MADFILM – a Multimodal Approach to Handle Search and Organization in a Movie Recommendation System

Pontus Johansson
Dept. of Computer Science
Linköping University
SE-581 83 Linköping, Sweden
ponjo@ida.liu.se

Abstract

The MADFILM multimodal movie information and recommendation system prototype addresses the information search and recommendation problem with a natural language interface, and the information organization problem with a direct manipulation interface. The two modalities are integrated to allow for coordinated and simultaneous interaction. This paper describes the design and implementation of the MADFILM system.

1 Introduction

The WIMP[1] metaphor in traditional graphical user interfaces (GUIs) aids users to organize and view information and content of applications and operating systems. One of the most well-known examples is the hierarchical organization of files and folders as directory trees in graphical file managers. This view of objects in an information space allows for well-known, highly functional, ways of managing objects spatially (e.g. selecting, moving, and deleting). There are, however, several operations where this spatial model falls short, and where other interaction modalities are more advantageous. For example, natural language (NL) interaction (Androutsopoulos et al., 1995) in on-going dialog allows for incremental refinement of information queries; detection and resolution of e.g. contradictions and verifications; and ambiguity resolution by clarification questions. Dialog also serves as an efficient and natural way to incrementally build a user model used for recommendations (Johansson, 2003).

Within the multimodal research community the combination and integration of different modalities has been of interest in order to develop interfaces where users can choose between interaction techniques that fit the current task at hand, and combine modalities to efficiently carry out tasks. A pioneer example is the Put-That-There system suggested by Bolt (1980). This paper presents a movie information and recommendation system that allows users to use NL dialog in combination with mouse gestures as input modalities to manage objects in a movie information space. The system is a multimodal extension of the ADFILM movie recommendation system (Johansson, 2003).

[1] Window-Icon-Menu-Pointer

This paper is organized as follows: Section 2 describes this project's initial user study and its implications for the system design in terms of the two input modalities, and motivations for combining them. Section 3 gives an overview of the system architecture and the UI design. The paper is concluded in Section 4.

2 Multimodal Interaction in the Movie Domain

The movie domain is an information space consisting of a plethora of film titles, actors, genres, directors, and plot information, etc. Most users are considered to know their way around – at least conceptually – in the domain. They have most likely seen several movies, and are able to articulate preferences about domain properties such as genres (Burke et al., 1997). MADFILM tries to utilize this by letting users convey preferences and ask for information using NL. The system is developed using an iterative development method, where each iteration benefits from real end-usage feedback (Degerstedt & Jönsson, 2001).

2.1 User Study

In order to get empirical usage data for the first iteration, a corpus of human-human dialogs was collected. The dialogs are between one dialog partner playing a movie *recommender* role, and one partner acting as a *customer* who is looking for movie recommendations. As information resource, the recommender has access to a laptop and Internet connection to a large movie information database. The customer is also allowed to look at the screen if s/he desires. Twenty-four dialogs were collected using 48 subjects, resulting in 7.5 hours of recorded material. All dialogs were recorded in a relaxed living room environment. The dialog collection forms the basis for the construction of use-case scenarios, which are designed through dialog distilling (Larsson et al., 2000). The use scenarios serve as a requirement specification for the system, and define what dialog features the MADFILM system should handle.

In the 2684 utterances (a mean of 112 utterances per dialog), several interesting features were found. Both participants take initiative and drive the dialog; the recommender actively asks the customer for preferences, whereas the customer initiative typically consists of information requests or volunteering preferences. Three categories of dialog phenomena of special interest can be identified in the domain of movie recommendation dialog. They are:

- **Object Manipulation**: Utterances referring to domain objects spatially (e.g. moving and organizing movies in lists).
- **Clarification**: Utterances initiating sub-dialogs (by either the recommender or the customer) in order to complete an information or preference request.
- **Global Focus Management**: Utterances utilizing the dialog history for *global* focus shifts. Global shifts imply a focus shift to a previously introduced concept not in the current focus.

Table 1 shows the selected three dialog categories we are interested in covering in the first iteration, and their proportions in the corpus.

Category	**Percentage of utterances**	**Sub-category**	**Percentage**
Dialog:	43.5%	Object manipulation:	12.0%
		Clarification:	16.7%
		Global Focus Management:	14.8%

Table 1: Dialog categories of interest in the corpus and how often they occur.

