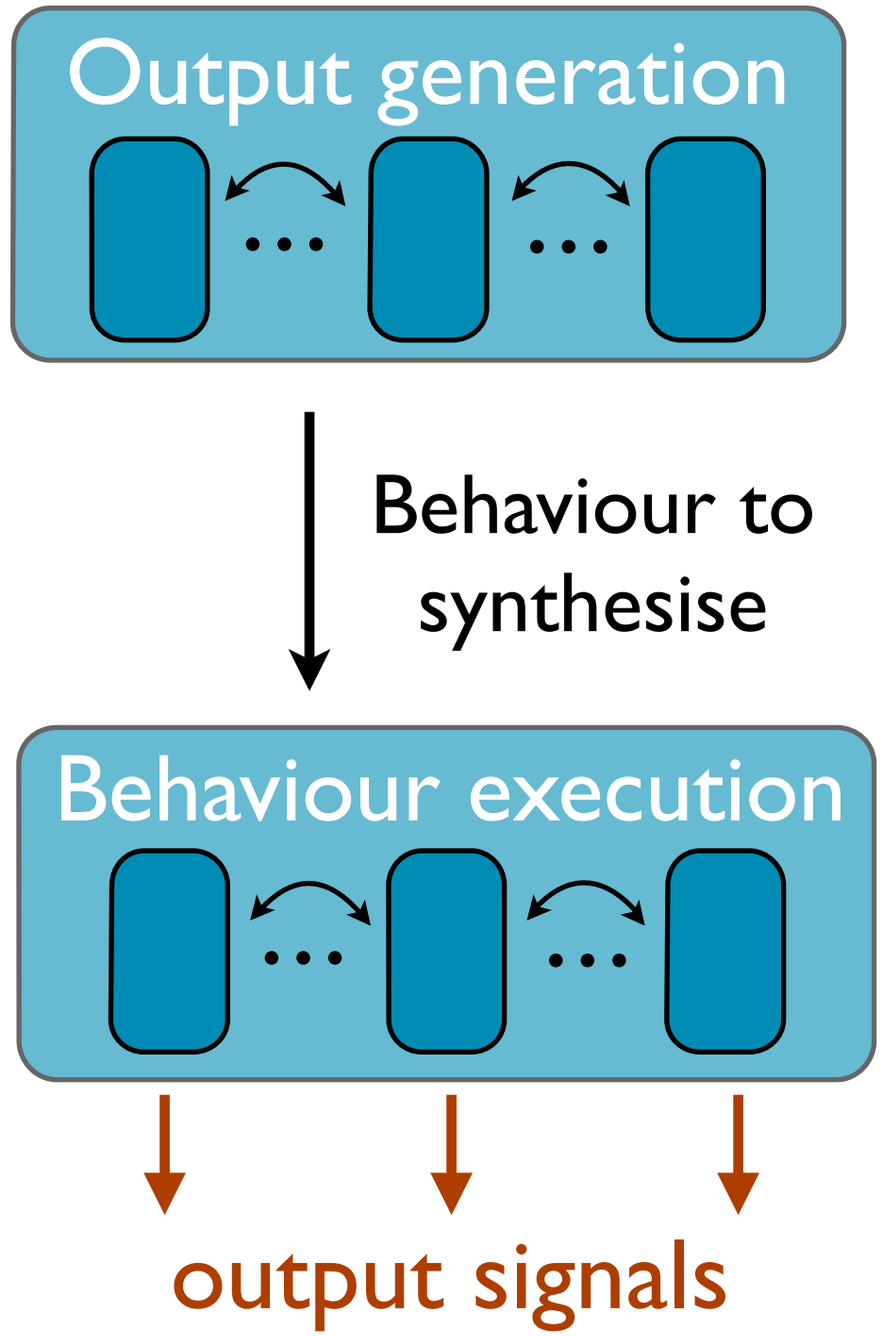
Input fusion

- *Merge* information arising from different sources
  - Content may be redundant, complementary, or conflicting
- Fusion stages:
  - *Early fusion*: combine coupled signals at feature level
  - *Late fusion*: construct cross-modal semantic content



