Multimodal Interaction

Christoph Amma & Felix Putze
08.11./15.11.2013
„Design and Evaluation of Innovative User Interfaces“
WS 2013/14
Overview

• What is multimodal interaction?
• Modalities
• Cognitive foundations
• How do people interact with multimodal systems?
• How can multimodal systems reduce recognition errors?
• Design guidelines
• Multiparty multimodal interaction
• Example real world systems
Literature

- Sharon Oviatt „Ten Myths of Multimodal Interaction“, Communications of the ACM, 1999
Motivation

- Few input/output possibilities in current WIMP interfaces
  - WIMP = window, icon, menu, pointing device
  - 2 input devices: keyboard, mouse (based on manual manipulation)
  - Input devices are normally not used in parallel
    → unintuitive, inefficient

- How does Human – Human Interaction work?
  - The mix of audio-visual signals is essential in communication
  - Haptics are important for interaction
  - All senses (including smell, taste) are relevant

- What is the idea of multimodal Interfaces?
  - Human – Human Interaction as Prototype for MMI
  - Use all five senses for interaction with humans
Design and Evaluation of Innovative User Interfaces

Put-That-There


2. Input to system is composed of
   - Gestures
   - Speech

3. Processing
   - Semantics via speech
   - Deictic terms resolved by coordinates of pointing gestures
   - Put that there → [Put] (32,53) (87,12)

4. Multimodality: The combination of speech and gestures offers new functionality!
Put-That-There

- Main paradigm:
  - Issue commands by speech
  - Give spatial information by gestures
  - Overlay pointing gesture with spoken deictic term

- Assumptions
  - Point & Speak is a „good“ task for multimodal interaction
  - Deictic term and gesture happen at the same time

- Critic
  - Very restricted grammar of actions (terms, time alignment)
  - No Context of Interaction (think of error resolution: „No, put it there“)
  - Deictic terms must be spoken in order to get coordinates
**Definition of Modalities**

- **Modality**: Mode or path of communication according to
  - Human senses, not a signal
  - Different types of information
  - Different types of interface devices

- **Human Senses**:
  - Sight, Touch, Hearing, Smell, Taste

<table>
<thead>
<tr>
<th>Description</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>Acoustic, (EMG, Vision, Ultrasound)</td>
</tr>
<tr>
<td>Motion, Pose</td>
<td>Vision, Acceleration, EMG</td>
</tr>
<tr>
<td>Haptics</td>
<td>Pressure Sensors, Force Sensors</td>
</tr>
<tr>
<td>Gaze</td>
<td>Vision, EOG</td>
</tr>
<tr>
<td>Inner States</td>
<td>EEG, EDA, Vision, Acoustic</td>
</tr>
</tbody>
</table>
Definition of MM-Interfaces

- There is no such thing as THE Definition:

- A multimodal HCI system is simply one that responds to inputs in more than one modality or communication channel (e.g. speech, gesture, writing and others) [James/Sebe]

- Multimodal interfaces process two or more combined user input modes (such as speech, pen, touch, manual gesture, gaze and head and body movements) in a coordinated manner with multimedia system output. [Oviatt]
How modalities differ: Speech vs. Gestures

Speech:
- Expressiveness
  - General purpose
  - Imprecise in spatial and action description domain

- Language dependent

- Path of Communication
  - Generated by vocal tract
  - Perceived by acoustics

Gestures:
- Expressiveness
  - Spatial content
  - Limited throughput

- Multi-cultural (to some extent)

- Path of Communication
  - Generated by movement
  - Perceived by sight
Speech

- Speech is the dominant modality in human communication
- Capturing by microphones
  - Close Speak
  - Far Field
- Automatic Speech Recognition (Speech-to-Text)

hello my name is christoph

- No semantic representation at this level
- High variability to express semantically equivalent meanings
Progress in Speech Recognition

Since 2002: Multilinguality (triggered by 9/11)
Information in Speech

Speech Recognition

Language Recognition

Speaker Recognition

Emotion Recognition

Accent Recognition

Words
Onune baksana be adam!

