Language Technology II: Language-Based Interaction

Beyond Spoken Dialogue Systems

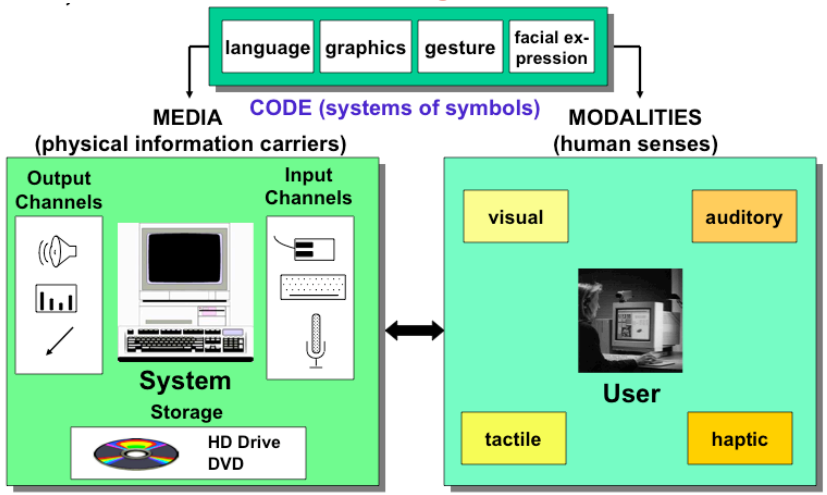
Ivana Kruijff-Korbayová
korbay@coli.uni-sb.de
www.coli.uni-saarland.de/courses/late2/

I have reused slides from presentations of W. Wahlster, M. Johnston and J. Cassell

Outline

- Multimodality and multimedia
- Multimodal Dialogue Systems for Mobile Environments
- Embodied Conversational Agents
  - Talking heads
  - Virtual animated characters
  - Talking robots

Interaction Using More Senses



(Wahlster, 2003)

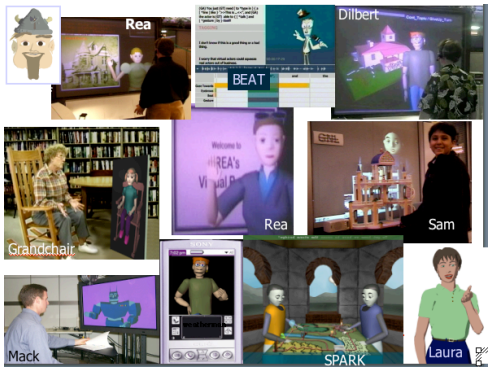
Multimodal and Multimedia Interfaces

- A multimodal/multimedia interface supports more than one mode/medium through which the users and applications programs can communicate
  - E.g., natural language, graphics, sketching, animation, gestures, menus, video, sounds ...
- Medium = physical carrier of information
  - E.g., acoustic vs. optical
- Mode = particular sign system
  - E.g., language vs. graphics
- Examples:
  - Text + graphics output on display = multimodal presentation
  - Speech + text output on display = monomodal multimedia presentation
  - Speech + DataGlove input = multimedia input

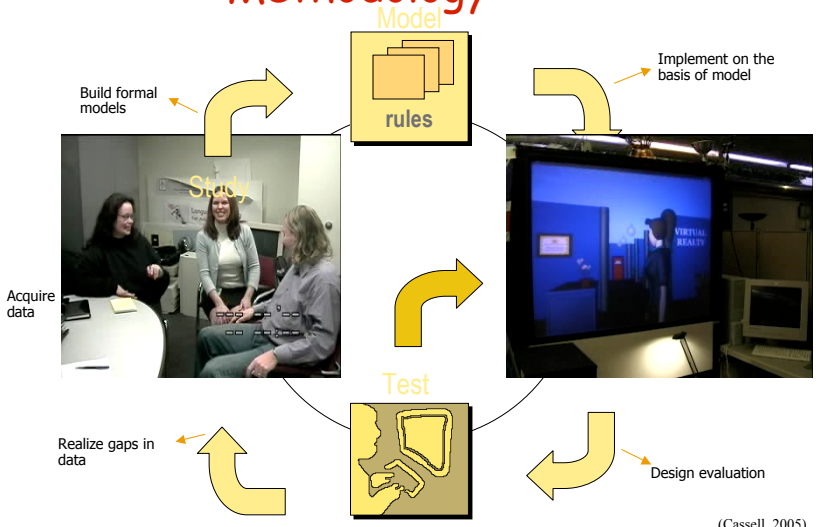
(Wahlster, 2004)

Anthropomorphic Interfaces

= interfaces which have a "persona", i.e. at least a face or a whole body often also called Embodied Conversational Agents (ECA)
  - Talking heads
  - Virtual animated characters
- Added aspects of social interaction



Methodology



(Cassell, 2005)

Posture Shifts mark the beginning of new discourse segments (Cassell et al., '01)

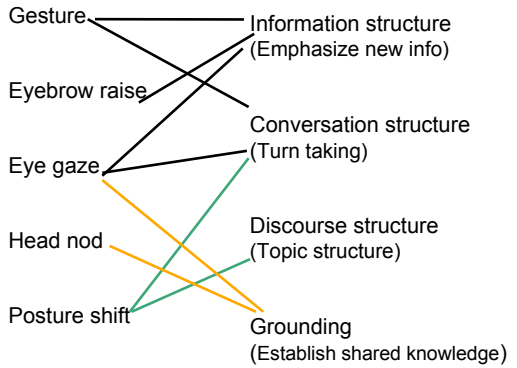
Looks towards the listener indicate that further grounding is needed (Nakano, et al. '02)

Gestures are more likely to occur with rhematic material than thematic material (Cassell et al. '94)

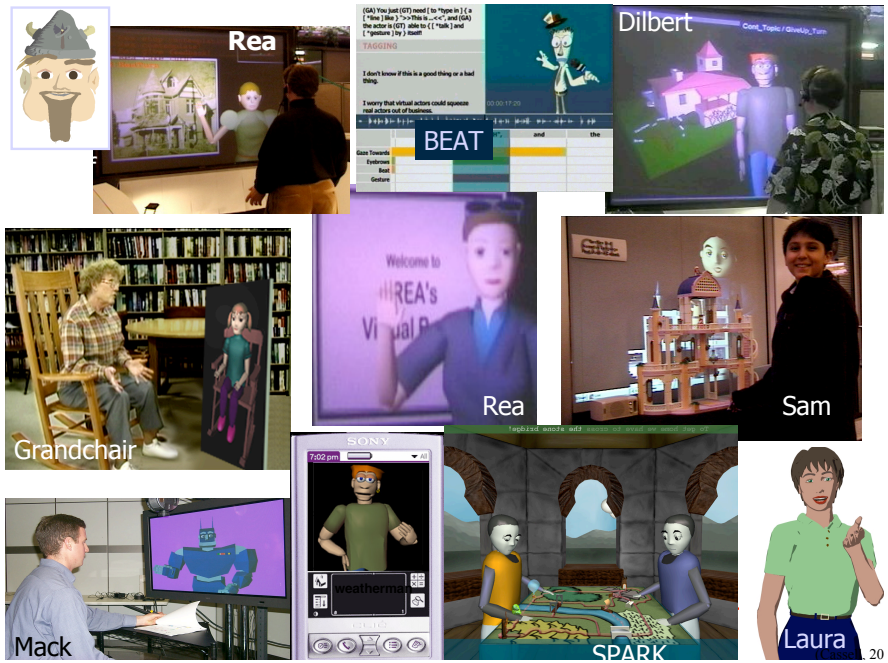
Small talk occurs before face-threatening discourse moves (Bickmore & Cassell, '02)

(Cassell, 2005)

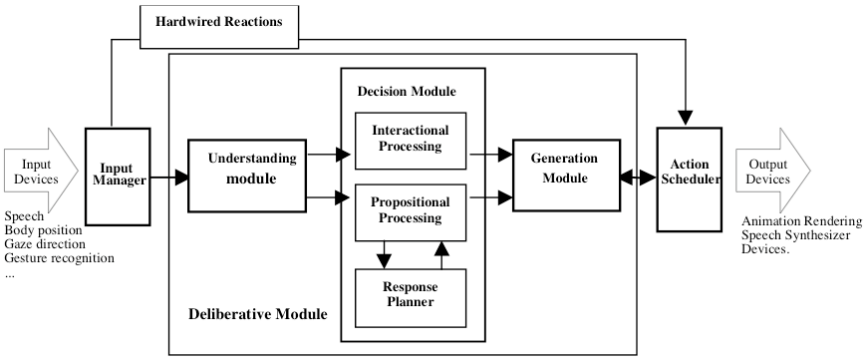
Relationship between Linguistic Structure & Behavioral Cues



