Multimodal Interaction

Lesson 9
Multimodal Coordination

Maria De Marsico
demarsico@di.uniroma1.it

Credits

Derived from:


• Niels Ole Bernsen, Laila Dybkjr. Multimodal Usability. Springer 2009

Multimodal architecture

- A multimodal system takes multimodal inputs from the fusion of different modalities (e.g. speech, sketch, gesture, etc.).

- Each modality is characterized by a different model (grammar) that describe the features of the particular modality and that must be implemented by the system developer.

Three kinds of components

- A modality recogniser (for each modality) that translates user input according to the proper grammar
  - for speech recognizer this may be an ASR module
- A module for multimodal input fusion that takes input from the recognisers for each modality and fuses their results into a complete semantic frame for the application.
- A module that accepts semantic frames from the application and provides user feedback through multimodal output (output fission).
Reminder

• During interaction, the user inputs *input* modalities to the system and the system outputs *output* modalities to the user.
• When *humans* communicate with each other, the input and output modalities are typically the *same*.
• This is called *input/output modality symmetry*.
• In human–system interaction, input/output modality *asymmetry* is the rule and symmetry the exception.
• Try to think of examples of both kinds.

Rewind: a semantic frame?

<table>
<thead>
<tr>
<th>Level Number</th>
<th>Name of Layer</th>
<th>Exchanged unit of Information</th>
<th>Text Example</th>
<th>GUI Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Goal</td>
<td>Real world concepts, external to computer</td>
<td>Remove section of my letter</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Task</td>
<td>Computer-oriented objects/actions</td>
<td>Delete 6 lines of edited text</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Semantics</td>
<td>Concrete objects, specific operations</td>
<td>Delete line no 27</td>
<td>Delete selected lines</td>
</tr>
<tr>
<td>4</td>
<td>Syntax</td>
<td>Sentences (1 or 2 dimensional sequences or layouts) of tokens</td>
<td>DELETE 27</td>
<td>Click to the left of the first line; while holding down the shift-key, click to the right of last line; select CUT in menu</td>
</tr>
<tr>
<td>3</td>
<td>Lexical</td>
<td>Tokens: smallest info-carrying units</td>
<td>DELETE</td>
<td>Click at left of first line</td>
</tr>
<tr>
<td>2</td>
<td>Alphabetic</td>
<td>Lexemes: primitive symbols</td>
<td>D</td>
<td>Click at (345,120)</td>
</tr>
<tr>
<td>1</td>
<td>Physical</td>
<td>&quot;Hard I/O&quot;, light, sound, movement</td>
<td>Press D-key</td>
<td>Press mouse button</td>
</tr>
</tbody>
</table>
The communication principle in the layered protocol model. A communication on level \( i \) of the model (indicated by the gray arrow) is realized by an exchange of information on level \( i-1 \). Both the user and the computer will have to translate between the two levels as indicated by the vertical arrows. The units of information exchanged at each level of the model are listed in the Table backwards.

**Does this remind you something?**

From a conceptual point of view... 7 is a magic number!

- Let us further remind the revisititation of Norman's cycle, especially for what concerns its central items.
Goals

Intention

Evaluation

waiting

Action specification

Mental Activity

Interpretation

Perception

Physical Activity

Execution

7 steps of user activities involved in computer-based tasks

Norman’s cycle revisited

• In the case of multimodal interactive cycles, Norman’s model of interaction may be reformulated as follows:

1. Establishing the goal.
2. Forming the intention.
3. Specifying the **multimodal** action sequence in terms of human output modalities.
4. Executing the **multimodal** action.
5. Perceiving the system state in terms of human input modalities.
6. Interpreting the system state.
7. Evaluating the system state with respect to the goals and the intentions.

... ... Action plan

**Specifying the multimodal action sequence:** the sequence of actions performed to accomplish the required task should be precisely stated at this stage.

Complexity of multimodal interaction appears for the first time in the cycle. Each multimodal action can be specified in terms of:

1. **Complementary human output sensory modalities** (i.e., multiple utterances at once form the action)
   and/or
2. **Alternative human output sensory modalities** (i.e., alternative, redundant utterances for the same action).

Some **unintentional** utterances: blood pressure, temperature, heartbeat, excretion, etc.