Language Name
Turkish

Speaker Name
Umut

Emotion
Angry

Accent
Istanbul

Topic ID:
Chemicals

Entity Tracking:
Istanbul

Acoustic Scene:
Bus Station

Discourse Analysis:
Negotiation

Motion and Posture

- Humans use their body as communication modality
  - Gestures (explicit and implicit)
  - Body Language
  - Focus of Attention
  - Activity

- Perception by Computers follows two main paradigms
  - Computer Vision (external)
  - Body-mounted Sensors

Smart Control Room Fraunhofer IOSB

Xsens Maven
### Vision vs. Body Mounted

<table>
<thead>
<tr>
<th>Vision (external)</th>
<th>Body mounted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensors:</strong></td>
<td><strong>Sensors:</strong></td>
</tr>
<tr>
<td>• One or more cameras</td>
<td>• Accelerometers, Gyroscopes, Magnetometer, (cameras)</td>
</tr>
<tr>
<td><strong>Pro:</strong></td>
<td><strong>Pro:</strong></td>
</tr>
<tr>
<td>• Unobstrusive</td>
<td>• Mobile</td>
</tr>
<tr>
<td>• No devices on users needed</td>
<td>• Relatively independent from environment</td>
</tr>
<tr>
<td>• Equivalent to human sight</td>
<td>• Direct motion data</td>
</tr>
<tr>
<td><strong>Cons:</strong></td>
<td><strong>Cons:</strong></td>
</tr>
<tr>
<td>• Environment dependent</td>
<td>• User needs to carry special devices</td>
</tr>
<tr>
<td>• Overlap of objects</td>
<td>• Less general data</td>
</tr>
<tr>
<td>• Stationary</td>
<td></td>
</tr>
</tbody>
</table>
Haptics

- Manipulation tasks in the real world require feeling of objects and dynamics
- Computers can perceive and simulate this by
  - Haptic interfaces (perception/simulation)
  - Tangible objects (perception/simulation)
  - Force sensors (perception)
Biophysiological Modalities

- Humans emit even more information through their body
  - Brain activity
  - Skin conductance
  - Temperature
  - Heart Rate
- Can be perceived by measuring electric signals on the body
  - EEG
  - EDA
- Reveal information on users cognitive and emotional state
  - Workload
  - Emotional state
  - Mood
  - Fatigue

→ Context information for multi-modal interfaces
Types of Multimodal Interfaces

- Perceptual, Interactive
  - Highly interactive
  - Rich, natural, effective interaction

- Attentive
  - Context-aware
  - Implicit

- Enactive
  - Communicate information that relies on active manipulation through the use of hands or body
  - e. g. tangible interfaces
Cognitive Foundations

- How do humans deal with perception and interaction?
- Processing takes place in “working memory”
- Modalities can be processed partially independently

→ Working memory capacity is increased by presenting information in different modalities
Working Memory

- The Term „Working Memory“ refers to a theoretical construct from cognitive psychology
- Working Memory – Active momentary memory
  - Mental representation of environment
  - Goal oriented active monitoring of information or behavior
- Model of Baddeley/Hitch:

![Diagram of Working Memory model](image)
- Central Executive: supervisory system controlling the flow of information from and to the its slave systems
- Phonological Loop: Deals with sound, language or phonological information
- Visuo-spatial Sketchpad: Visual Information and planning of spatial movement
- Episodic Buffer: linking of information across domains (audio, visual) with time sequencing
Working Memory and MMI

- Example: Reducing Cognitive Load by Mixing Auditory and Visual Presentation Modes [Mousavi et al. 95]
  - Geometry task with additional given symbolic information
  - Split Attention Effect (visual/visual) increases cognitive load
  - Humans can process audio/visual information in parallel and independently to a certain extent

→ Multimodal Interfaces can make use of this effect
Innovation Cycle

- Remember the innovation cycle
  - Formative Evaluation
    - Identify the basic usage patterns of multimodal interaction
  - Design and System implementation
    - What components do we need in a multimodal system
  - Summative Evaluation
System Design Overview

- **Speech**
- **Gesture**
- **Input Modality Fusion**
- **Output Modality Fission**

**Strategy**

- **Synchronization**
  - Preprocessing
  - Feature Extraction
  - Recognition

- **Synthesizing**
  - Synchronization

- **Fusion**

- **Fission**

**Speech**

**Visual**
Formative Evaluation

- Do we need a formative evaluation?
- Do people really want to interact multimodally with machines?
- Will they prefer it?
- Can we simply offer more modalities and they will be used?
- What kind of information will be expressed by which modality?
- Can we simply plug our unimodal systems together?

