Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

Dialogue systems

Joakim Gustafson
Centrum för Talteknologi, KTH
Tal, musik och hörsel, KTH

Agenda

• Introduction to speaker
• Spoken Dialogue Systems
• Research SDSs
• Commercial SDSs
• SDS Components
• An example SDS

Introduction
Joakim Gustafson

1987-1992 Electrical Engineering program at KTH
1992-1993 Linguistics at Stockholm University
1993-2000 PhD studies at CTT/KTH
2000-2007 Senior researcher at Telia R&D department
2007- Assistant Professor CTT/KTH

My research field: spoken dialogue systems

Multi-disciplinary field
Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

**CTT - Centrum för talteknologi**

Research areas
- Speech production
- Speech perception
- Communication aids
- Multimodal speech synthesis
- Speech recognition
- Speaker verification
- Spoken conversational speech
- Interactive dialogue systems

---

**The KTH speech group**

- Early days

Gunnar Fant and OVE i 1953

---

**Course Offering**

<table>
<thead>
<tr>
<th>Name</th>
<th>Nr</th>
<th>Period</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speech Technology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2F111</td>
<td>J3-A,B,F4</td>
<td>3-4</td>
<td></td>
</tr>
<tr>
<td>2F112</td>
<td>Media 3-4</td>
<td>3-5</td>
<td></td>
</tr>
<tr>
<td>2F113</td>
<td>Media 3-4</td>
<td>2-4</td>
<td></td>
</tr>
<tr>
<td>2F1212</td>
<td>D4, E4, F4</td>
<td>4-4</td>
<td></td>
</tr>
<tr>
<td>2F1213</td>
<td>Media 3-4, E4, D4</td>
<td>4-5</td>
<td></td>
</tr>
<tr>
<td>2F1402</td>
<td>D4-5, E4-5</td>
<td>1-4</td>
<td></td>
</tr>
<tr>
<td><strong>Music Technology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2F1410</td>
<td>Media 3-4, E4, D4</td>
<td>2-5</td>
<td></td>
</tr>
<tr>
<td>2F1411</td>
<td>e&amp;N</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>2F1412</td>
<td>e&amp;N</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>2F1413</td>
<td>e&amp;N</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>2F1414</td>
<td>D1-2</td>
<td>3-4</td>
<td>5</td>
</tr>
<tr>
<td>2F1415</td>
<td>D1-2</td>
<td>3-4</td>
<td>5</td>
</tr>
</tbody>
</table>

---
Spoken Dialogue Systems

The spoken dialogue system vision

An interface that allows speakers to interact with a computer using spontaneous, unconstrained speech

Talking machines as we know them from movies...

- Star Trek
- HAL 9000 (2001)
- C3PO (Star Wars)
...and since a couple of years from real systems

How do we want speaking machines to behave?

...but to a limit?

How are Spoke Dialogue Systems perceived?

- People may use more than one metaphor to make sense of human-computer interaction
  - The interface metaphor – The dialogue system as an alternate means of interfacing with a computer
  - The human metaphor – The computer as a person with human-like conversational abilities
- One is not necessarily better than the other!
Example #1: interface metaphor

- Stockholm Transport (public transport) Journey planner
- First released as web service (forms)
- Later released as telephony based speech service
- Frequent user, knows order of system prompts
- Clues suggesting use of interface metaphor
  - Strictly controlled input: place names, times, yes/no (or lack thereof)
  - Strictly limited output: slot filler requests, verifications, journey information
  - Input and output follows the web interface closely
  - User refrains from social utterances and feedback

Example #2: human metaphor

- KTH experiment on responsive grounding
- Partial Wizard-of-Oz
- Naïve user
- Near-perfect system understanding (WoZ)
- Fast interaction, backchannels, fragmentary grounding
- Clues suggesting use of human metaphor
  - greetings, socializing
  - user gives verbal feedback to system utterances
  - user expresses opinions
  - long user utterances
  - very high confidence in system's conversational abilities

