Multimodal Interaction & Interfaces

Gabriel Skantze
gabriel@speech.kth.se
Department of Speech, Music and Hearing

First some introduction to the topic, then some introduction to the course
Who am I?

• MSc in Cognitive Science (1996-2000)
  – Linköping University
  – Computer Science, Psychology, Linguistics
  – HCI, Human Factors, AI, NLP
• Voice User Interface Designer (2000-2002)
  – Pipebeach AB, Stockholm
• PhD in Speech Communication (2002-2007)
  – Error Handling in Spoken Dialogue Systems
• Present: Researcher at KTH/TMH
  – Incremental processing
  – Human-robot interaction
History of the Graphical User Interface

- In the beginnings: Punch cards (18th century)
- The Command Line Interface (1950s)
- The GUI: NLS (1960s) developed at SRI
  - Display, Keyboard, Mouse
  - Multiple windows
- Alto personal computer (1973) developed at Xerox PARC
  - Desktop metaphor, WIMP (windows, icons, menus, pointing)
  - WYSIWYG
- Apple Macintosh (1984)
- X Window System (1980s)
- Microsoft Windows 3.0 (1990)
Multimodal interaction

Milo in Project Natal for MS Xbox 360
Multimodal interfaces

Technology in Project Natal for MS Xbox 360
What are Multimodal Interfaces?

- Humans perceive the world through senses.
  - Touch, Smell, Sight, Hearing, and Taste
  - A **mode** = Communication through one sense

- Computers process information through modes
  - Keyboard, Microphone, Camera etc.

- Multimodal Interfaces try to combine several different modes of communicating: Speech, gesture, sketch ...
  - Use human communication skills
  - Provide user with multiple modalities
  - Multiple styles of interaction
  - Simultaneous or not
Other distinctions

• “Modality” is a fuzzy concept
• Language modality vs Action modality
  (Bos et al., 1994)
  – Indirect vs Direct manipulation

• Fine-grained distinctions:
  – Visual: Graphics, Text, Simulation
  – Auditory: Speech, Non-verbal sounds
Potential Input Modalities

- Pointing, Pen, Touch
- Motion controller
  - Accelerometer, Gyro
- Speech
  - or other sounds...
- Body movement/Gestures
- Head movements
  - Facial expression, Gaze
- Positioning
- Tangibles
- Digital pen and paper
- Brain?
- Biomodalities?
  - Sweat, Pulse, Respiration
- Taste? Scent?
Multimodal interaction

Potential Output Modalities

• Visual:
  – Visualization
  – 3D GUIs
  – Virtual/Augmented Reality

• Auditory:
  – Speech
  – Embodied Conversational Agents (ECAs)
  – Sound

• Haptics (tactile)
  – Force feedback
  – Low freq. bass
  – Pain

• Taste? Scent?
Strict Multimodality

- **Strict modality redundancy:**
  - All user actions should be possible to express using each modality
  - All system information should be possible to present in each modality

- **Motivation:**
  - Flexibility, predictability
  - “Design for all”

- **Problems:**
  - Modalities are good for different things, complement each other
  - Too limiting?
Multimodal vs. Multimedia

- Multimedia – more than one mode of communication is output to the user
  - An example is a sound clip attached to a presentation.
  - Media channels: Text, graphics, animation, video: all visual media
- Multimodal – Computer processes more than one mode of communication.
  - An example is the combined input of speech and touch in new mobile phones
  - Sensory modalities: Visual, auditory, tactile, ...
- Multimedia: subset of Multimodal Output
A Multimodal System

**Input**
- **Senses**
  - **Auditory:** Speech, Intonation
  - **Visual:** Facial expression, Body language, Gestures, Gaze
  - **Touch:** Tabs, pads, devices
  - *(Scent)*
  - *(Taste)*

**Cognition**
- **Interpretation / Modality Fusion**
- **Context:** World geometry, Application, Activity
- **Memory:** Grammar, Semantics, History
- **Personal Attribution:** User Configuration

**Runtime Framework**
- **Behaviors**
- **Generation**
- **Synthesis**

**Output**
- **Feedback**
  - **Auditory:** Speech, Sounds
  - **Visual:** Agents/Avatars, Environment, Virt. HCI entities
  - **Touch:** Force feedback, Low freq. Bass, Electrodes, Physical augmentations
  - *(Scent)*
  - *(Taste)*
Early vs. Late Modality Fusion

**Late Fusion**

- Speech
- Pen
- Speech Recognition
- Gesture Recognition
- Modality Fusion

**Early Fusion**

- Speech
- Pen
- Speech Recognition
- Gesture Recognition
- Modality Fusion
Why Multimodal Interaction?

