SCXML, Multimodal Dialogue Systems and MMI Architecture

Kristiina Jokinen and Graham Wilcock
University of Tampere / University of Helsinki

Departure point

- Background in
  - XML-based language processing
  - SCXML as a basis for voice interfaces
  - Cooperative dialogue management
  - Multimodal route navigation
- Interest in how the MMI architecture supports
  1) Fusion of modalities
  2) Incremental presentation
  3) Design of cooperative interaction

Limitations of Interactive Systems

- Mainly speech-based interaction
- Static interaction
- Task-orientation

From Limitations to Advanced Issues

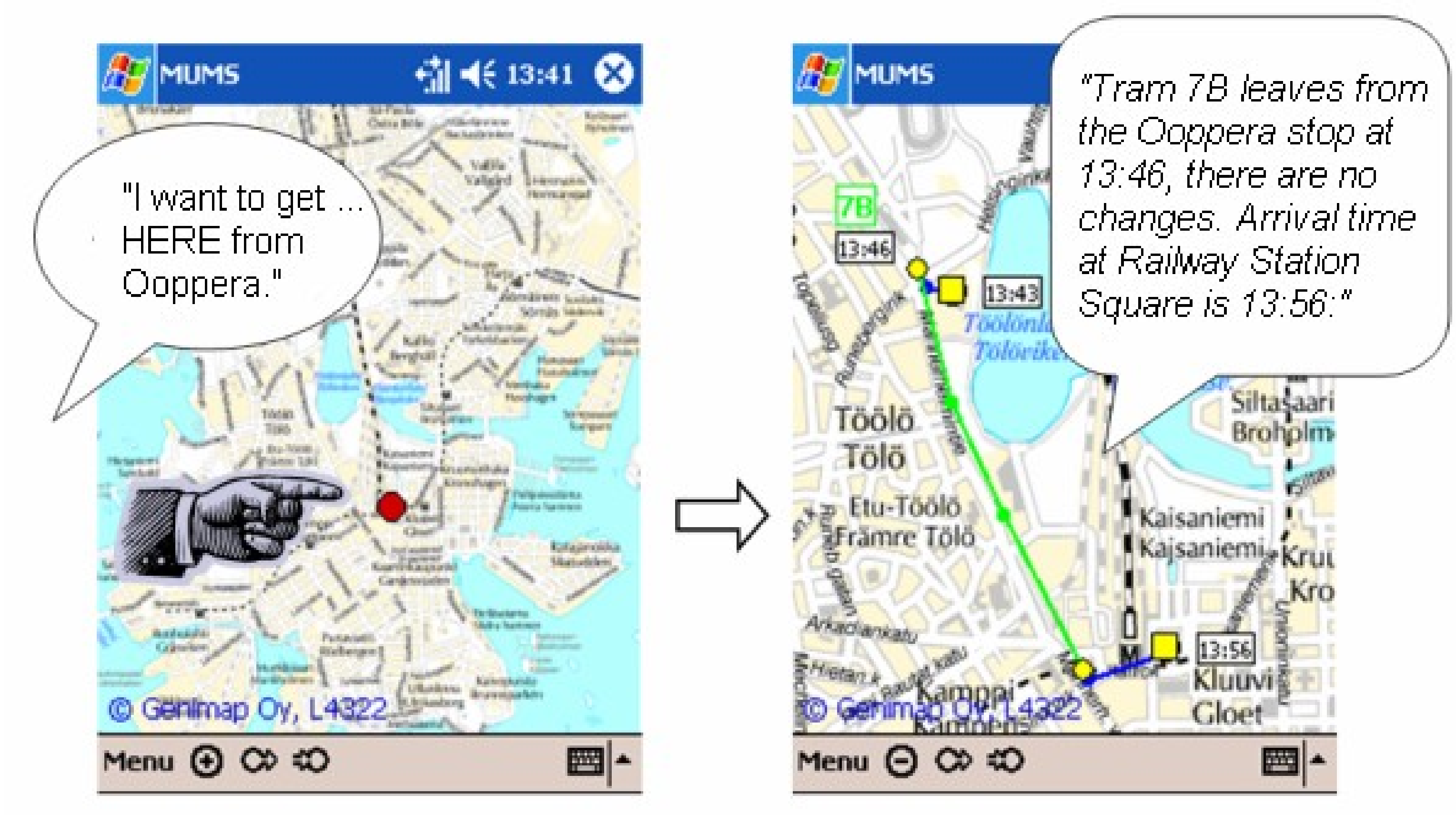
- Mainly speech-based interaction
  - Multimodality
- Static interaction
  - Adaptation
- Task-orientation
  - Human conversations
  - Non-verbal communication