The beauty of speech

1. Works in hands free situations
2. Works in eyes free situations
3. Works when other interfaces are inconvenient:
4. Works where disabilities make other interfaces useless
5. Works with common hardware, e.g. a telephone
6. Efficient information transfer (as far as humans are concerned).
Speech application Domains

- Information retrieval systems
  - e.g. train time table information or directory inquiries
- Ordering
  - ticket booking, buying music
- Command control systems
  - home control ("turn the radio off") or voice command shortcuts ("save")
- Dictation

The beauty of speech (cont)

7. Reasoning.
8. Problem solving.
9. Naturalness
10. Easy-of-use
11. Flexibility
12. Error handling and hedging
13. Mutual adaptation enables error handling and more efficient information transfer

Application Domains (cont)

- Games and entertainment
  - Games can take advantage of the more social aspects of spoken dialogue
- Co-ordinated collaboration
  - The task of controlling or over-viewing complex situations requires flexibility and efficiency.
- Expert systems
  - Diagnose and help systems that need to reason about facts and goals and may benefit from natural dialogue
- Learning and training
  - Naturalness, flexibility, and robustness are attractive features in training environments
Multimodal interfaces

• User and system interact by speech and other modalities.

Why multimodality?

• Different modalities suited for different things
• Increased flexibility
• New kinds of terminals
• Facilitates error handling
• Graphical output can confirm spoken input

Research Spoken Dialogue Systems
Classic research dialogue systems

- Historic research systems
  - Voyager (1989)
  - ATIS (1992)
  - SUNDIAL (1993)
  - TRAINS (1996)
- Application
  - Philips Train Information (1995)
- Large Efforts
  - Communicator
  - Verbmobil
  - SmartKom

The Nordic Scene

- Stockholm, Sweden
  - KTH (Waxholm, August, AdApt, Higgins)
  - Telia (Resebokning, NICE)
- Linköping, Sweden
  - IDA (LINGUIST, Ornithology, Movie Finder)
- Göteborg, Sweden
  - Lingvistik (TRIDI, GODIS, TALK)
  - Chalmers (State-chart xml)
- Aalborg, Odense Denmark
- Helsinki, Tampere Finland
- Trondheim, Norway

The Waxholm Project (92-95)

- Tourist information
  - Stockholm archipelago
  - time-tables, hotels, hostels, camping and dining possibilities.
- Mixed initiative dialogue
  - speech recognition
  - multimodal synthesis
- Graphic information
  - pictures, maps, charts and time-tables.
There are lots of boats from Stockholm to Waxholm.

Information about the restaurants in Waxholm.

I am looking for boats to Waxholm.

The Waxholm system.

The August kiosk at Kulturhuset 98.

Thank you too.

• Swedish spoken dialogue system for public use
• 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
People who live in glass houses

Yes, it might be that we will!

I come from the department where...er...I don't really have a surname...er...I'm Joakim Gustafson from the KTH's Department of Speech, Music and Hearing.

Yes, it might be that we will!

Strindberg was born in 1849.

Over a million people live in the Stockholm area.

What do you do for a living?

The Royal Institute of Technology! Strindberg was married three times!

Secondarily, HIGGINS is intended to provide simple information about the immediate surroundings.

The primary domain of HIGGINS is city navigation for pedestrians.

