Multimodal Database Access on Handheld Devices

Elsa Pecourt and Norbert Reithinger
DFKI GmbH
Stuhlsatzenhausenweg3
D-66123 Saarbrücken, Germany
{pecourt,reithinger}@dfki.de

Abstract

We present the final MIAMM system, a multimodal dialogue system that employs speech, haptic interaction and novel techniques of information visualization to allow a natural and fast access to large multimedia databases on small handheld devices.

1 Introduction

Navigation in large, complex and multidimensional information spaces is still a challenging task. The search is even more difficult in small devices such as MP3 players, which only have a reduced screen and lack of a proper keyboard. In the MIAMM project[1] we have developed a multimodal dialogue system that uses speech, haptic interaction and advanced techniques for information visualization to allow a natural and fast access to music databases on small scale devices. The user can pose queries in natural language, using different dimensions, e.g. release year, genre, artist, or mood. The retrieved data are presented along this dimensions using various visualization metaphors. Haptic feedback allows the user to feel the size, density and structure of the visualized data to facilitate the navigation. All modalities are available for the user to access and navigate through the database, and to select titles to be played.

The envisioned end-user device is a handheld Personal Digital Assistant (PDA, see figure 1) that provides an interface to a music database. The device includes a screen where data and system messages are visualized, three force-feedback buttons on the left side and one combined scroll wheel/button on the upper right side, that can be used to navigate on the visualized data, as well as to perform actions on the data items (e.g. play or select a song), a microphone to capture spoken input, and speakers to give audio output. Since we do not develop the hardware, we simulate the PDA using a 3D model on a computer screen, and the buttons

Figure 1: The PDA simulator with the terrain visualization of the database

by means of Phantom devices[2] that allow the user to touch and manipulate virtual objects.

In the rest of this paper, we will first give an overview of the visualization metaphors, the MIAMM architecture, and a short description of its interface language. Then we will demonstrate its functionality using an example dialogue. For more details on the MIAMM system and its components see (Reithinger et al., 2004).

2 Visualization metaphors

The information from the database is presented on the device using metaphors of real world objects (cf. *conceptual spaces* (Gärdenfors, 2000)) so as to provide an intuitive handling of abstract concepts. The lexicon metaphor, shown in figure 2 to the left, presents the items alphabetically ordered in a rotary card file. Each card represents one album and contains detailed background information. The time-

[1] http://www.miamm.org

[2] http://www.sensable.com

Figure 2: Visualizations

line visualization shows the items in chronological order, on a "rubber" band that can be stretched to get a more detailed view. The wheel metaphor presents the items as a list on a conveyor belt, which can be easily and quickly rotated. Finally, the terrain metaphor (see figure 1) visualizes the entire database. The rendering is based on a three layer type hierarchy, with genre, sub-genre and title layers. Each node of the hierarchy is represented as a circle containing its daughter nodes. Similarities between the items are computed from the genre and mood information in the database and mapped to interaction forces in a physical model that groups similar items together on the terrain. Since usually albums are assigned more than one genre, they can be contained in different circles and therefore be redundantly represented on the terrain. This redundancy is made clear by lines connecting the different instances of the same item.

3 The MIAMM prototype

The MIAMM system uses the standard architecture for dialogue systems with analysis and generation layers, interaction management and application interface (see figure 3). To minimize the reaction delay of haptic feedback, the visual-haptic interaction component is decoupled from other more time-consuming reasoning processes. The German experimental prototype[3] incorporates the following components, some of which were reused from other projects (semantic parser and action planning): a speaker independent, continuous speech recognizer converts the spoken input in a word lattice; it uses a 500 word vocabulary, and was trained on a automatically generated corpus. A template based semantic parser for German, see (Engel, 2004), interprets this word lattice semantically. The multimodal fusion module maintains the dialogue history and handles anaphoric expressions and quantification. The action planner, an adapted and enhanced version of (Löckelt, 2004), uses non-linear regression planning and the notion of *communicative games* to trigger and control system actions. The visual-haptic interaction manager selects the appropriate visualization metaphor based on data characteristics, and maintains the visualization history. Finally, the domain model provides access to the MYSQL database, which contains 7257 records with 85722 songs by 667 artists. Speech output is done by speech prompts, both for spoken and for written output. The prototype also includes a MP3 Player to play the music and speech output files. The demonstration system requires a Linux based PC for the major parts of the modules written in Java and C++, and a Windows NT computer for visualization and haptics. The integration environment is based on the standard Simple Object Access Protocol SOAP[4] for information exchange in a distributed environment.