MUMS - MUltiModal navigation System

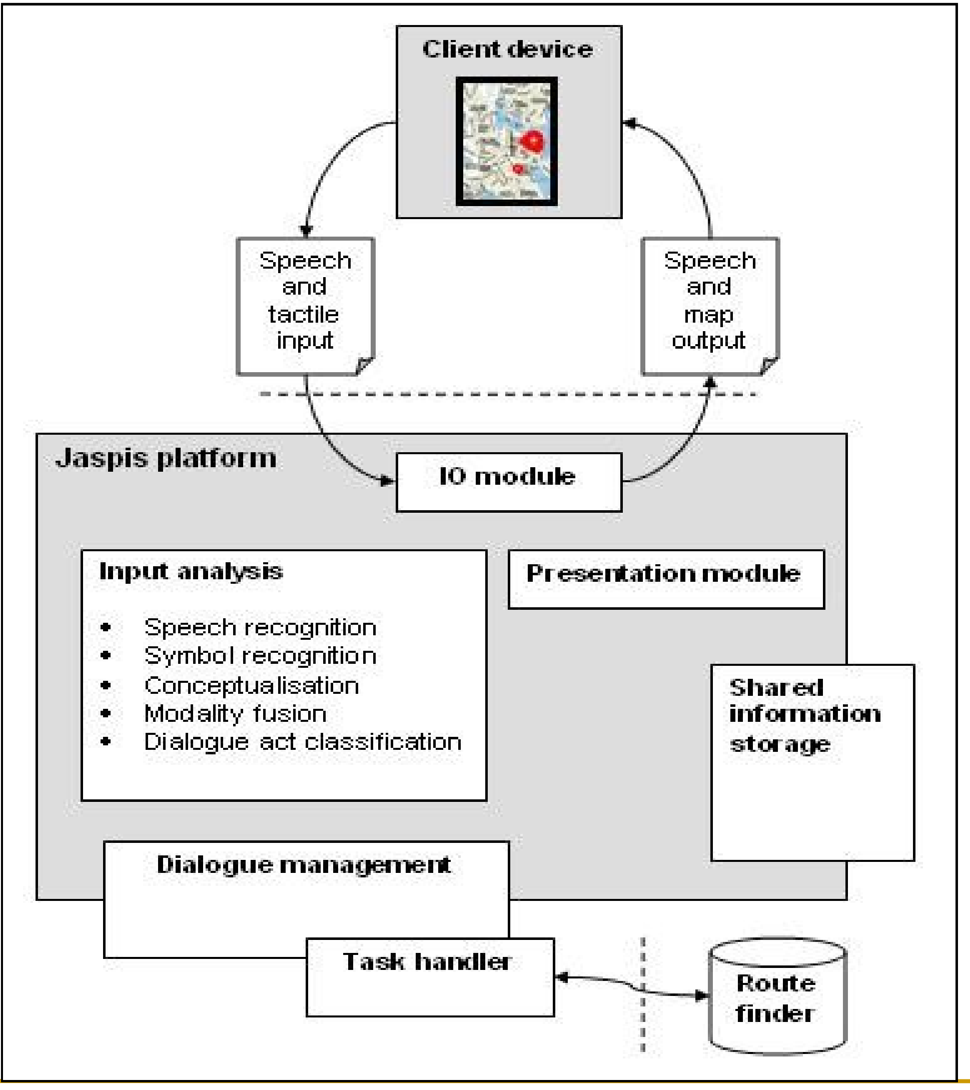
- Speech and tactile interface on a PDA
- Helsinki public transportation
- Target: mobile users who wish to find their way around
- Hurtig & Jokinen 2006, 2005; Hurtig 2005; Jokinen & Hurtig 2006; Jokinen 2007

MUMS Video

MUMS interaction



MUMS - MUltiModal navigation System



Input Fusion (T. Hurtig)

Speech recognition (N-best)

Symbol recognition (N-best)