→ We need to perform user studies to find out
The 10 Myths of MMI

1. If you build a multimodal system, users will interact multimodally
2. Speech and pointing is the dominant multimodal integration pattern
3. Multimodal input involves simultaneous signals
4. Speech is the primary input mode in any multimodal system that includes it
5. Multimodal language does not differ linguistically from unimodal language
6. Multimodal integration involves redundancy of content between modes
7. Individual error-prone recognition technologies combine multimodally to produce even greater unreliability
8. All users’ multimodal commands are integrated in a uniform way
9. Different input modes are capable of transmitting comparable content
10. Enhanced efficiency is the main advantage of multimodal systems
Do we act multimodally?

- Do users interact multimodally?
- Do they prefer it?
- Is it more efficient?
- Do users make less errors?

When we look at interaction as a series of commands:
- Do users use multimodal commands?
- Do they need less time to issue commands?
- How often do users make errors (construct wrong commands)?
Study Design:

- Interactive map table with pen and speech input
- People were free in their use of speech and pen input
- Tasks performed during experiment
  - Real estate selection
  - Sim City like updating of urban area

The original study of Oviatt did not use Microsoft Surface
User Study

- Wizard-of-Oz study:
  - People were told, they use a well developed computer system
  - System was simulated in parts by human operators
  - Questionnaire revealed that people believed it
  - Debriefing of test persons after study

- 18 subjects

- Factorial Design
  - Multimodal (pen + speech), unimodal pen, unimodal speech
  - Map structure: high, low

- 2 tasks for each combination
- 12 tasks per subject → 216 recorded tasks
Input Modalities

- Pen, unconstrained usage, for example
  - Cursive handwriting or printing
  - Gestures
  - Symbols
  - Pointing

- Speech
  - Natural language

- Practice:
  - Users completed tasks by using only one modality to show their functional equivalence
  - Users completed tasks by choosing freely between modalities
  - Users were encouraged to speak and write naturally
Analysis

- Statistical measures computed from data
  - Length of utterance
  - Disfluencies and self-corrections
  - Task completion time
  - Self reported and observed preferences
- Statistical tests performed to verify significance of results
Results

- Clear performance advantage of multimodal interaction in spatial domains
  - Faster
  - Less error prone (“add a hotel east, uhm west of the station”)
  - Less complex linguistic expressions $\rightarrow$ easier automatic recognition

- In numbers
  - 10% faster task completion
  - 36% fewer task-critical content errors
  - 50% fewer spontaneous disfluencies
  - 95-100% preference for multimodal interaction

$\rightarrow$ These numbers refer to spatial domains
Integration and Synchronization

- How do humans integrate and synchronize different modalities?
- A more detailed view on the data of the tabletop study
- Only the 72 multimodal tasks are considered
- Goals:
  - When do users use multimodal commands and when unimodal ones?
  - Are there differences in linguistic features of MMI in contrast to unimodal interaction
  - How are spoken and written modes integrated?
  - How are they synchronized?
Multimodal vs Unimodal

- How often did users use multimodal commands?
  - 19% of all commands were expressed multimodally
  - 17.5% unimodal with pen
  - 63.5% unimodal with speech

- User commands were classified in categories:
  - Spatial location commands: add, move, modify, ...
  - Selection commands: zoom, label, delete, query
  - General action commands: constraints, overlay, locate, print, scroll, ...

- Selection commands were most likely unimodally (see next slide), because object was already in focus from:
  - Previous dialogue context (e.g. user adds object and then deletes it)
  - Visual context (e.g. user zooms on a house, then queries it)
Multimodal vs Unimodal

- Percentage of commands that users expresses multimodally

> Multimodal input is used, but only for certain types of commands

From Oviatt 1999, Ten Myths of Multimodal Interaction
Does multimodality influence individual modalities?
For example, do we talk differently when interacting multimodally?

