Dialog systems

Professor Joakim Gustafson

CV for Joakim Gustafson

1987-1992 Electrical Engineering program at KTH

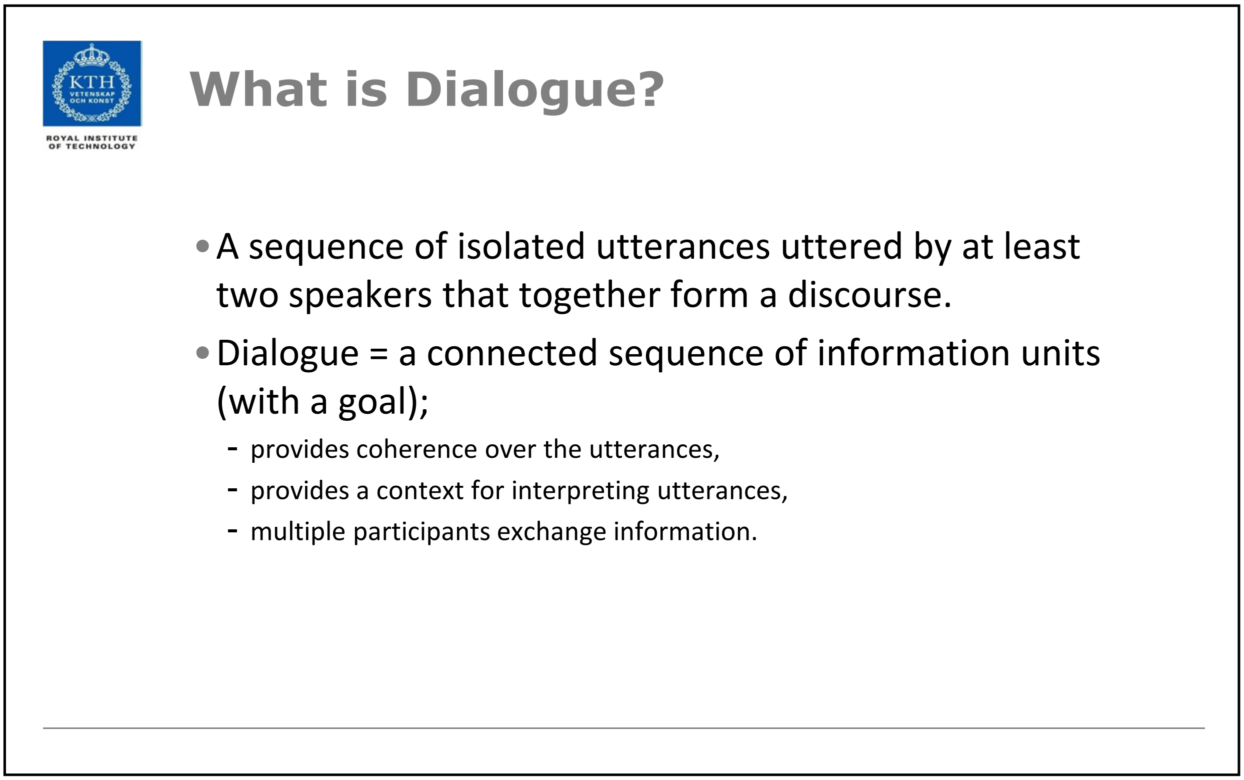
1992-1993 Linguistics at Stockholm University

1993-2000 PhD studies at KTH

2000-2007 Senior researcher at Telia R&D department

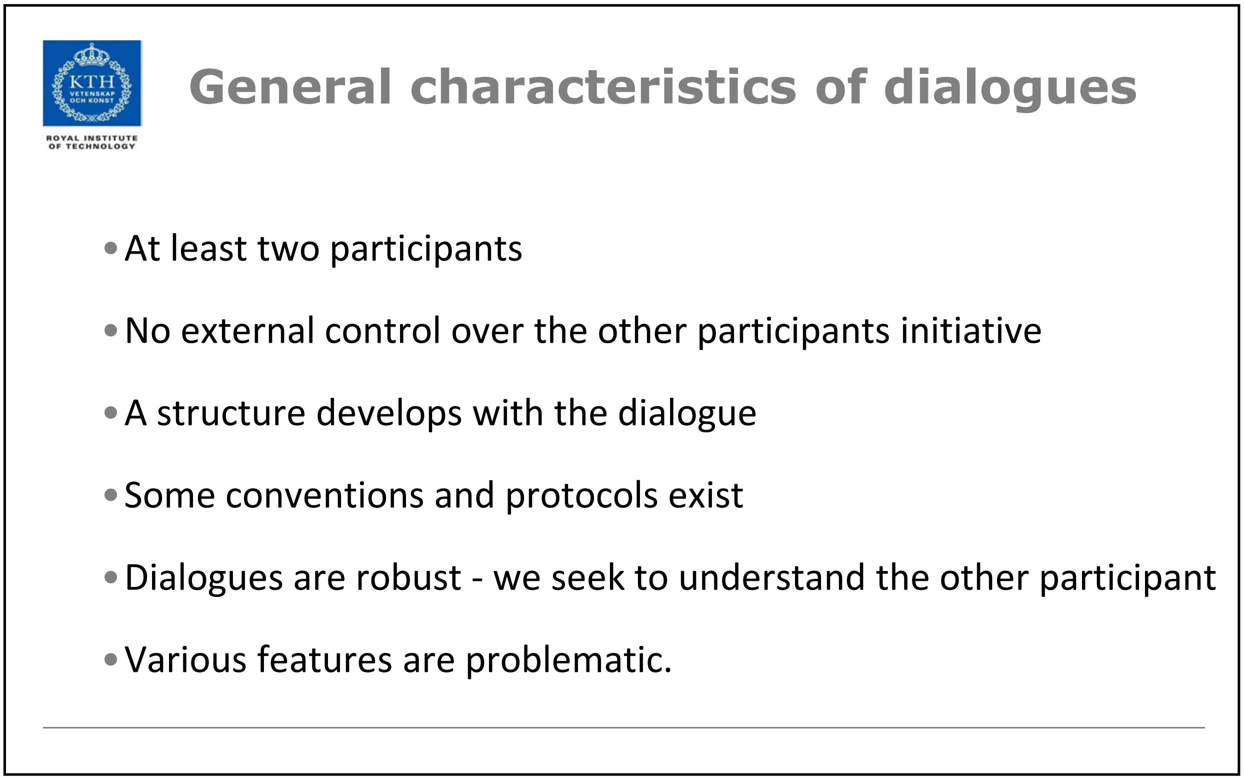
2007- Future faculty position at KTH

2013 – Professor, Head of the Speech Group



What is Dialogue?

- A sequence of isolated utterances uttered by at least two speakers that together form a discourse.
- Dialogue = a connected sequence of information units (with a goal);
  - provides coherence over the utterances,
  - provides a context for interpreting utterances,
  - multiple participants exchange information.



General characteristics of dialogues

- At least two participants
- No external control over the other participants initiative
- A structure develops with the dialogue
- Some conventions and protocols exist
- Dialogues are robust - we seek to understand the other participant
- Various features are problematic.

Different types of dialogue

- Conversations
  - Informal (spoken) interaction between two individuals
  - **Main Goal:** development and maintenance of social relationships
- Task-oriented Dialogues
  - Possibly formal multimodal interaction
  - **Main Goal:** perform a given task
- Natural Dialogues:
  - Occur between humans
- Artificial Dialogues:
  - At least one of the participant is a computer

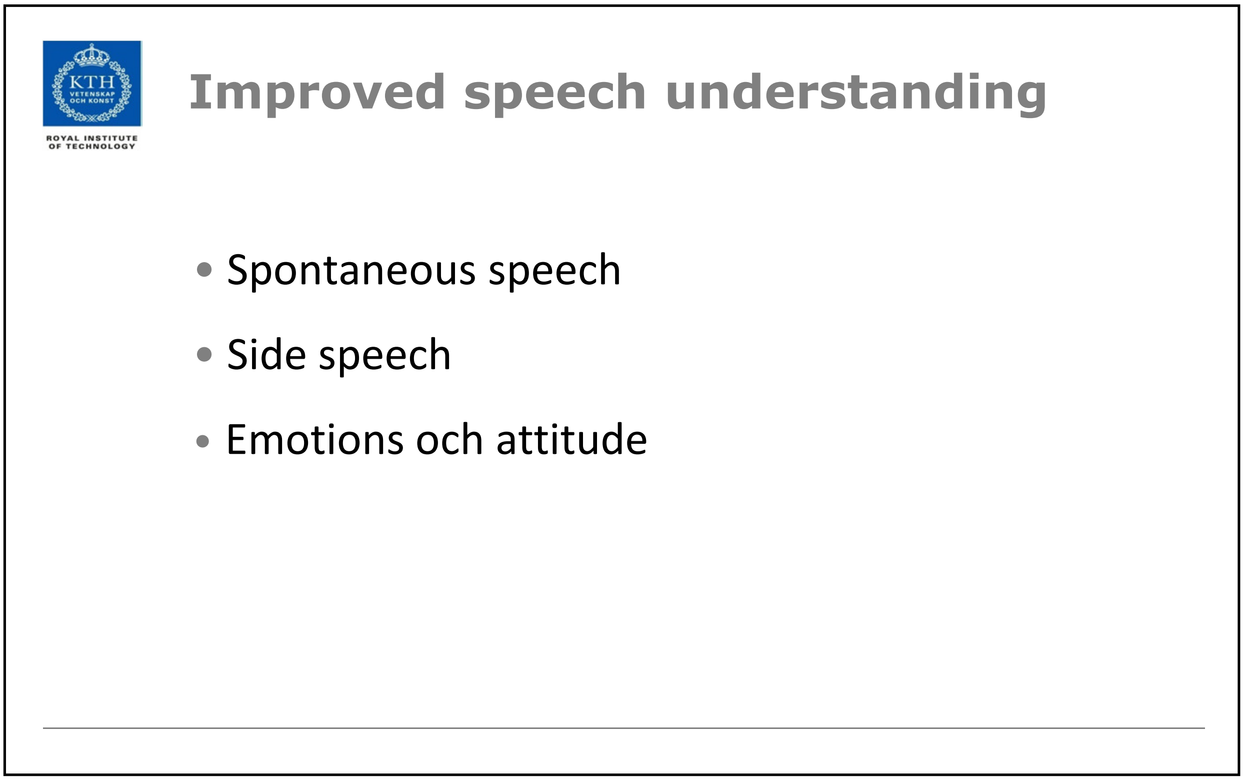
Vision: artificial dialogues 2001

Did we reach it today?

HUR GÅR DET EGENTLIGEN? DU HAR SUTTIT OCH GASTAT HÄR EN TIMME!
I'M SORRY, I DIDN'T CATCH THAT?
HÅLL KÄFT! DEN HÄR RÖST-AKTIVERADE DATORN ÄR UT-VECKLINGSSTÖRD! DET HÄR ÄR DET SVÅRASTE JAG NÅGONSIN GJORT!
THANK YOU FOR CALLING AMTRAK, HAVE A NICE DAY! ›KLICK‹
1353

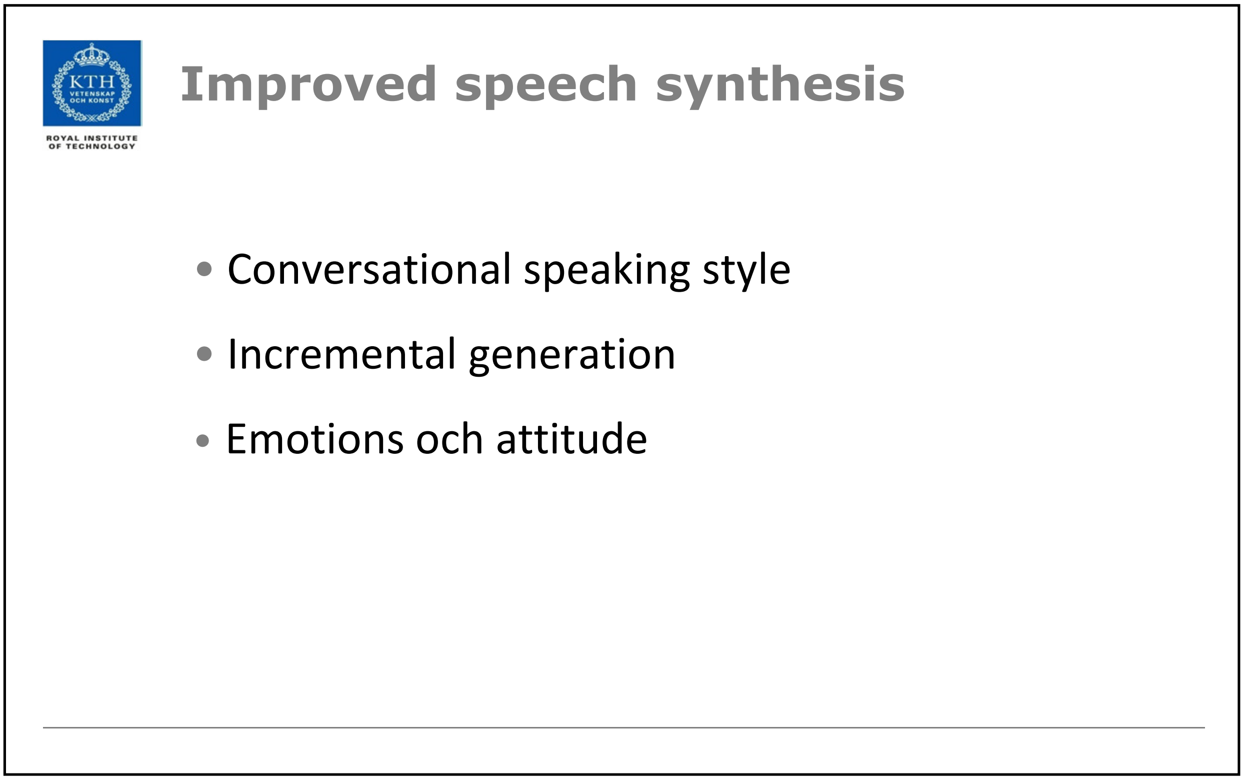
What can be improved?