1. Produce legal concept and symbol combinations
2. Weight combinations
3. Select the best candidate in a given dialogue context

Chosen user input

Phase 1

Speech: ”.. here no I mean here from the Operahouse …”
Tactile:

- Find all input combinations by pairing concepts with symbols
- In the example above, there are 3 possible combinations which maintain the order of input
  - Pair: {pointing, ”from the Operahouse”} could also be in accordance with the user’s intention

User command representation

command:
- score: 615.0
- start: 0
- end: 8190
- constituent_1:
  - text: here
  - concept: loc_tact
  - start: 2567
  - end: 3100
  - score: 899.0
  - symbol:
    - type: point
    - score: 1000
    - coords: (x, y)
    - start: 2701
    - end: 2701
- constituent_2: ...
- constituent_n: ...

Phase 2

- Calculate the weight of each concept-symbol pair
- Classification parameters:
  - Overlap
  - Proximity
  - Quality and type of concept and symbol
- These weighted pairs are used to calculate the final weight of each combination (-> N-best list of inputs)

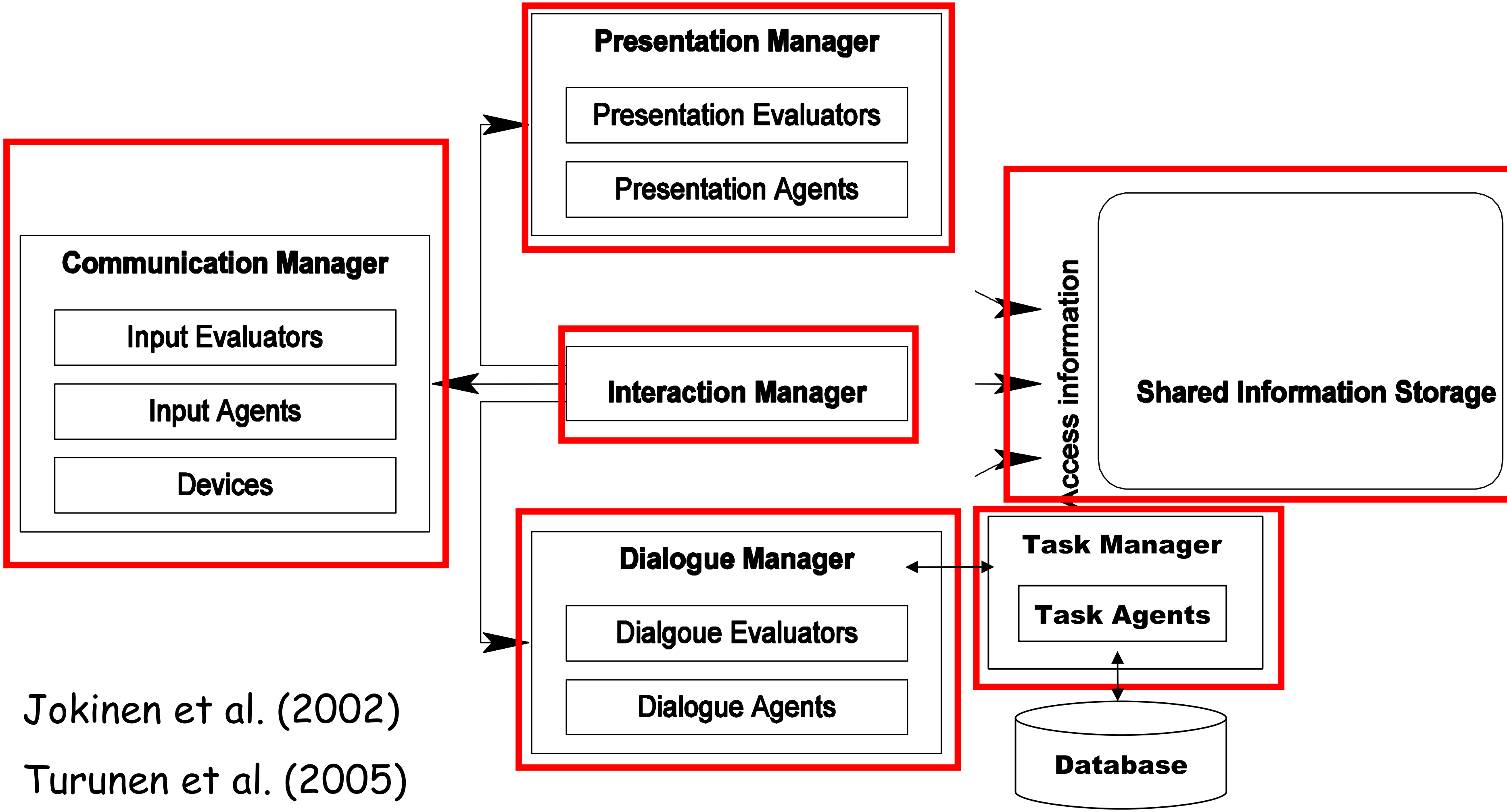
Phase 3

- Anticipate the type and context of the user’s next utterance
- Dialogue Manager chooses the best fitting candidate from the N-best list

Issues in Input Fusion

- Recognition of the user's pen gestures (point, circle, line) and their relation to speech events
- Temporal disambiguation
- Representation of information (use EMMA!)
- Natural interaction
  - Human interaction modes (how gestures and speech are usually combined: compatible, complementary, contradictory)
  - Use of gestures in spatial domains vs. information-based domains
  - Flexible change in tasks