Output fission

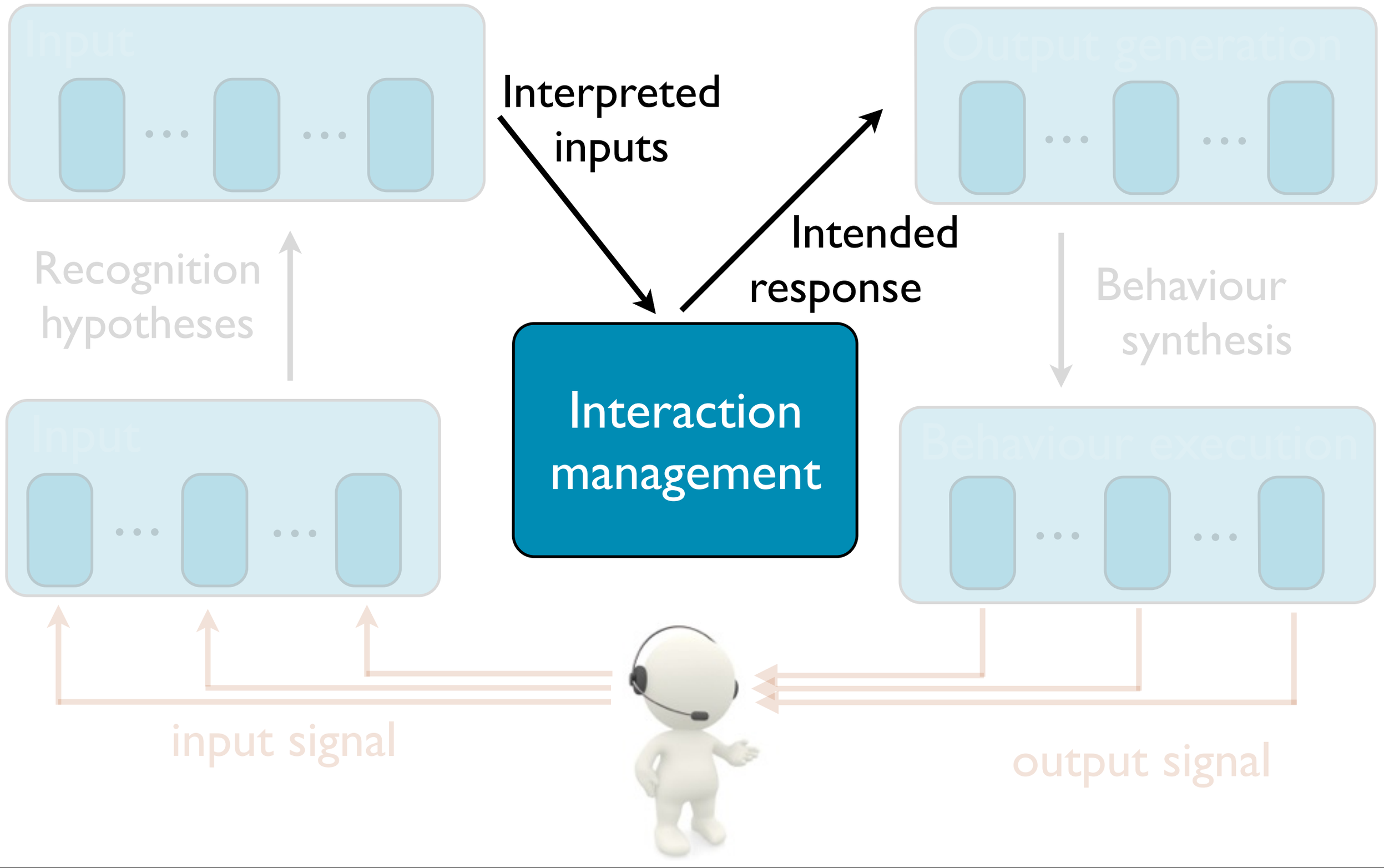
- *Distribute* a given output over the set of available modalities
  - Find the best way to convey the content or behaviour
- Processing steps:
  - Message construction
  - Modality selection
  - Output coordination