In unimodal speech:
- Sentences followed S-V-O order
- Location part almost always in the end
- Example: “I want to see the photo of the house on the southwest end of Reward Lake, please”

In multimodal speech:
- Telegraphic command language V-O
- Location part in form of gesture normally in the beginning
- Example: [circles house] “Show photo”
Multimodal Integration Patterns

• How were the different modalities integrated to commands?
The multimodal commands consisted of

• Point & Speak  14%
• Draw & Speak  86%

→ Relying on point & speak (put-that-there) is an artificial limitation

• Deictic Terms (there, here, that, ...)
  • Often not even articulated
  • Overlap of gesture and speech in only 25%

→ The simple put-that-there processing is not adequate
Multimodal Integration Patterns

• Synchrony of multimodal commands
  • Sequential integration in more than 50%
  • Pen input precedes speech input

→ System Designers need to take this into account
  → We cannot rely on simultaneous signals
More Myths of MM Interaction

1. If you build a multimodal system, users will interact multimodally
2. Speech and pointing is the dominant multimodal integration pattern
3. Multimodal input involves simultaneous signals
4. Speech is the primary input mode in any multimodal system that includes it
5. Multimodal language does not differ linguistically from unimodal language
6. Multimodal integration involves redundancy of content between modes
7. Individual error-prone recognition technologies combine multimodally to produce even greater unreliability
8. All users’ multimodal commands are integrated in a uniform way
9. Different input modes are capable of transmitting comparable content
10. Enhanced efficiency is the main advantage of multimodal systems
Redundancy Of Content

- Do humans use multiple modalities to express the same thing?
- Spoken and gestural modes contributed complementary information
  - Subject, Verb and Object of an utterance were spoken
  - Locative information were likely to be written/gestured
- Contrastive functionality of modes
  - Information was rarely duplicated in both modes

→ If we offer more modalities, semantics will not be duplicated
→ We need to correctly process all modalities to avoid loosing information
→ We cannot simply use both modes to improve our unimodal recognizer
Interpersonal Differences

- Do all users integrate multimodal input the same way?
- We look at the simultaneous vs. sequential integration pattern (speech/gesture) for different users

Different users used different patterns, but were consistent during the session

⇒ The system should try to adapt to the users pattern
Recognition Errors

- Multimodal user interfaces typically contain at least one “recognizer” module for each modality, for example:
  - Speech-to-text system
  - Gesture recognition
  - Gaze direction recognition

- Recognizers make errors! Why?
  - Ambiguities humans resolve by semantics
  - Indefinite signals (even humans can not interpret everything right)
  - Environmental influences
Problems with real world scenarios

• In field environments typically involve
  • Variable noise levels in the signal (acoustic noise, lighting, ...)
  • Adaptation of users to noise
  • Social interchange
  • Multitasking and interruption of tasks
  • Increased cognitive load
  • Human performance errors

→ Big differences to in lab training data
  • In case of speech, speakers change their speaking style in real world:
    – Miscommunication → hyperarticulation
    – Emotional state in interpersonal tasks
    – Surrounding noise

→ Training data does not fit anymore
→ This happens in short periods of time, making it even harder

→ Drop of recognition rate
Dealing with errors

- Can we build multimodal system, which have better recognition rates than their unimodal parts?

- Two strategies
  - User centered error resolution
  - Architecture based error resolution
User centered error resolution

Can multimodal systems change the behavior of users, so their input gets less error prone?