(Cassell, 2005)



REA's Architecture



Intelligent MM Interfaces

Coexistence of input and output in different media and modes ≠ *Effective user interface*

(Wahlster, 2004)

Limitations of Current MM Technology

- Although multimedia technology is so popular, that virtually everyone is using it in some form, the technology is still in its infancy:
  - Canned presentation segments
  - Hardwired presentation sequences
  - Very limited user adaptation
  - Inadequate media coordination
  - No unanticipated presentation possible
  - No follow-up questions possible
  - Very limited user interaction
  - No high-level authoring tools
  - No content-based reuse strategies
  - No inference services
  - No representation of the content of the presentation
  - No formal semantics/pragmatics of presentations

(Wahlster, 2004)

MM Interfaces for Public Kiosks

- Since the introduction of ATMs in the 70's, public kiosks have been deployed to provide users with a broad range of of information and services
- BUT: Majority have rigid system-initiative graphical interfaces with user input by touch or keypad
  - Thus, they can only support simple tasks for able-bodies users
- To support more complex tasks for a broader range of users, kiosks will need to provide a more flexible and natural user interface
  - MM interfaces provide naturalness and flexibility
  - E.g., Gustafson et al. 1999 (August), Narayan et al. 2000 (MVPQ), Raisamo 1998, Lamel et al. 2002 (MASK), Wahlster et al. 2003 (SmartKom Public), Cassell et al. (MACK)

(Johnston, 2004)

Burning Research Issues in Multimodal Dialogue Systems

- **Multimodality**: from alternate modes of interaction towards **mutual disambiguation and synergistic combinations**
- **Discourse Models**: from information-seeking dialogs towards **argumentative dialogs and negotiations**
- **Domain Models**: from closed world assumptions towards **the open world of web services**
- **Dialog Behaviour**: from automata models towards a **combination of probabilistic and plan-based models**

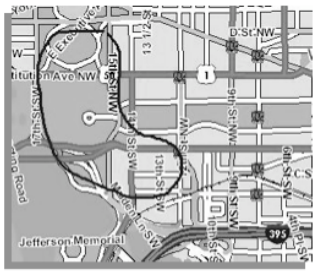
Intelligent MM Interfaces

Coexistence of input and output in different media and modes $\neq$ *Effective user interface*

- From alternate modes of interaction to composite multimodality
- Careful coordination of different media and modes in a coherent and cooperative dialogue is required

Composite Multimodality: Input

- Composite input:
  - Enabling users to provide a single contribution (turn) which is optimally distributed over the available input modes
    e.g., speech + ink "zoom in here"

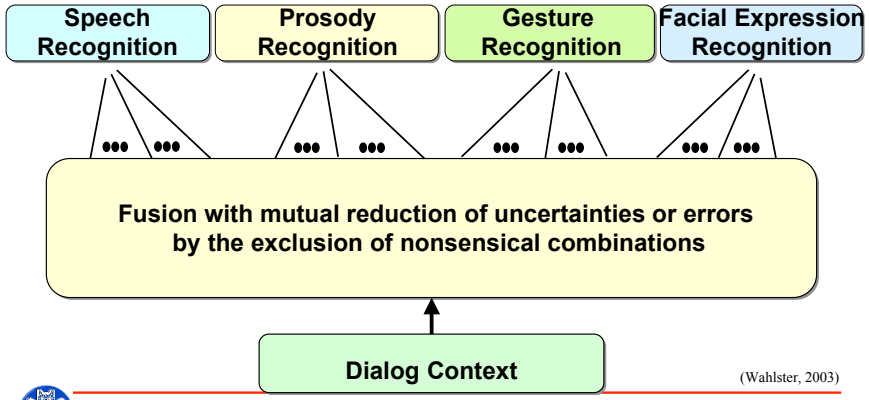


- Motivation
  - Naturalness
  - Certain kinds of content within a single communicative act are best suited to particular modes, e.g.,
    - Speech for complex queries or constraints, reference to objects currently not visible or intangible
    - Ink/gesture for selection, indicating complex graphical features
  - Empirical studies
    - Task performance and user preference advantages (Oviatt et al. 1999)
    - Compensation for errors (Oviatt 1999; Bangalore & Johnston n 2000)

(Johnston, 2004)

Composite Multimodality: Input Fusion

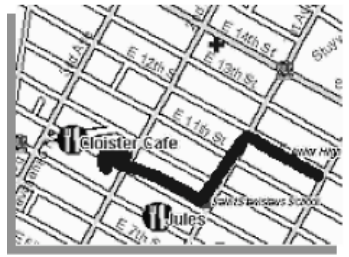
Mutual disambiguation and synergistic combinations: semantic fusion of multiple modalities in dialog context helps to reduce ambiguity and errors



(Wahlster, 2003)

Composite Multimodality: Output

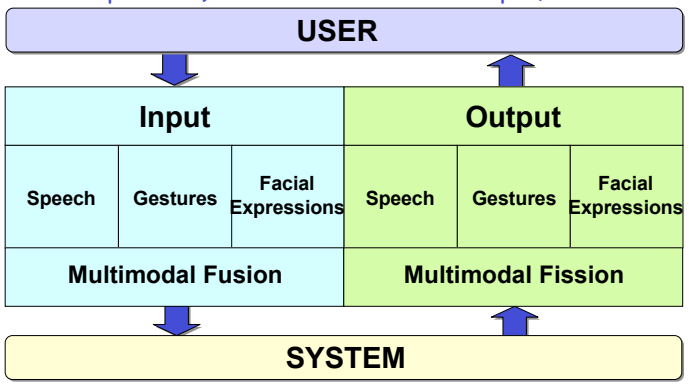
- Composite output:
  - Allowing for system output to be optimally distributed over the available output modes, e.g.,
    - High level summary in speech, details in graphics: "Take this route across town to the Cloister Café"
    - Multimodal help providing examples for the user: "To get the phone number for a restaurant, circle one like *this* and say or write *phone*." (Hastie et al. 2002)
  - Output should be dynamically tailored to be maximally effective given the situation and user preferences
- Same motivation as for multimodal input



(Johnston, 2004)

Full Symmetric Multimodality

Symmetric multimodality means that all input modes (speech, gesture, facial expression) are also available for output, and vice versa.



The modality fission component provides the inverse functionality of the modality fusion component.

Challenge: A dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own.

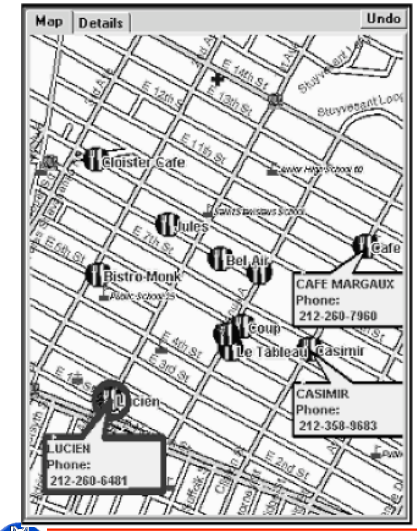
(Wahlster, 2003)

MATCH: Multimodal Access to City Help

- Interactive city guide and navigation for information-rich urban environments
  - Finding restaurants and points of interest, getting info, subway routes for New York and Washington, D.C.
- Composite input and output
  - Speech, ink, graphics
- Mobile (standalone on a tablet or distributed WLAN)
- MATCHKiosk (deployed at AT&T visitor center in DC)
  - Social interaction
  - Also printed output

(Johnston, 2004)