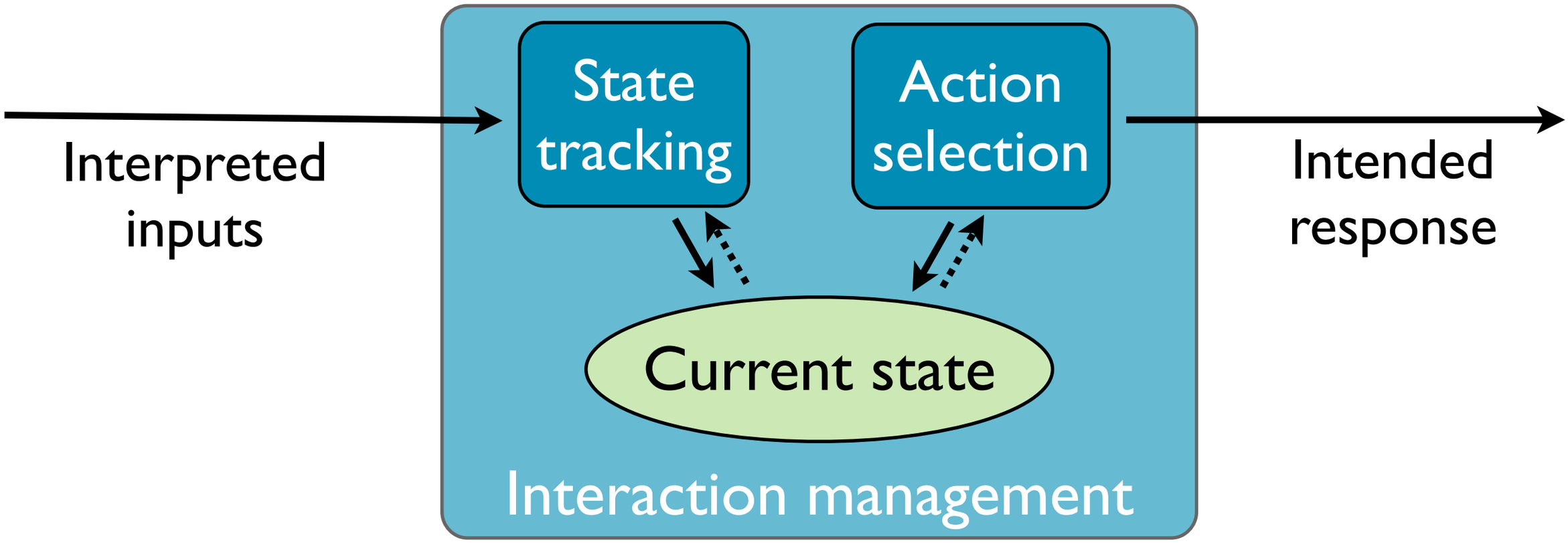
Outline of the lecture

1. What is a multimodal system?
2. Multimodal architectures
3. **Interaction management**
4. Conclusion

Multimodal architecture



Interaction management



Tasks:

1. *Track* the current state of the interaction given the inputs and past history
2. *Decide* on the best action(s) to perform

