Seminar: "Advances in Human Computer Interaction"

Topic:

W3C Multimodal Interaction Activity

Lecturer: Michael Kanonik
Advisor: Martin Rumpler
Chair: Wolfgang Burgard
Introduction

Multimodal interaction provides the user with multiple modes of interfacing with a system beyond the traditional keyboard and mouse input/output.

The most common interface combines:

- visual modality (e.g. a display, keyboard, and mouse) with
- voice modality (speech recognition for input, speech synthesis and recorded audio for output).
- pen-based
- haptic
Introduction

The **Multimodal Interaction Activity** is an initiative from [W3C](https://www.w3.org) aiming to provide standards (mostly [XML](https://www.w3.org/XML)) to support [Multimodal Interaction](https://multimodal.w3.org) scenarios on the [Web](https://www.w3.org).  

**What the Group doesn’t do:**
- work outside of the standardization process
- GUI interface usability
- tools for application development
Multimodal Interaction Framework

Identifies and relates markup languages used to describe information required by components and for data flowing among components.

Level of abstraction above an architecture.

- major components for every multimodal system.
- each component represents a set of related functions.
Multimodal Interaction Framework
Input Components
Input Components

*Recognition component*
Captures natural input from the user and translates the input into a form useful for later processing

*Interpretation component*
Further process the results of recognition components. Each interpretation component identifies the "semantics" intended by the user

*Integration component*
Combines the output from several interpretation components
Output Components

Figure 3: Output Components
Output Components

**Generation component**

The generation component determines which output mode or modes will be used for presenting information from the interaction manager to the user.

**Styling component**

This component adds information about how the information is "layed out."

**Rendering component**

The rendering component converts the information from the styling component into a format that is easily understood by the user.
Multimodal Interaction Framework

*Interaction manager*

Logical component that coordinates data and manages execution flow from various input and output modality component interface objects.

*Session component*

Provides an interface to the interaction manager to support state management, and temporary and persistent sessions for multimodal applications.

*System and Environment component*

Enables the interaction manager to find out about and respond to changes in device capabilities, user preferences and environmental conditions.
Use cases

A **use case** is a technique to capture the functional requirements of a system.

Use cases describe the interaction between a primary system actor—the initiator of the interaction—and the system itself, represented as a sequence of simple steps.

Devices that are used in various use cases can be classified from the point of view of “thickness.”
Use cases

**Thin client**
A device with little processing power or capabilities that can be used to capture user input (microphone, touch display, stylus, etc.) as well as nonuser input, such as GPS

**Thick client**
A device with powerful processing capabilities, such that most of the processing can occur locally. E.g.: PDA or notebook

**Medium client**
A device capable of input capture and some degree of interpretation; the processing is distributed in a client/server or a multidevice architecture
Use cases - example displayed map

User points to a position on a displayed map and speaks: "What is the name of this place?"

*Human user*
- Points to a position on a map and says, "What is the name of this place?"

*Speech recognition component*
- Recognizes the words

*Mouse recognition component*
- Recognizes the x-y coordinates of the position

*Speech Interpretation component*
- Converts the words into an internal notation.

*Pointing interpretation component —*
- Converts the x-y coordinates of the position into an internal notation.

*Integration component*
- Integrates the internal notation for the words "What is the name of this place?" with the internal notation for the x-y coordinates.
Use cases - example displayed map

User points to a position on a displayed map and speaks: "What is the name of this place?"

*Interaction manager component*

Stores the internal notation in the session object.
Request → a database request
Request → database management system
Returns value - “Bodensee”.
Response → internal notation in the session object

Sends response to the generation component.
Use cases - example displayed map

User points to a position on a displayed map and speaks: "What is the name of this place?"

**Generation component**
- Access the Environment component - voice and graphics modes are available.
- Result as 2 complementary modes: voice and graphics.
- Internal notation representing "Lake Bodensee" → voice styling component
- Internal notation representing the location on a map → graphics styling component.

**Voice styling component**
- Converts the internal notation representing "Lake Bodensee" into SSML.

**Graphics styling component**
- Converts the internal notation representing the "Lake Bodensee" location on a map into HTML notation.

**Voice rendering component:**
- Converts the SSML notation into acoustic voice for the user to hear.

**Graphics styling component:**
- Converts the HTML notation into visual graphics for the user to see.
Multimodal architectures (MA)

Properties for consistency with multimodal interaction framework:

1. CONTAINS A SUBSET OF THE COMPONENTS. MA contains 2 or more input & output modes

2. COMPONENTS MAY BE PARTITIONED AND COMBINED

3. COMPONENTS ARE ALLOCATED TO HARDWARE DEVICES Centralized, client-server or distributed MA

4. COMMUNICATION SYSTEMS ARE SPECIFIED

5. DIALOG MODEL IS SPECIFIED
Form fill-in user wants to make a flight reservation with his mobile device while he is on the way to work.