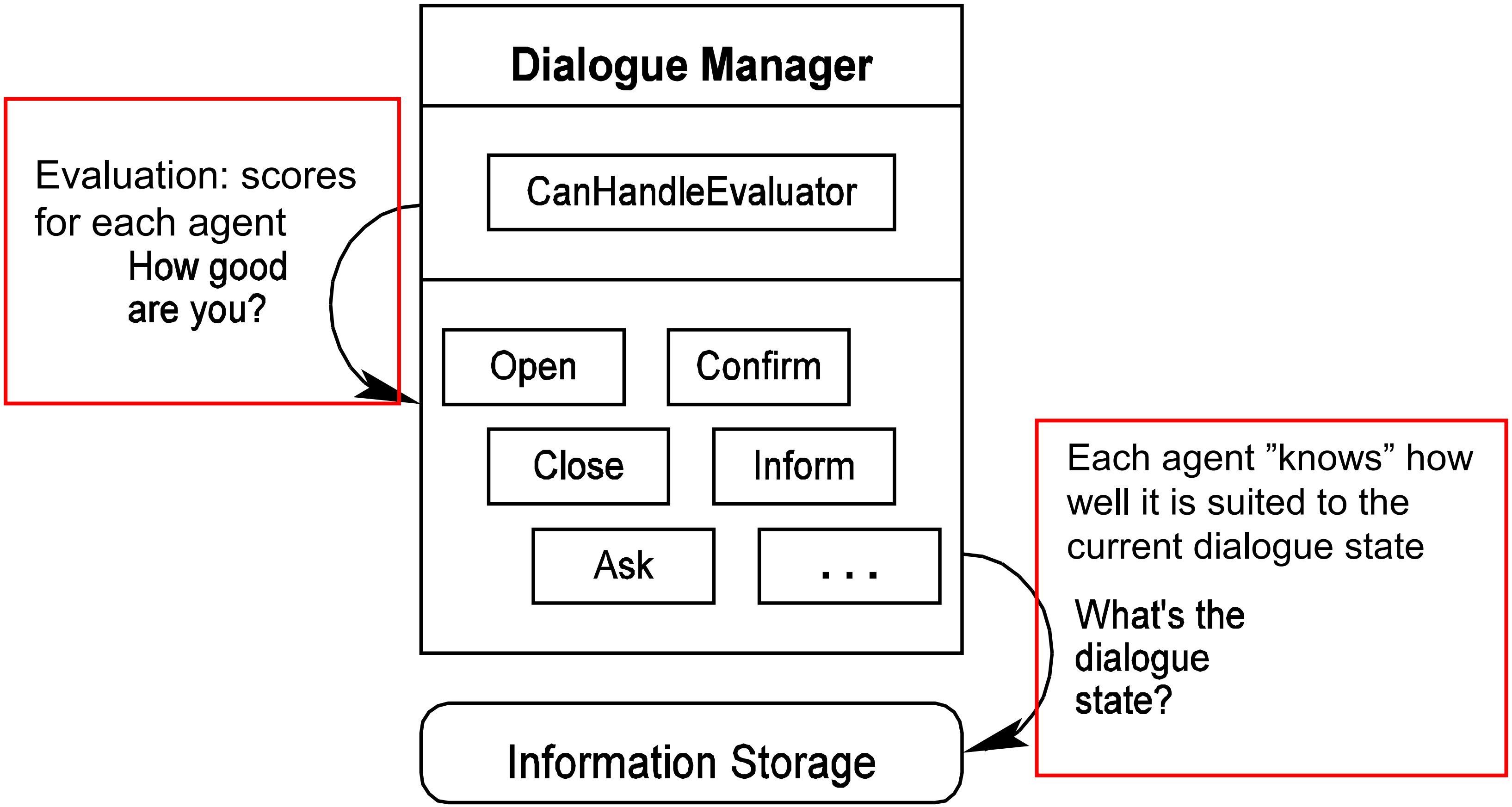
Interact system /Jaspis architecture