As Table 1 suggest, these three suggested categories cater for 43.5% of dialog-related phenomena. Parts of the remaining 56.5% not covered by this taxonomy are still catered for in the system, and include dialog properties that are so to speak "common" in information-providing dialog systems. A major portion is for example using the dialog history for *local* focus management. This is allowed for in the system interaction and the mechanisms for handling this is based on previously developed information-providing dialog systems (e.g. Johansson et al., 2002). There are, however, phenomena in the corpus not covered by the system including dialog properties inherent in human-human dialog that are not suitable for human-machine communication (e.g. various forms of feedback, irony and jokes). In sub-sequent iterations, we will collect human-machine corpora based on real system usage to minimize these features and provide more realistic human-machine dialogs.

Utterances are also categorized based on their content as belonging to one of the three categories *task*, *communication management*, and *irrelevant* (see Table 2).

Category	**Percentage of utterances**	**Sub-category**	**Per-centage**	**Sub-category**	**Per-centage**
Task:	79.3%	Information:	28.6%	Quantification:	8.9%
				Atomic:	19.7%
		Preference:	50.7%	Quantification:	18.3%
				Atomic:	32.4%
Communication management:	14.5%				
Irrelevant:	6.2%				

Table 2: Utterance content taxonomy consisting of three categories.

There are two task sub-categories in this movie recommendation domain: *information* and *preference* requests. Both types of requests can be further divided into *quantification* and *atomic* requests.

The quantification sub-category refers to utterances where something is said about a set of objects (movies). Quantifications are of varying complexity and can include negations. This is an important feature which can be beneficial for recommendation systems employing NL dialog as interaction technique, since it allows users to convey preferences such as: "*I like Bruce Willis but I can't stand his comedies.*" This is an example of a preference quantification that is hard to capture in a purely graphical user interface (Johansson, 2003).

The atomic sub-category refers to utterances about exactly one movie. Atomic preference utterances (i.e. "ratings" of movies) are important in order to establish a user model for the collaborative filtering recommendation engine. Examples of atomic requests (of both information and preference types) are showed in Table 3.

Communication management consists of a wide variety of phenomena, e.g. various forms of feedback. This category is – despite its importance in communication – intentionally left without finer sub-categorization since it is anticipated that the content of this category will differ in human-machine dialogs compared to human-human dialogs. It should also be noted that the boundaries between categories are "fuzzy", as one utterance may belong to e.g. both task and communication management. In ambiguous cases, we opt to look at functions related to task. This explains why the communication management percentage is relatively low compared to the task percentage.

Irrelevant utterances, e.g. reporting system malfunctions to the experiment conductor during a dialog session, are put in a separate category which is ignored when modeling the sub-language of this domain.

The analysis shows that object management in the movie information space takes two distinct shapes. First, it can be viewed as a *search problem*, where users frequently want to find out properties (such as actors, directors, and genres) about a specific title; or they want to find all titles in a specific genre directed by their favorite director, etc. Second, we can view the interaction as an *organization problem*, where users want to organize and save selections (e.g. maintaining a "to-see-list" consisting of recommended titles, or throwing bad recommendations in a "trashcan"). MADFILM tries to accommodate both search and organization in the domain by allowing users to interact using NL in ongoing dialog with the system, and direct manipulation of objects on-screen.

2.2 Natural Language and Dialog

The distilled dialogs form the basis for constructing dialog flow representations, as well as grammar and lexicon coverage. As the corpus analysis suggest, there are a significant number of clarification sub-dialog initiations. They range from clarifying both information requests and preference conveying. For example, a common phenomenon is that customers often pose information requests in order to be able to answer the recommender's movie preference requests, as in the example in Table 3.

Dialog	Task category
R1: Have you seen Memento?	*Atomic preference request*
C1: Who is starring in it?	*Atomic information request*
R2: Memento stars these actors. <shows a list of actors>	
C2: Is it about that guy who lost his memory?	*Atomic information request*
R3: Yeah, exactly	
C3: Yeah, that's a good movie	

Table 3: Sample human-human dialog from the corpus showing interleaved information requests in a clarification sub-dialog. (R = Recommender, C = Customer).

Allowing for this sort of interaction in the system enhances usability, but also requires robust focus management in order for the system to use C3 as answer to R1, and extract the positive rating of Memento in C3 to the user model. Utterances such as C1 are classified as a search problem and can successfully be handled with an information-providing dialog system, (cf. Johansson et al., 2002).

The basis for the dialog model and NL coverage for MADFILM is the distillation of the collected dialog corpus. In its current version, the system's dialog model accommodates one of the constructed use-case scenarios.