- Speech understanding
- Speech synthesis
- Dialogue behavior



Improved speech understanding

- Spontaneous speech
- Side speech
- Emotions och attitude



Improved speech synthesis

- Conversational speaking style
- Incremental generation
- Emotions och attitude

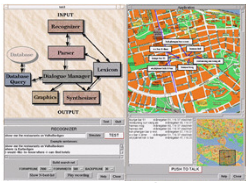
Improved dialogue behaviour

- Common sense
- Improved error message
- Longer memory span

Our multimodal dialogue systems

92-95 Waxholm

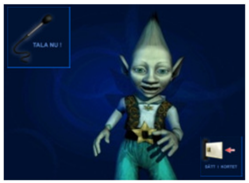
96 Olga



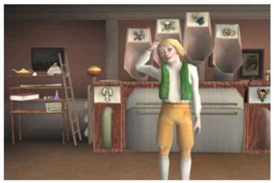
97-98 Gulan

98-99 August

99-02 AdApt



02-03 Pixie


03-05 NICE

04-06 Higgins

07-09 MonAMI

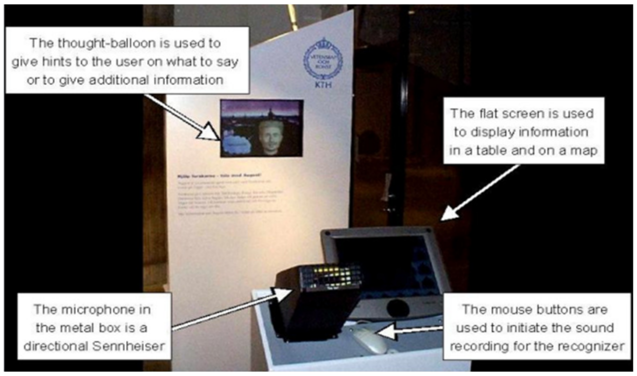
10-12 IURO

12-15 GetHomeSafe

11- FurHat

Waxholm: the first Swedish spoken dialogue system (1994)

August a public system (1998)


The thought-balloon is used to give hints to the user on what to say or to give additional information
The flat screen is used to display information in a table and on a map
The microphone in the metal box is a directional Sennheiser
The mouse buttons are used to initiate the sound recording for the recognizer

- Swedish spoken dialogue system for public use in kulturhuset
- Animated agent named after August Strindberg
- Speech technology components developed at KTH
- Designed to be easily expanded and reconfigured
- Used to collect spontaneous speech data

A walk-up-and-speak user

The HIGGINS system (2004)

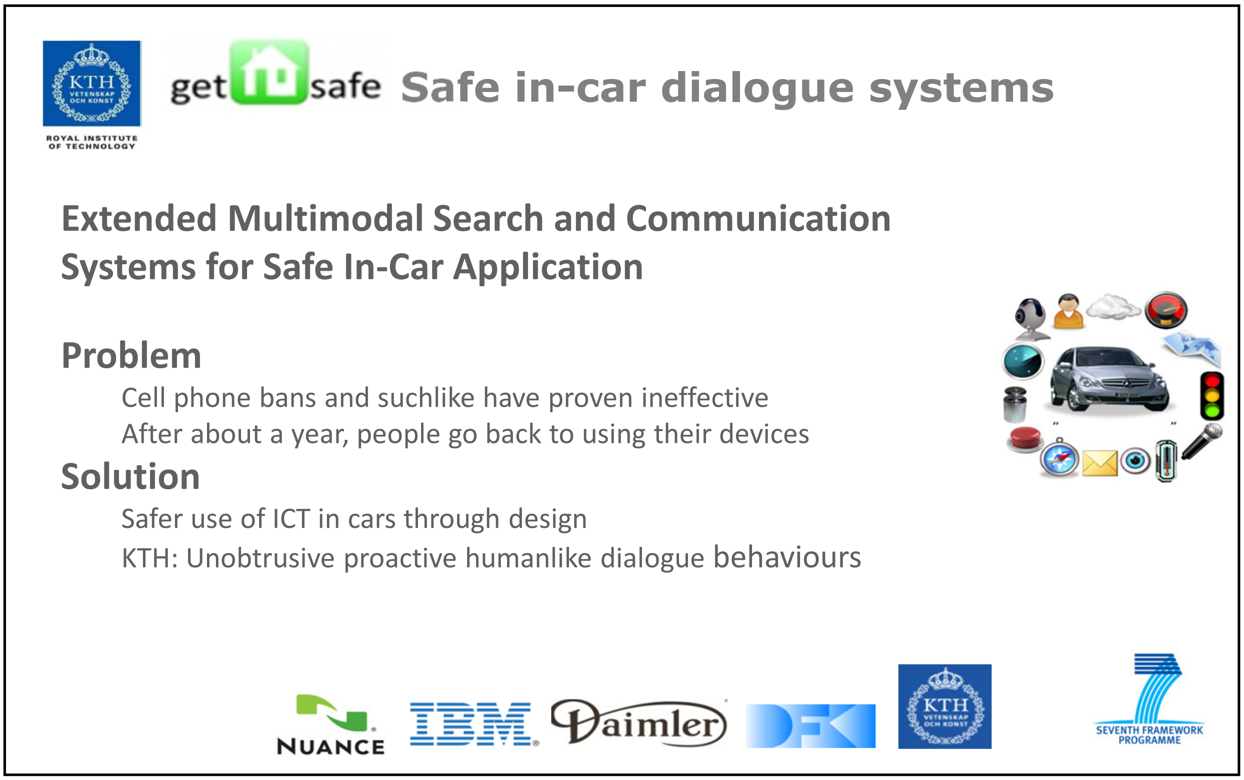
- The primary domain of HIGGINS is city navigation for pedestrians.
- Secondarily, HIGGINS is intended to provide simple information about the immediate surroundings.

Higgins demo movie

iURO> Interactive Urban Robot (2012)

Research issues - How can a robot:

- Build a model of a city using sensory input?
- Obtain information by talking with humans?
- Talk with several humans at the same time?



get safe Safe in-car dialogue systems

Extended Multimodal Search and Communication Systems for Safe In-Car Application

Problem

Cell phone bans and suchlike have proven ineffective

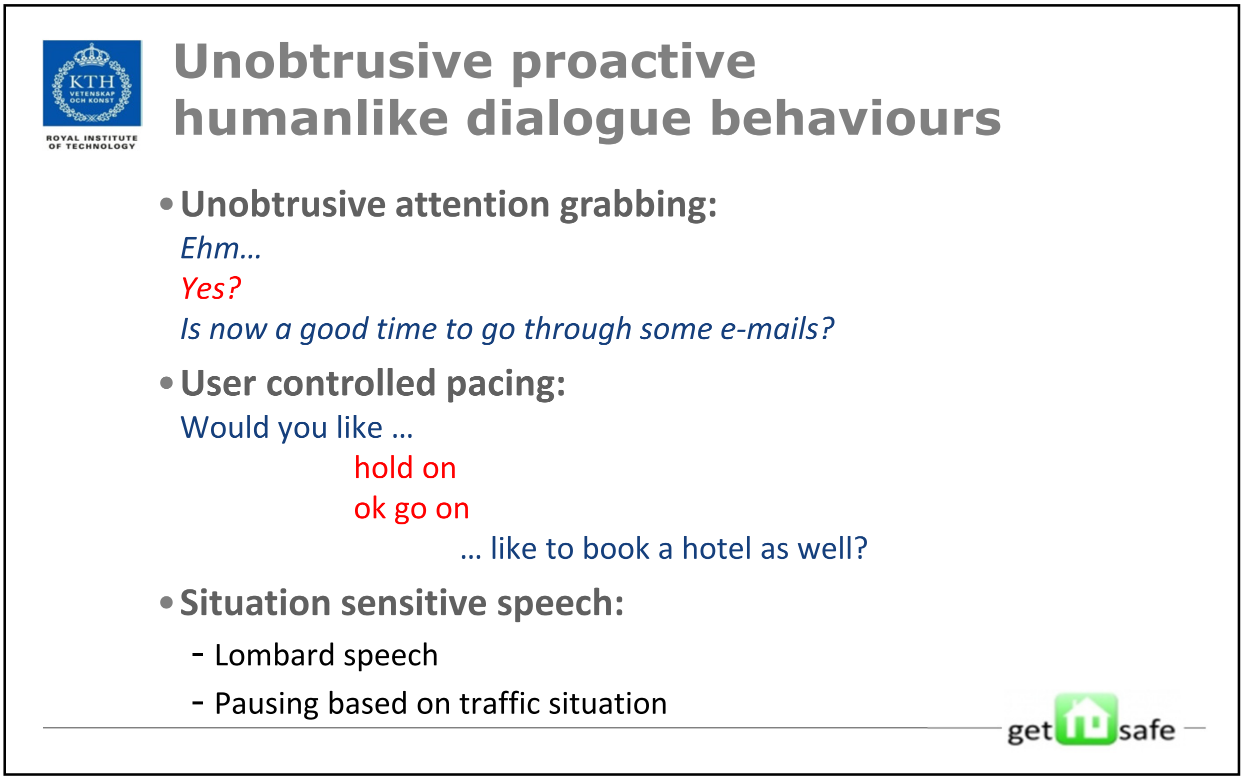
After about a year, people go back to using their devices

Solution

Safer use of ICT in cars through design

KTH: Unobtrusive proactive humanlike dialogue behaviours

KTH VETENSKAP OCH KONST
ROYAL INSTITUTE OF TECHNOLOGY
NUANCE
IBM
Daimler
DFKI
SEVENTH FRAMEWORK PROGRAMME



Unobtrusive proactive humanlike dialogue behaviours

- **Unobtrusive attention grabbing:**

  Ehm…

  Yes?

  Is now a good time to go through some e-mails?

- **User controlled pacing:**

  Would you like …

  hold on

  ok go on

  … like to book a hotel as well?

- **Situation sensitive speech:**

  - Lombard speech
  - Pausing based on traffic situation

KTH VETENSKAP OCH KONST
ROYAL INSTITUTE OF TECHNOLOGY
get safe

Commercial dialogue systems

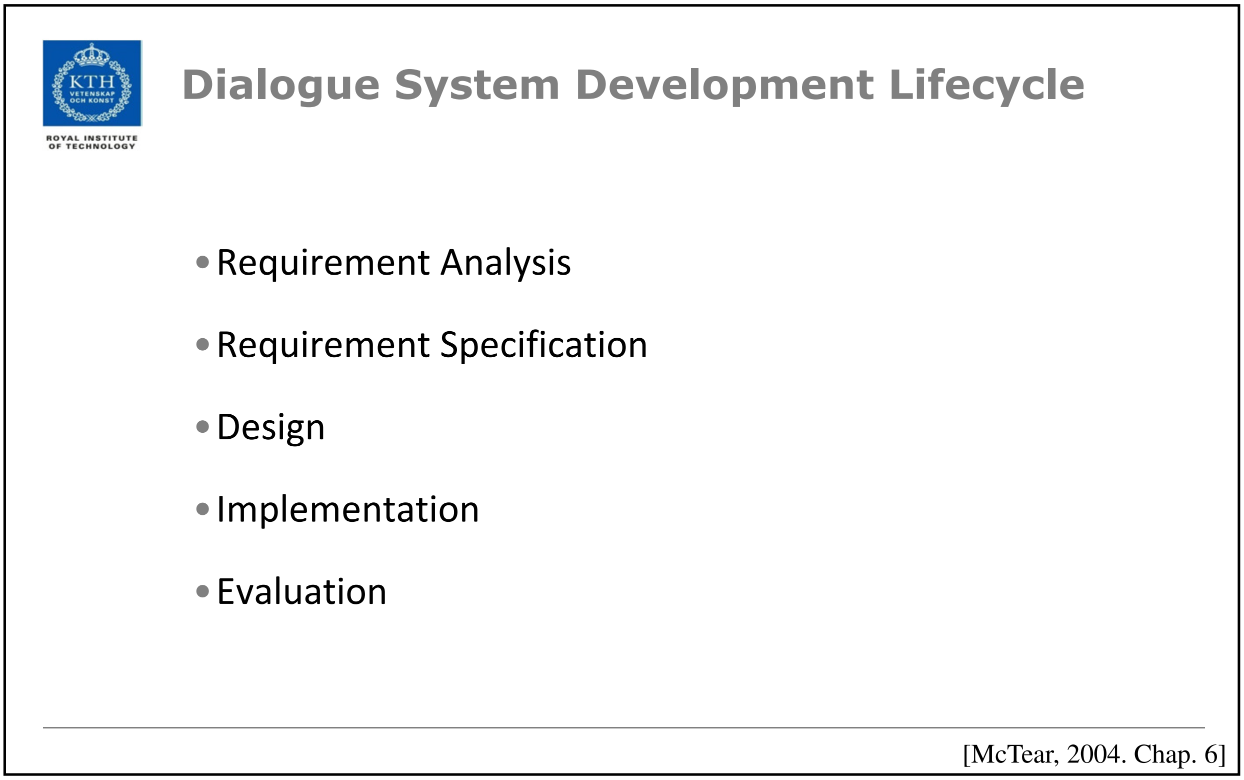
- Information retrieval
  - SJ, SL , WAMI toolkit
- Call routing
  - Telia, ComHem
- Mobile assistant
  - Siri (apple)
- Technical support
  - SpeechCycle, Voiceprovider

Problem solving from SpeechCycle

- **Services**
  - Broadband support Agent
  - Video support Agent
  - IP Telephony Agent
- **Features**
  - Open prompts with free speech ASR
  - References to user's previous attempts to fix problem
  - Answer to free prompts
  - Error identification
  - Memory of earlier turns
  - Encouraging the user to stay