Jokinen et al. (2002)

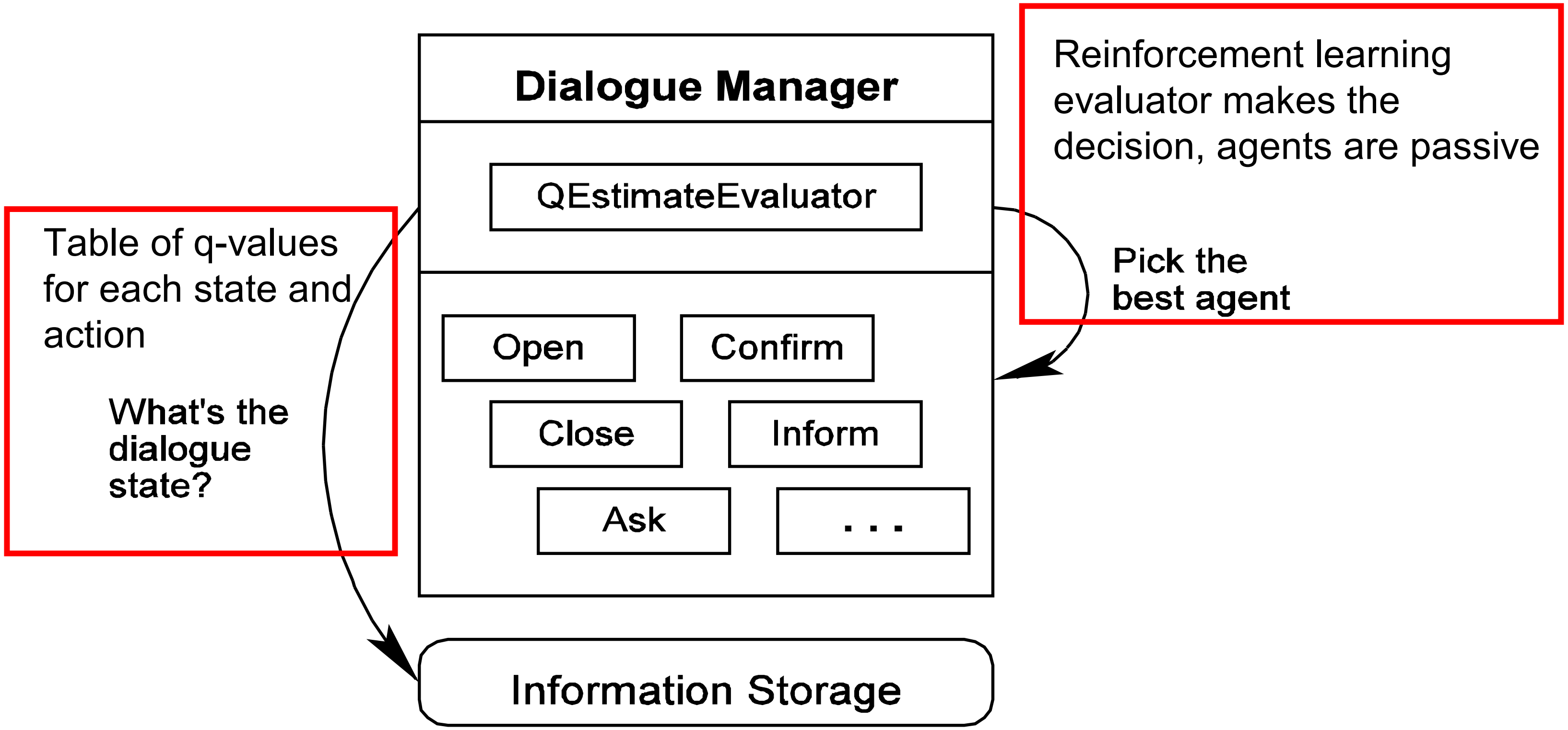
Turunen et al. (2005)