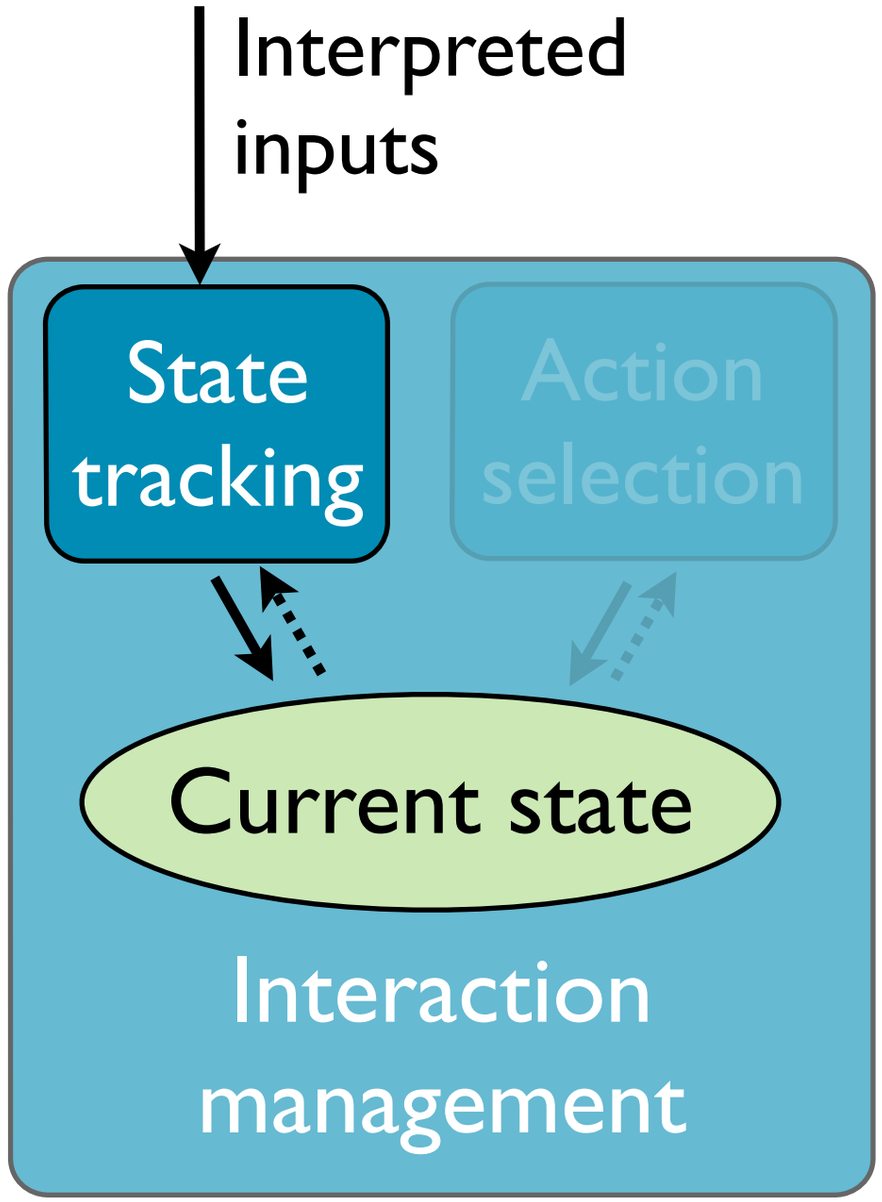
Challenges for multimodal systems

1. Tracking the interaction state
2. Deciding *when* to talk
3. Deciding *what* to do/say
4. End-to-end evaluation

Tracking the interaction state

The *interaction state* can be difficult to track:

- Numerous state variables (user & task models, history, external environment)
- Multiple, asynchronous streams of observations
- High levels of uncertainty
- Stochastic action effects



State tracking methods

- Allow state variables to be *partially observable* (e.g. POMDP models)
- Rely on *structural assumptions* and *abstraction* methods to avoid combinatorial explosion of state space
- Use *approximate inference* to ensure state tracking can be done in real-time

[J. Hoey et al. (2005), "POMDP models for assistive technology", *AAAI]*
[J. Williams (2007), "Using Particle Filters to Track Dialogue State", *ASRU]*