Higgins målsättning -demo

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

The HIGGINS system (03-)

Strindberg was married three times!
**The Commercial Speech Applications**

**Access Convergence & Speech**

**Commercial Applications of Speech technology**

- Dictation
  - Medical
  - Legal
  - Windows Vista (95-99% without training?)
- Customer care center automation
  - Call Routing
  - Technical support
- Multimodal mobile applications
  - Information search
  - Content download
  - Entertainment
Problem solving

- Voxify
  - Wyndham hotel reservation
  - Continental airline tickets
- SpeechCycle
  - Broadband support Agent
  - Video support Agent
  - IP Telephony Agent

Some nice dialogue features of Speechcycle’s Agents

- 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

SDS Components
Spoken dialogue processing

The Components of Spoken Dialogue Systems

- Automatic Speech Recognition (ASR)
- Natural Language Understanding (NLU)
- Dialogue Management (DM)
- Natural Language Generation (NLG)
- Text-To-Speech synthesis (TTS)
- Multimodal modules

Speech Recognizer

Purpose: convert speech to text
Input: digital speech
Output: String(s) or lattice that represent the most probable utterance(s)
Considerations: Vocabulary size, Grammar and Speech type
Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

Speech Recognition
- Identifiera talstämman
  - Mått på spektrum på talssignaler 100
  - Matcha de identifierade stämmorna mot ett
  - Konstruerar en serie av de mest
  - Följer med självständighet

What makes speech recognition difficult?
- Mått på talssignaler 100
  - Matcha de identifierade stämmorna mot ett
  - Konstruerar en serie av de mest
  - Följer med självständighet

Svenska dialekter
"Flyget, tåget och bilbranschen tävlar om länissenhet och folkelikt garnit."
Statistically based recognition

Results 2005
(earlier result in parenthesis)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Speech type</th>
<th>Lexicon size</th>
<th>Word Error Rate (%)</th>
<th>Human Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digit string (phone)</td>
<td>spontaneous</td>
<td>30</td>
<td>0.3</td>
<td>0.009</td>
</tr>
<tr>
<td>Resource Management</td>
<td>read</td>
<td>1000</td>
<td>3.6</td>
<td>0.1</td>
</tr>
<tr>
<td>ATIS</td>
<td>spontaneous</td>
<td>2000</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Wall Street Journal</td>
<td>read</td>
<td>4000</td>
<td>6.6</td>
<td>1</td>
</tr>
<tr>
<td>Radio News</td>
<td>speech</td>
<td>4000</td>
<td>13.5 (20)</td>
<td>-</td>
</tr>
<tr>
<td>Switchboard (phone)</td>
<td>conversation</td>
<td>10000</td>
<td>19.3 (18)</td>
<td>4</td>
</tr>
<tr>
<td>Call Home (phone)</td>
<td>conversation</td>
<td>10000</td>
<td>30 (50)</td>
<td>-</td>
</tr>
</tbody>
</table>

Natural Language Understanding

**Purpose:** produce meaning representation from ASR output

**Input:** String(s) lattice of words

**Output:** Meaning representation

**Considerations:** Type of grammar (Finite-state, Full parse or Word-spotting)
Phrases and syntax

I want to take the boat

Robust Utterance Analysis

- Robust interpretation
  - Using grammar to automatically detect non-expected words between and inside phrases
  - Performs better than keyword-spotting for detecting erroneous content-words

Pragmatics/world knowledge

S: There is a flight at 8.10.
U: Isn’t that usually delayed?

- Flights can be delayed
- To be delayed is bad
- “Usually” -> recurring in time. How often is “usually”? 
- Is this particular flight delayed a lot?
- What would that mean for connecting flights?
**Dialogue Management**

*Purpose:* decide the next system action  
*Input:* a meaning representation from NLU  
*Output:* High-level communicative goal(s)

---

**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 (what to say)  
    - Information Packaging (how much to say)  
    - Turn-taking (who can speak/interrupt)

---

**Knowledge Sources for Dialogue Systems**

- **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:
    - Grammar, Dictionaries, Corpora
- **User Model:**
  - Information about user that could be relevant to the dialogue
    - Age, gender, spoken language, location, preferences, etc.  
    - Dynamic information: the current mental state (belief, desires, intentions)
**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

**Interaction control**

- Conversation
  - Exchange of information
  - Control of the exchange of information
- Turn-taking
  - Control of the ‘floor’
- Feedback
  - Perception, attention, understanding, attitude...
- Dialogue systems needs both turn-taking and feedback