- Users typically choose the appropriate less error prone modality for the specific task by themselves

- Multimodal Interaction can reduce complexity of speech
  - if another modality suited for the task is at hand
  - For example gestures in spatial domains

- In case of errors, users try to resolve it by using different modalities
  - It has been shown, that this is faster and more reliable than for example respeaking
Architectural based error resolution

- Mutual Disambiguation
  - Do not take the best result of each recognizer
  - Look at the multimodal result tuples and choose most probable tuple
  - Example (Speech and Gesture)

\[
\text{N-Best list speech} \\
1. Zoom out \\
2. Show all \\
3. Show info
\]

\[
\text{N-Best list gesture} \\
1. Arrow \\
2. Line \\
3. Circle
\]

Choose 3\textsuperscript{rd} entry in both cases because only this combination has high probability: [Circle] Show info
Design Guidelines

- Can we extract guidelines from the results of the studies?
  1. Users do like to interact multimodally with artificial systems
  2. Multimodal Interaction is preferred in spatial domains
     → We should offer & expect more than Point & Speak
  3. Multimodal Interaction changes the way we talk
     → We need to adapt our speech processing components
  4. Speech and Gesture are synchronized but not simultaneous
     → Be prepared for differently synchronized integrations
  5. We cannot assume redundancy of content
     → We need to process modalities in an integrated manner
  6. Use multimodality for better system error characteristics
     - Through expecting simplified speech
     - Through fusion of modalities to clear out uncertainty
     - Through offering the right modality for the right task
Exam dates

• We now have two full days for exams:
  • 27.02.2014
  • 04.04.2014

• Choose your preferred one and contact our secretary Frau Scherer
  • helga.scherer@kit.edu

• First come, first serve!!
Human Discourse

- Until now, we focused on humans using multimodal interfaces.
- What about interfaces that act like other humans by resembling human conversational patterns?
What properties must such interfaces have?

- They would have to respect the rules of human conversation
  - Contextual understanding
  - Sentence constructions
  - Body-language, mimics, turn taking behavior

→ We gently leave out the problem of semantics here

- Turn taking and floor control
  - Humans do not talk all at the same time (at least most of the time)
  - Normally no explicit coordination is necessary
  - We align our utterances in a way, that allows fluent conversation

→ Can our interfaces do that too?
Floor and Turn in Discourse

Sacks et al., *A Simplest Systematics for the Organization of Turn-Taking for Conversation*, Language, Vol. 50:4, 1974

- Investigates the role of speech in turn taking behavior
- Turn taking in conversations can be described by
  - Turn-Constructional Unit (TCU)
    - Complete semantic entity
    - May be a complete sentence, a phrase, or just a word
    \(\rightarrow\) Completion results in a *transition relevance place (TRP)*
  - Transition Relevance Place
    - Possible point in time, where floor control changes
  - Turn-Allocation Component
    - Current speaker selects next speaker
    - Another speaker grabs the floor (self-selection)
Example Conversation

Seen death of a salesman last night?

Never seen it?

No

(pause)

No

Ever seen it?

Yes, watched it some years ago.

A horrible piece of art

Turn Allocation Component:

- current speaker selects next
- self select
Turn Management Rules

- **Transitions**
  - Commonly no or only slight gap or overlap
  - Rarely huge gap or overlap

- **Order of turns varies**
  - Bias towards speaker, that had the floor just before: A → B → A

- **Selection of next speaker at transition relevance points**
  - Speaker stays the same by keeping the turn
  - Speaker selects next speaker by himself
  - Speaker ends his turn without assigning a new speaker
  - Other speaker self selects next turn
Current selects next speaker

- Selection of next speaker by current floor controller
  - Question/Answer
  - Complaint/Denial
  - Compliment/Rejection
  - Challenge/Response
  - Request/Grant

- Commonly contains address part to select next speaker, e.g.
  - Name
  - Gaze

- Every Turn-Constructional-Unit may be transformed to selecting a speaker
  - „You know?“, „Don’t you agree?“
  - Also used as exit technique from current turn
Self Selection

• Next speaker self selects the floor and grabs it

• In case of open transition relevance point
  • First starter gets the floor

• In case of an ending turn constructional unit
  • Next floor is taken over by starting utterance before TCU has ended

• TCUs can also be interrupted in the middle

• Commonly started by
  • well, but, and, so, ...
  • Allow fast floor grabbing
Multimodal Floor Control

- After the first phone conference with 4 participants, you know: *Turn taking is a multimodal mechanism*