Figure 6. Form Fill-in Example
Multimodal architectures (MA) Example

Properties for consistency with multimodal interaction framework:

1. A SUBSET OF THE COMPONENTS
   The architecture contains a subset of the components of the W3C Multimodal Interface Framework, including GPS and Ink

2. COMPONENTS MAY BE PARTITIONED AND COMBINED
   Speech recognition component has been partitioned into two components, one which will be placed on the client and the other on the server

3. COMPONENTS ARE ALLOCATED TO HARDWARE DEVICES
   Components in pink are allocated to the client and the components in green are allocated to the server.

4. COMMUNICATION SYSTEMS ARE SPECIFIED
   Communications system is SIP, session initiation protocol

5. DIALOG MODEL IS SPECIFIED
   User provides input to fields such as "Travel Origin:", "Travel Destination:", "Leaving on date", "Returning on date"
Representing Users' Intent - EMMA

Standardization becomes more important as the users' inputs and their representations increase in complexity.

It is important for multimodal applications to have a common format for representing results.

XML specification- Extensible MultiModal Annotation (EMMA) for representing the intentions reflected in a user's input.

EMMA document can be considered to hold three types of data:

1. Instance data
2. Data model
3. Annotations
Extensible MultiModal Annotation

• **Instance data**
  Representation of the user's input.
  Application-specific markup corresponding to input information.
  Meaningful to the consumer of an EMMA document.
  Instances - application-specific and built by input processors at runtime.

  EMMA document may hold more than one instance.

• **Data model**
  Constraints on structure and content of an instance.
  Data model - typically pre-established by an application, and may be implicit

• **Metadata**
  Annotations associated with the data contained in the instance.
Extensible MultiModal Annotation

EMMA version of the semantic interpretation result for:
"I want a medium coke and three large pizzas with pepperoni and mushrooms"

<drink>
  <liquid> coke </liquid>
  <drinksize> medium </drinksize>
</drink>

<pizza>
  <number> 3 </number>
  <pizzasize> large </pizzasize>
  <topping length="2">  
    <item index="0"> pepperoni </item>
    <item index="1"> mushrooms </item>
  </topping>
</pizza>
Extensible MultiModal Annotation

EMMA version of the semantic interpretation result for:
"I want a medium coke and three large pizzas with pepperoni and mushrooms"

<emma:emma version="1.0" xmlns:emma="http://www.w3.org/2003/04/emma"> <emma:interpretation id="int1" >
  <drink>
    <liquid> coke </liquid>
    <drinksize> medium </drinksize>
  </drink>
  <pizza>
    <number> 3 </number>
    <pizzasize> large </pizzasize>
    <topping length="2">
      <item index="0"> pepperoni </item>
      <item index="1"> mushrooms </item>
    </topping>
  </pizza>
</emma:emma>
Extensible MultiModal Annotation

EMMA version of the semantic interpretation result for:
"I want a medium coke and three large pizzas with pepperoni and mushrooms"

Annotations added:
<emma:emma version="1.0" xmlns:emma="http://www.w3.org/2003/04/emma"> <emma:interpretation id="int1" confidence="0.6" tokens="I want a medium-sized coca cola and three large pizzas with pepperoni and mushrooms" >
  <drink>
    <liquid> coke </liquid>
    <drinksize> medium </drinksize>
  </drink>
  <pizza>
    <number> 3 </number>
    <pizzasize> large </pizzasize>
    <topping length="2">
      <item index="0"> pepperoni </item>
      <item index="1"> mushrooms </item>
    </topping>
  </pizza>
</emma:emma>
Digital ink InkML

More electronic devices with pen interfaces are becoming available for entering and manipulating information.

MMIWG created a standard specification for an XML representation of digital ink InkML.

Fundamental data element in an InkML file is the `<trace>`. A trace represents a sequence of contiguous ink points—e.g., the X and Y coordinates of the pen's position.
Digital ink InkML

• 5 traces in input ("h", "e", "l", "l", "o")
• trace is represented as a series of points.

Traces in InkML
<ink>
  <trace>0 0 9 14 8 28 7 42 6 56 6 70 8 84 8 98 8 112 9 126 10 140 13 154 14 168 17 182 18 188 23
  174 30 160 38 147 49 135 58 124 72 121 77 135 80 149 82 163 84 177 87 191 93 205
  </trace>
  <trace>130 155 144 159 158 160 170 154 179 143 179 129 166 125 152 128 140 136 131 149 126
  163 124 177 128 190 137 200 150 208 163 210 178 208 192 201 205 192 214 180
  </trace>
  <trace>227 50 226 64 225 78 227 92 228 106 228 120 229 134 230 148 234 162 235 176 238 190
  241 204 </trace>
  <trace>282 45 281 59 284 73 285 87 287 101 288 115 290 129 291 143 294 157 294 171 294 185
  296 199 300 213
  </trace>
  <trace>366 130 359 143 354 157 349 171 352 185 359 197 371 204 385 205 398 202 408 191 413
  177 413 163 405 150 392 143 378 141 365 150
  </trace>
</ink>
Conclusion

W3C Multimodal Interaction Working group is working on specifications that support multimodal interaction on the web.

• Multimodal interaction offers ease of use benefits over uni-modal interaction.

• MI Framework - describes the framework for multimodal interaction and identifies components for multimodal systems.

• EMMA - Representing Users' Intent.

• InkLM - Representing digital ink.
End

Thank You!

Questions, remarks?