**An interaction control experiment**
Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

The Telia Customer Care Line
• 14 million calls/year
• A number of areas/families
  – Fixed telephony, Mobile telephony, Mobile data, Broad band,
    IP telephoni, IP-TV...
• A number of reasons for calling
  – Billing, ordering, problem solving, support, information

Simulation (Wizard-of-Oz)

User

Human operator

The point of WoZ simulation
• Data for speech recognition
• Typical utterance patterns
• Typical dialogue patterns
• Click-and-talk behaviour
• Hints at important research problems
An in-service WoZ recording

- Ten real customer care agent were used as wizards (in-service WoZ collection)
- Five weeks of recordings resulted in 42,000 transcribed and tagged calls

The WOZ recording environment

- Customer care agent were used as wizards:
  - Listened to the customer (acting as ASR)
  - Generate system answers via a prompt piano (acting as DM)
  - Routing the call to the appropriate agent (acting as call router)
- All calls were logged and recorded
- The wizard could route to themselves in problematic cases:
  - Less customer irritation
  - Human error handling dialogues
  - Used to collect interviews

Potential benefits of making a call routing SDS more human-like

- The callers are more likely to behave as if they were speaking to a human operator
- The caller utterances become longer and less like menu speak
- More information is made available to the call routing classifier
- It should be possible to perform more fine grained call routing
The re-design of the prompts

- The prompts should make the system more responsive (indicating listening, thinking)
- Feedback and grunts (mmm, okej) were added to be used to encourage the user to keep on describing their reason for calling
- Utterances as response to channel checks (hello) and social utterances like greeting (hi) were added

Wizard-of-Oz interface 2

Example dialogue

SYS: välkommen till telia vad kan jag hjälpa dig med? (Hello)
KUND: ja hej [PAUS] ehh [PAUS] hör du mig (Greeting) (Channel)
SYS: ja hallo (Channel)
KUND: ja [PAUS] jo eh jag har problem med en räkning (Description)
SYS: mm (Feedback)
KUND: jag har EHH anmält att jag har ingen telefon och jag har fått en eh [PAUS] en räkning på ett abonnemang (Description) (Feedback)
SYS: okej (Feedback)
KUND: och jag har ringt flera flera gånger och försökt förklara det att jag betalar inte så långt jag inte får telefonen inkopplad (Description) (Question) (Answer)
SYS: gäller det din hemtelefon? (Question)
KUND: ja (Answer)
SYS: mm dröj lite så ska jag köppla dig (Reading)

Natural Language Generation

**Purpose:** produce an output string to be shown/spoken to the user
**Input:** Representation from DM
**Output:** String for TTS (possibly marked for prosody, etc.)

Utterance Generation

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

Text-to-Speech Synthesis

**Input:** text string (with tags)
**Purpose:** speak string to user
**Considerations:** Human-like, Flexibility
What is speech synthesis?

- Recorded human speech
  - Words and phrases
  - Fix vocabulary
- Systematically recorded fragments of speech
  - One fixed speaker
  - Record the phoneme transitions “diphones”
- Parametric synthesis (artificial)
  - Formant synthesis
  - Articulatory synthesis
- Multimodal synthesis (talking head)

Text-to-speech (TTS)

“abcd”

Text Preprocessing

- Sentence end detection (semicolon, period – ratio, time and decimal point, sentence ending respectively)
- Abbreviations (e.g. – for instance). Changed to their full form with the help of lexicons
- Acronyms (I.B.M – these can be read as a sequence of characters, or NASA which can be read following the default way)
- Numbers (Once detected, first interpreted as rational, time of the day, dates and ordinal depending on their context)
- Idioms (e.g. “in spite of”, “as a matter of fact” – these are combined into single FSU using a special lexicon)
Grapheme-to-phoneme conversion