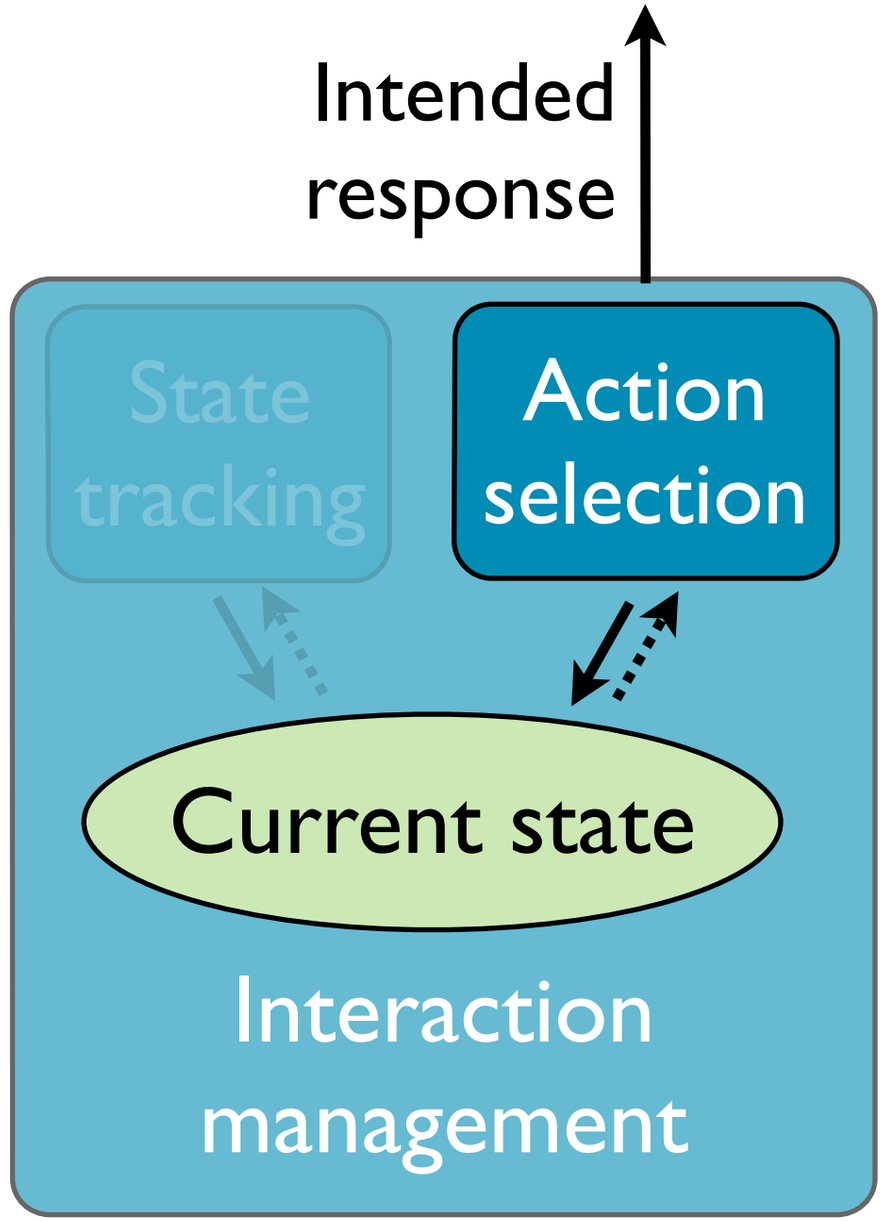
Deciding when to talk

- When should the machine *take the turn* in face-to-face interaction?
  - Combination of both verbal (syntax, prosody) and non-verbal factors (gaze, gestures, etc.)
- Statistical models to predict *when* the current speaker will end its turn
  - Sequential probabilistic modelling (e.g. CRFs) with multimodal features

[I. de Kok & D. Heylen (2009), "Multimodal End-of-Turn Prediction in Multi-Party Meetings", *ICMI*]

Deciding what to do/say

- Multimodal systems must coordinate multiple tasks in parallel
  - Engagement, communicative behaviour, physical actions, etc.
  - Tasks may be decomposed in a hierarchical manner
- How to decide on the best behaviour to execute?



[Simon Keizer et al. (2013), "Training and evaluation of an MDP model for social multi-user human-robot interaction", *SIGDIAL*]

Action selection methods

- Optimisation of multimodal policies via *reinforcement learning*
- *Temporal abstraction* can be used to capture hierarchical tasks
- Reward function can be harder to design in multimodal settings
  - Exploit social signals to infer rewards?

[V. Rieser & O. Lemon (2009), "Learning Human Multimodal Dialogue Strategies", *NLE*]

[H. Cuayáhuitl & N. Dethlefs (2012), "Spatially-Aware Dialogue Control Using Hierarchical Reinforcement Learning". In *ACM Transactions on Speech and Language Processing*]

End-to-end evaluation

- For applications with clear-cut tasks, standard metrics of task success & efficiency can be extended to multimodal settings
  - But the empirical effects of each modality on the interaction are often hard to measure
- However, many interaction domains do not have a single, predefined task
  - Naturalness & likability may be more important

[F. Schiel (2006), "Evaluation of Multimodal Dialogue Systems", *SmartKom*. Springer]
[D. Traum et al. (2004), "Evaluation of multi-party virtual reality dialogue interaction", *LREC]*

Outline of the lecture

1. What is a multimodal system?
2. Multimodal architectures
3. Interaction management
4. **Conclusion**

Take-home messages

1. **Multimodal systems** provide users with more than one communication channel
2. They offer many advantages in terms of *robustness*, *usability*, and *adaptivity*
3. But they need to address non-trivial engineering challenges:
   - Multimodal *fusion* and *fission*
   - Complex *interaction models*

Questions?