Integration Patterns during Multimodal Interaction

Anurag Gupta, Tasos Anastasakos

Motorola Labs
1301 E. Algonquin Road, Schaumburg, IL 60196, USA
{Anurag.Gupta, Tasos.Anastasakos}@motorola.com

Abstract
The development of multimodal interfaces and algorithms for multimodal integration requires knowledge of integration patterns that represent how people use multiple modalities. We analyzed multimodal interaction with three different applications. Semantic analysis revealed that multimodal inputs can exhibit cooperation other than complementary and redundancy. Analysis of the relationship between tasks and multimodality revealed a positive correlation between the type of a task and multimodal interaction. In addition, analysis revealed that referential expressions can contain more information than simple deictic terms such as instructions to select or modify the referents.

1. Introduction
A lot of research is being carried out to develop multimodal interfaces that can provide reliable and fault-tolerant interaction within real world application scenarios [6]. Contemporary literature has looked at the joint use of input modes such as speech and gesture from a linguistic perspective [5], as well as the role of gesture in discourse and thought [3]. Significant literature is available on user studies on the multimodal integration and synchronization patterns during combined speech and pen-gesture based interaction [5]. Oviatt [5] also performed an analysis of the relationship between task and multimodality. The focus of the prior user studies has been on the analysis of the temporal characteristics of multimodal inputs, and the patterns describing the use of multiple modalities during interaction.

Most of the previous literatures have assumed that the content in multiple modalities is either redundant or complementary. Currently there is little work available on the relationship between the semantic content of multimodal inputs during human computer interaction, although such work is crucial to guiding the design of suitable algorithms for the combination of multimodal inputs. In addition, the dominant multimodal phenomena of cross-modal referential expressions, i.e., referencing input provided in one modality with input provided in another modality, has not been studied in detail within the context of multimodal interaction.

We conducted a user study to analyze multimodal integration patterns. The goal of this user study was to understand the semantic relationships between multimodal inputs and to understand the content of cross-modal referential expressions. The following tasks were critical part of the study: (a) Observe semantic integration patterns, i.e., understand the semantic relationships expressed by a user between inputs provided in multiple modalities, (b) Observe modality integration patterns, i.e., the modalities used to complete a particular task and the relationship between tasks and multimodality, and (c) Analyze the content of cross-modal referential expressions and the associated multimodal inputs that specify the referents. In section 2, we provide details of the applications, tasks, and procedure used during this study, and in section 3, we describe its results.

2. Method

2.1. Applications
We used three applications for the user study. The first application was an in-car navigation system [2], which creates routes to user specified destinations, guides the driver along a route offering turn-by-turn directions, and provides entertainment services such as reading the latest news. Users can interact with the application using speech, handwriting, and gesture. The second application was Media organizer, which offers a multimodal interface for storing and classifying media files such as pictures and songs. The purpose of the application is to allow the user to organize their collection of pictures and songs and retrieve them using a combination of speech, handwriting and gesture inputs. The third application was Multimodal Airline Travel Information System (MATIS) [1], which provides a multimodal form-filling interface to book flights and to obtain information on flights. During form filling, the system allows users to provide inputs using keyboard, speech and handwriting. We used the ‘Wizard-of-Oz’ technique to simulate multimodal interaction capabilities for the Media organizer and the in-car navigation systems, i.e., people using the system believed that they were interacting with a real system, while actually there was a human operator controlling the application and engaging in a dialogue to get the required inputs to complete a task. The users could interact using a combination of speech, handwriting, and gestures. For example, in the navigation system, a user could indicate an area by a gesture, or by speaking or writing the name of a suburb.

2.2. Participants and tasks
Twenty-two subjects participated in the user study. Of these fourteen were native speakers of English. Thirteen subjects were males and nine were females. They represented a broad spectrum of age from 12 years to 48 years. Eighteen of them were professionals and eight of them had prior experience of using a speech-based application. Each subject was required to perform fourteen tasks. We provided the subjects with a scenario to complete each task, for example, going shopping while picking up children on the way. We demonstrated the ability to complete each task using both multimodal and unimodal interaction.

The tasks required to be performed on the in-car navigation system were: (1) Zoom the map around a point or area – e.g. “zoom in here”, (2) Pan the map in a particular
direction – e.g. “move the map this way”, (3) Create a route between two or more locations – e.g. “Create a route from here to there”, (4) Find petrol stations within an area – e.g. “Find the closest petrol station in this area”, (5) Find hotels within an area – e.g. “Find a Comfort Inn within this area”, (6) Calculate distance between two locations – e.g. “How far from here to here?”, (7) Query for bus information – e.g. “When is the next bus from this bus stop?”