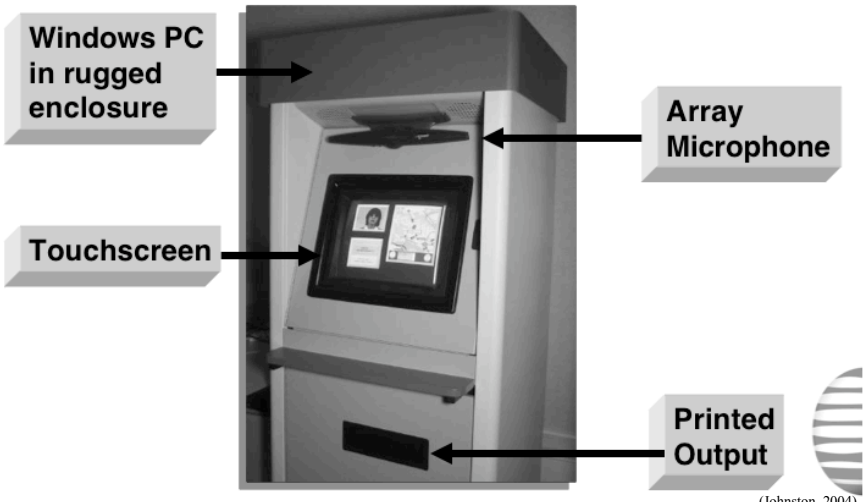
MATCH



- Finding restaurants
  - Speech: "show inexpensive italian places in chelsea"
  - Multimodal: "cheap italian places in this area"
  - Pen: 
  - Getting info: "numbers for these"
  - Subway routes: "how do I get here from Broadway and 95th street"
  - Pan/zoom map: "Zoom in here"

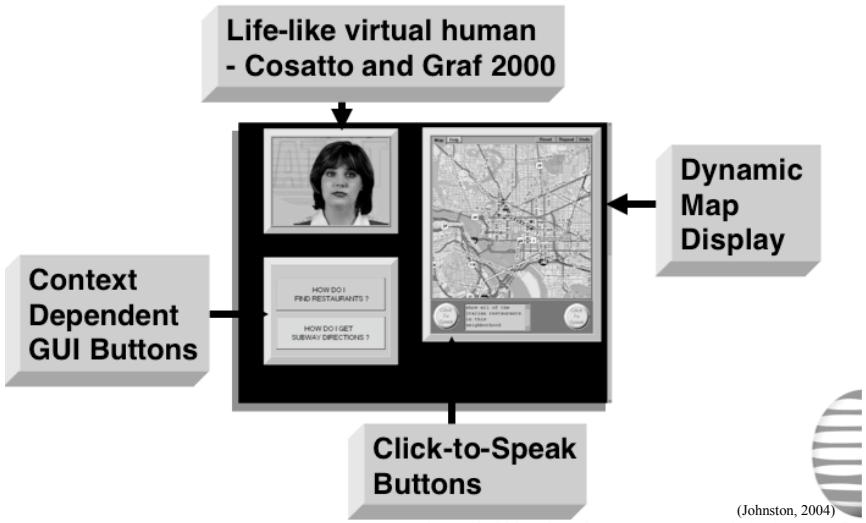
(Johnston, 2004)

MATCHKiosk



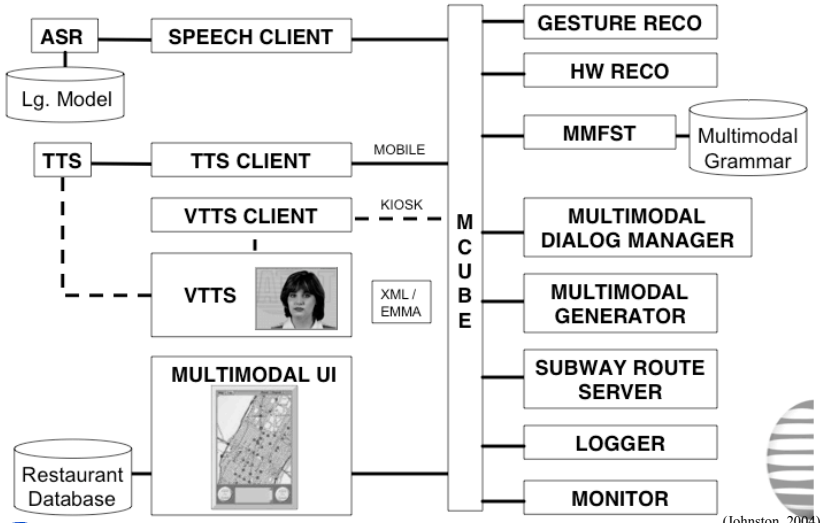
(Johnston, 2004)

MATCHKiosk

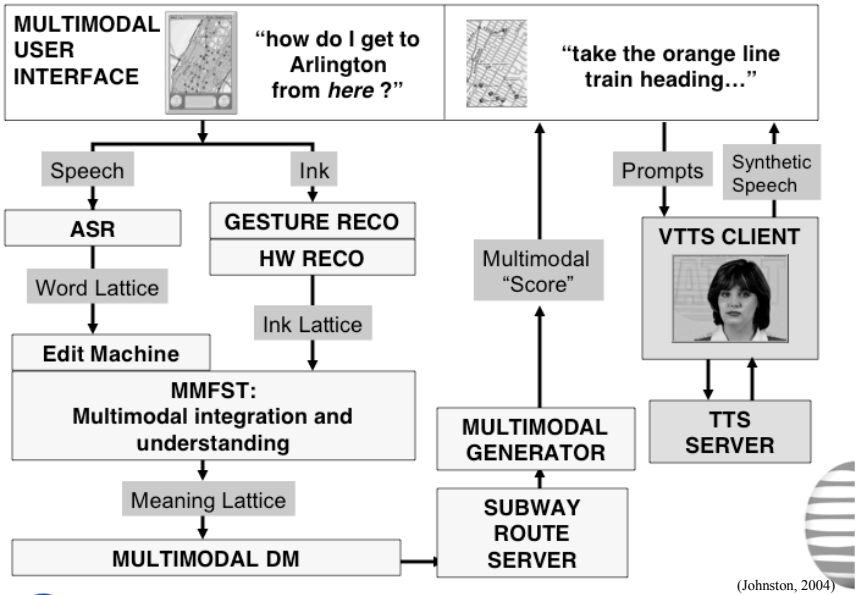


(Johnston, 2004)

MATCH Architecture

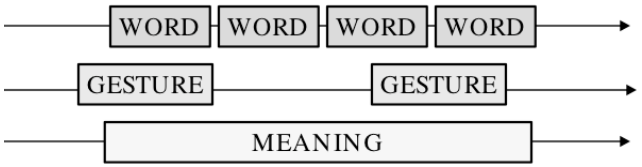


(Johnston, 2004)



(Johnston, 2004)

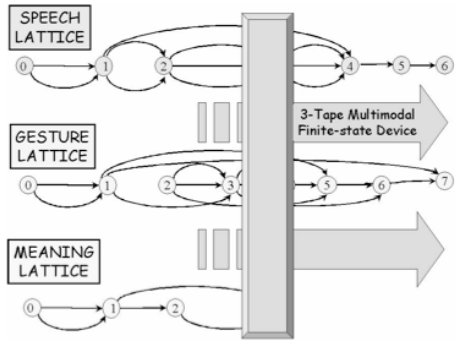
Multimodal Understanding



- Associate word sequence + gesture sequence with meaning
  - Late integration: first compute meaning of word sequence and meaning of gesture sequence, then "merge" the meanings, e.g., (Johnston 1998), SmartKom: overlay (Alexandersson&Becker 2001; Pfleger 2002)
  - Early integration: compute meaning of a composite word+gesture sequence: MMFST (Johnston&Bangalore 2000,2004)
    - Enables compensation for errors in individual modes

MMFST: Early Multimodal Integration

- Speech and gesture parsing, multimodal integration, and understanding in single MM grammar model
  - (Johnston&Bangalore 2000,2004)
  - Compiled to efficient finite state device
    - G:W transducer aligns speech and ink
    - G_W:M transducer takes a composite alphabet of speech and gesture symbols and outputs meaning
  - Compiled from a declarative multimodal CFG (terminals are triples W:G:M = Words:Gestures:Meaning)
- Robust, efficient
- Enables compensation for errors



(Johnston, 2004)

MMFST: Example

"Phone numbers for these three restaurants"

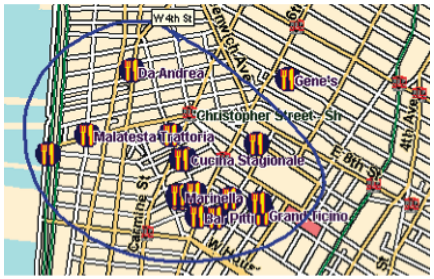
S	→	eps:eps:<cmd> CMD eps:eps:</cmd>
CMD	→	phone:eps:<phone> numbers:eps:eps for:eps:eps DEICTICNP eps:eps:</phone>
DEICTICNP	→	DDETPL eps:area:eps eps:selection:eps NUM RESTPL eps:eps:<restaurant> eps:SEM:SEM eps:eps:</restaurant>
DDETPL	→	these:G:eps
RESTPL	→	restaurants:restaurant:eps
NUM	→	three:3:eps

Johnston et al. (2004)

User-Tailored Generation

"compare these restaurants"

- User-tailored summaries, comparisons or recommendations can be generated using a model of user preferences



Compare-A: *Among the selected restaurants, the following offer exceptional overall value. Uguale's price is 33$. It has excellent food quality and good decor. Da Andrea's price is 28$. It has very good food quality an good decor. John's Pizzeria's price is 20$. It has very good food quality and mediocre decor.*

Compare-B: *Among the selected restaurants, the following offer exceptional overall value. Babbo's price is 60$. It has superb food quality. Il Mulino's price is 65$. It has superb food quality. Uguale's price is 33$. It has excellent food.*

Johnston et al. (2004)

The Need for Going Mobile

Broadband mobile Internet access technologies via UMTS or mobile hotspots pave the way for a wide spectrum of added-value web services.

but: the user must input more and more complex commands to specify his information needs.