A user may move an object in the interaction scene by speaking and pointing at (gesturing) the new object location (complementary modalities).
Then, instead of gesturing (s)he may want to gaze at the new location on the interface where the object should be moved (alternative modality).
Execution

**Executing each multimodal action**: at this stage, each human modality used to specify an action is translated into corresponding interaction modes.

Each action is executed through
1. **Complementary modes**
   or
2. **Alternative modes**

Text, speech, Braille, mimicking, eye/motion capture, haptics, bio-electrical sensoring are examples of modes used to translate human output modalities into the system input language.

When the execution of the whole sequence of multimodal actions is complete, the system reaches a new state and communicates it to the user again exploiting (possibly multiple) interaction modes, such as speech synthesis, display, haptic/tactile feedback, smell rendering and so on.

Perception ...

**Perceiving the system state**: at this stage, the evaluation phase of the cycle begins.

Depending on the combination of system output modes, the user may perceive the new state through multiple input sensory modalities, such as visual, auditory, tactile, and (in some revolutionary interfaces) even smelling and tasting.
An example of timing
(from the system described in Hank Liao. Multimodal Fusion)

The algorithm captures multimodal inputs and increments a counter for every recognition process that is started, and decrements it every time one reports ending.

When the counter is decremented to zero, a timer is started depending whether the end recognition message came from the gesture recogniser or speech. This is necessary because the speech recogniser also has an ep.EndSeconds length to ensure the proper detection of the end of speech, whereas the gesture recogniser lacks this delay.

The ep.EndSeconds parameter is set at 2.0 seconds, the FUSION TIMEOUT SPEECH timeout interval at 1000 milliseconds, whereas the FUSION TIMEOUT GESTURE is set at 2000 milliseconds; this was found to work adequately but could be tuned to decrease latency.

The Windows timer ID_FUSION_TIMEOUT is set at the timeout interval for the modality. When it times out, as indicated by the EndOfRec in Figure, the FusionCallback function is called which performs the final fusion routines, checks for complete frames, finds the best complete command frame, and whether to accept and send this frame to the application or to send a rejection.

If while the timer is timing out another start recognition event is received, the timer is reset.

Maria De Marsico - demarsico@di.uniroma1.it
Why timing is important

• **Multimodal interaction patterns** refer to the possible **relationships** between the inputs representing a multimodal production.
• Relationships may hold for example in the **temporal**, **spatial** or **semantic** domain but ...
• ... temporal relationships, in particular, are **crucial** for correctly interpreting multimodal interaction, in both human-to-human and human-to-computer communication.
• Multimodal input fusion systems rely on such knowledge to validate or refuse possible fusion of inputs.

Maria De Marsico - demarsico@di.uniroma1.it

Why timing is important

• The **qualitative** and **quantitative** aspects of the temporal relationships have been analyzed in several researches in order to provide a better design of multimodal input fusion modules, but ...
• ... also to progress the fundamental understanding of human communication.

Maria De Marsico - demarsico@di.uniroma1.it
Reminder: time relations

More timing relations

- Both overlapping and non-overlapping interactions can also be divided in other six types of cooperation between modalities
- Complementarity
- Concurrency
- Equivalence
- Redundancy
- Specialization
- Transfer

(Martin, 1997)
Formal notations

• Martin proposes formal notations to define more precisely these types of cooperation.
• They aim at stating explicitly the parameters of each type of cooperation and the relation between these parameters which is subsumed by the type of cooperation.
• He considers the case of input modalities (human towards computer).

Formal definition of modality

• A modality is a process receiving and producing chunks of information.
• A modality M is formally defined by:
  o $E(M)$ the set of chunks of information received by M
  o $S(M)$ the set of chunks of information produced by M
Transfer

- When several modalities cooperate by transfer, this means that a chunk of information produced by a modality is **used by another** modality.
- Transfer is commonly used in hypermedia interfaces when a mouse click provokes the display of an image.
- In information retrieval applications, the user may express a request in one modality (speech) and get relevant information in another modality (video) (but this is rather a matter of asymmetry).