The tasks required to be performed on the MATIS were: (1) Fill flight search form by entering the ‘From’ and ‘To’ fields, and optionally the ‘Day’ field, (2) Book a flight by selecting a flight – e.g. “Book this flight”.

The tasks required to be performed on the Media Organizer system were: (1) Play songs – e.g. “Play this song”, (2) View pictures – e.g. “Show this, this and these”, (3) Create a new folder – e.g. “Add a new directory here”, (4) Move objects into a folder – e.g. “Move these into this folder”, (5) Search songs from an artist – e.g. “Which are the songs from Madonna?”

### 2.2.1. Task classification

We classified the individual tasks into five categories in order to identify any significant change in the multimodal interaction based on the type of task. The categories and their associated tasks were as follows: (1) Location – Create a route, Calculate distance, and Move objects, (2) Selection – Play, View, and Book a flight, (3) Search - Find petrol stations, Find hotels, and Search songs, (4) Control – Zoom, Pan, and Create a new folder, (5) Query – Fill flight search form, and Query for bus information

### 3. Results

We observed 1545 interactions during this user study. Users demonstrated a strong preference for interacting multimodally since 992 out of the 1545 instances had multimodal inputs. During the post-experiment interviews, 19 of the 22 participants expressed preference to interact multimodally. The majority of the participants, or 21 of the 22, felt that it was very natural for them to interact using multimodal means. One of the participants felt overwhelmed by the use of multiple modalities. All the participants liked the ability to use both handwriting and speech. They felt that it was much easier to point to an object than to describe it using a gesture. This shows a clear preference of the users to specify the task (e.g., to create a route) using speech and the task’s parameters (e.g., source and destination) through other multimodal means. The results also show that if both buttons and spoken inputs are available to issue a command, users prefer to use spoken inputs.

### 3.2. Modality integration patterns

Considering the modalities used in a multimodal turn, there were four possible combinations (ignoring the order of use) – ‘speak and gesture’, ‘speak and write’, ‘write and gesture’, and ‘speak, write and gesture’. The rates of occurrence of these combinations within the 992 multimodal interactions were - ‘speak and gesture’ in 70%, ‘speak and writing’ in 4.5%, ‘write and gesture’ in 16%, and ‘speak, write and gesture’ in the remaining 9.5%. Table 1 shows the modality integration pattern for the individual task categories during multimodal interaction. We observed that the task category influences the modality pattern used, but overall there was a distinct preference for use of ‘speak and gesture’ pattern.

Using all the available modalities in a turn was the least used pattern. Table 2 shows the distribution of the first modality used within a turn for each task category. Users initiated tasks in most of task categories except Control and Query categories using speech. Query and Control tasks were started more often with a gesture than with speech.

#### Table 1: Modality use patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Location</th>
<th>Selection</th>
<th>Search</th>
<th>Control</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speak &amp; gesture</td>
<td>86%</td>
<td>68%</td>
<td>56%</td>
<td>100%</td>
<td>46%</td>
</tr>
<tr>
<td>Speak and write</td>
<td>2%</td>
<td>2%</td>
<td>18%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Write and gesture</td>
<td>12%</td>
<td>20%</td>
<td>10%</td>
<td>0%</td>
<td>31%</td>
</tr>
<tr>
<td>Speak, write and gesture</td>
<td>0%</td>
<td>10%</td>
<td>16%</td>
<td>0%</td>
<td>23%</td>
</tr>
</tbody>
</table>

#### Table 2: Distribution of the first modality used within a turn

<table>
<thead>
<tr>
<th>Modality</th>
<th>Location</th>
<th>Selection</th>
<th>Search</th>
<th>Control</th>
<th>Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>74%</td>
<td>56%</td>
<td>64%</td>
<td>43%</td>
<td>0%</td>
</tr>
<tr>
<td>Gesture</td>
<td>16%</td>
<td>28%</td>
<td>34%</td>
<td>57%</td>
<td>100%</td>
</tr>
<tr>
<td>Handwriting</td>
<td>10%</td>
<td>16%</td>
<td>2%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

### 3.3. Semantic integration patterns

We analyzed the semantic relationships between the content provided in multiple modalities. In general, the inputs
provided in multiple modalities were complementary. For example, a user of the navigation system said, “I want to go here,” and pointed to a location on a map. The pointing gesture provides the location description while the speech provides the command description. In some cases, users provided redundant inputs in multiple modalities. For example, a user of the navigation system clicked on a button to zoom on the map and said, “Zoom In”. In this case, the information in multiple modalities is redundant, since they do not provide any extra information. Redundant information can be helpful in disambiguation when modalities generate ambiguous interpretations since the likelihood of redundant information being correct is higher than the likelihood of non-redundant information being correct.