Developing Spoken Dialogue System

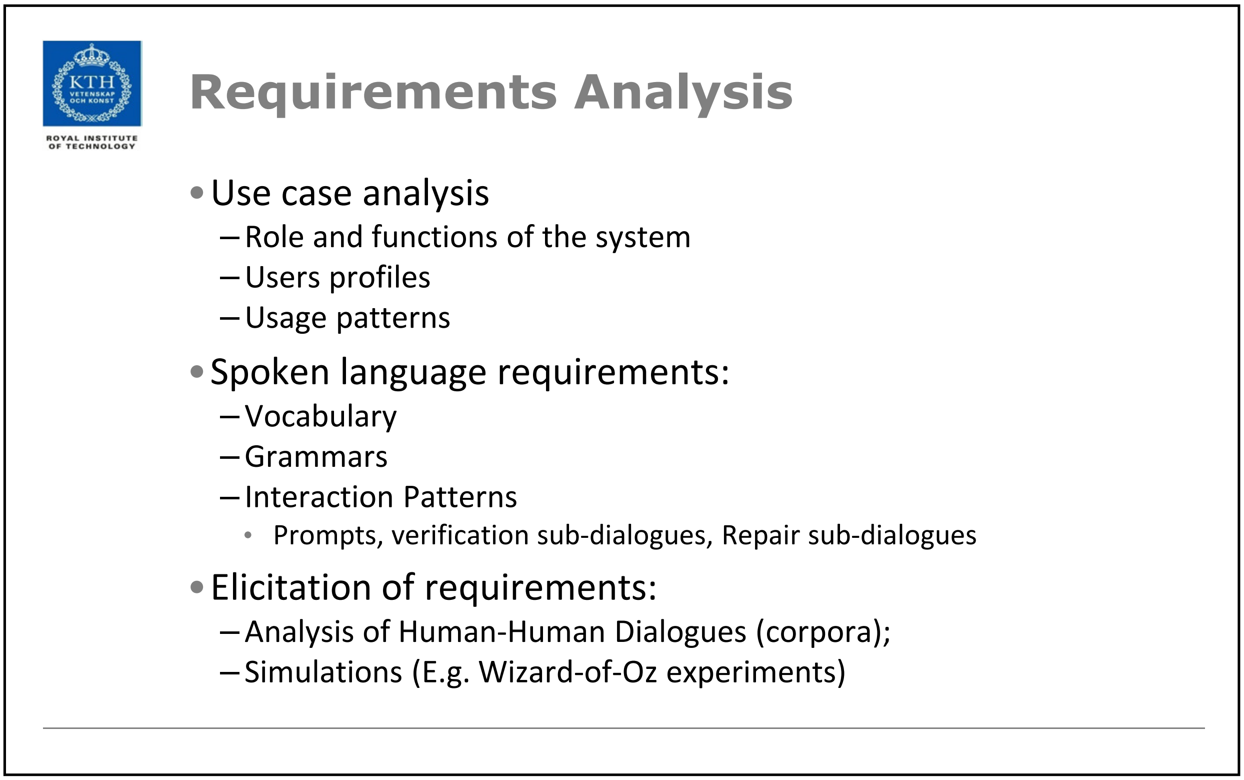
23



Dialogue System Development Lifecycle

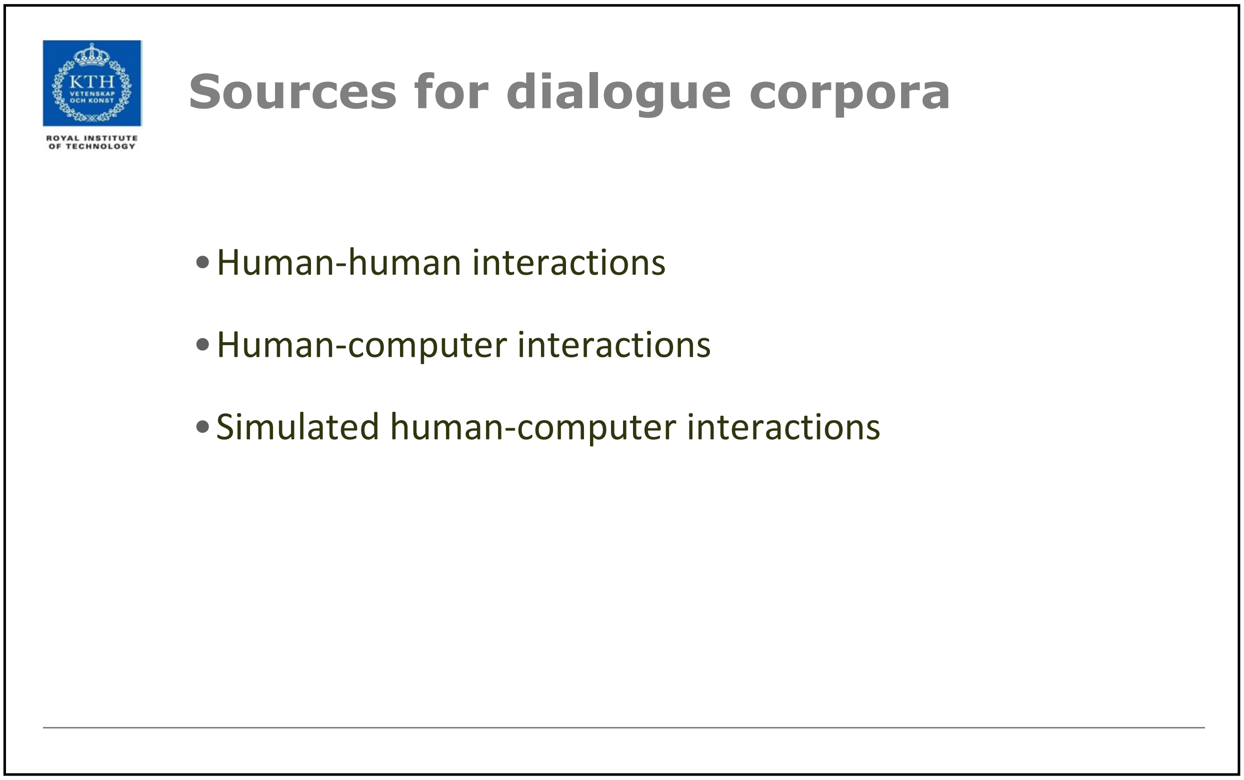
- Requirement Analysis
- Requirement Specification
- Design
- Implementation
- Evaluation

[McTear, 2004. Chap. 6]



Requirements Analysis

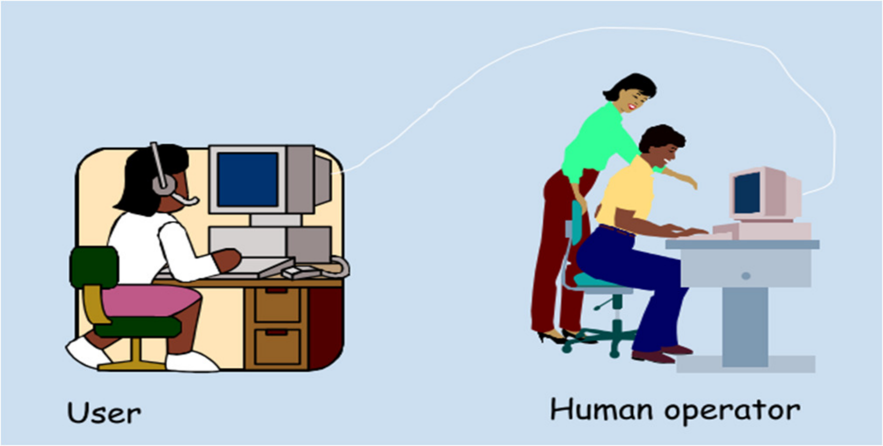
- Use case analysis
  - Role and functions of the system
  - Users profiles
  - Usage patterns
- Spoken language requirements:
  - Vocabulary
  - Grammars
  - Interaction Patterns
    - Prompts, verification sub-dialogues, Repair sub-dialogues
- Elicitation of requirements:
  - Analysis of Human-Human Dialogues (corpora);
  - Simulations (E.g. Wizard-of-Oz experiments)



Sources for dialogue corpora

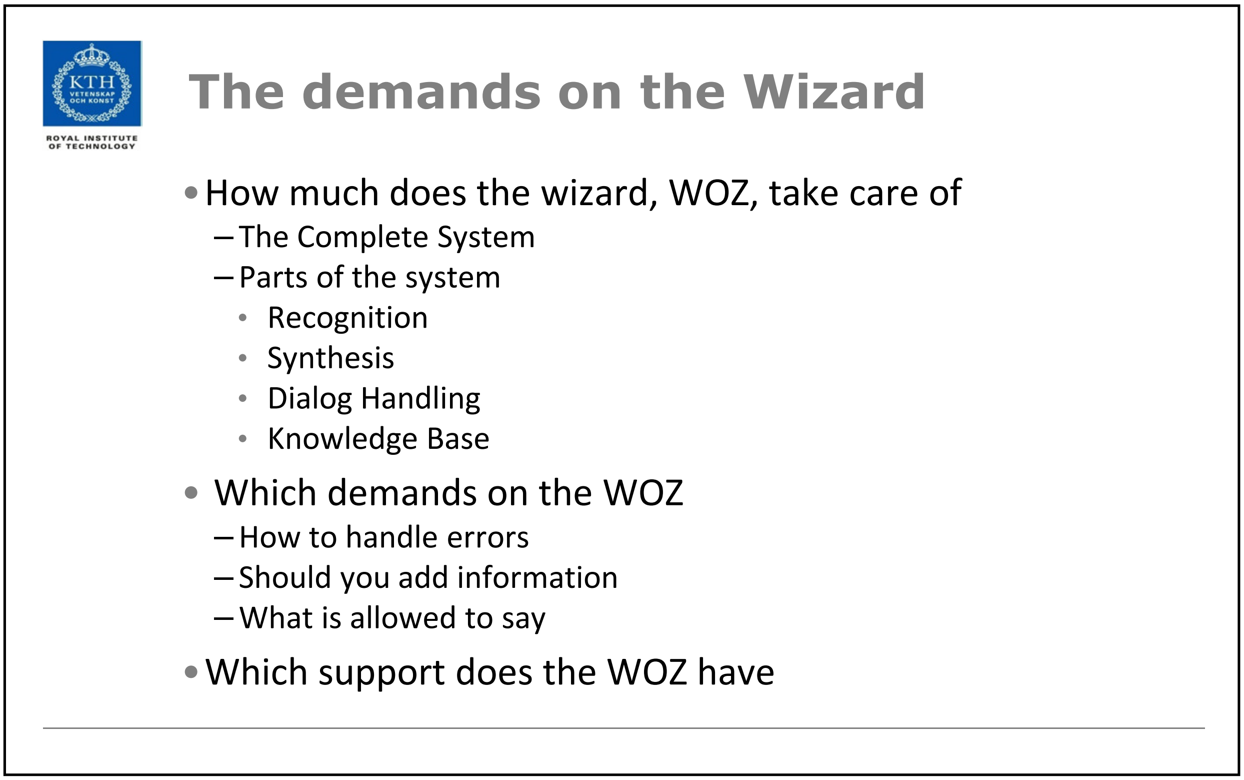
- Human-human interactions
- Human-computer interactions
- Simulated human-computer interactions

Simulation (Wizard-of-Oz)



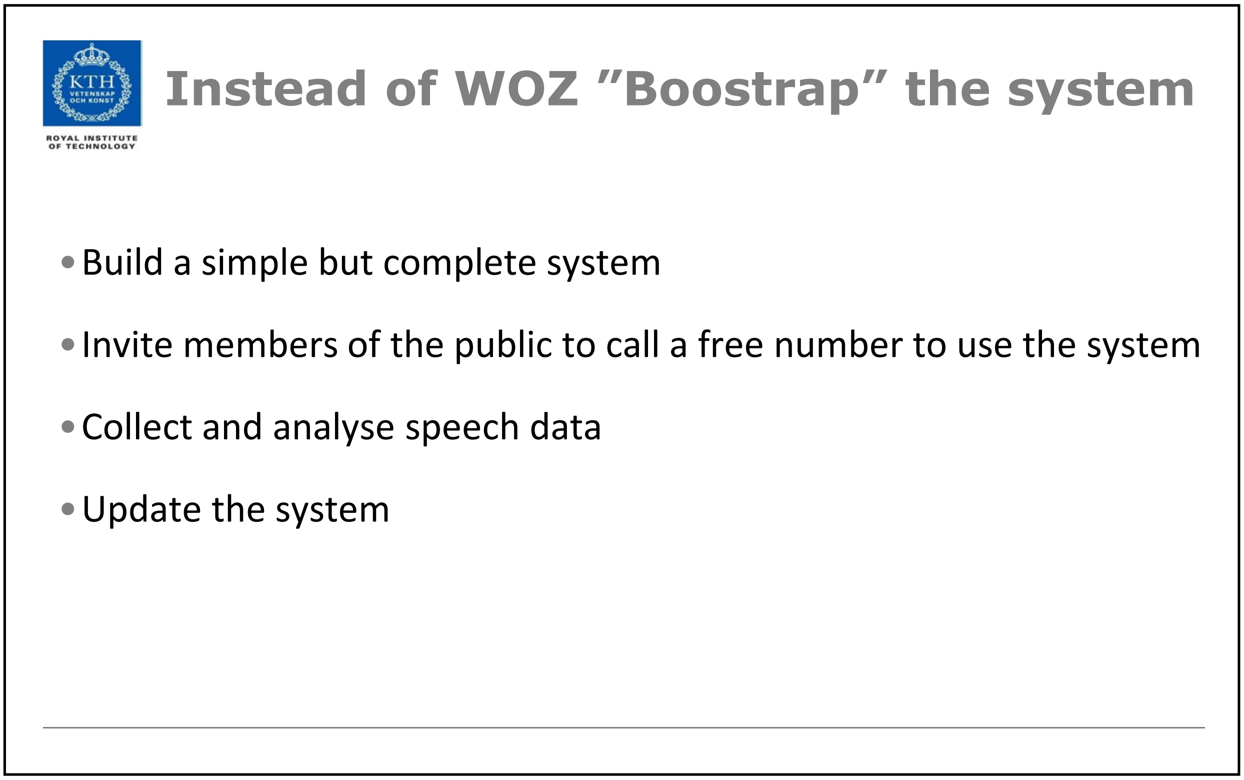
The point of WoZ simulation

- Data for speech recognition
- Typical utterance patterns
- Typical dialogue patterns
- Click-and-talk behaviour
- Hints at important research problems



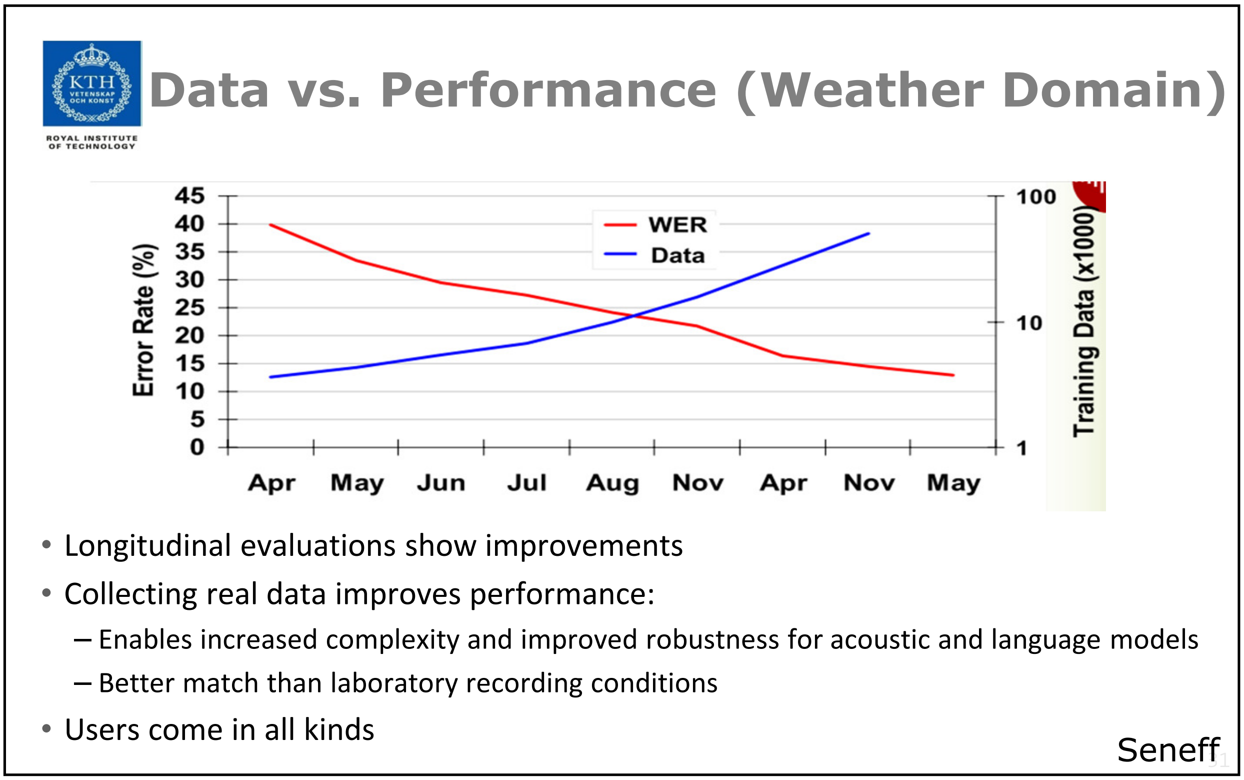
The demands on the Wizard

- How much does the wizard, WOZ, take care of
  - The Complete System
  - Parts of the system
    - Recognition
    - Synthesis
    - Dialog Handling
    - Knowledge Base
- Which demands on the WOZ
  - How to handle errors
  - Should you add information
  - What is allowed to say
- Which support does the WOZ have