- Dictionary:
  - Store a maximum of phonological knowledge into a lexicon.
  - Compounding rules describe how the morphemes of dictionary items are modified.
  - Hand-corrected, expensive
  - The lexicon is never complete; needs out of vocabulary pronunciators, transcribed by rule.
- Rules:
  - A set of letter to sound [grapheme to phoneme] rules.
  - Words pronounced in such a particular way that they have their own rule are stored in exceptions directory.
  - Fast & easy, but lower accuracy
- Machine learning:
  - Cart tree
  - Analog pronunciation

Sound generation approaches

- Parametric synthesis (Synthesis by Rule)
  - Formant synthesis or articulatory synthesis
  - Cognitive approach of the phonation mechanism
  - Speech is produced by mathematical rules that formally describe the influence of phonemes on one another
  - Flexible, but lower quality - today
- Synthesis by Concatenation
  - Diphones or larger units (unit selection)
  - Limited knowledge of the data to be handled
  - Elementary speech units are stored in a database and then concatenated and processed to produce the speech signal
  - One speaker

Text-to-Speech Synthesis (TTS) Evolution

- Poor intelligibility, Poor Naturalness
- Good Intelligibility, Poor Naturalness
- Good Intelligibility, Good Naturalness

- LPC-Based Synthesis
- Bell Labs; Joint Speech Research Unit; MIT; AT&T; T&B; Nokia
- Good in Japan; CNET in France; BT in England; ASENT Labs (1988); LBS in Belgium

- Unit Selection Synthesis
- Good in Japan; CNET in France; BT in England; ASENT Labs (1988); LBS in Belgium

- HMM Synthesis
- Good in Japan; CNET in France; BT in England; ASENT Labs (1988); LBS in Belgium

- Formant Synthesis
- Bell Labs; Joint Speech Research Unit; MIT; AT&T; T&B; Nokia
- Good in Japan; CNET in France; BT in England; ASENT Labs (1988); LBS in Belgium

Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

Source filter theory

Frequency: [Diagram]

Time: [Diagram]

Source (glottal flow)  Filter (shape of vocal tract)  Radiation (lips)

Formant synthesis (1959-1987)

- Haskins, 1959
- KTH – Stockholm, 1962
- Bell Labs, 1973
- MIT, 1976
- MIT-talk, 1979
- Speak ‘N spell, 1980
- BELL Labs, 1985
- Dec talk, 1987

Articulatory parameters

- Jaw opening
- Lip rounding
- Lip protrusion
- Tongue position
- Tongue height
- Tongue tip
- Velum
- Hyoid
Concatenative synthesis

- Diphone synthesis
  - Svensk (Mbrola)
- Unit selection
  - Svensk: weather (CSTR) good, less good

Diphone synthesis

- Mid-phone is more stable than edge:

<table>
<thead>
<tr>
<th>w</th>
<th>eh</th>
<th>d</th>
<th>b</th>
<th>eh</th>
<th>n</th>
</tr>
</thead>
</table>

  Time (s)

Diphone Synthesis

- Well-understood, mature technology
- Problems:
  - Signal processing still necessary for modifying durations
  - Source data is still not natural
  - Units are just not large enough, can’t handle word-specific effects, etc
**Unit Selection Synthesis**

- Natural data solves problems with diphones
  - Diphone databases are carefully designed but:
    - Speaker makes errors
    - Speaker doesn't speak intended dialect
    - Requires database design to be right
  - If it's automatic
    - Labeled with what the speaker actually said
    - Articulation, schwa, fags are natural
- "There's no data like more data"
  - Lots of copies of each unit mean you can choose just the right one for the context
  - Larger units mean you can capture wider effects

---

**Unit Selection Summary**

- **Advantages**
  - Quality is far superior to diphones
  - Natural prosody selection sounds better
- **Disadvantages:**
  - Quality can be very bad in places
  - HC problem: mix of very good and very bad is quite annoying
  - Synthesis is computationally expensive
  - Can't synthesize everything you want:
    - Diphone technique can move emphasis
    - Unit selection gives good (but possibly incorrect) result