In some cases, we observed that users tried to replace the information in one modality with information provided in another modality. For example, a user said, “Find similar hotel but in Sydney” and pointed to a hotel. The user’s intention was to replace the value of the ‘location’ attribute of the gestured hotel with ‘Sydney’. In some cases, the inputs supplied in one modality were logically related, and the logical relationship was expressed in another modality. For example, a user while interacting with the Media Organizer system said, “Show this and this picture.” In this example, ‘and’ relationship is expressed in speech, while the objects themselves are provided using gesture. In another example, a user said, “Create route from here to there not going through here”, and pointed to three locations, each corresponding to a spoken deictic term. The user’s intention was to not to use the given waypoint on the route created.

We also observed a pattern to aggregate certain content. For example, a user of the Media Organizer system said, “Show these” and made a single gesture that provided a number of referents. The user’s intention was to aggregate the referents together and execute the same task, i.e., to show them, on each of them.

### 3.4. Referencing in multimodal systems

Referential expressions were present in 80% of the observed multimodal interaction instances. Our observations indicate that referential expressions have a number of characteristics, which need to be handled during multimodal integration. As shown in Table 3, referential expressions can be either singular or plural, defining the concept of aggregation number, which specifies the number of referents required by a referential expression. A singular referential expression requires a single referent, while a plural referential expression requires multiple referents. A plural referential expression can be either definite or indefinite. A definite plural referential expression requires a predetermined number of referents such as two or three, while an indefinite referential expression requires an undetermined number of referents. We observed that 99% of the singular referential expressions were associated with deictic terms such as “this”, “that”, “here”, “there”. Indefinite referential expressions were associated with terms such as “these”, “those”, etc. Define plural referential expressions were associated with terms such as “these two”, “those three” that specified the number of required referents. Considering the referential expressions observed in the multimodal interaction instances, 80% were singular, 14% were definite plural, and the remaining 6% were indefinite plural.

#### Table 3: Aggregation in referential expressions

<table>
<thead>
<tr>
<th>Type</th>
<th>Expression Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular</td>
<td>“Play this” or “Go here”</td>
</tr>
<tr>
<td>Definite plural</td>
<td>“Play these two” or “Show these three”</td>
</tr>
<tr>
<td>Indefinite plural</td>
<td>“Move these”</td>
</tr>
</tbody>
</table>

Figure 1 shows the histogram of explicit aggregation number provided within referential expressions made in the multimodal instances. The aggregation numbers in plural referential expressions are usually small (one, two, or three). When the number of referents grows large (four or more), the users use an indefinite referential expression.

#### Figure 1: Frequency of required referents

### 3.4.1. Aggregation and associated gesture

Analysis of the multimodal instances having indefinite referential expressions revealed that 86% of indefinite referential expressions were associated with a single selection gesture providing a number of referents, and the remaining 14% had multiple gestures each providing one or more referents. Considering the multimodal instances having definite plural referential expressions, 85% had a single selection gesture that provided all the necessary referents, 8% had multiple gestures providing the necessary number of referents, and 7% did not have adequate number of referents.

#### 3.4.2. Multiple references

16% of multimodal instances had multiple referential expressions within them, e.g., the sentence, “I want to go from here to here”, contains two singular referential expressions. 65% of multimodal instances with multiple referential expressions had multiple singular referential expressions, 25% had a mixture of singular and plural referential expressions, and 10% had multiple plural referential expressions. In 65% of the multimodal instances having multiple referential expressions, there was a logical relationship between the expressions - 45% were related using the “and” relation, 40% using the “or” relation, and the rest using a mix of “and”, “or” and “not”.