PDAs and smartphones with tiny keyboards and mice are useless for mobile settings.

→ Multimodal Dialogue Systems for Mobile Systems

(Wahlster, 2003)

Multimodal Mobile Devices

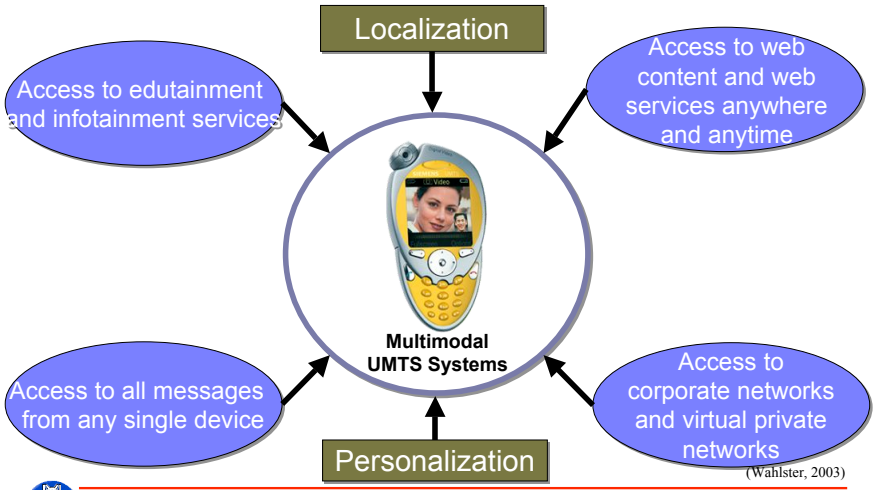
Today's Cell Phone **Third Generation UMTS Phone**

Verbmobil
Speech only

SmartKom
Speech, Graphics and Gesture

(Wahlster, 2003)

Intelligent Interaction with Mobile Internet Services



(Wahlster, 2003)

Mobile Access to an Edutainment System

Mobile Dialogue with a Virtual Tourist Guide for the Heidelberg Castle

Location-adaptive Query Interpretation

Cooperation: DFKI, EML, DFG, FhG

(Wahlster, 2003)

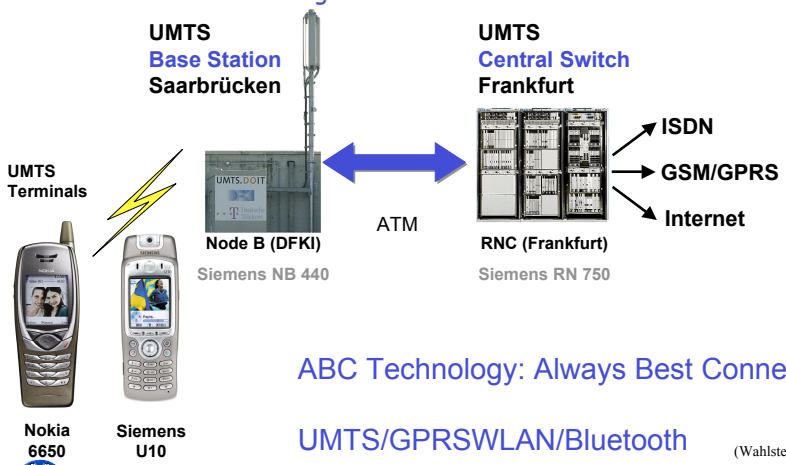
Spoken Dialogues with the Car Navigation System

Pervasive Computing

- conversational access to information and services anytime / anywhere
- event-driven interaction rather than traditional "well-formed" dialogues
  - task-switching
  - interruption
- deployment in robust environments
- adaptive and multimodal
  - needs of user
  - different environments

Going Mobile

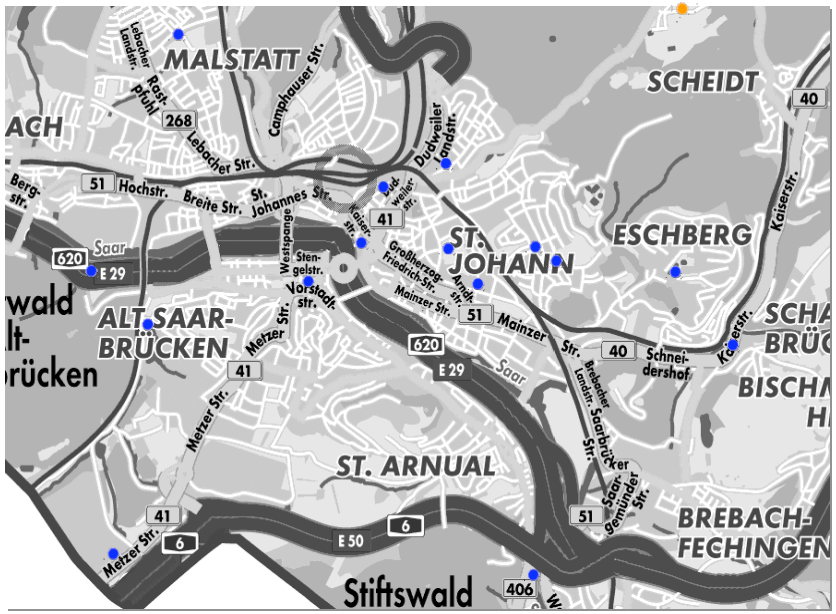
DFKI operates the first German Test Site for Innovative 3G/4G Broadband Mobile Technologies



ABC Technology: Always Best Connected

UMTS/GPRSWLAN/Bluetooth

(Wahlster, 2003)



(Wahlster, 2003)

CityShopper's Added-Value Mobile Service

Please let me know, when I pass a shop selling batteries.

- **CityShopper sends a note to the user or activates an alarm as soon as the user approaches a shop that offers of an item on his active shopping list.**
- **CityShopper's spatial alarm can be combined with:**
  - **route planning and navigation**
  - **temporal and spatial optimization of the shopping tour**

(Wahlster, 2003)

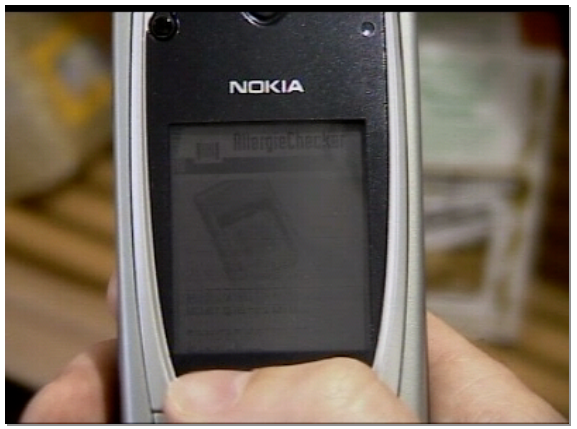
Spatial Alarm by CityShopper

A Location-based Active Shopping List based on UMTS
(Cooperation of DFKI with Telekom)

(Wahlster, 2003)