Instead of WOZ ”Boostrap” the system

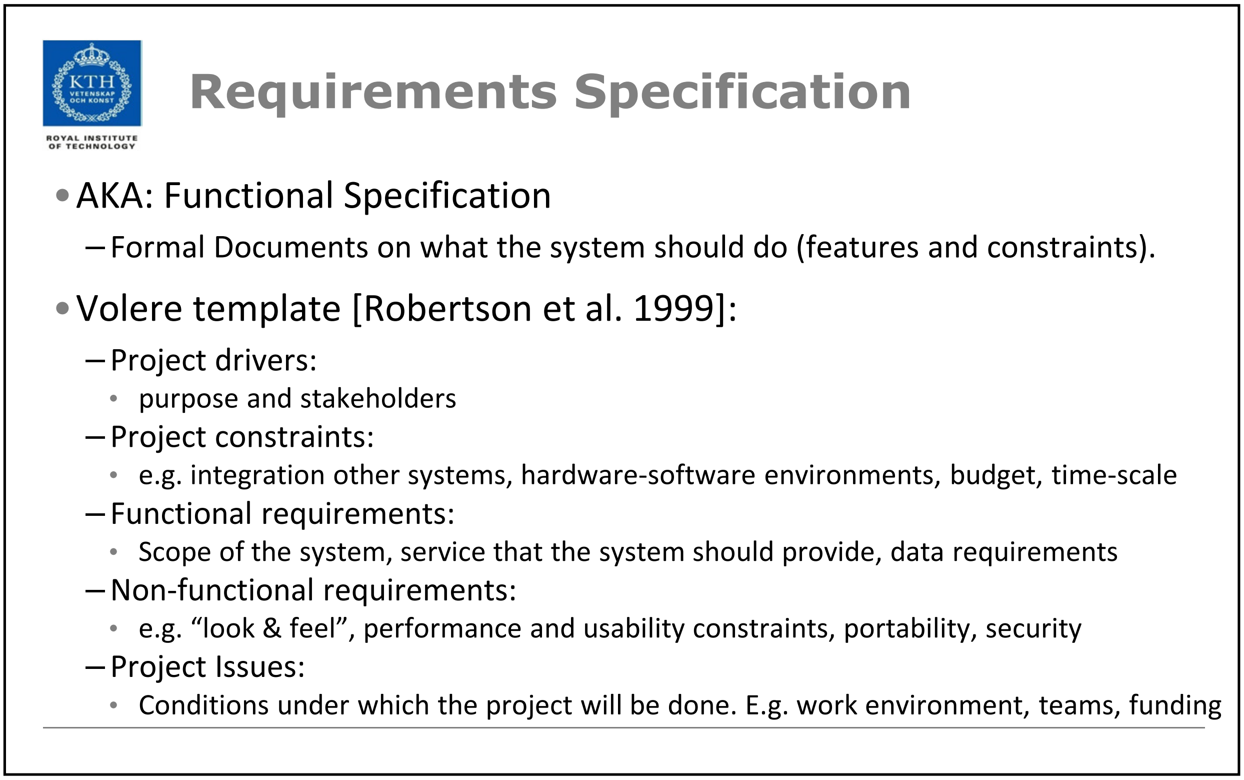
- Build a simple but complete system
- Invite members of the public to call a free number to use the system
- Collect and analyse speech data
- Update the system



Data vs. Performance (Weather Domain)

- Longitudinal evaluations show improvements
- Collecting real data improves performance:
  - Enables increased complexity and improved robustness for acoustic and language models
  - Better match than laboratory recording conditions
- Users come in all kinds

Seneff



Requirements Specification

- AKA: Functional Specification
  - Formal Documents on what the system should do (features and constraints).
- Volere template [Robertson et al. 1999]:
  - Project drivers:
    - purpose and stakeholders
  - Project constraints:
    - e.g. integration other systems, hardware-software environments, budget, time-scale
  - Functional requirements:
    - Scope of the system, service that the system should provide, data requirements
  - Non-functional requirements:
    - e.g. “look & feel”, performance and usability constraints, portability, security
  - Project Issues:
    - Conditions under which the project will be done. E.g. work environment, teams, funding

ROYAL INSTITUTE OF TECHNOLOGY

Design

- Architecture that implements the selected Dialogue Model
- Task Ontology
- Dialogues and Protocols
  - Linked to the Task's entities, Prompts, Input recognition grammars, System Help
- Dialogue Control Strategies:
  - Initiative, Confirmation, Error handling
- Overall Consistency:
  - Personality, Timing, Consistent dialogue structure, Tend to not surprise the user

ROYAL INSTITUTE OF TECHNOLOGY

Implementation

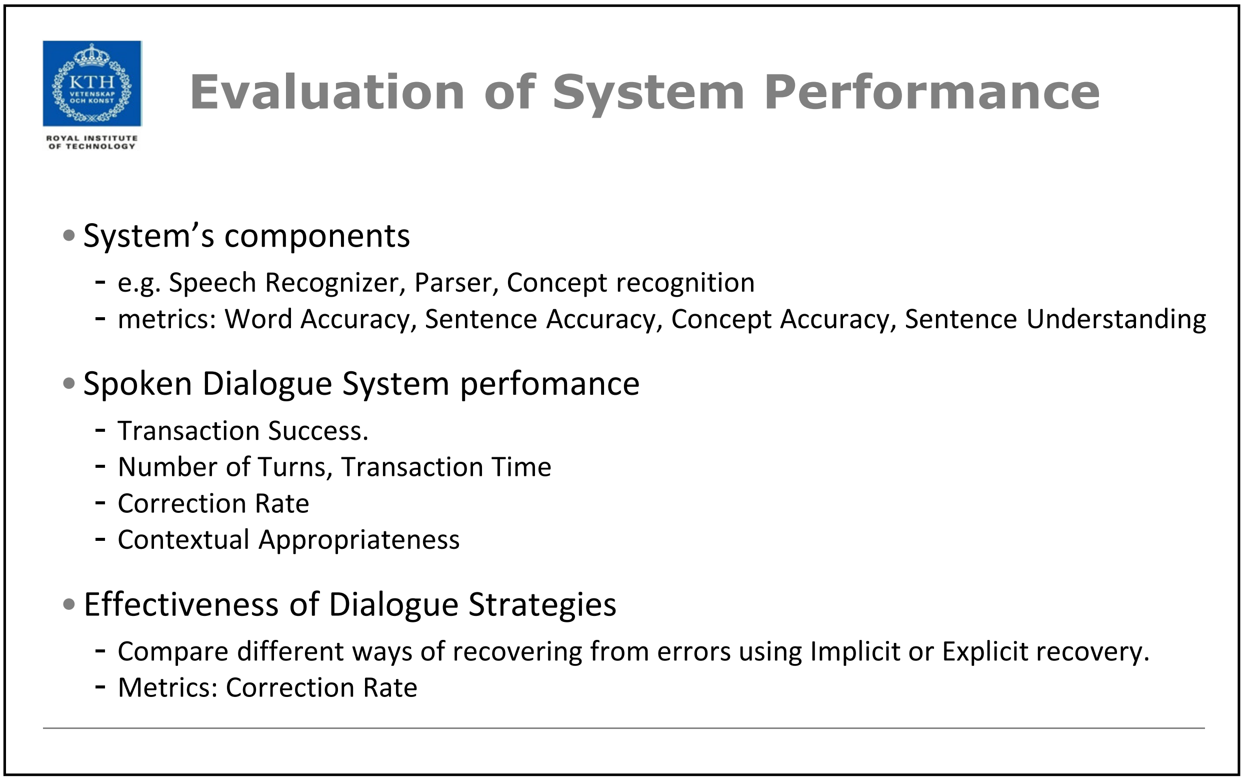
- Chose the appropriate programming tools:
  - Rapid Dialogue Prototypers:
- Web-based platforms:
  - VoiceXML (W3C)
  - EMMA (W3C, multimodal)
  - SALT (Microsoft .NET Speech SDK)
  - X+V (Opera + IBM)

Testing

- Integration testing:
  - Test interfaces between modules of the system.
  - Check if the dialogue components correctly communicate with the task components.
- Validation testing:
  - Check whether the system meets the functional and non-functional requirements.
    - E.g. all functionalities are (correctly) implemented
- System testing:
  - Check performance, reliability and robustness:
  - Performance measures: transaction time, success, and overall reliability

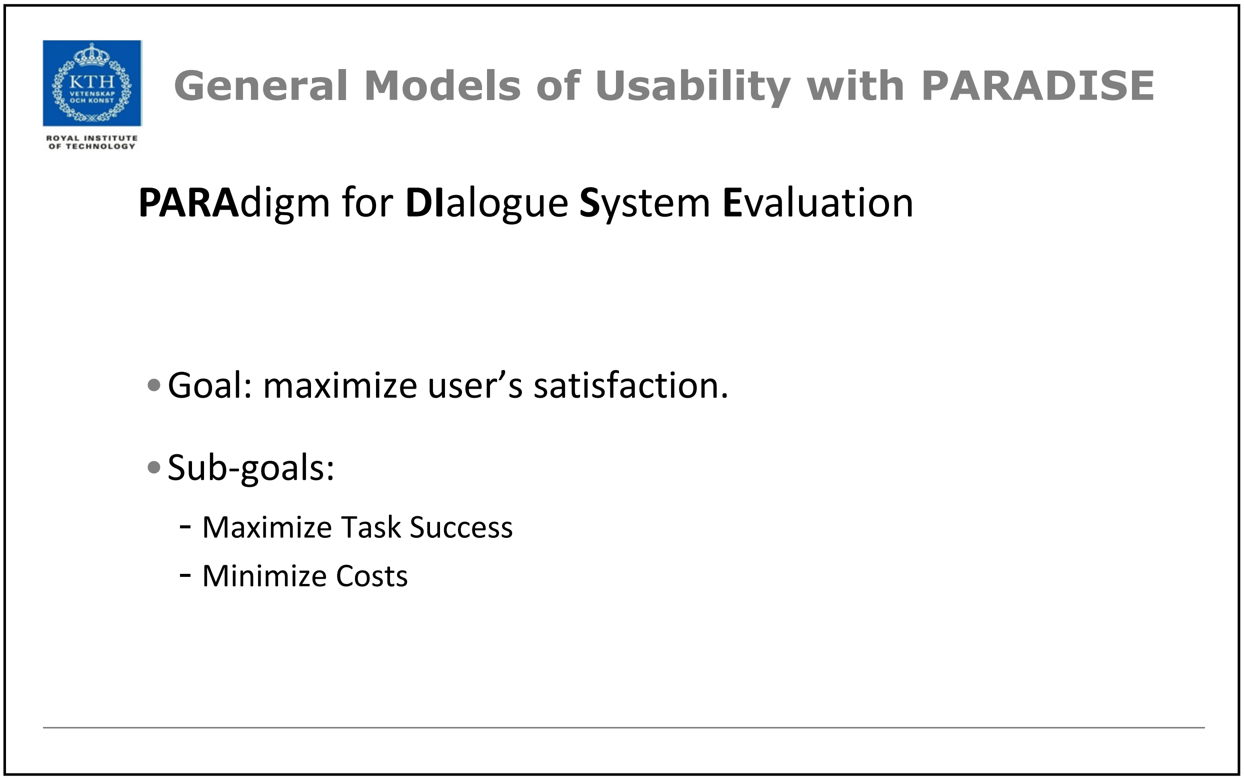
Evaluation

- Different than testing:
  - **Testing:** identify differences between expected and actual live performance
  - **Evaluation:** examine how the system performs when used by actual users
- Observational studies:
  - The user interact with the system according to a set of “scenarios”.
  - Scenarios provide either precise instructions on each stage or can be more open-ended
- Analysis of User Acceptance:
  - Qualitative measure of how the user likes the system
  - Based on a questionnaire with pre-defined statements to be scored in a 1 to 5 scale



Evaluation of System Performance

- System’s components
  - e.g. Speech Recognizer, Parser, Concept recognition
  - metrics: Word Accuracy, Sentence Accuracy, Concept Accuracy, Sentence Understanding
- Spoken Dialogue System perfomance
  - Transaction Success.
  - Number of Turns, Transaction Time
  - Correction Rate
  - Contextual Appropriateness
- Effectiveness of Dialogue Strategies
  - Compare different ways of recovering from errors using Implicit or Explicit recovery.
  - Metrics: Correction Rate



General Models of Usability with PARADISE

PARAdigm for **DI**alogue **S**ystem **E**valuation

- Goal: maximize user’s satisfaction.
- Sub-goals:
  - Maximize Task Success
  - Minimize Costs

Paradise User Satisfaction

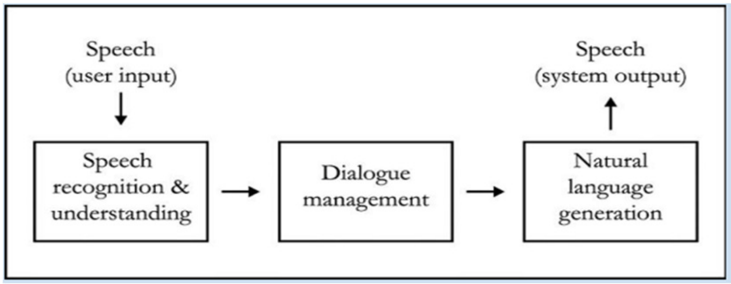
- I found the system easy to understand in this conversation. (TTS Performance)
- In this conversation, I knew what I could say or do at each point of the dialogue. (User Expertise)
- The system worked the way I expected it to in this conversation. (Expected Behaviour)
- Based on my experience in this conversation using this system to get travel information, I would like to use this system regularly. (Future Use)