#### 3.4.3. Deictic, anaphoric and contextual references

Within multimodal interaction, deictic expressions in one modality (e.g., speech) refer to entities provided in another modality (e.g., gesture). Anaphoric expressions refer to objects mentioned previously during the interaction. Anaphora can extend across multiple turns or occur in the same turn. For example, in the phrase ‘John came in and he lit a fire’, ‘he’ can be interpreted as referring to ‘John’. Consider phrases spoken in two turns by a user: “I want to go to Central Hotel”, and “How many rooms does the hotel have?” In the second utterance, the expression ‘the hotel’ refers to the ‘Central Hotel’ mentioned in the previous turn.
We calculated that 90% of the references in the observed multimodal instances were deictic, and 10% of the references were anaphoric and anaphora extended across multiple turns. We observed in 5% of the instances containing referential expressions, the referential expression had a meaning within the application and the user did not provide an explicit referent. For example, ‘here’ in the map-application meant the current location. In another 5% of multimodal interaction instances, users did not provide any referents explicitly using other modalities but they wanted the application to select appropriate referents from their visual context. For example, a user of the Media Organizer system said, “Show these”, when a number of picture icons were displayed in the application’s GUI. The user’s intent was for the system to display the pictures which were part of the visual context.

3.4.4. Selection and modification

A characteristic of multimodal referential expressions was the use of selection criterion that required the system to select referents with specific attribute values. For example, a user of the Media Organizer system said, “Show pictures from Yosemite out of these” and selected a few photos. This interaction requires the system to select among the gestured pictures, only the pictures that were taken at Yosemite, i.e., the value of the ‘location’ attribute of the selected pictures must be ‘Yosemite’. 4% of the multimodal instances having a referential expression had an associated selection criterion. Users associated selection criterion mostly, or 70%, with indefinite referential expressions. The remaining 30% of selection criterion were associated with definite plural referential expressions. Users mentioned the selection criterion mostly, or 80%, in the modality used to provide the referential expression. The remaining 20% of the time, users expressed the criterion in another modality. For example, a user of the Media Organizer said, “Which are the songs from the artist out of these?” and wrote “Eminem”.

In addition, users included within referential expressions instructions to modify values of certain attributes of the referents. For example, a user said, “Find hotels similar to this but with four stars”. For this interaction, the navigation system needs to replace the ‘Rating’ attribute of the referred hotel with the value of ‘4’ before executing the hotel search query. In 90% of the interactions that contained expressions to modify the value of a referent’s attribute, users provided the modified values directly within the modality used to provide the referential expression. The remaining 10% of instances, users provided another referent containing the attribute’s desired value. For example, a user of the navigation system said, “Find hotels similar to this but in this area”, and gestured twice on the map to provide a hotel and to select an area.

4. Discussion

Similar to Oviatt’s [5] findings, users preferred to interact multimodally. The user’s choice to interact either multimodally or unimodally was influenced by the task being accomplished by the user. Knowledge of a given task can influence the likelihood of receiving unimodal or multimodal inputs, and affects the decision about when to integrate multimodal inputs to generate multimodal interpretations.

Similar to Oviatt’s [5] and Xiao’s [7] findings, multimodal input usually provided either complementary or redundant information. In addition, we observed that a number of other relationships were expressed between multimodal inputs such as logical connections between objects provided in other modalities. This significantly affects the algorithms used in the integration of multimodal inputs.

In contrast to the results reported in [5], we observed that most multimodal constructions involving speech, or 80%, contained a spoken deictic term. However, for multimodal constructions involving handwriting and gesture only 30% contained a written deictic term. Multimodal architectures designed for speech and gesture based applications can make use of the deictic term to assist in interpreting the referent. Our results indicate that during multimodal interaction users are likely to provide more than simple deictic terms such as ‘here’ and referential expressions can include instructions to select or modify the referents. A generic multimodal system must be able to interpret those referential expressions, resolve multiple references, and select, aggregate, and modify referents through discourse and visual contexts.

5. Conclusion

This study has provided further understanding of semantic integration patterns during multimodal interaction. The study clearly indicates a positive correlation between a task and the use of multimodal interaction, and the modality integration pattern used to specify the task. In addition, based on our observation, we have generated a detailed description of the possible content of referential expressions during multimodal interaction. The results of this study will be useful in guiding the design of future multimodal architectures and integration algorithms that will support complex, intuitive, and robust multimodal interaction for real work tasks. It has already inspired the design of a multimodal integration component, which we have applied to multiple applications such as in-car and handheld navigation systems.

6. References