---

**Current trend – machine learning**

- Problems with Unit Selection Synthesis
  - Can't modify signal
  - (mixing modified and unmodified sounds bad)
  - But database often doesn't have exactly what you want
- Solution: HMM (Hidden Markov Model) Synthesis
  - Won the last TTS bakeoff.
  - Sounds unnatural to researchers
  - But naive subjects preferred it
  - Has the potential to improve on both diphone and unit selection.
Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

HMM synthesis

- A speech synthesis technique based on
- HTK (Hidden Markov Model Toolkit)
- Developed by the HTS working group
- at the Department of Computer Science Nagoya
- Institute of Technology
- Interdisciplinary Graduate School of Science and
- Engineering Tokyo Institute of Technology
- http://hts.sp.nitech.ac.jp

Examples of HMM Synthesis

- Unit selection (Roger)
- HMM (Roger)

- Unit selection (Nina)
- HMM (Nina)

Multimodal Synthesis (talking heads)

Input: tagged text string
Purpose: support synthesis with lip movements
Considerations: Synchronization, Human-like?
Why Talking Heads?
- The most natural form of communication that we know of is face-to-face interaction
- A virtual talking head can improve information transfer
  - Through expression, gaze, head movements
  - Through lip-cheek and tongue movements

Tasks of an Animated Agent
- Provide intelligible synthetic speech
- Indicate emphasis and focus in utterances
- Support turn-taking
- Give spatial references (gaze, pointing etc)
- Provide non-verbal back-channeling
- Indicate the system’s internal state

Applications
- Improved speech synthesis
- Human-Computer Interface in spoken dialogue systems
- Aid for hearing impaired
- Educational software
- Stimuli for perceptual experiments
- Entertainment: games, virtual reality, movies etc.
Techniques for facial animation

- Talking head synthesis requires a signal model and a control model. The signal model may be:
  - Video-based (2D)
    - Enables realistic reproduction
    - Lacking flexibility
  - Muscle based (3D)
    - Highly flexible
    - Difficult to gather anatomical data
    - Computationally intensive (although less of a problem today...)
  - Direct parameterisation/morphing based (3D)
    - Good compromise between flexibility and realism
    - Simple data acquisition using optical methods

Direct parameterisation

- 3D-modelling
- Deformation through high-level parameters
- Different possible parameterisations:
  - Articulatory oriented
  - MPEG-4 (low-level)

Articulatory oriented direct parameterisation

- High-level parameterisation tailored for visual speech animation
- Parameter set includes
  - Jaw opening
  - Lip rounding
  - Bilabial closure
  - Labiodental closure
- Parameters are normalized relative to spatial targets
MPEG-4 direct parameterisation

- Original purpose: model based video coding
- The standard defines a generic face object
- 84 feature points (FPs)
- 68 facial animation parameters (FAPs)
- FAPs are normalized relative to distances in the face
- Expressed in FAPU (FAP unit)

Measuring talking faces

- Cyberware scanning
  - High definition 3D shape and texture
  - Static
  - Expensive equipment
- Motion capture
  - Captures points in 3D
  - Dynamic
  - Starting to become affordable
- Video analysis
  - Can provide texture, contours and points
  - Inexpensive hardware
  - Typically less accurate than motion capture

Collection of audio-visual databases

- Eliciting technique: information seeking scenario
- Focus on the speaker who has the role of information giver
- The speaker seats facing 4 infrared cameras, a digital video-camera, a microphone. The other person is only video recorded.
Recording and model

Emotional synthesis

Interactions: emotion and articulation
(from AV speech database – EU/PF_STAR project)
Datadriven facial synthesis with MPEG4 model