The communication between the modules uses a declarative, XML-schema based representation lan-

[3]There are also French and English versions of the system. The modular architecture facilitates the replacement of the language dependent modules.

[4]http://www.w3.org/TR/SOAP/

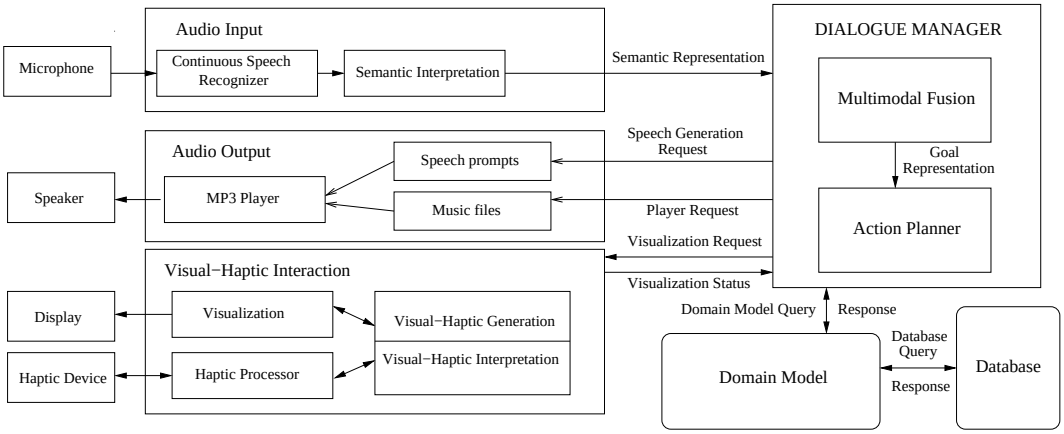

Figure 3: MIAMM architecture

guage called MMIL (Romary and Bunt, 2002). This interface specification accounts for the incremental integration of multimodal data to achieve a full understanding of the multimodal acts within the system. Therefore, it is flexible enough to handle the various types of information processed and generated by the different modules. It is also independent from any theoretical framework, and extensible so that further developments can be incorporated. Furthermore it is compatible with existing standardization initiatives so that it can be the source of future standardizing activities in the field[5]. Figure 4 shows a sample of MMIL representing the output of the speech interpretation module for the user's utterance "*Give me rock*".

4 An example

To sketch the functionality of the running prototype we will use a sample interaction, showing the user's actions, the system's textual feedback on the screen and finally the displayed information. Some of the dialogue capabilities of the MIAMM system in this example are, e.g. search history (S2), relaxation of queries (S3b), and anaphora resolution (S5). At any moment of the interaction the user is allowed to navigate on the visualized items, zoom in and out for details, or change the visualization metaphor.

U1: *Give me rock*
S1a: `I am looking for rock`
S1b: displays a terrain with rock albums
U2: *I want something calm*
S2a: `I am looking for calm rock`
S2b: displays list of calm rock albums
U3: *I want something from the 30's*
S3a: `I am looking for calm rock 1930-1939`
S3b: `I could only find albums of the adjacent years`
displays list of calm rock albums of the 40's
U4: *What about the 50's*
S4a: `I am looking for calm rock 1950-1959`
S4b: displays a map with rock albums
U5: selects ALBUM with the haptic buttons
Play this one
S5a: `Playing ALBUM`
S5b: MP3 player starts

We will show the processing details on the basis of the first utterance in the sample interaction *Give me rock*. The speech recognizer converts the spoken input in a word graph in MPEG7. The semantic parser analyzes this graph and interprets it semantically. The semantic representation consists, in this example, of a `speak` and a `display` event, with two participants, the `user` and `music` with constraints on its genre (see figure 4).