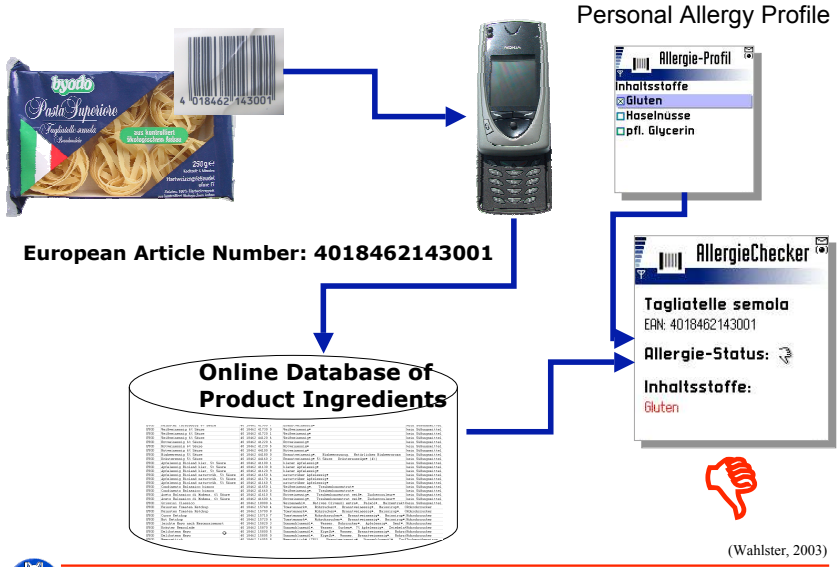
Mobile AllergyChecker

Mobile Bar Code Analyzer and Broadband Personalized Allergy Checker
(cooperation of DFKI with Mineway)



(Wahlster, 2003)

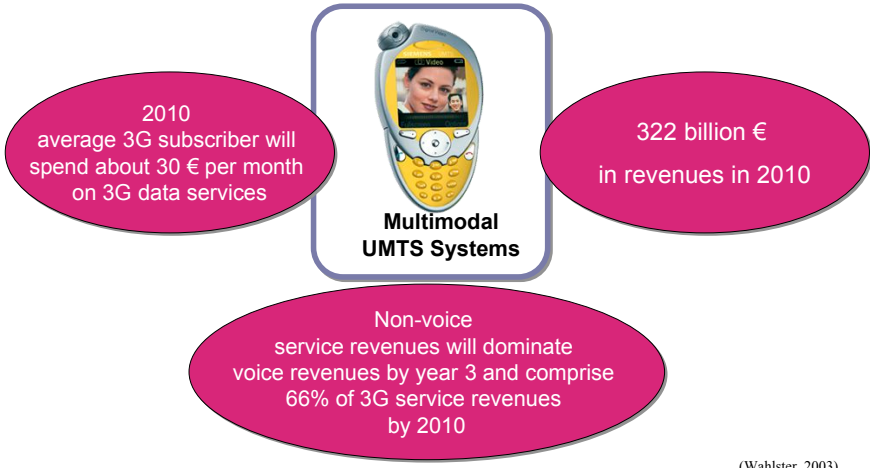
Real-Time Video-based Product Identification and Personalized Allergy Check



(Wahlster, 2003)

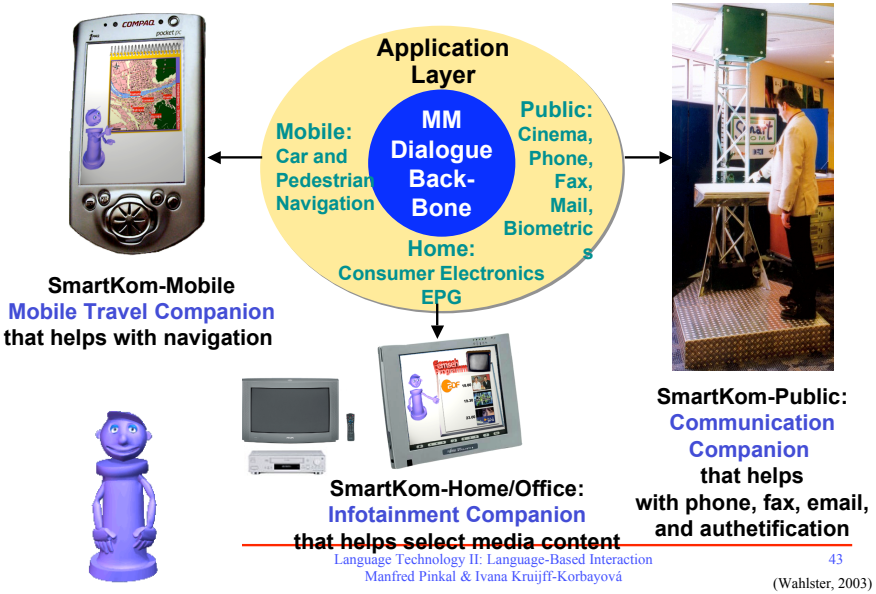
Market studies predict:

Cumulative revenues of almost 1 trillion €
from launch of mobile 3G services until 2010



(Wahlster, 2003)

SmartKom



SmartKom-Mobile
Mobile Travel Companion
that helps with navigation

SmartKom-Public:
Communication Companion
that helps
with phone, fax, email, and authetification

SmartKom-Home/Office:
Infotainment Companion
that helps select media content

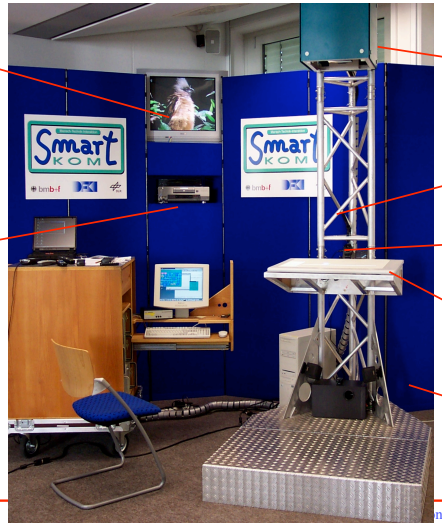
(Wahlster, 2003)

SmartKom

SmartKom-Home: DFKI Collaboration with Sony und Philips

(Wahlster, 2003)

SmartKom-Public (Kiosk Version)



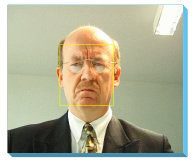
(Wahlster, 2003)

Using Facial Expression Recognition for Affective Personalization

Processing ironic or sarcastic comments

(1) Smartakus: Here you see the CNN program for tonight.

(2) User: That's great.



(3) Smartakus: I'll show you the program of another channel for tonight.

(2') User: That's great.

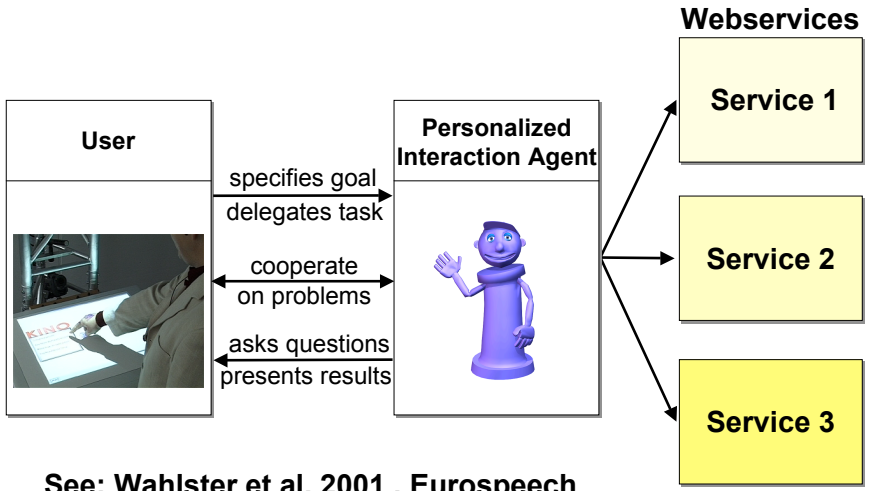
(3') Smartakus: Which of these features do you want to see? (Wahlster, 2003)

SmartKom

	Input by the User	Output by the Presentation agent
Speech	+	+
Gesture	+	+
Facial Expressions	+	+

(Wahlster, 2003)

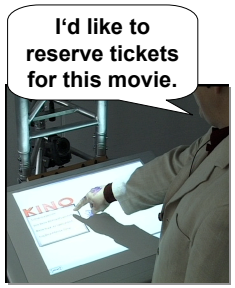
SmartKom



See: Wahlster et al. 2001 , Eurospeech

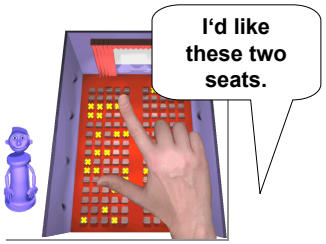
(Wahlster, 2003)