Evaluation metrics

- Dialog Efficiency Metrics:
  - Total elapsed time, Time on task, System turns, User turns, Turns on task, time per turn for each system module
- Dialog Quality Metrics:
  - Word Accuracy, Sentence Accuracy, Mean Response latency
- Task Success Metrics:
  - Perceived task completion, Exact Scenario Completion, Scenario Completion
- User Satisfaction:
  - Sum of TTS performance, Task ease, User expertise, Expected behaviour

Dialogue Management Design Issues

Spoken dialogue processing



Speech (user input)

Speech recognition & understanding

Dialogue management

Natural language generation

Speech (system output)

ROYAL INSTITUTE OF TECHNOLOGY

Automating Dialogue

- To automate a dialogue, we must model:

User's Intentions

- Recognized from input media forms (multi-modality)
- Disambiguated by the Context

System Actions

- Based on the task model and on output medium forms (multi-media)

Dialogue Flow

- Content Selection
- Information Packaging
- Turn-taking

ROYAL INSTITUTE OF TECHNOLOGY

Design Issues in Dialogue Management

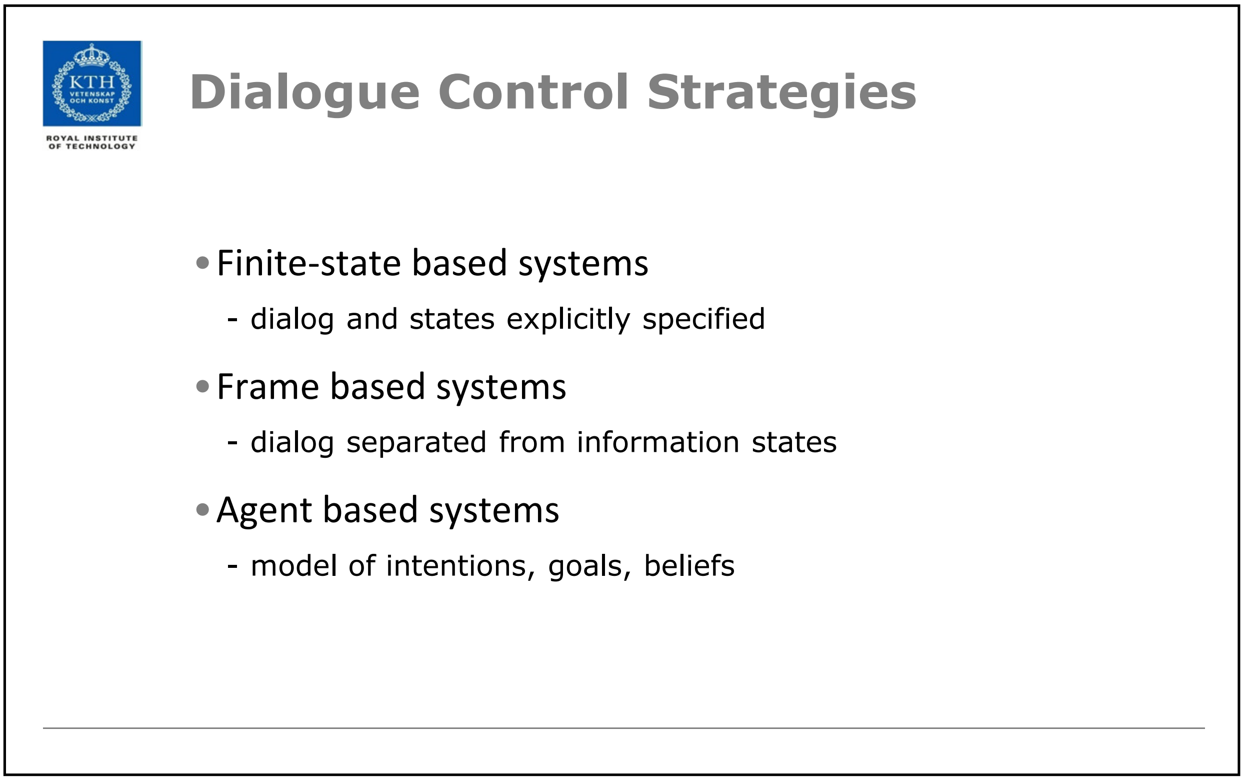
- Choice of dialogue initiative:
  - System-directed, User-directed, Mixed-initiative
- Choice of dialogue control strategies:
  - Finite-state, frame-based, agent-based, …
- Design of system prompts.
- Choice of Grounding strategies:
  - Integration with external Knowledge Sources and Systems
- Persona design
  - Choice of voice talent or speech synthesizer
  - Influences prompt wording, dialogue strategies
  - Choice of interface or human metaphor

Dialogue Initiative

- System Directed (easiest, “finite states automaton”)
  - S: Welcome to the information service
  - S: Please state point of departure
  - U: Stockholm
  - S: Please state your destination
  - U: Waxholm
  - S: When do you want to depart
  - …
- User Directed (harder, “frame filling”)
  - S: What can I do for you?
  - U: I would buy a ticket from Stockholm to Waxholm, departure at noon, one way

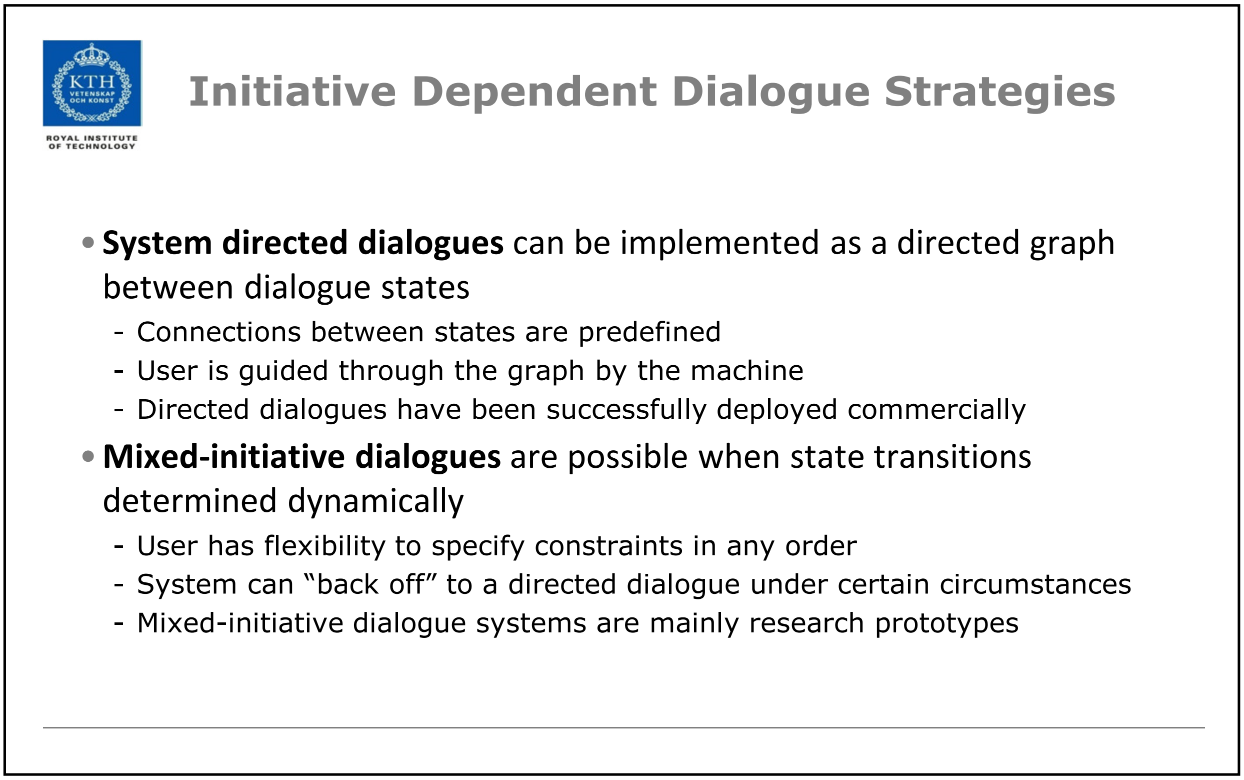
Mixed Initiative

- System is supposed to control the dialogue, but user can take control at times providing more information:
  - S: Welcome to the information service
  - U: Can you tell me when is the next boat to Waxholm
  - S: Are you leaving from Stockholm?
  - U: I want to go from Grinda today.
  - S: There is a boat every hour at 17 minutes past the hour
- The user can also change the task or perform two or more tasks at the same time:
  - S: Welcome to the SmartHome. What would you like to do?
  - U: Switch on the radio.
  - S: Which channel?
  - U: Well, I changed idea, now turn off the lights and select CNN on TV.



Dialogue Control Strategies

- Finite-state based systems
  - dialog and states explicitly specified
- Frame based systems
  - dialog separated from information states
- Agent based systems
  - model of intentions, goals, beliefs



Initiative Dependent Dialogue Strategies

- **System directed dialogues** can be implemented as a directed graph between dialogue states
  - Connections between states are predefined
  - User is guided through the graph by the machine
  - Directed dialogues have been successfully deployed commercially
- **Mixed-initiative dialogues** are possible when state transitions determined dynamically
  - User has flexibility to specify constraints in any order
  - System can “back off” to a directed dialogue under certain circumstances
  - Mixed-initiative dialogue systems are mainly research prototypes

The tasks of the dialogue manager

- update the dialog context
- provide a context for sentence interpretation
- coordinate other modules
- decide what information to convey to the user (when)

Knowledge Sources for dialogue managers

- Context:
  - Dialogue history
  - Current task, slot, state
- Ontology:
  - World Knowledge model
  - Domain model
- Language:
  - Models of conversational competence:
    - Turn-taking strategies
    - Discourse obligations
  - Linguistic models:
    - Grammars, Dictionaries, Corpora
- User Model:
  - Age, gender, spoken language, location, preferences, etc.
  - Dynamic information: the current Mental State

Decisions where data-driven methods would help

- **task-independent behaviors** (e.g., error correction and confirmation behavior)
- **task-specific behaviors** (e.g., logic associated with certain customer-care practices)
- **task-interface behaviors** (e.g., prompt selection).

Machine learning algorithms that have been used for DM

- Markov Decision Processes (MDP)
- Partially Observable MDPs (POMDPs)
- Reinforcement learning
- HMMs
- Stochastic Finite-State Transducers
- Bayesian networks

Problems with machine learning

- Hard to find enough data
- Solved by simulated users
- Not an optimal solution..

Utterance Generation methods

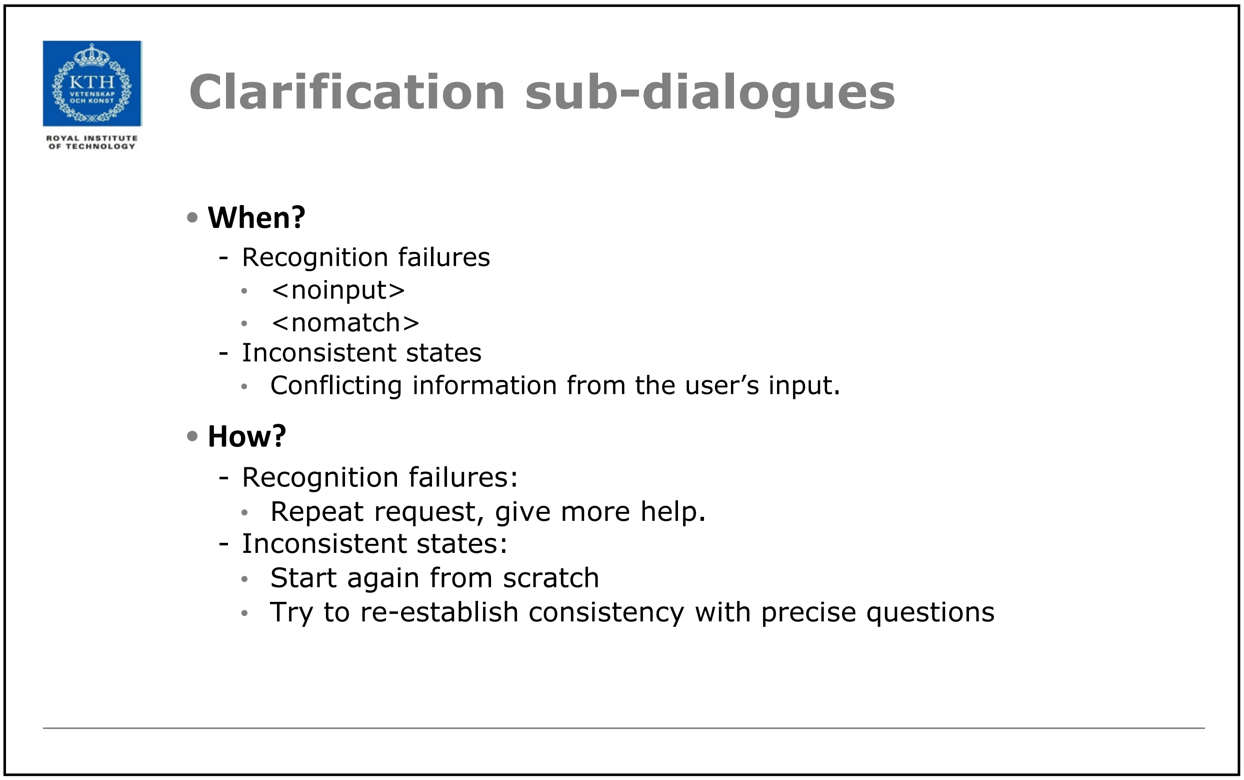
- Predefined utterances
- Frames with slots
- Generation based on grammar and underlying semantics

System Utterances

- The output should reflect the system’s vocabulary and linguistic capability
- Prompts should not be too long (short time memory 7+-2)
- Good error messages