- **Gaze**
  - „Mutual Gaze Break“ pattern: mutual gaze between consecutive floor holders A and B, broken by next floor holder B

- **Gestures**
  - Gestures are used to hold floor during silence
  - Gestures indicate attempt to grab floor during turn of prev. holder
  - Deictic gestures assign floor to next holder

- **Prosodic Features**
  - Silence yields floor
  - Decrease of loudness yields floor
  - Lengthening of final syllable marks end of floor
  - Rise of intonation is a feature for floor grabbing
Floor Control for Multimodal Systems

- If we want systems, which are capable to communicate naturally with humans, these systems must
  - Know, when to express information in a discourse
  - Know, how to express, they are finished
  - Decide what to do, when somebody else interrupts

→ Multimodal systems must manage floor control
  - Recognize or predict transition points
  - Interpret the meaning of interruptions
  - Give up control when necessary

→ We need real data for the formative and summative evaluation and for the training of the system
If we want to develop and evaluate algorithms we need real world example data

A data corpus with human discourse behavior is needed

In this case: synchronized audio/video recordings of meetings

Raw data must be annotated either manual or automatically:
  • Speech to text
  • Gestures
  • Gaze
  • Communication floors

VICE Corpus, see Chen et al. *A Multimodal Analysis of Floor Control in Meetings*, LNCS 4299/2006, 2006
Speech Annotation

- Word transcriptions by human listener
- Synchronization to audio with ASR (Automatic-Speech-Recognition)
- Sentence Unit annotation (SU expresses a speaker’s complete thought or idea) and could be a:
  - Statements
  - Questions
  - Backchannel
  - Incomplete
Gesture Annotation

- Gestures were categorized into types:
  - Metaphoric – representing abstract idea
  - Iconic – representing a concrete idea
  - Emblematic – symbol with semantic meaning without parallel speech
  - Deictic – pointing
  - Beat – rhythmic up/down movements in sync with speech

- Excluded from the annotations were gestures of types:
  - Fidgeting movements: tapping, touching clothe, ...
  - Instrumental movements: holding a cup, sorting paper, ...
Gaze Annotation

- **Time annotation**
  - Major saccades: Periods of no specific fixation during fixations

- **Spatial annotation by object of fixation**
  - Other people
  - Non-human entities (board, paper, ...)
  - Personal objects
  - Neutral space
Floor Control Annotations

- Annotation of floor events
  - Control – Main communication stream
  - Sidebar – Concerning subfloors that have split from main floor
  - Backchannel – Confirmation of floor controller (e.g. “yeah”)
  - Challenge – Attempt to gain floor control
  - Cooperative – Like backchannel but with propositional content
  - Other – e.g. self talk

- Enforce consistency with sentence units (SU’s)
Transition Annotation

- The shift of communication floor control might happen according to the following types of transitions
  - Keep – current holder keeps the floor
  - Change – clear transition with gap in between
  - Overlap – clear transition with overlap
  - Stop – holder gives up → floor open to all
  - Self select – not yielded by prev. owner → floor overtake by new owner

- Occurrence frequencies of transition types are highly dependent on type, content and participants of conversations
  - Hierarchy of participants
  - Number of different positions
  - Future impact of conversation
Analysis: Speech

• Backchannel SU’s range between 25-30%
  → Frequent pattern, might also be expected by humans in MMI
  → MM-Interface must recognize and eventually generate backchannel information

• Discourse Markers (DM)
  • A DM is a word or phrase that functions as structuring unit in speech
  • Example: actually, now, see, basically, I mean, well, you know
  • Often used to mark a boundary in discourse
  • Frequently used in challenges and floor beginnings in the study
Analysis: Gaze and Gesture

• Previous floor holder gazing at next floor holder
  • In general, this assumptions holds
  • Exceptions: Visual focus of participants on other objects (paper, board)

• Mutual gaze break
  • Frequent pattern with same exceptions than above

• Floor giving gestures
  • Frequent deictic pointing on next floor holder

• Floor taking gestures
  • Often incorporates holding the hand like in school
  • Objects (pen, ...) are used to underpin floor grab attempt
So how can we build a system that is capable of detecting floor control mechanisms?
   • When does control of floor shift?
     → When does it seem to be a good point in time to grab the floor?