SmartKom



User Input:
Speech and Gesture

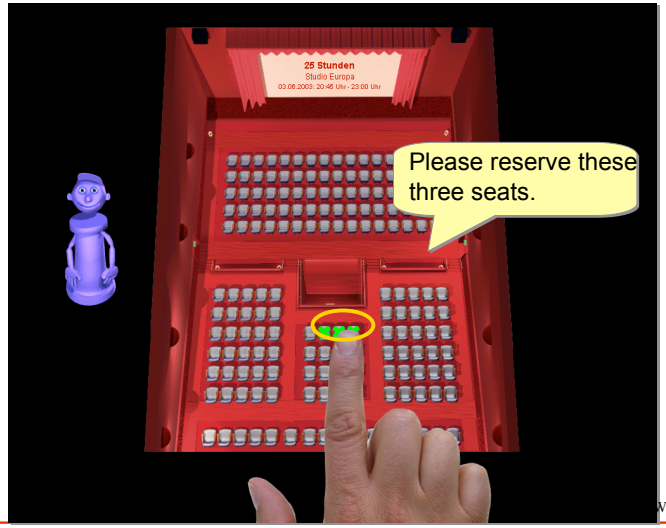
Smartakus Output:
Speech, Gesture and Facial Expressions



User Input:
Speech and Gesture

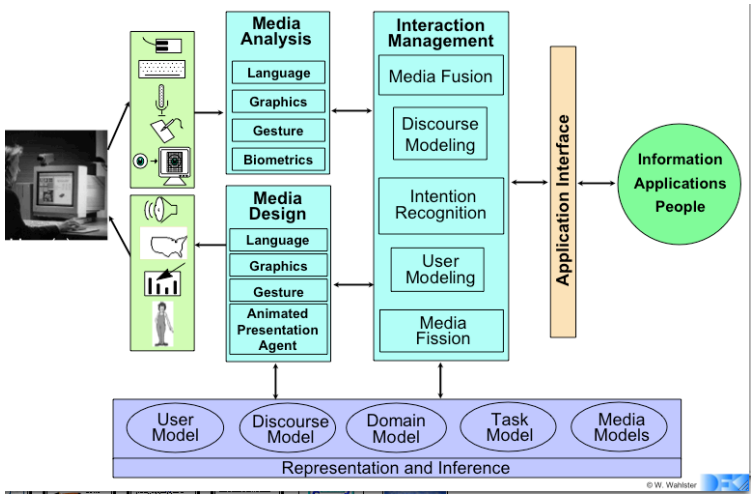
(Wahlster, 2003)

SmartKom

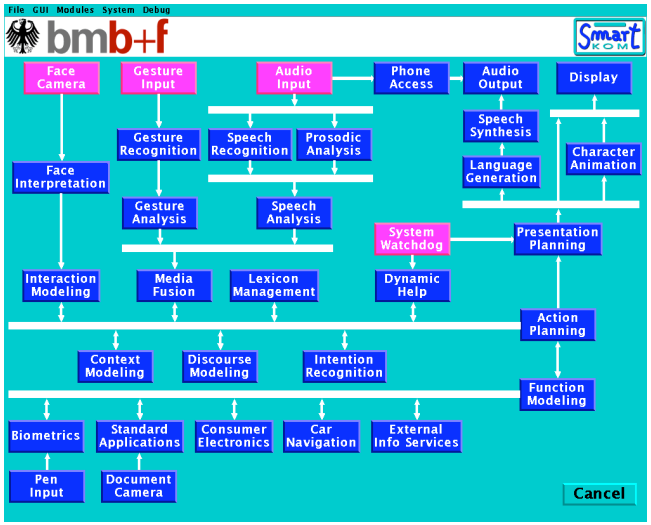


(Wahlster, 2003)

SmartKom Architecture

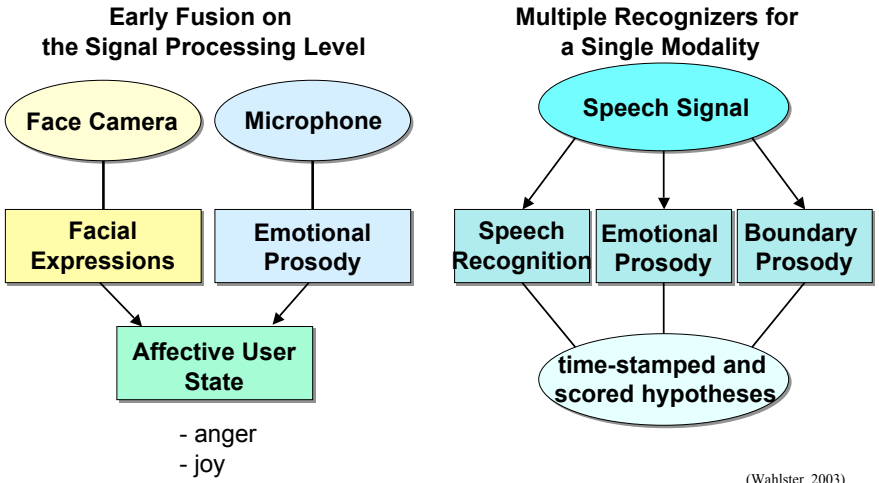


The High-Level Control Flow of SmartKom



(Wahlster, 2003)

Fusing Symbolic and Statistical Information in SmartKom



(Wahlster, 2003)

Computational Mechanisms for Modality Fusion and Fission

Modality Fusion	Modality Fission
Unification Overlay Operations	**Ontological Inferences**
	Planning Constraint Propagation

M3L: Modality-Free Semantic Representation

(Wahlster, 2003)

M3L Representation of an Intention Lattice Fragment

```
<intentionLattice>
[...]
  <hypothesisSequences>
   <hypothesisSequence>
    <score>
     <source>   acoustic   </source>
     <value>    0.96448    </value>
    </score>
    <score>
     <source>   gesture    </source>
     <value>    0.99791    </value>
    </score>
    <score>
     <source>   understanding   </source>
     <value>    0.91667         </value>
    </score>
    <hypothesis>
     <discourseStatus>
      <discourseAction>  set  </discourseAction>
      <discourseTopic><goal>  epg_info  </goal></discourseTopic>
     [...]
     <event id="dim868">
      <informationSearch id="dim869">
       <pieceOfInformation>
        <broadcast id="dim863">
         <avMedium>
          <avMedium id="dim866">
           <avType>  featureFilm  </avType>
            <title>   Enemy of the State   </title>
           [...]
   </hypothesisSequence>
[...]
  </hypothesisSequences>
</intentionLattice>
```

I would like to know more about this

Confidence in the Speech Recognition Result

Confidence in the Gesture Recognition Result

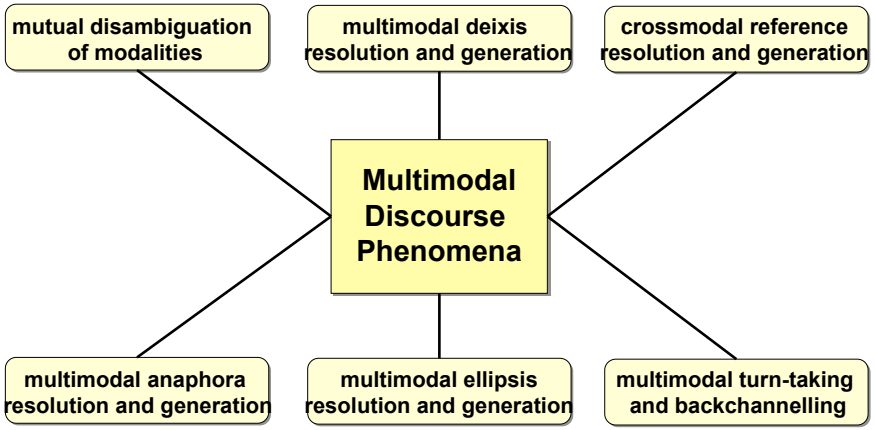
Confidence in the Speech Understanding Result