Heuristic Agent Selection



Adaptive Agent Selection

Kerminen and Jokinen (2003)



- Agent selection by managers compares to action selection by autonomous agents
- Use reinforcement learning to learn appropriate actions

Presentation of information

- Presentation of route instructions
  - Appropriate size of information at any given time
  - Take user's knowledge and skill levels into consideration
- Incremental representation of information
  - user can zoom in and out both verbally and on the map
- Allow users to give feedback on their understanding:
  - answer to an explicit question ("Did you say the Opera stop?", "Was it this one?")
  - acknowledge each item separately (system-initiative)
  - continue the interaction with an appropriate next step ("Give me the next piece of information") (user-initiative)
  - subtle verbal and non-verbal signals in the speech (variation of pronunciation together with the length of the following pause can signal wish to continue rather than the end of one's turn)

MUMS Example Dialogue

U: Uh, how do I get from the Railway station ... uh…

S: *Where would you like to go?*

U: Well, there!   + <map gesture>

S: *Tram 3B leaves Railway Station at 14:40, there is one change. Arrival time at Brahe Street 7 is 14:57.*

U: When does the next one go?

S: *Bus 23 leaves Railway Station at 14:43, there are no changes. Arrival time at Brahe Street 7 is 15:02.*

U: Ok. Navigate.

S: *Take bus 23 at the Railway Station at 14:43.*

U: Navigate more.

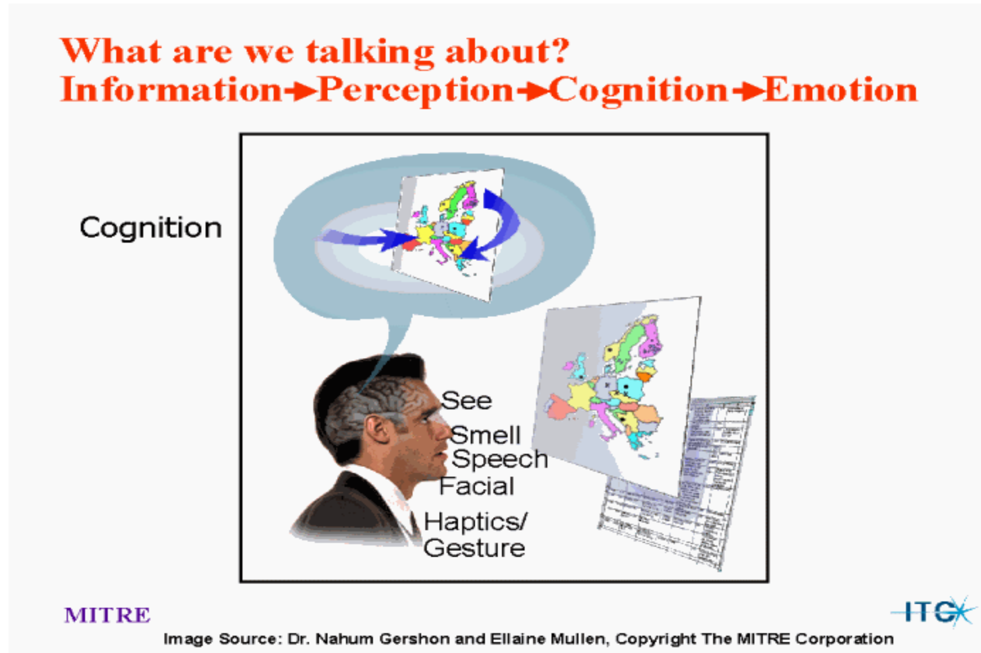
S: *Get off the bus at 14:49 at the Brahe Street stop.*

U: Navigate more.