An example SDS

The NICE system
Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

Background

- NICE was a EU project 2002-2005
- Five Partners
  - Liquid Media, Sweden
  - Scansoft (initially Philips, now Nuance), Germany
  - Limsi, France
  - TeliaSonera, Sweden
  - Nislab, Denmark
- Two systems were developed
  - a Swedish Fairytale game
  - an English HC Andersen story-teller

The NICE system setup

A computer game with multimodal dialogue

- It puts new & different demands on speech and dialogue technology
  - young users
  - a game lasts for many hours
  - many kinds of dialogue: problem solving, negotiation, information-seeking, socializing, command & control, multi-party
  - the setup is multimodal and asynchronous
- Many things are easier than for 'normal' dialogue systems
  - low penalty for recognition errors
  - voluntary suspension of belief
  - gamers already know all the clichés
- $ Might pay off big some day $
Research challenges

- Conversational Fairytale characters
  - Different personalities and goals
  - Autonomous
  - Reactive
  - Aware of the user, the environment and each other
- Conversational multimodal output generation
- Game-progression VS user control
- Understanding children’s multi-modal input

Character design in Nice

- Cloddy Hans looks stupid, but friendly – Karin looks a bit angry and smart
- Cloddy Hans is a classical friendly helper character
- The tasks the users have to solve are centered around what is seen in the scenes
- All characters have emotional and attention states
- The system features characters with different roles (e.g., helper and gatekeeper)

The fairy-tale game

- Two scenes – the room and the fairy-tale world
  - Scene one – Multimodal training session put-that-there
  - Scene two – Multi-party negotiation dialogue
- Three characters
  - Cloddy Hans – slow but friendly Helper character
  - Karen – fast and sullen Gatekeeper character
  - Thumblina – silent supporting character with attitudinal gestures
Inside a fairy-tale brain

- **World model:** Interrelated (Java) objects representing things, characters, and their relationships.
- **Discourse history:** Who said what when?
- **Agenda:** Goals, actions and their causal relationships.

Scene scripting

There is only one "current scene"
Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

**Characters are autonomous**

![Character diagram]

**Message-passing**

![Message-passing diagram]

**Event-based dialogue**

![Event-based dialogue diagram]
Dialogsystem: Joakim Gustafson
Tal, musik och hörsel

The Animation system

Animation handler actions used to
"pick it up"+selection of lamp

System Module (Message) | Animation Handler Request | Animation Renderer Request
--- | --- | ---
GR(StartOfGesture) | AttentionTo(GestureStart) | 1) RaisedEyebrows
GR(EndOfGesture) | LookAt(Lamp) | 2) Play(Animation, TextToSpeech(Emote HeadTurnLeft(amount))
| | | 3) TurnTo(LookAtCamera)
| | | 4) Play(Animation, Emote(HeadTurnLeft(amount)), Eyebrow(frown))
| | | 5) Play(Sound(synthesis_file))
NLP(Select(Lamp)) | TakeTurnGesture | 6a) DM(perform(PickUp(Lamp)))
| | 6b) DM(convey(tell(Will you pick up the lamp?))
| | 6c) DM(convey(tell(I will pick up the lamp))))
Attentional and emotional output behaviors

- Emotional display
  - neutral
  - surprise
  - anger
  - happiness
- State-of-mind gestures
  - joy
  - listening
  - thinking
  - sad
- Turn regulation feedback gestures
  - attention
  - giving turn
  - taking turn

Body gestures and physical actions

- Specific body movements
  - grimace
  - whispering
  - waving
  - Billy's arm
- General movements (single body part)
  - head
  - hands
  - eyes
  - right
  - arm side
- Physical actions
  - on it
  - turn it
  - execute

Signs of social involvement and engagement

- Users form alliances with either Cloddy Hans or Karin
- User show repent when being accused of deceit
- Users lie, make ironic, sarcastic and humorous remarks
- Users react to the character’s mood and use politeness markers
- Users lecture Cloddy Hans while making reference to common dialogue history.
The End