The multimodal fusion module receives this representation, updates the dialogue context, and passes it on to the action planner, which defines the next goal on the basis of the propositional content of the top event (in the example event `id1`) and its object (in the example participant `id3`). In this case the user's goal cannot be directly achieved because the object to display is still unresolved. The action planner has to initiate a database query to acquire the required information. It uses the constraint on the genre of the requested object to produce a database query for the domain model and a feedback request for the visual-haptic interaction module. This feedback message (S1a in the example) is sent to the user while the database query is being done, providing thus implicit grounding. The do-

[5]The data categories are expressed in a RDF format compatible with ISO 11179-3

```
<component>
  <event id="id0">
    <evtType>speak</evtType>
    <speaker>user</speaker>
    <addressee>system</addressee>
    <dialogueAct>request</dialogueAct>
  </event>
  <event id="id1">
    <evtType>display</evtType>
  </event>
  <participant id="id2">
    <objType>user</objType>
    <refType>1PPDeixis</refType>
    <refStatus>pending</refStatus>
  </participant>
  <participant id="id3">
    <objType>music</objType>
    <genre>rock</genre>
    <refType>indefinite</refType>
    <refStatus>pending</refStatus>
  </participant>
  <relation
    source="id3"
    target="id1"
    type="object"/>
  <relation
    source="id1"
    target="id0"
    type="propContent"/>
</component>
```

Figure 4: MMIL sample

main model sends the result back to the action planner who inserts the data in a visualization request.

The visual-haptic interaction module computes the most suitable visualization for this data set, and sends the request to the visualization module to render it. This component also reports the actual visualization status to the multimodal fusion module. This report is used to update the dialogue context, that is needed for reference resolution. The user can now use the haptic buttons to navigate on the search results, select a title to be played or continue searching.

5 Conclusions

The MIAMM final prototype combines speech with new techniques for haptic interaction and data visualization to facilitate access to multimedia databases on small handheld devices. The final evaluation of the system supports our initial hypothesis that users prefer language to select information and haptics to navigate in the search space. The visualizations proved to be intuitive (van Esch and Cremers, 2004).

Acknowledgments

This work was sponsored by the European Union (IST-2000-29487). Thanks are due to our project partners: Loria (F), Sony Europe (D), Canon (UK), and TNO (NL).

References

Ralf Engel. 2004. Natural language understanding. In Wolfgang Wahlster, editor, *SmartKom - Foundations of Multi-modal Dialogue Systems*, Cognitive Technologies. Springer Verlag (in Press).

Peter Gärdenfors. 2000. *Conceptual Spaces*. MIT Press.

Markus Löckelt. 2004. Action planning. In Wolfgang Wahlster, editor, *SmartKom - Foundations of Multi-modal Dialogue Systems*, Cognitive Technologies. Springer Verlag (in Press).

Norbert Reithinger, Dirk Fedeler, Ashwani Kumar, Christoph Lauer, Elsa Pecourt, and Laurent Romary. 2004. Miamm - a multimodal dialogue system using haptics. In Jan van Kuppevelt, Laila Dybkjaer, and Niels Ole Bersen, editors, *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. Kluwer Academic Publications.

Laurent Romary and Harry Bunt. 2002. Towards multimodal content representation. In *Proceedings of LREC 2002, Workshop on International Standards of Terminology and Linguistic Resources Management*, Las Palmas.

Myra P. van Esch and Anita H. M. Cremers. 2004. User evaluation. MIAMM Deliverable D1.6.