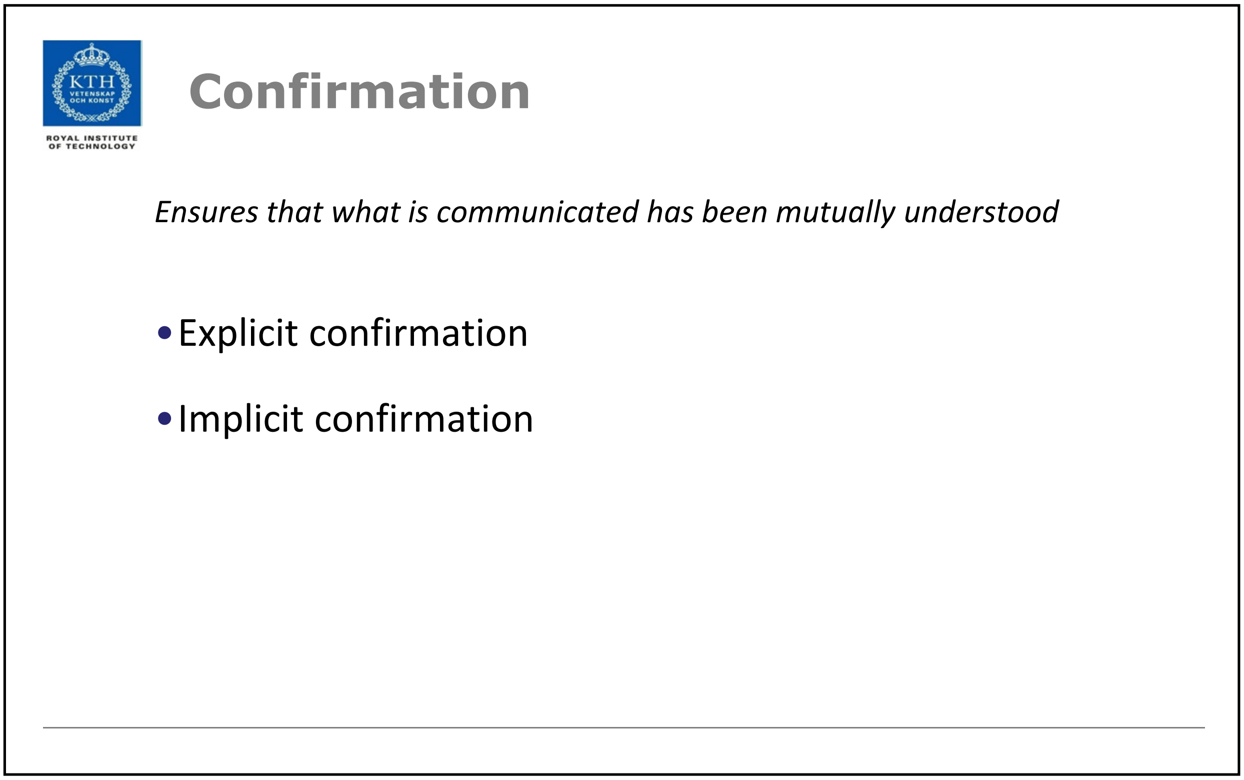
Grounding

- How to give feedback to the user?
  - Find the right trade-off in being too explicit or too implicit.
- Why?
  - The system cannot guarantee that the user’s input has been accurately understood:
  - The user might request information or ask to perform actions that are not available in the system



Clarification sub-dialogues

- **When?**
  - Recognition failures
    - <noinput>
    - <nomatch>
  - Inconsistent states
    - Conflicting information from the user’s input.
- **How?**
  - Recognition failures:
    - Repeat request, give more help.
  - Inconsistent states:
    - Start again from scratch
    - Try to re-establish consistency with precise questions



Confirmation

Ensures that what is communicated has been mutually understood

- Explicit confirmation
- Implicit confirmation

Explicit confirmation

- YN-question repeating what has been understood by the system
  - “So you want to go to waxholm?”
- It might become boring if made all the time!
- Increases transaction time
- Makes the system appear unsure
- Solution: delay the confirmation
  - “So you want a ticket from Grinda to Waholm for tomorrow at 15h34?”

Implicit confirmation

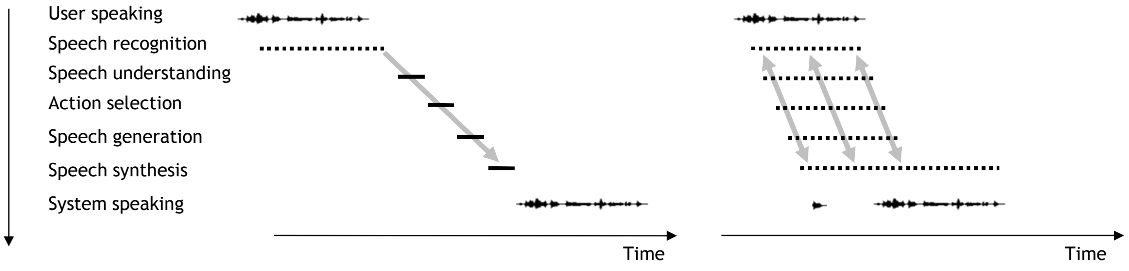
- Embed the confirmation in ordinary requests
  - “When do you want to travel from Grinda to Waxholm?”
- Problems
  - The user may not pay attention to the confirmation and only focus on the request
- Start a repair dialogue is more difficult
  - “I did not want to go to Waxholm at seven thirty”?!?

FurHat – the social robot

What would make a speech interface intuitive?

- The user should not have to learn:
  - What to say
  - How to speak
  - When to speak
- The system should support the user by:
  - Clearly showing when it will speak
  - Saying things that reflects what it understands

Incremental dialogue processing

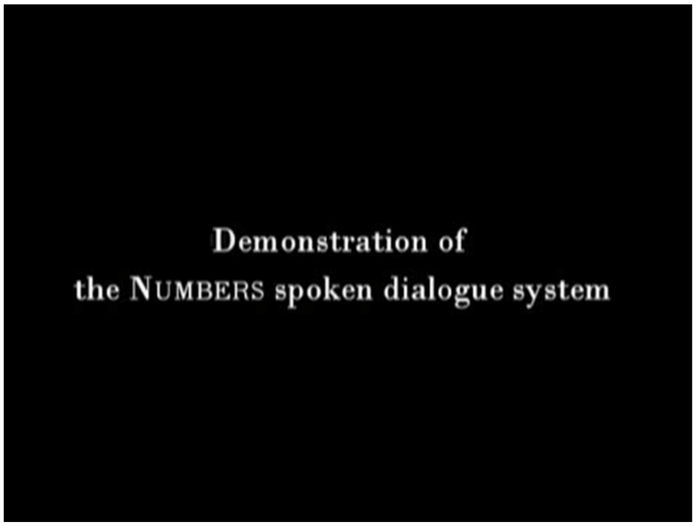

User speaking
Speech recognition
Speech understanding
Action selection
Speech generation
Speech synthesis
System speaking

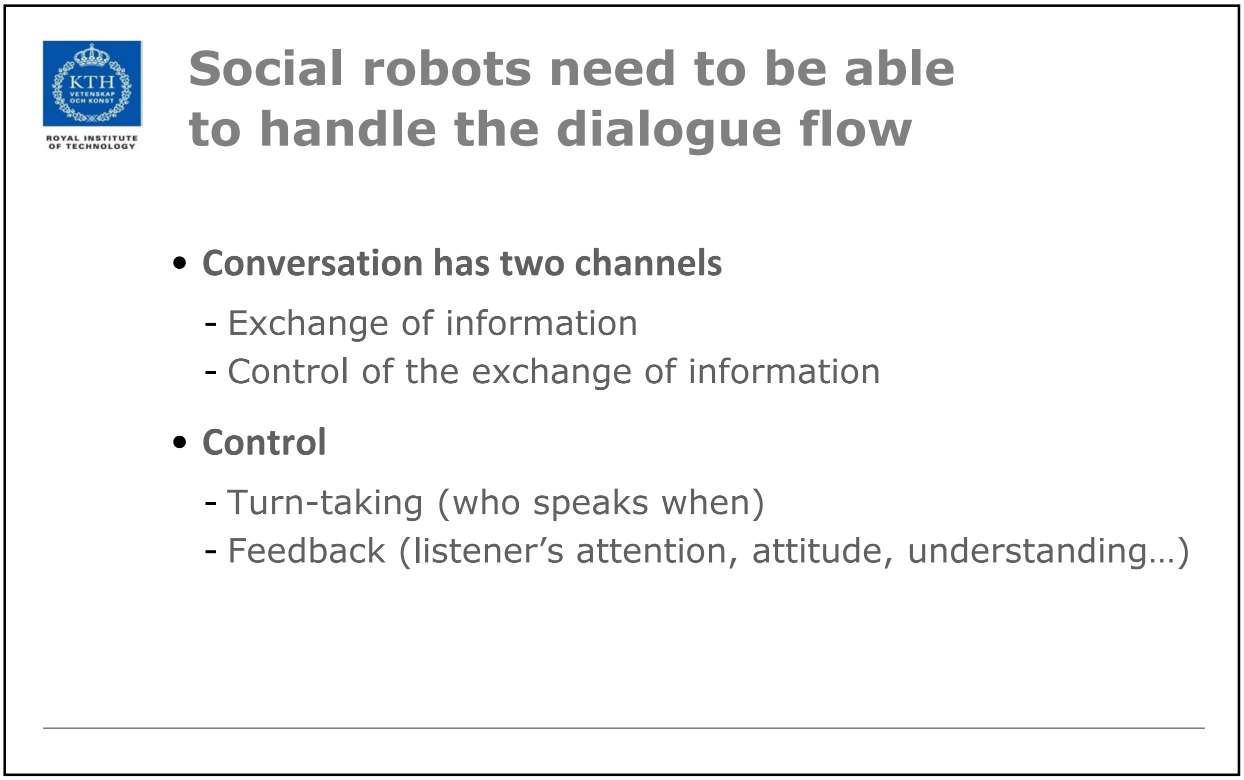
Non-incremental processing

Incremental processing

Skantze & Schlangen (2009) "*Incremental dialogue processing in a micro domain*", Proceedings of European Chapter of the Association for Computational Linguistics

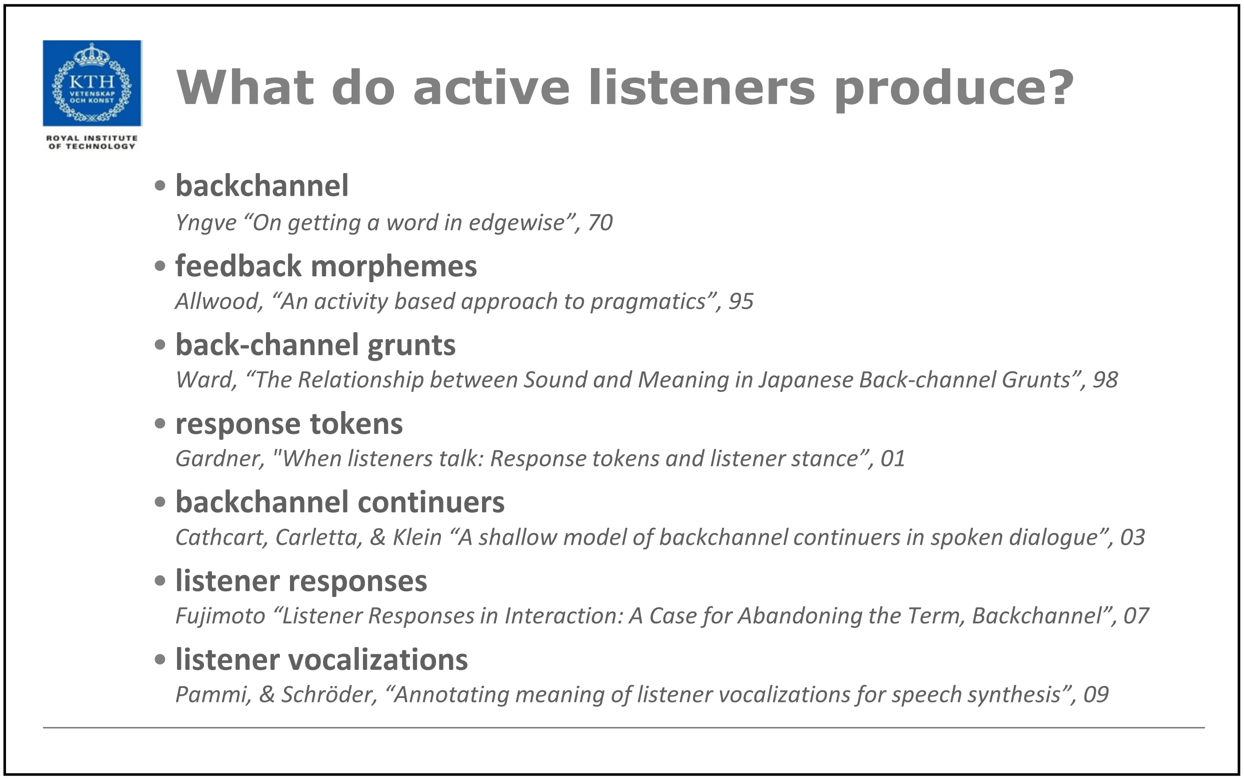
The Numbers dialogue system


Demonstration of
the NUMBERS spoken dialogue system



Social robots need to be able to handle the dialogue flow

- **Conversation has two channels**
  - Exchange of information
  - Control of the exchange of information
- **Control**
  - Turn-taking (who speaks when)
  - Feedback (listener's attention, attitude, understanding…)



What do active listeners produce?

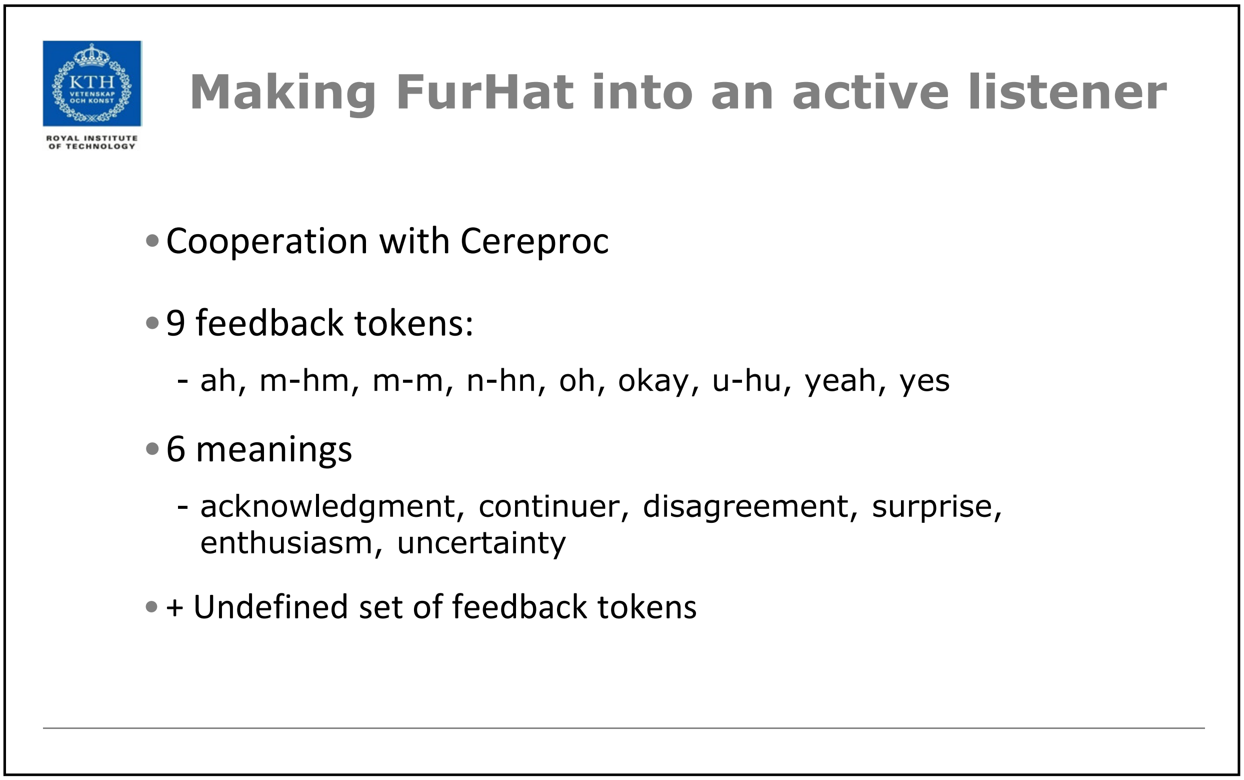
- **backchannel**
  Yngve "On getting a word in edgewise", 70
- **feedback morphemes**
  Allwood, "An activity based approach to pragmatics", 95
- **back-channel grunts**
  Ward, "The Relationship between Sound and Meaning in Japanese Back-channel Grunts", 98
- **response tokens**
  Gardner, "When listeners talk: Response tokens and listener stance", 01
- **backchannel continuers**
  Cathcart, Carletta, & Klein "A shallow model of backchannel continuers in spoken dialogue", 03
- **listener responses**
  Fujimoto "Listener Responses in Interaction: A Case for Abandoning the Term, Backchannel", 07
- **listener vocalizations**
  Pammi, & Schröder, "Annotating meaning of listener vocalizations for speech synthesis", 09

