Multimodal Dialogue Systems

Lecturer: Homeira Quraischy
Advisor: Martin Rumpler
Chair: Saied Tehrani
Outline

- Introduction
- Needs for multimodality
- Multimodal Dialogue Systems
  - Multimodal fusion
  - QuickSet and SmartKom
- Summary & Outlook
Introduction

Most dialogue systems can either be

- Text based
- Speech based

Multimodal
Needs for multimodality - motivation

- Omnipresence of information
  - Internet
  - E-Mail
  - Multimedia
- Mobile access to these information
- Powerful devices
  - SmartPhones
  - PDA
  - Mobile phones
- Restricted user interfaces
  - Keyboard
  - Touch
Needs for multimodality - what

- Modalities refer to human senses
  - Vision
  - Audition
  - Touch
  - Smell
  - Taste

"Multimodal systems process two or more combined user input modes such as speech, pen, touch, manual gestures, gaze, and head and body movement – in a coordinated manner with multimedia system output." S. Oviatt

- Modeling interpersonal communication
- User-centric
Needs for Multimodality - why

- Advances of multimodal systems
  - Flexible
    *Choice of different modalities*
  - Adaptive
    *Environmental conditions*
  - Improved error handling
  - Supports user’s preferred interaction

- But: Challenges
  - Fusion/multimodal integration
  - Greater complexity of architecture
  - Many interdependencies with other technologies
Put-that-there

- One of the earliest multimodal systems ~1980
  - At MIT by Bolt
  - (limited) speech and gesture input

- Users can manipulate 2D-objects on a screen
  - Create, move, make etc.

- Words like „there“ and „this“ are like selected objects
  e.g. User: „Move this [+] to there [+]“

- Watchband at wrist as sensor
Put-that-there – Media room

Move that to there
Multimodal fusion

- Combination and integration of unimodal inputs
  - User’s intention
  - One single representation

- Synchronizes recognition and analysis components

- Goals
  - Reduce uncertainty
  - Mutual disambiguation

- Two types
  - Early fusion
  - Late fusion
Early & late fusion

- For signals that are highly dependent e.g. lip movement + speech
- Combines input signals at early stages

- Used for less coupled input modes e.g. speech + gesture
- Each input has its recognizer
- Input signals do not need to occur simultaneously
- Recognizers create time stamps for each input mode
QuickSet
QuickSet – System description

- Built in ~1997 at OGI in Oregon
- Agent-based (OAA) and collaborative multimodal system
- Developed for map-based applications and military trainings
- Scaleable from a hand-held to a wall-sized format
- Input modes
  - Speech
  - Pen-Gesture
- Use of late fusion
QuickSet - Szenario

- Users can draw
  - Points
  - Lines
  - Areas …

- *pen-drawing*
  User: *burn line*

- System draws commands
QuickSet Architecture
QuickSet – Pen-Gesture recognizer

Neural Network Model
- normalizes size
- centers in a 2D image

Hidden Markov Model
- smoothes
- re-samples

Estimations of both models compute probability for each possible gesture

Challenge: ambiguity
QuickSet – Multimodal integration

- Use of Typed Feature Structures (TFS)
- Command represented by constituents
- Three steps
  - Temporally
    - Time stamps
    - Integration within a short time window
    - Speech arrives after interval, gesture will interpreted unimodally
  - Statistically
    - Takes cross product of probabilities of modes and
    - Derives multimodal probability in final n-best list
  - Semantically
    - Use of TFS
QuickSet - Semantically integration

- **TFS:** Collection of feature-value pairs

- **Value of feature**
  - Feature structure
  - Attribut
  - Value

- **Unification over TFS**
  - Determines consistency
  - Combines to a single result
  - But rules out contradictory inputs

- **Composite structure is formed**

\[
\text{Attrib}_1: \  \text{Val}_1
\]
\[
\text{Attrib}_2: \  \text{Val}_2
\]
\[
\text{Attrib}_3: \quad \text{Attrib}_4: \  \text{Val}_4
\]

\[
\text{Type}_2
\]
QuickSet – Unification over TFS

User: Red barbed wire

Gesture Recognizer

Speech Recognizer

Result
SmartKom – System description

- German Federal of Education and Research (BMBF) Project
- Duration 1999 - 2003
- Consists of three scenarios
  - Mobile (Main scenario)
  - Public
  - Home&Office
- Face-to-face dialogue with Smartakus
- Interaction
  - Speech
  - Gesture
  - Facial expression
- Full symmetric multimodality
SmartKom - Architecture
SmartKom - Fusion

- **Gesture analyzer**
  sends ranked list of probably referenced objects

- **Language understanding**
  provides a set of different hypotheses

- **Modeler knowledge**
  provides information about current location

- **Discourse modeling**
  provides information about discourse context for undefined referring expression
SmartKom – Input synchronization (I)

- Mechanism for combining gesture with utterances
- Inputs obtain time stamps
- Gestures are combined with speech at a certain time span

**Speech input arrives**
- $t_{after} \ (1.5 \text{ sec})$ and $t_{before} \ (1.5 \text{ sec})$ speech
- If no gestures between $t_{start\_of\_speech} - t_{before}$ and $t_{end\_of\_speech} + t_{after}$

**speech-only input**

$\begin{align*}
& t_{start\_of\_speech} - t_{before} \\
& t_{before} \ (1.5 \text{ sec}) \\
& t_{start\_of\_speech} \\
& t_{end\_of\_speech} \\
& t_{end\_of\_speech} + t_{after} \\
& t_{after} \ (1.5 \text{ sec})
\end{align*}$
**SmartKom – Input synchronization (II)**

**Gesture input arrives**
- Processing time of speech recognizer to be considered
- Gesture will be stored

**Solution:**
- With timeout
  - Fusion gets message from speech recognizer when it's working; timeout will start
  - Speech recognizer sends result

**Drawbacks:**
- Update of recognizer would lead to errors
- Delays within the communication channels
SmartKom – Input synchronization (III)

- Without timeout
  - If no input, speech recognizer sends idle messages (every 500 ms)
  - Idle messages obtain time stamps, independent from delays
  - Fusion sends gestures satisfying \( t_{\text{end of gesture}} + t_{\text{before}} < t_{\text{idle}} \)

- Drawbacks:
  - Increased communication overhead
  - Increased latency
User: *I want to go from [ ] to [ ]*

Output of NLU comprises two **refProps** of type *LOCATION*

Gesture recognition sends recognized hypotheses

**Fusion**
- Integrates all possible hypotheses
- Sends a set of scored multimodal intention hypotheses
SmartKom – Encircling gesture

- Specialization of pointing gesture
- User selects with on continuous gesture
- Weighted lists of captured objects

Occured problems + solutions

P1: Number of selected seats doesn't correlate with spoken ones
S1: Mutual disambiguation by specifying desired number in utterance

P2: Selected seats are in different rows
S2: Fusion applies set of plausibility test as many seats as possible should belong to same row

I'd like these two seats.
Summary & Outlook

- New way of interaction

- Most multimodal dialogue systems have integrated speech + gesture

- Fusion and challenges

- Integration of other modalities:
  - Facial expression, prosody (e.g. for affect recognition)
  - Smell (e.g. smell-based interfaces)
Thank you for your attention !!!

Questions are welcome