Planning Act

Object Reference

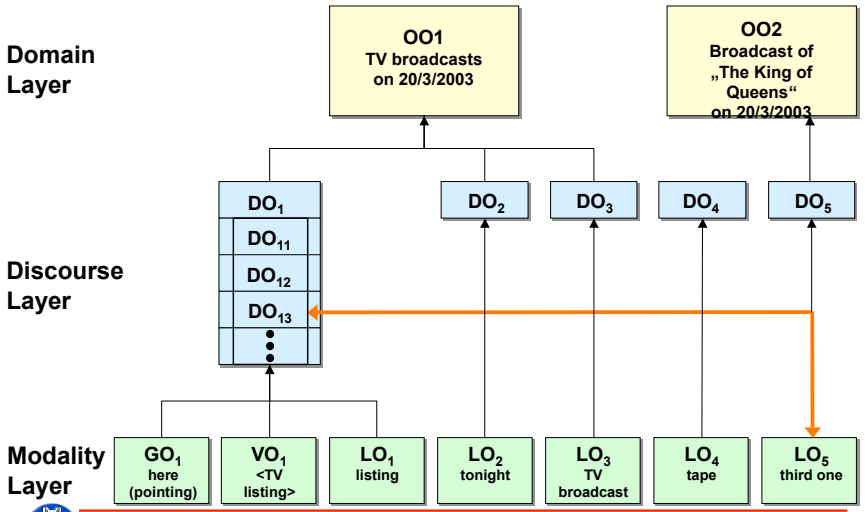
Multimodal Discourse Phenomena

Symmetric multimodality is a prerequisite for a principled study of many discourse phenomena.



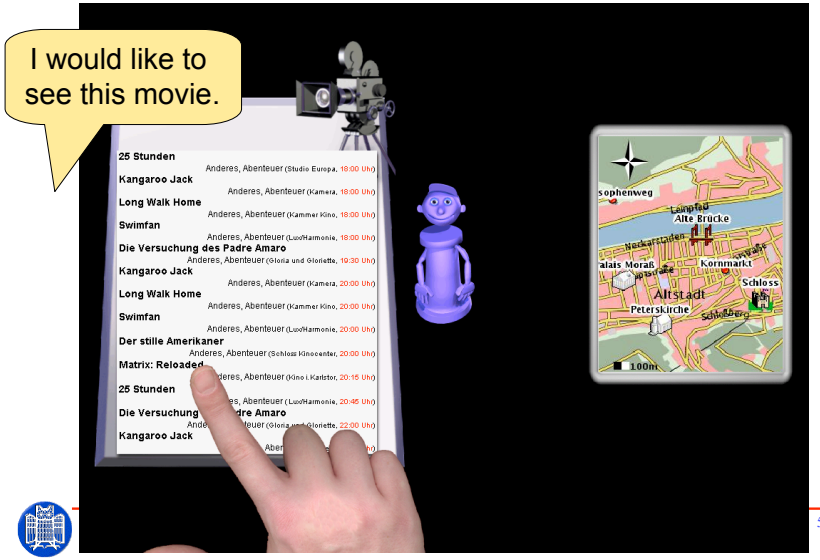
(Wahlster, 2003)

Three-Tiered MM Discourse Model



(Wahlster, 2003)

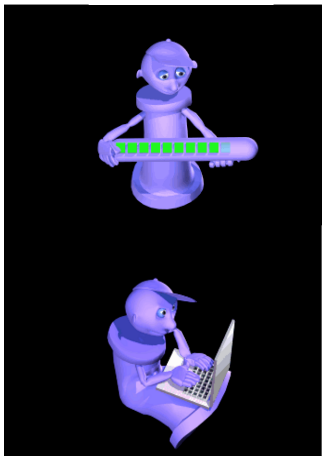
Reference Resolution based on a Symbolic Representation of the Smart Graphics Output



Generating Maps, Animations and Information Displays on the Fly

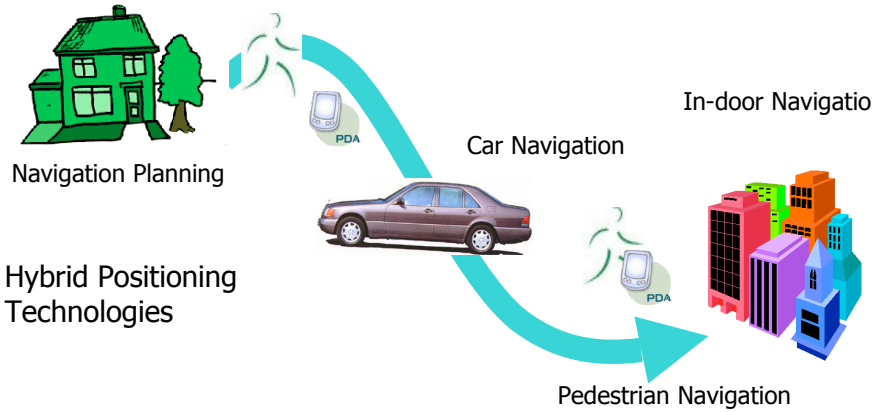
Synchronization of Map Update and Character Behaviour

Adaptive Perceptual Feedback on the System State



(Wahlster, 2003)

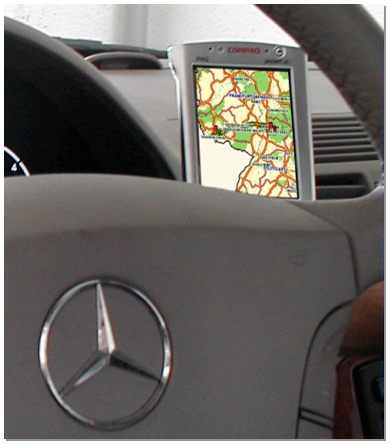
Seamless Navigation Services



(Wahlster, 2003)

Getting Driving and Walking Directions via SmartKom

SmartKom can be used for Multimodal Navigation Dialogues in a Car



User: I want to drive to Heidelberg.

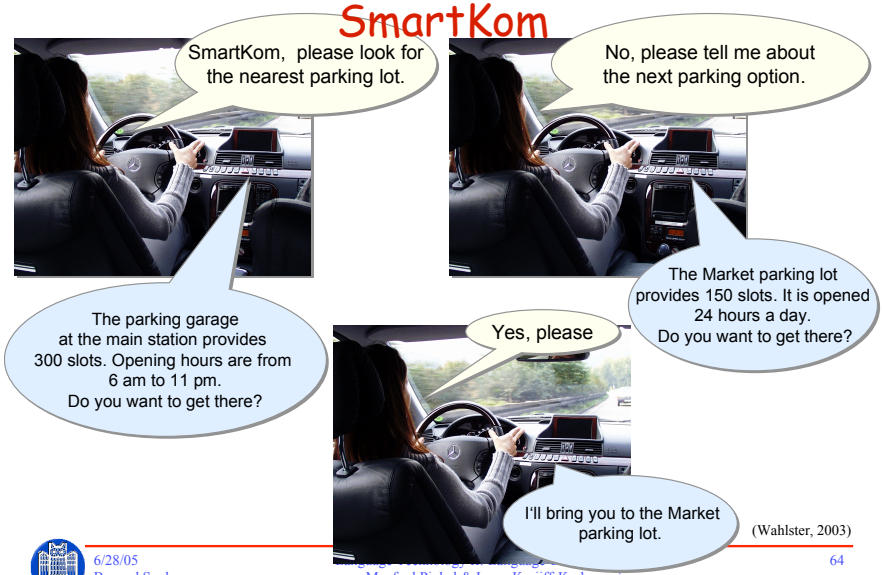
Smartakus: Do you want to take the fastest or the shortest route?

User: The fastest.

Smartakus: Here you see a map with your route from Saarbrücken to Heidelberg.