Transfer

- Two modalities M1 and M2 cooperate by transfer when a chunk of information produced by M1 can be used by M2 after translation by a transfer operator tr which is a parameter of the cooperation.

\[ \text{transfer}(M1, M2, tr): tr(S(M1)) \subseteq E(M2) \]
Specialization

• When modalities cooperate by specialization, this means that a specific kind of information is always processed by the same modality.
• Specialization is not always absolute and may be more precisely defined.
• Example: in several systems, sounds are somehow specialized in errors notification (forbidden commands are signaled with a beep).
  o It is a modality-relative specialization if sounds are not used to convey any other type of information.
  o It is a data-relative specialization if errors only produce sounds and no graphics or text.
  o When there is a one-to-one relation between a set of information and a modality, we will speak of an absolute specialization.
• The choice of a given modality should add semantic information and hence help the interpretation process.
• When a modality is specialized, it should respect the specificity of this modality including the information it is good at representing.
• Example: in reference interpretation, the designation gesture aims at selecting a specific area and the verbal channel provides a frame for the interpretation of the reference: categorical information, constraints on the number of objects selected.

\[ \text{specialization}(M, I, \{Mi\}) = I = S(M) \land \forall i \ I \not\subseteq S(M) \]

• Stronger version

\[ \text{specialization}(M, I, \{Mi\}) = I = S(M) \land \forall i \ I \cap S(M) = \emptyset \]
Equivalence

• When several modalities cooperate by equivalence, this means that a chunk of information may be processed as an alternative, by either of them.
• Equivalence also enables adaptation to the user by customization: the user may be allowed to select the modalities he prefers.
• The formation of accurate mental models of a multimodal system seems dependent upon the implementation of such options over which the user has control.
• Equivalence $\leftrightarrow$ Alternative.

Equivalence

• Two input modalities $M_1$ and $M_2$ cooperate by equivalence for the production of a set $I$ of chunks of information when each element $i$ of $I$ can be produced either by $M_1$ or $M_2$. An operator $eq$ controls which modality will be used and may take into account user’s preferences, environmental features, information to be transmitted...

$$\text{equivalence}(M_1, M_2, I, eq):$$
$$\forall i \in I, \quad \exists e_1 \in E(M_1) \land \exists e_2 \in E(M_2) \mid i = eq(M_1, e_1), (M_2, e_2))$$
Redundancy

• If several modalities cooperate by redundancy, this means that the same information is processed by these modalities.

• Example: if the user types "quit" on the keyboard or utters "quit", the system asks for a confirmation. But if the user both types and utters "quit", the systems interpret this redundancy to avoid a confirmation dialogue thus enabling a faster interaction by reducing the number of actions the user has to perform.

Redundancy

• Two input modalities M1 and M2 cooperate by redundancy for the production of a set I of chunks of information when each element i of I can be produced by an operator re merging a couple (s1, s2) produced respectively by M1 and M2.

• The operator re will merge (s1, s2) if their redundant attribute has the same value and a criterion crit is true.

• A chunk of information has several attributes. For instance, a chunk of information sent by a speech recognizer has the following attributes: time of detection, label of recognized word, recognition score.

• The redundant attribute of two modalities plays a role in deciding whether two chunks of information produced by these modalities is redundant or complementary.
Redundancy

\[\text{redundancy} \ (M_1, M_2, i, \text{redundant\_attribute, crit}):
\forall i \in I, \exists s_1 \in S(M_1), \exists s_2 \in S(M_2),
\text{redundant\_attribute} (s_1) = \text{redundant\_attribute} (s_2) \land
i = \text{re}(s_1, s_2, \text{crit})\]

Complementarity

- When several modalities cooperate by complementarity, it means that different chunks of information are processed by each modality but have to be merged.
- First systems enabled the "put that there" command for the manipulation of graphical objects.
- Example: if the user wants to create a radio button, he may type its name on the keyboard and select its position with the mouse. These two chunks of information have to be merged to create the button with the right name at the right position.
- Complementarity may enable a faster interaction since the modalities can be used simultaneously and convey shorter messages which are moreover better recognized than long messages.
- Complementarity may also improve interpretation, for example a graphical output is sufficient for an expert but need to be completed by a textual output for a novice.
- An important issue concerning complementarity is the criterion used to merge chunks of information in different modalities. The most classical approaches are to merge them because they are temporally coincident, temporally sequential or spatially linked.
- Two types of behavior may feature complementarity. In the "sequential" behavior, the user would by example utter "what are the campsites at" and then select a town with the tactile screen. In the "synergistic" behavior, the user would utter "Are there any campsites here?" and select a town with the tactile screen while pronouncing "here". The second behavior seems more common.
Some notes about complementarity

- Modalities cooperating by complementarity may be specialized in different types of information.
- In the example of a graphical editor, the name of an object may be always specified with speech while its position is specified with the mouse.
- Modalities cooperating by complementarity may be also be equivalent for different types of information. As a matter of fact, the user could also select an object with the mouse and its new position with speech ("in the upper right corner").