Advantages over GUI and Unimodal systems:

• Natural/realism: Making use of more (appropriate) senses
• New ways of interacting
• Flexible: Different modalities excel at different tasks
• Wearable Computers and Small devices:
  — Usable Keyboard Typing Devices hard to use.
• Helps the Visually/Physically Impaired
• Faster, more efficient: Higher bandwidth is possible
• Robust: Mutual disambiguation of recognition errors
• Multimodal interfaces are more engaging
Why? Natural

Human – human protocols
Initiating conversation, turn-taking, interrupting, directing attention, ...

Human – computer protocols
Shell interaction, drag-and-drop, dialog boxes, ...

Based on real world interaction

- Use more of users’ senses
- Users perceive multiple things at once
- Users do multiple things at once
  - e.g., speak and use hand gestures, body position, orientation, and gaze
Pointing and speaking

Early example: Put-that-there (1980)
Multimodal interaction control

Comparing **Push-to-talk** with **Head pose tracking**
Multimodal interaction control

- System directed, recognized
- Tutor directed, ignored
- Tutor directed, recognised

Push-to-talk and Head pose tracking.
Why? Virtual Realism

Making use of more senses:
• Vision
• Sound
• Haptics

Important in simulated training
Why? Flexibility

User may choose the mode of input

Output through different modalities
Flexibility in referring to apartments

• Deictic:
  – “How much does it cost?” *(clicking on an apartment)*

• Descriptions:
  – “How much does the red apartment cost?”
  – “How much does the apartment at Karlavägen 108 cost?”

• Anaphora:
  – “How much does it cost?” *(local anaphora)*
  – “How much did the apartment we spoke about before cost?” *(global anaphora)*
Why? Robustness – Modality Switching

- The user should be promoted to use the least error-prone means of expression.
- Different modalities and means of expression could be more or less error prone for different users.
- The user should be promoted to alternate means of expression when errors occur. (Oviatt 1996)
Why? Robustness – Modality Fusion

Audio Feature Extraction

Audio-Visual Speech Recognition

Visual Feature Extraction

Recognition in speaker-dependent mode

- visual only
- audio only
- AV-early
- AV-late
- Multi-stream
Why? Flexibility: MonAMI Reminder

**Output**
- Embodied conversational agent (phone, screen)

**Input**
- Digital pen & paper
- Speech
Unifying speech, pen and web

- SMS notifications
- Web GUI
- Google Calendar
- Handwriting
- Embodied conversational agent

Example: 15.00 Meeting with Sara at Wayne's Coffee
Why? Easier on small devices

Example: Google Voice Search

Output
• Screen

Input
• Touch screen
• Speech
• Accelerometer
• Proximity meter
• Positioning
Why? Impairment support

Speech synthesis for non-vocal persons
Course overview
Course overview

- 10 Lectures
- 4 Laboratory exercises
- 1 Project
- 3 Assignments
- 2 Seminars
- 2 Visits
Another view...
Lectures

1. Introduction to multimodal interfaces
2. Mixed Reality
3. Tabletops, Tangibles and Tracking
4. Gesture-based interfaces
5. Sound in interaction
6. Speech technology interfaces
7. Multimodal speech interfaces
8. Haptic interfaces
9. Haptic interfaces
10. Issues in combining modalities in human-computer-interfaces
Laboratory exercises

1. Visual Interfaces
2. Gestures and sounding objects
3. Multimodal speech interfaces
4. Haptic interfaces

*Please do the preparatory exercises!*
Project

• Group project on multimodal interfaces
  – Explore new ways of using (one or more) modalities
  – Explore how to combine modalities
  – Implementation and/or evaluation
  – Compare to what others have done

• 3 persons per project
• ~2 weeks of work
Project instructions

1. Find two partners to do the project with.
2. Select a topic suitable for about two weeks work.
   - Extension of the lab exercises, a user evaluation, a replication of an experiment reported in the literature, a new interface etc. Combine theory and technology
3. Discuss your ideas with the teachers.
   - Is the equipment you need available? Is the project feasible within the time limits?
4. Register your project group at Bilda
5. Submit the project plan as Assignment 2 (November 21)
6. Do your project at home or use KTH lab facilities
   - Arrange with your supervisor if you need assistance.
7. Present at the project seminar (December 13)
8. Finalize the report and submit (January 5)
   - Check project requirements and grading criteria on the home page!
The project report should answer the following questions:

• What did you do?
• How did you do it?
• What results came out of your project?
• How did you evaluate them?
• What background and specific explanation do you need to provide so that people can understand?
• What has been done earlier in this area? How do earlier studies compare to yours?
In order to pass, the report should:

• be 6-12 pages in length, if a 12 pt font size is used.
• contain an Abstract giving a summary of the paper in no more than 200 words.
• have an Introduction that cites relevant previous work in the area and outlines why the topic is of general interest.
• describe clearly the work that has been performed in the project.
• link the theory of multimodal human-computer interaction to the work performed.
• include a Bibliography of no less than 5 relevant scientific citations.
Assignments

1. Summarize and review a recent scientific paper (2-3 students/group)
   Deadline: November 7

2. Project pre-study (3 students/group)
   Deadline: November 21

3. Personal assessment and reflection
   Deadline: January 5
Assignment 1

- In what way is the interface innovative compared to a similar traditional interface for the same type of task?
- Are there any other similar innovative interfaces that this work could be compared with?
- Is the proposed interface suitable/optimal for the given task? What are the strengths and weaknesses?
- Has the interface been properly evaluated in experiments (are the methods and results that are presented sound and interesting)? Why or why not?
- Is there a commercial, industrial, scientific or entertainment potential in the interface/application? Would you like to use it yourself?
- How could the work be improved or continued?
In order to pass, the review should:

• be submitted on time: Not later than **November 7**
• contain a summary (of about 200 words) that is significantly different from the original abstract.
• contain a review of at least 3 pages (excluding the summary) with 12 pt text showing that the students have considered the method and results critically.
• demonstrate that the students make adequate use of the theory of the course when reviewing the paper.
Assignment 2

You should write a description and specification for your project, consisting of the following headings:

1. Title of the project
2. Supervisor
3. Background
4. Aims and delimitations
5. Set up
   - If applicable with an illustrating Figure
   - Describe the Human-Computer Interface and interaction
   - The different technological components of your system and how they are communicating etc.
6. Evaluation scheme
7. Suggested project plan
8. Responsibilities within the project group
9. Time plan
10. Risk analysis
11. Related work
Assignment 2: requirements

In order to pass, the specification should:

• be submitted on time: Not later than November 21st.

• contain the 11 parts outlined above.

• be at least 4 pages with 12 pt text, with the emphasis on parts 3-6 and 9-11.
Assignment 3

You should write a personal assessment and self-reflection answering the following questions:

• What and how did you contribute to the work?
• Was this in accordance with "Responsibilities within the project group" of the project plan? If not, why?
• Are you satisfied with your own contribution? Why or why not? What are you most (least) satisfied with?
• What was the major learning outcome for you of the project work? What did you learn? Why?
Assignment 3: requirements

In order to pass, the Assignment 3 should:

• be submitted on time: Not later than January 5th.

• contain justified answers to the four questions above.

• be 1 page with 11-12 pt text (not more or less).
Seminars

1. Discuss scientific reviews
   November 30

2. Project presentations
   December 13
Visits

- Simulation Center, Karolinska Sjukhuset
  November 26 (9-12)
  One hour per group, fill in the doodle on the website

- Tobii Eye Tracking
  November 16 (15-17)
Requirements & Grades

• Required
  – The 4 Laboratory exercises
  – The 2 Seminars
  – The 3 Assignments
  – The Project (report)

• Grades
  – Grades from A-F on the Assignments and the Project report will be used to assign the final grade, with a higher weight for the project.
Course literature

• Article compilation
  – About Human-Computer Interfaces
  – Application-oriented
  – Available on bilda.kth.se in pdf
• Find your own...

Recommended further reading:
• Shneiderman & Plaisant: Designing the User interface
• Maragos, Potaminaons, Gros: Multimodal Processing and Interaction
Teachers

**Five teachers from CSC:**

- **Speech interfaces:** Gabriel Skantze
- **Visual interfaces & augmented reality:** Alex Olwal
- **Gesture-based interfaces:** Anders Friberg
- **Sound in interaction:** Kjetil Falkenberg Hansen
- **Haptic interfaces:** Eva-Lotta Sallnäs
Bilda & Computer account

- We will use bilda.kth.se for submission and correction of the assignments and the project.
- You need a kth.se account or a special bilda account
- How many of you do not have a kth.se account?
- You will do computer exercises at CSC.
- You need a Windows account at nada.kth.se and you need an access card to the computer rooms.
  - Access card from Kortexpeditionen, Osquldas väg 6
  - Windows account from Delfi, Osquars backe 2
Speech project proposals
Program using available API:s

- **Microsoft ASR & TTS**
  - Available in English Windows Vista & 7
  - .NET (C#, VB, etc)
- **WAMI toolkit (MIT)**
  - Javascript
- **Nuance Café**
  - VoiceXML
- **CSLU toolkit**
- **FaceAPI (head tracking)**
  - C++
Proposal: Evaluation

Example: How good is Read Out Loud in Acrobat Reader?

Acrobat Reader now comes with a functionality that reads the text with (a quite low-quality) TTS. How useful is that? Test how much different listeners understand of different texts when they are read by the software.