Some functions of listening feedback

- **continuer**
  keeps the floor open for the current speaker to continue speaking
- **acknowledgment**
  shows one has heard the current speaker
- **assessment**
  evaluates the talk of the current speaker
- **display understanding/non-understanding**
  shows whether one has comprehended the speaker
- **agreement/disagreement**
  shows support/non-support of what the speaker has said
- **interest or attentive signal**
  displays interest and engagement in what the speaker has said

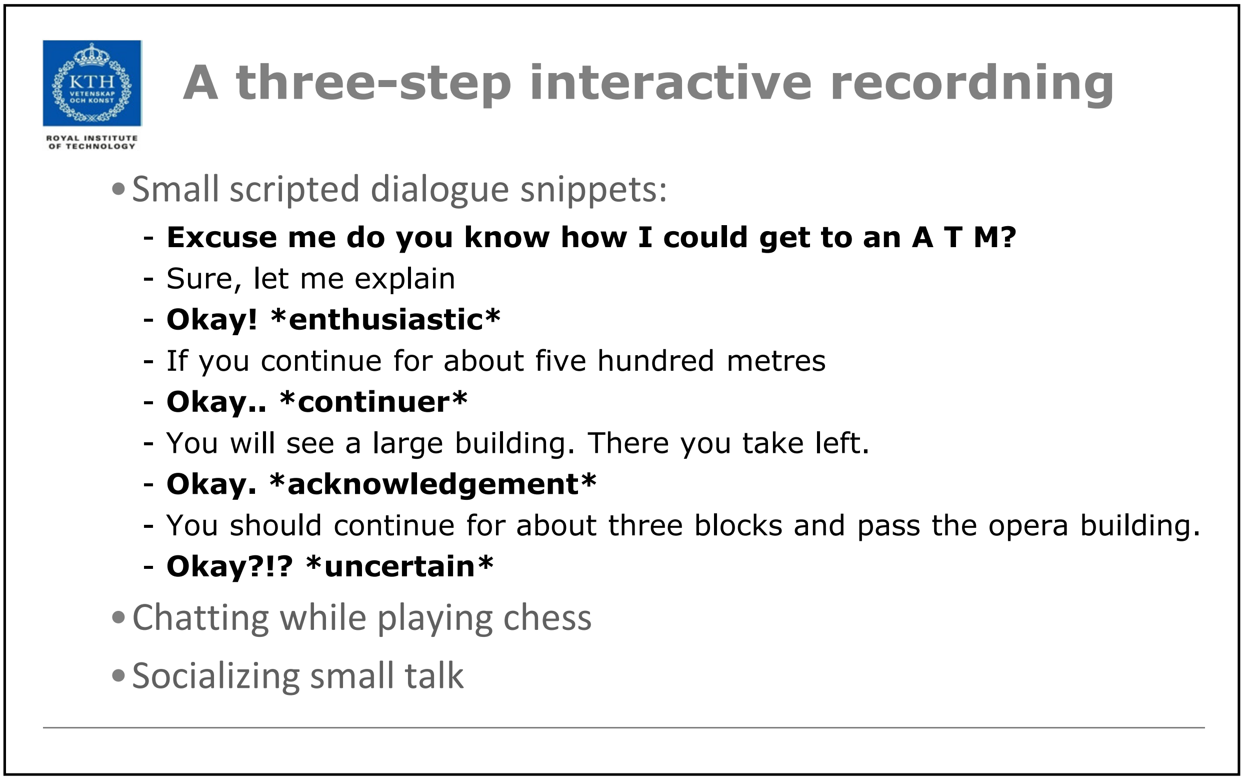
Prosody gives feedback its meaning

- Prosodic features
  - Loudness
  - Height and slope of pitch
  - Duration
  - Syllabification
  - Duration
  - Abruptness of the ending
- Examples of meanings
  - A falling pitch completion or low involvement
  - A bisyllabic “mhm” *often functions as continuer*
  - *A fall-rise pitch on a monosyllabic mm functions as continuer*
  - *A rise is associated with heightened involvement and interest*



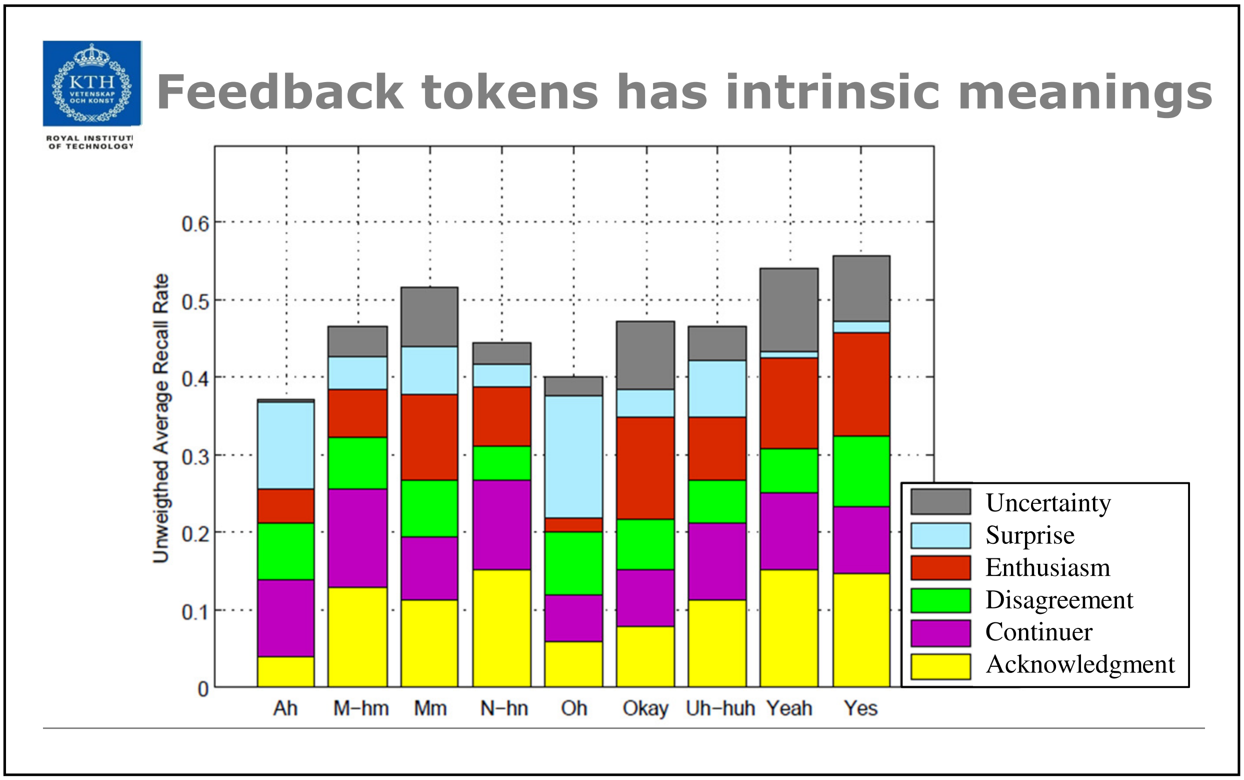
Making FurHat into an active listener

- Cooperation with Cereproc
- 9 feedback tokens:
  - ah, m-hm, m-m, n-hn, oh, okay, u-hu, yeah, yes
- 6 meanings
  - acknowledgment, continuer, disagreement, surprise, enthusiasm, uncertainty
- + Undefined set of feedback tokens

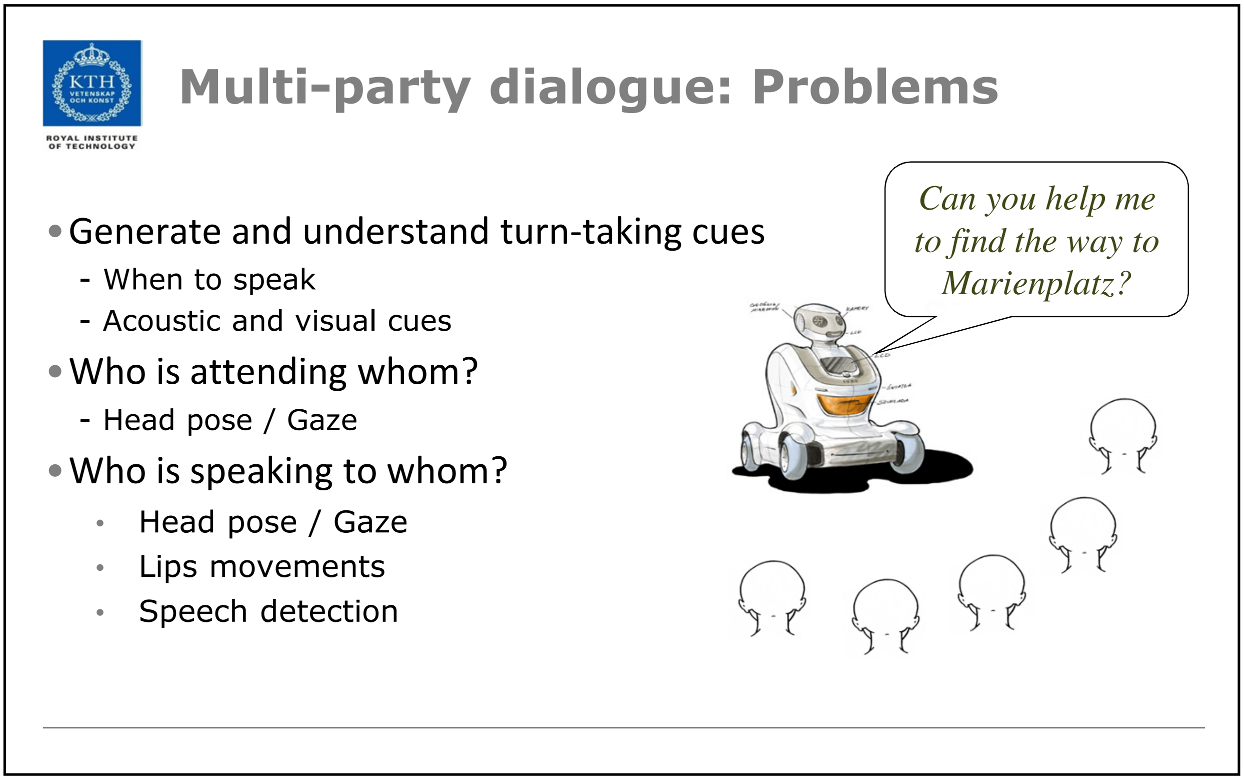


A three-step interactive recordning

- Small scripted dialogue snippets:
  - **Excuse me do you know how I could get to an A T M?**
  - Sure, let me explain
  - **Okay! *enthusiastic***
  - If you continue for about five hundred metres
  - **Okay.. *continuer***
  - You will see a large building. There you take left.
  - **Okay. *acknowledgement***
  - You should continue for about three blocks and pass the opera building.
  - **Okay?!? *uncertain***
- Chatting while playing chess
- Socializing small talk



Feedback tokens has intrinsic meanings

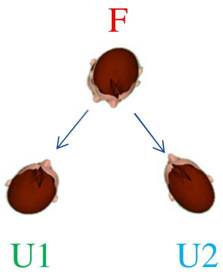


Multi-party dialogue: Problems

- Generate and understand turn-taking cues
  - When to speak
  - Acoustic and visual cues
- Who is attending whom?
  - Head pose / Gaze
- Who is speaking to whom?
  - Head pose / Gaze
  - Lips movements
  - Speech detection

Multi-party dialogue example

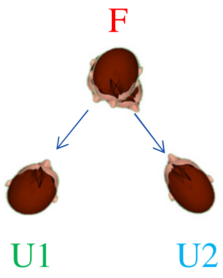
F: Hi there. Could you perhaps help me?
U1: Yes [S: **yes**]
F: I have some questions for you. When do you think robots will beat humans in football?
U1: Maybe in 10 years
[D: that see **in 10 years**]
F: That soon! Could you elaborate on that?
U1: Well, they are already pretty good at it
[D: while they are already predicted owners]
F: Yeah… I have another question for you.
Would you like robots to be like humans?
U2: Yes, absolutely [D: **yes** that see]
F: Could you just wait a second?
F: I'm sorry, where were we.
Would you like robots to be like humans?
U1: Sure [S: **sure**]