Complementarity

- Two input modalities M1 and M2 cooperate by complementarity for the production of a set I of chunks of information when each element i of I can be produced by an operator co merging a couple (s1, s2) produced respectively by M1 and M2.
- The process co will merge (s1, s2) if their redundant attribute does not have the same value and a criterion crit is true:

```latex
complementarity (M_1, M_2, I, redundant_attribute, crit):
\forall i \in I, \exists s_1 \in S(M_1), \exists s_2 \in S(M_2),

\text{redundant_attribute} (s_1) \neq \text{redundant_attribute} (s_2) \land

i = co(s_1, s_2, crit)
```
Concurrency

- when several modalities cooperate by concurrency, it means that different chunks of information are processed by several modalities at the same time but must not be merged.
- This may enable a faster interaction since several modalities are used in parallel.

Not the only classification

<table>
<thead>
<tr>
<th>1. Type of relation</th>
<th>2. What it does</th>
<th>3. Co-ordination</th>
<th>4. Aimed at user groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complementarity</td>
<td>Several modalities necessary to express a single communicative act</td>
<td>Tight</td>
<td>Same</td>
</tr>
<tr>
<td>Addition</td>
<td>Add up different expressiveness of different modalities to express more information</td>
<td>Loose</td>
<td>Same or different</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Express partly the same information in different modalities</td>
<td>Tight</td>
<td>Same or different</td>
</tr>
<tr>
<td>Elaboration</td>
<td>Express partly the same information in different modalities</td>
<td>Tight</td>
<td>Same or different</td>
</tr>
<tr>
<td>Alternative</td>
<td>Express roughly the same information in different modalities</td>
<td>Loose</td>
<td>Same or different</td>
</tr>
<tr>
<td>Stand-in</td>
<td>Fail to express the same information in a less apt modality</td>
<td>None</td>
<td>Same or different</td>
</tr>
<tr>
<td>Substitution</td>
<td>Replace more apt modality/modalities by less apt one(s) to express the same information</td>
<td>None</td>
<td>Special</td>
</tr>
<tr>
<td>Conflict</td>
<td>The human system cannot handle modality addition</td>
<td>Tight</td>
<td>None</td>
</tr>
</tbody>
</table>

From: N. Ole Bernsen, L. Dybkjær. Multimodal Usability. Springer 2009
Fusion and fission

- In order to enable a natural dialogue between users and computer systems, in multimodal systems the two main challenges to face are:
  o to combine and integrate information from different input modalities (**fusion** process) and
  o to generate appropriate output information (**fission** process)
In order to allow that the multimodal human–computer communication process takes place successfully, the actions that the user expresses through the human output modalities have to be acquired by the system through the computer input modalities, and the human input channels of the user have to be able to perceive and understand feedback from the computer output channels.


Multimodal fusion

- A GUI requires **atomic** and **unambiguous** inputs (such as the selection of an element by mouse or the insertion of a character by keyboard).

- Multimodal interaction involves **several simultaneous** inputs that have to be **recognized** and opportunely **combined** by managing the uncertainty of inputs through probabilistic techniques.

- The process of integrating information from various input modalities and combining them into a complete command is called **multimodal fusion**.
Multimodal timing

• As we have seen, in a multimodal interaction temporal constraints of inputs have to be taken into account and consequently it requires a time-sensitive architecture and the recording of time intervals of each modalities.

Multimodal fission

• In a GUI the output messages are conveyed to the user through a single medium (the graphical display), whereas in a multimodal system a way of disaggregating outputs through the various channels has to be found in order to provide the user with consistent feedback.

• This process is called multimodal fission, in contrast with multimodal fusion
Readings


- W3C. Multimodal Interaction Requirements - W3C NOTE 8 January 2003: http://www.w3.org/TR/mmi-reqs/


- Jean-Claude MARTIN. Towards "intelligent" cooperation between modalities. The example of a system enabling multimodal interaction with a map. http://perso.limsi.fr/Individu/martin/ijcai/article.html