S: *Walk 200 meters in the direction of the bus route. You are at Brahe Street 7.*

Multimodal Communication

- Human communication research
  - Perception: sensory info to higher level representations
  - Control: manipulation and coordination of information
  - Cognition
- Modality = senses employed to process incoming information



Mark Maybury, Dagstuhl Multi-Modality Seminar, 2001

Communicative Competence in DS

Jokinen, K. Rational Agents and Speech-based Interaction (2008, Wiley and Sons)

- Physical feasibility of the interface
  - Enablements for communication
  - Usability and transparency
  - Multimodal input/output, natural intuitive interfaces
- Efficiency of reasoning components
  - Speed
  - Architecture
  - Robustness

Communicative Competence in DS

- Natural language robustness
  - Linguistic variation
  - Interpretation/generation of utterances
- Conversational adequacy
  - Clear up vagueness, confusion, misunderstanding, lack of understanding
  - Non-verbal communication, feedback
  - Adaptation to the user

Summary

- Fusion:
  - Early vs late
  - Combining modalities that may support, complement or contradict each other
- Architecture and learning of interaction strategies
- Presentation
  - Different user interests and needs
- Effect of the modalities on the user interaction
  - Speech presupposes communicative capability
  - Tactile systems seem to benefit from speech as a value-added feature
  - Communicative competence

Thanks!

References

- Hurtig, T., Jokinen, K. 2006. Modality Fusion in a Route Navigation System. Proc. Workshop on Effective Multimodal Dialogue Interfaces EMMDI-2006. January 29, Sydney, Australia.
- Hurtig, T. 2005. Multimodaalisen informaation hyödyntäminen reitinopastusdialogeissa (Utilising Multimodal Information in Route Guidance Dialogues). Master's Thesis (in Finnish).
- Hurtig, T., Jokinen, K. 2005. On Multimodal Route Navigation in PDAs. Proc. 2nd Baltic Conference on Human Language Technologies HLT'2005. April 5, Tallinn, Estonia.
- Jokinen, K. 2007. Interaction and Mobile Route Navigation Application. In Meng, L., A. Zipf, and S. Winter (eds.) *Map-based mobile services - usage context, interaction and application*, Springer Series on Geoinformatics.
- Jokinen, K., Hurtig, T. 2006. User Expectations and Real Experience on a Multimodal Interactive System. *Proceedings of the Interspeech 2006*, Pittsburgh, US.
- Jokinen, K., Kerminen, A., Kaipainen, M., Jauhiainen, T., Wilcock, G., Turunen, M., Hakulinen, J., Kuusisto, J., Lagus, K. (2002). Adaptive Dialogue Systems - Interaction with Interact, *3rd SIGdial Workshop on Discourse and Dialogue*, July 11-12, 2002, Philadelphia, U.S. pp. 64 – 73.
- Kerminen, A., Jokinen, K. 2003. Distributed Dialogue Management in a Blackboard Architecture. *Proceedings of the EACL Workshop Dialogue Systems: interaction, adaptation and styles of management*, Budapest, Hungary. pp. 55-66.
- Turunen, M., Hakulinen, J., Räihä,K-J., Salonen, E-P., Kainulainen, A., Prusi, P. 2005. An architecture and applications for speech-based accessibility systems. IBM Systems Journal, Vol. 44, No 3, 2005.

Design a dialogue system...

- Requirements:
  - Travel planner for one-time visitor and a frequent user
  - Agent-based architecture
  - Speech interaction
  - Maintains dialogue history
  - Has a user model
  - Task model
- (practical exercise at the Elsnet Summer School 2007)

Results: 5 groups => 5 designs

- Differences along the lines:
  - Modularity of architecture: emphasis on different agents
  - Granularity of modules: task composition
  - Speech processing: prosody, emotional speech recognition
  - Dialogue history: evolution model vs. user model
  - User model: user profile (configuration) databases vs. conceptual modelling vs. distributed among other components
  - Task model: task ontology vs. dialogue manager
  - Generation of system responses: planning vs. templates
  - Reasoning components: elaborated pragmatic inferences vs. more shallow (hard-coded?) relations

Shared features of the 5 systems

1. Extract various information from the user and process it in detail
2. Parallel processing; provide correct dialogue behaviour time-wise
3. Take pragmatic aspects into account on several levels; user model scattered in different parts of the system; fine tuning of the system utterances
4. Adaptation and adaptability
5. Adapt speech models and provide different output modalities depending on user expertise