And if we have built one, how can we evaluate it?
   • How can we measure the technical quality of the system?
   • How can we tell if it is better than another system?
   • How can we tell if it is better than a trivial system?
Lei Chen, Mary P. Harper, “Multimodal Floor Control Shift Detection”, ICMI 2009, Best Paper Award

- Try to detect floor shift after each Sentence Unit (SU)
  \[ \rightarrow \] 2-class classification problem \{keep, shift\}

- Operating on the introduced meeting data corpus
  - 3170 SUs
  - 1301 Keep, 688 Change, 1181 None (backchannel, ...)

- We want an on-line recognition system, that means
  - Can be used in real time scenarios \[ \rightarrow \] performance is essential
  - Classification only on past data
Feature Extraction

- Based on analysis of conversations and floor control mechanisms we need to define features, which our system will operate on.

- Lexical Features
  - Word n-grams: given word token $w_i$ at position $i$, take:
    
    \[
    \{w_{i-2}, w_{i-1}, w_i\} \quad \{w_{i-1}, w_i\} \quad \{w_i\}
    \]

  - POS n-grams: same as above with part of speech instead of words

- Prosodic Features
  - Duration
  - Energy
  - Other’s speaker overlap
  - $F_0$ related
Feature Extraction

- Gesture Features
  - Holder’s gesturing time
  - Other’s gesturing time

- Gaze Features
  - Holders Gaze Time – gazing at other participants
  - Others Gaze Time – others gazing at floor holder
  - Number of mutual gazes
Feature Extraction

- Consider only a time window before current point in time
  - This is essential to do runon recognition

- Compute the features on this window
  - Lexical features
  - Prosodic features
  - Gesture features
  - Gaze features
Statistical Modeling

- Our statistical model can compute

\[ P(E | F) \]

\[ E \in \{ \text{keep, change} \}, F \text{ denotes the feature vector} \]

- We then simply classify by computing

\[ \hat{E} = \arg\max_E P(E | F) \]

- Parameters of the model must be trained to fit the data
Data Setup

• We need the data we have to train, optimize and evaluate our model

• Why not just use all of our data for each of these task
  ➔ We would overfit our models

• Imagine the following gesture recognizer:

Model parameters might represent exactly the 10 samples leading to perfect recognition results on the same data.
Data Setup

- Use disjoint sets for training, optimization and evaluation

Training data (Train Set) → Optimization data (Development Set) → Evaluation data (Test Set)

- Optimize Parameters of model, avoids overfitting
- Independent test set to avoid bias on dev set

→ Gives good estimate on performance, still there is a bias towards experiment conditions
We need metrics to quantify the recognizer performance

- **Recognition Rate**
  \[
  REC = \frac{\text{#correct classified samples}}{\text{#samples in total}}
  \]

- **Classification Error Rate**
  \[
  CER = \frac{\text{#incorrect classified samples}}{\text{#samples in total}}
  \]

- **Error Rate for Floor Shift Detection**
  \[
  ERR = INS + DEL
  \]
  
  \[
  INS = \frac{\text{#incorrect insertions of change events}}{\text{#total of change events}}
  \]

  \[
  DEL = \frac{\text{#incorrect deletions of change events}}{\text{#total of change events}}
  \]

- Can be \(\geq 100\%\)

<table>
<thead>
<tr>
<th>Reference:</th>
<th>SU1</th>
<th>SU2</th>
<th>SU3 / SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>System:</td>
<td>SU1 / SU2</td>
<td>SU3</td>
<td>SU4</td>
</tr>
<tr>
<td></td>
<td>INS</td>
<td>DEL</td>
<td></td>
</tr>
</tbody>
</table>
Floor Control Evaluation

- Usage of CER and the self-defined metric ERR

- We need to compare the results to something

- A reference system is called a baseline system
  - Usually an old system
  - If no old system exists, some imaginary classifier ("always A", "chance")

- Baseline System
  - Always hypothesize a “keep” at every SU ending
    - Results in 100% ERR (corpus independent)
    - Results in 34,5% CER (corpus dependent)
Results