(Wahlster, 2003)

Spoken Navigation Dialogues with SmartKom



(Wahlster, 2003)

Salient Characteristics of SmartKom

• Seamless integration and mutual disambiguation of multimodal input and output on semantic and pragmatic levels

• Situated understanding of possibly imprecise, ambiguous, or incom-plete multimodal input

• Context-sensitive interpretation of dialog interaction on the basis of dynamic discourse and context models

• Adaptive generation of coordinated, cohesive and coherent multimodal presentations

• Semi- or fully automatic completion of user-delegated tasks through the integration of information services

• Intuitive personification of the system through a presentation agent

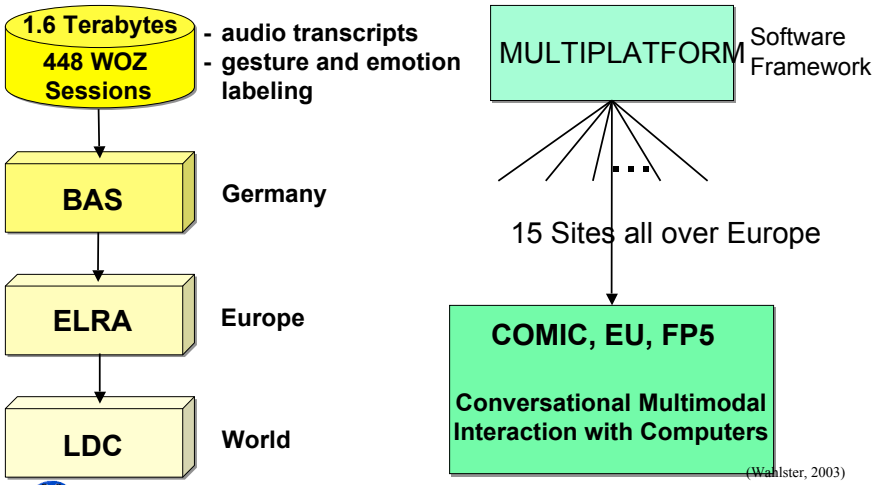
(Wahlster, 2003)

SmartKom Keypoints

- Various types of unification, overlay, constraint processing, planning and ontological inferences are the fundamental processes involved in SmartKom's modality fusion and fission components.
- The key function of modality fusion is the reduction of the overall uncertainty and the mutual disambiguation of the various analysis results based on a three-tiered representation of multimodal discourse.
- A multimodal dialogue sytsem must not only understand and represent the user's input, but its own multimodal output.

(Wahlster, 2003)

SmartKom's Impact on Software Tools and Resources for Research on Multimodality



(Wahlster, 2003)

Ambient Intelligence

=

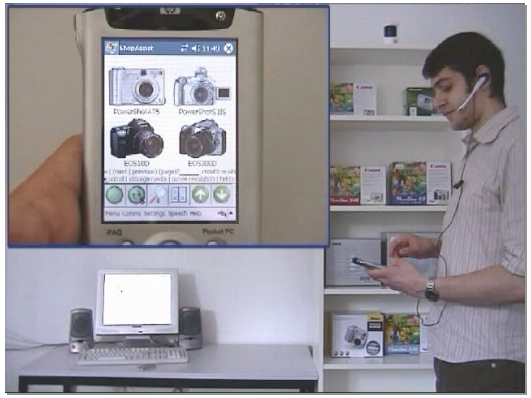
Embedding Artificial Intelligence in Everyday Objects and Environments

Key Characteristics of Ambient Intelligence:

Embedded	Many networked devices are integrated into the environment
Situated	These devices can recognize situational context
Personalized	They can be tailored towards your needs and affects
Adaptive	They can change in response to you and your task
Pro-active	They try to anticipate your plans and intentions

Mobile Shop Assistant

Automatic Comparison Shopping with Ambient Intelligence



Multimodal Fusion of Speech, Gesture and Physical Actions (Intra- and Extra-Gestures)

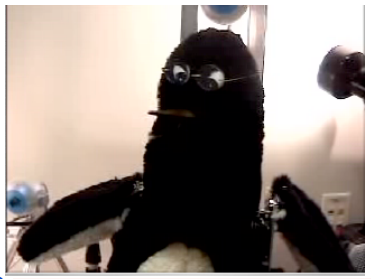
Conversational Robots

- Some examples:
  - WITAS (CSLI)
    - Unmanned aerial vehicle: start, land, locate & follow objects
    - Collaborative activities, multitasking
  - Godot (Edinburgh)
    - Navigation, learning natural language descriptions of objects
  - MEL (MERL)
    - Hosting
    - Egagement behaviors (initiate, maintain and disengage)
    - Modeling of look-tracking (head movement)
  - COSY project (DFKI)
- Language + vision + oter sensory input + gestures
- Activities situated in physical world

Conversational Robots

WITAS

GODOT

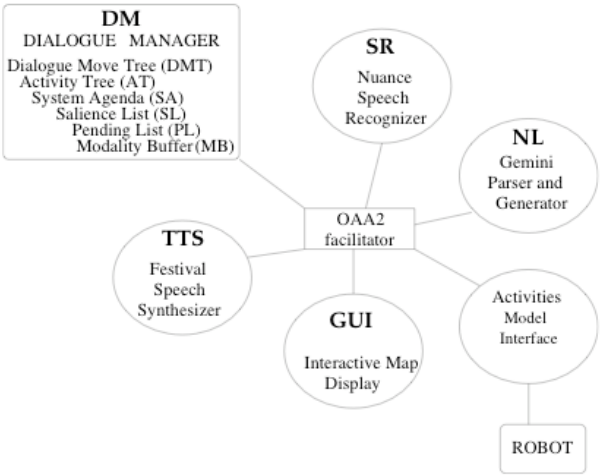


MEL

WITAS

Multi-modal Utterances	Dialogue Move
Operator (O) : *Our job is to look for a red car*	Command (Joint Activity)
UAV (U) : *Ok. I am looking for one.*	Report (Confirm Activity)
O : *Fly here please [+click on map]*	Command (Deictic)
U : *Okay. I will fly to waypoint one*	Report (Confirm Activity)
U : *Now taking off and flying there.*	Report (Current Activity)
O : *Stop that. Go to the tower instead.*	Command, Revision
U : *I have cancelled flying to waypoint one. I will fly to the tower.*	Report (Activity status)
O : *What are you doing ?*	Wh-question (Current Activity)
U : *I am searching for a red car and flying to the tower*	Answer (Current Activity)
O : *What will you do next ?*	Wh-question (Planned Activity)
U : *I have nothing planned.*	Answer(Planned Activity)
U : *I see a red car on main street* [display on map, show video images], *Is this the right car ?*	Report, Yn-question (Activity)
O : *Yes, that's the right car*	Yn-answer (Positive)
U : *Okay. I am following it .*	Report (Current activity)

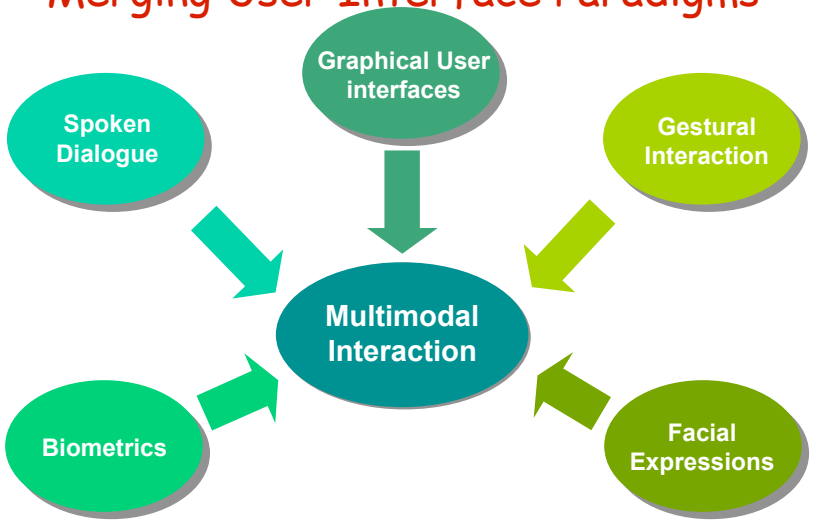
WITAS



Wrapping Up

- Multimodal and anthropomorphic interfaces
- Symmetric multimodality
- Composite multimodality
  - Fusion
  - Fission
- Adaptivity, personalization, collaboration (activity & pro-activity)
- Anthropomorphic objects
- Robots

Merging User Interface Paradigms



(Wahlster, 2003)

References

- M. Johnston et al. "MATCH: An architecture for Multimodal Dialogue Systems." In Proc. Of the 40th Annual Meeting of ACL. pp. 376-383. 2002.
- N. Pfleger et al. "Robust Multimodal Discourse Processing." In Proc. Of DiaBruck. pp. 107-114. 2003.
  SmartKom website: http://www.smartkom.org/
- J. Cassell. "More than just another pretty face: Embodied conversational interface agents." Communications of the ACM 43(4): 70-78.2000.
- Oliver Lemon, Alexander Gruenstein, and Stanley Peters, "Collaborative Activities and Multi-tasking in Dialogue Systems." Traitement Automatique des Langues (TAL), special issue on dialogue, 43(2): 131 - 154, 2002.
- Sidner et al. "Egagement by Looking:Behaviors for Robots when collaborating with people." In Proc. Of DiaBruck. pp. 123-130. 2003.