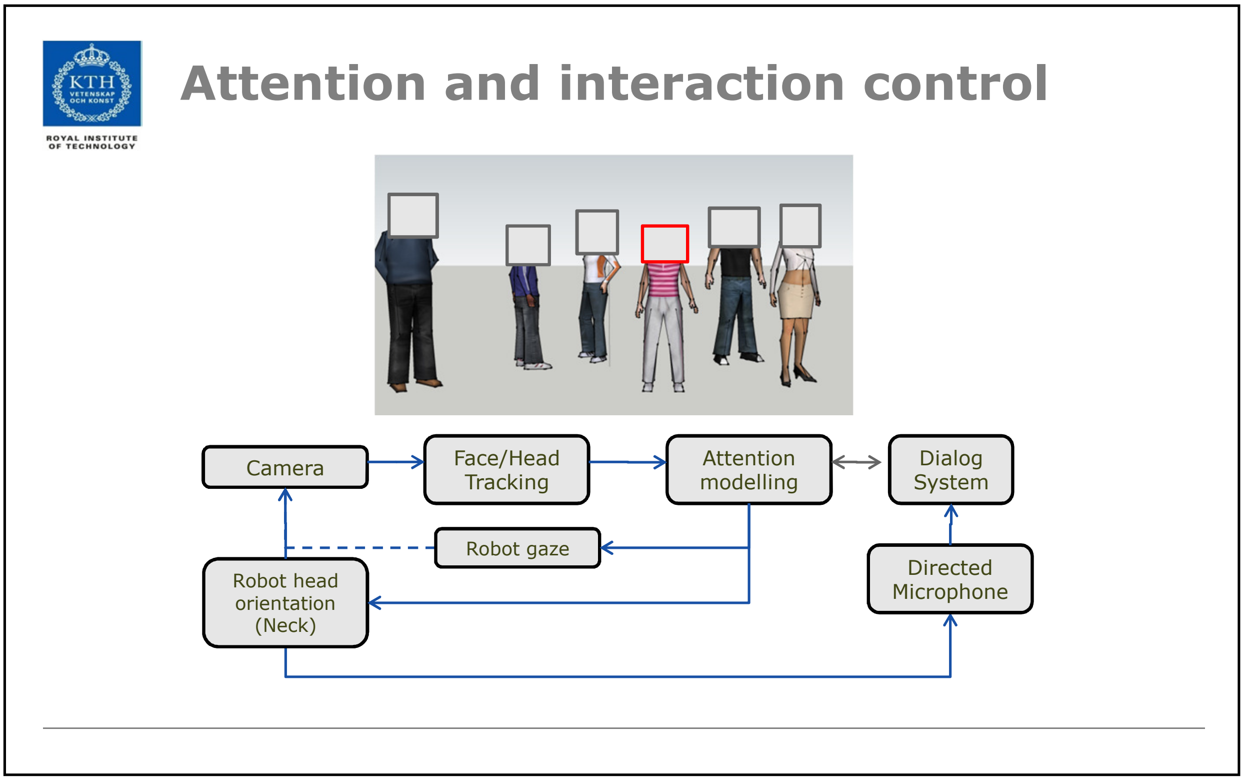
Skantze' s presention at Dialogverkstad 2012 in Stockholm

Multi-party dialogue cont…

F: I'm sorry, where were we.
Would you like robots to be like humans?
U1: Sure [S: **sure**]
F: Yeah?
F: Do you agree?
U2: No [S: **no**]
F: Mhm. Could you say something more about that?
U2: Well, I want robots to be like robots
[D: I want robots to be like Reynolds]
F: Yeah…
F: I have another question for you. When do you think robots will serve hamburgers at McDonalds?
U1: Do I need an umbrella?
F: What do you think, you are in London!
Maybe you have another question for me?



Skantze' s presention at Dialogverkstad 2012 in Stockholm



Attention and interaction control

Camera

Face/Head Tracking

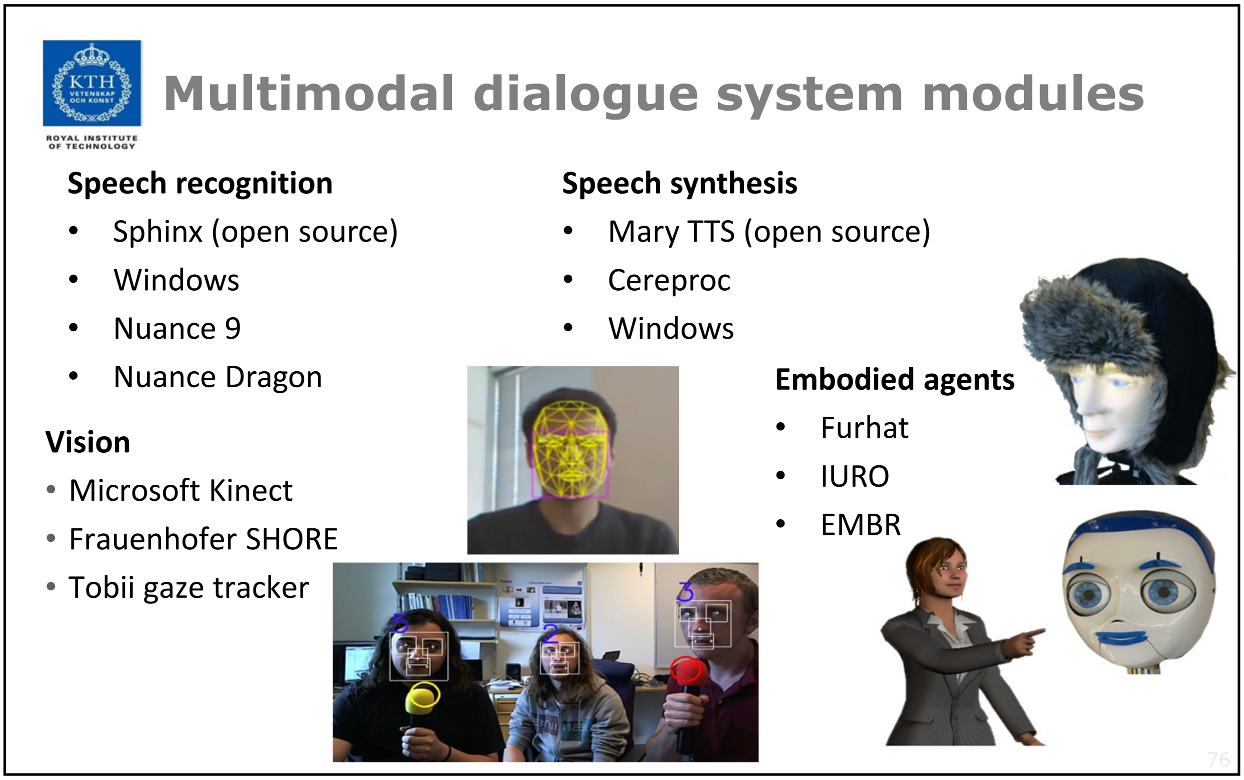
Attention modelling

Dialog System

Robot gaze

Robot head orientation (Neck)

Directed Microphone



Multimodal dialogue system modules

Speech recognition

- Sphinx (open source)
- Windows
- Nuance 9
- Nuance Dragon

Speech synthesis

- Mary TTS (open source)
- Cereproc
- Windows

Vision

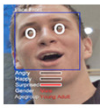
- Microsoft Kinect
- Frauenhofer SHORE
- Tobii gaze tracker

Embodied agents

- Furhat
- IURO
- EMBR

Cognitive load and attention sensors

Tobii eye trackers: gaze point, eye position, pupil size



Shore face tracker: head position and direction,
mouth and eye opening,
age and emotions

Affectiva Q Sensor: skin conductance, temperature,
3-axis accelerometer

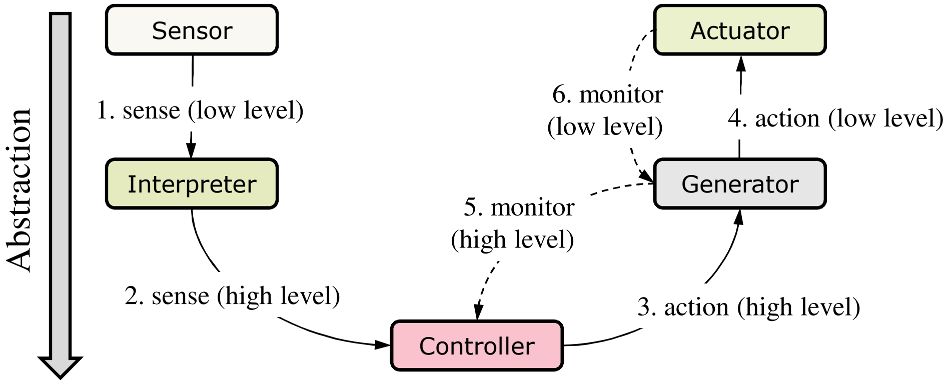
Microcone microphone: 6 track recordings in
6 different directions

Garmin Forerunner: heart rate

Ambu Inductance Belt: lung volume (variation)

getsafe

Modules and events



Abstraction

Sensor

1. sense (low level)

Interpreter

2. sense (high level)

Controller

3. action (high level)

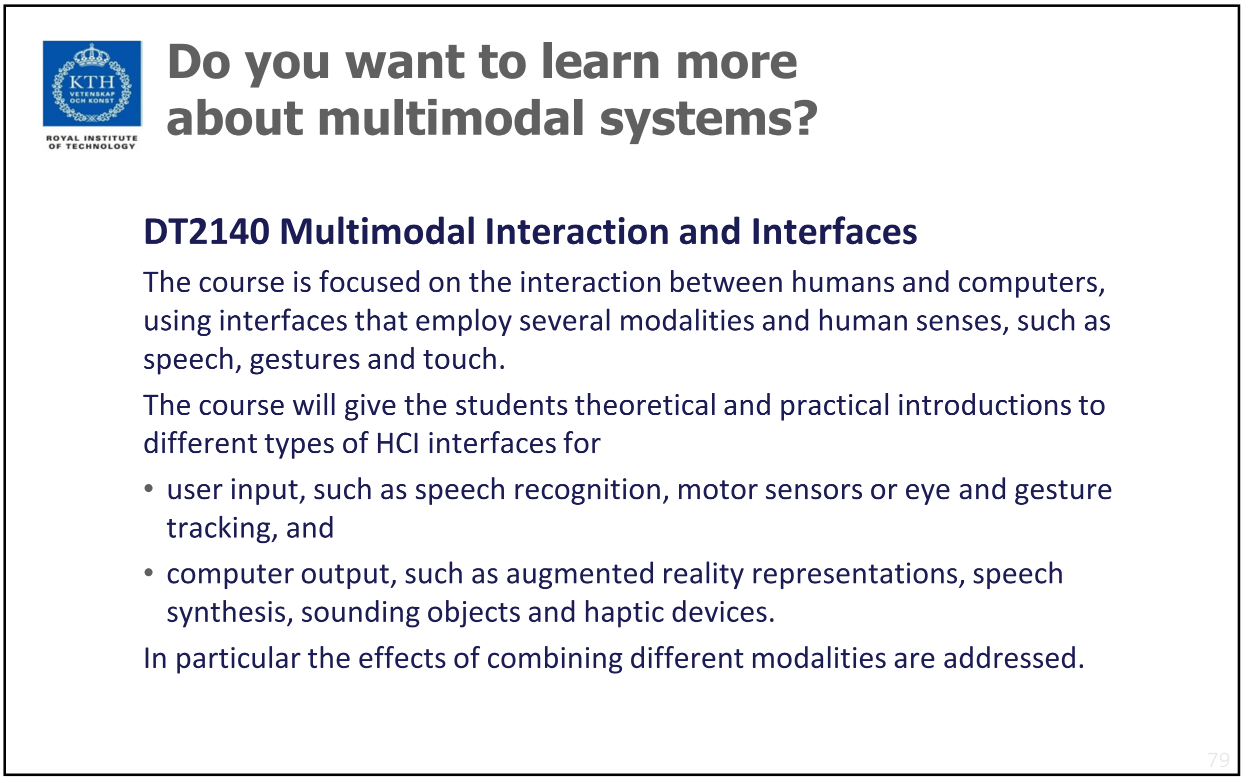
Generator

4. action (low level)

Actuator

5. monitor
(high level)

6. monitor
(low level)



Do you want to learn more about multimodal systems?

DT2140 Multimodal Interaction and Interfaces

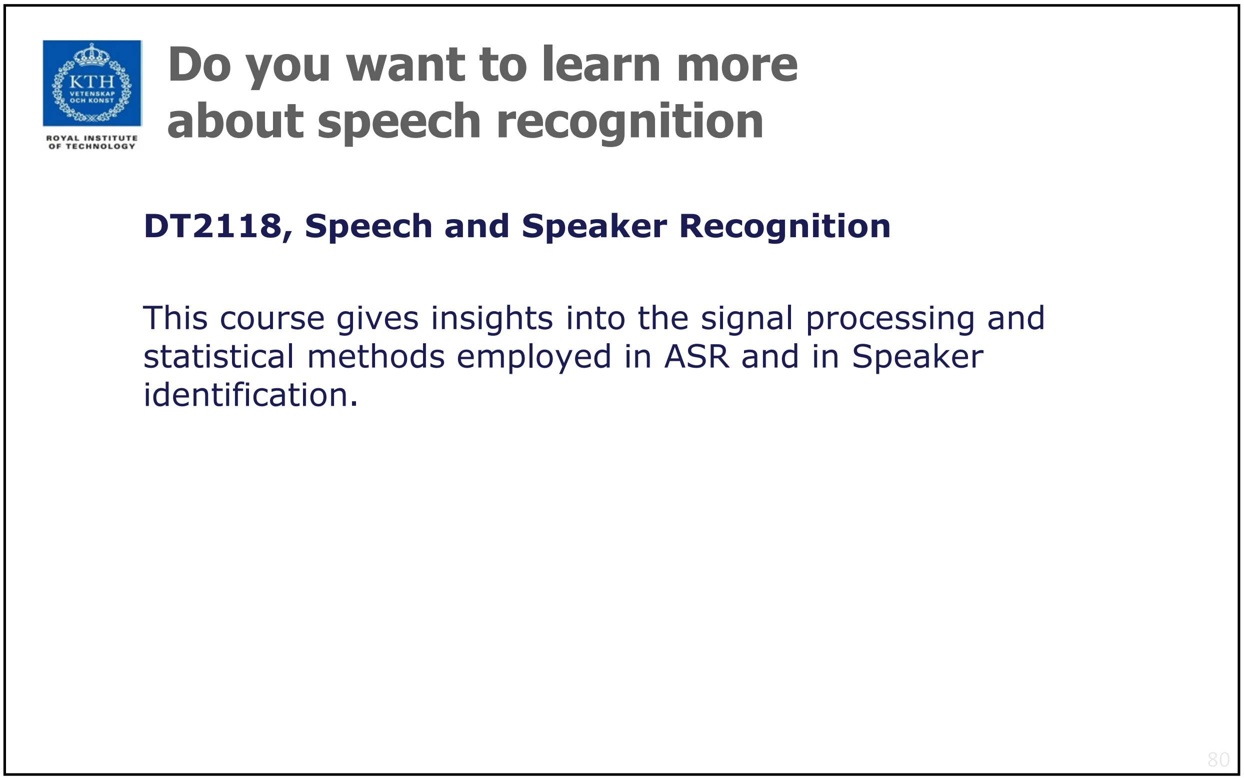
The course is focused on the interaction between humans and computers, using interfaces that employ several modalities and human senses, such as speech, gestures and touch.

The course will give the students theoretical and practical introductions to different types of HCI interfaces for

- user input, such as speech recognition, motor sensors or eye and gesture tracking, and
- computer output, such as augmented reality representations, speech synthesis, sounding objects and haptic devices.

In particular the effects of combining different modalities are addressed.

79



Do you want to learn more about speech recognition

DT2118, Speech and Speaker Recognition

This course gives insights into the signal processing and statistical methods employed in ASR and in Speaker identification.

80