• Unimodal vs. Multimodal Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>DEL (%)</th>
<th>INS (%)</th>
<th>ERR (%)</th>
<th>CER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>100.00</td>
<td>0.00</td>
<td>100.00</td>
<td>34.48</td>
</tr>
<tr>
<td>ME Speech</td>
<td>58.37</td>
<td>26.70</td>
<td>85.07</td>
<td>29.35</td>
</tr>
<tr>
<td>ME Visual</td>
<td>67.50</td>
<td>21.72</td>
<td>89.22</td>
<td>30.78</td>
</tr>
<tr>
<td>ME Multimodal</td>
<td>54.73</td>
<td>26.04</td>
<td>80.77</td>
<td>27.86</td>
</tr>
<tr>
<td>CRF Speech</td>
<td>63.85</td>
<td>18.41</td>
<td>82.26</td>
<td>28.38</td>
</tr>
<tr>
<td>CRF Visual</td>
<td>67.99</td>
<td>20.07</td>
<td>88.06</td>
<td>30.38</td>
</tr>
<tr>
<td>CRF Multimodal</td>
<td>56.55</td>
<td>19.90</td>
<td>76.45</td>
<td>26.37</td>
</tr>
<tr>
<td>AdaBoost Speech</td>
<td>51.91</td>
<td>27.53</td>
<td>79.44</td>
<td>27.40</td>
</tr>
<tr>
<td>AdaBoost Visual</td>
<td>66.34</td>
<td>19.57</td>
<td>85.90</td>
<td>29.63</td>
</tr>
<tr>
<td>AdaBoost Multimodal</td>
<td>50.58</td>
<td>23.22</td>
<td>73.80</td>
<td>25.46</td>
</tr>
</tbody>
</table>

• Classifier Combination

→ Multimodality gives significant better performance
→ Performance in general is still rather low, but
→ We have no idea, which performance would be good
Results

• So what do we learn from this study?
• Recognizing floor control mechanisms is an inherently multimodal task
• It is possible to predict floor shifts, but it is a hard task

• What to do, if we reliably could? Our systems could
  • choose the right moment to output information
  • Indicate their output request in human like form
  • Know if another person interrupts or uses the backchannel

• Open Questions
  • No consideration of semantic information. How to include semantics?
  • We need to reliably recognize speech, gesture, gaze and SU’s
  • Will people accept such behavior from systems?
Example System

Bohus, Horvitz, *Facilitating Multiparty Dialog with Gaze, Gesture, and Speech*, ICMI2010

- Virtual agent engages in game with multiple persons
- Model of turn-taking mechanisms
- Handcrafted models for feature classification
- Features include
  - Speech
  - Gaze
  - Gestures
Example System: Architecture

- **Architecture**
  - Sensing: tracking conversational dynamics including
    - Source and target of utterances
    - Floor state, actions and intentions
  - Turn taking decisions: Decide
    - When to contribute
    - What Floor action to perform
  - Dialog Management: Generate appropriate semantic output to the given semantic situation
  - Behavior Control: Render floor management actions into a stream of verbal and non-verbal behaviors with the content provided by the dialog manager
Example System: Behavior

- **Floor Actions**
  - Hold
  - Release
  - Take
  - Null

- **Evaluation**
  - 60 participants interacted with the system
  - Different group sizes
  - No control group, no factorial design! But
  - Computation of statistical measures (turn shaping scores)
  - Questionnaire

  $\rightarrow$ Users rated turn-taking abilities favorably
  $\rightarrow$ E.g. 80% of release actions resulted in turn take of the addressee
A final view on MMI

- Multimodal Interaction is a large research area
- We covered two main topics
  - Foundations of multimodal MMI with a focus on speech and gesture
  - Turn taking mechanisms in multiparty discourse
- Key concepts to remember
  - We cannot always believe our intuition on how interaction will function
    - We need to find out by performing well designed user studies
  - We need to take a close look at how human – human communication and interaction works before we can build systems that resemble this behavior
    - Costly collection and annotation of real world data is necessary