2.3 Direct Manipulation

Several of the utterances in the Object Manipulation category concern organization of movie titles to a list of recommended movies (the "to-see-list"). Both the recommender and the customer refer to the list, and they cooperate to incrementally add items to it. This is done by writing items from the recommender's screen listings down on a piece of paper. The recommender also maintains separate lists as memory aids on actors, genre, and director preferences the customer might have conveyed. These memory aids serve as points of global focus, and accommodate the global dialog history utterances found in the dialog collection (see Table 1). Visual representations on the screen for various movies and persons or other concepts that have been introduced, can thus be referred to by simply pointing at them.

The organization problem in this domain can be compared to the file management approach (see Section 1), where object representations can be manipulated using gestures, such as pointing, selecting, and drag-and-drop techniques. The gestures considered are indexical, and identifies objects merely by indicating their location, as opposed to iconic and symbolic gestures (Streit, 1999). Users can thus refer to graphical movie representations and areas on-screen by simply pointing at them with a suitable interaction device (currently a regular mouse). Movie objects can be moved to different areas in the interface such as the to-see-list or the trashcan (see Figure 2).

2.4 Combining Natural Language and Direct Manipulation

In normal situations deictic gestures are used together with NL, as it provides additional information, concerning properties or additional location information that helps to identify the object (Streit, 1999). The collected dialogs support this since several utterances contain spatial references (see previous section).

The search and organization tasks are supported by NL and gestures respectively, implying that considering the modalities separately allows users to address the search and organization tasks in two different ways. However, integrating the modalities lets the user use combinations in order to carry out these tasks more naturally; similar to the way the human dialog participants built their interaction in the user study (see Table 4).

Two important implications of the corpus analysis for the modality integration are the Object Manipulation and the Global Focus Management categories (see Table 1). Customers frequently use NL in combination with the various lists discussed above. The lists also connect to the dialog since the items seem to function as memory aids to global focus "access points". The focus shifts are an important part of the recommendation dialog, as the example in Table 4 shows.

Dialog	**Action**	**Focus**
R1: What sort of movies do you like?		
C1: A drama please	Recommender notes "Category: Drama"	Category: drama
R2: Do you have any favorite actors?		
C2: I like the star in Gladiator		Title: Gladiator Actor: wanted
R3: That's Russell Crowe.	Recommender notes "Actor: Russell Crowe"	Actor: Russell Crowe
C3: Who directed Gladiator?		Title: Gladiator Director: wanted
R4: Ridley Scott. Have you seen G. I. Jane?		Director: Ridley Scott Title: G. I. Jane
C4: Yeah, that's a good one! I like war movies.		Category: War
R5: Then I think you'd like Black Hawk Down. It's a war/drama film by him.		*combining focus in R4, C4, and C1*.
C5: Sounds good. Write that down please.	Recommender notes "Black Hawk Down" on the to-see-list	Title: Black Hawk Down
R6: I think you might like A Beautiful Mind also. That's a drama starring	Recommender points at the actor note from R3.	*combining focus in C1 and R3*.

Russell Crowe.		
C6: I think I wanna put that down as well.	Recommender notes "A Beautiful Mind" on the to-see-list	Title: A Beautiful Mind

Table 4: Sample dialog showing global focus points and their usage in the recommendation task. (R = Recommender, C = Customer).

As this example shows, the lists help the participants to establish important focus points for future reference, (e.g. R3). By switching between these foci, eventually the recommender can perform a recommendation, which – if the customer agrees – can be put on the to-see-list (e.g. C5 and C6).

Previous studies report that users naturally use and switch between speech and pointing, which leads to better performance, faster error recovery, and less frustration (Cohen, 1998; Oviatt, 1999). We differentiate between the following classes of multimodal input (W3C, 2000):

1. **Sequential multimodal input**: Input is processed in a sequential order, with no integration of the modalities.
2. **Uncoordinated simultaneous multimodal input**: Simultaneous input processed in random order and not integrated.
3. **Coordinated simultaneous multimodal input**: Exploiting multimodality to the full, integrating the multimodal input to one unified representation.

The coordinated combination allows the user to get a higher flexibility and accommodates individual interaction styles (e.g. talkative users might rely more on NL and dialog, whereas less talkative users – or users in environments where using voice is not suitable – may rely more on pointing).

The corpus analysis suggests that interaction using NL and gestures in the coordinated mode can take three principal forms in the movie domain, as shown in Table 5.

Modality	Search task	Organization task
NL	”Who directed Alien?”	”Put Star Wars on the to-see-list please”
Object manipulation	–	[Selects the Star Wars icon and moves it to the To-See-List]
NL and Object manipulation combined	”Who directed this one?” [points at the Alien movie icon]	”Put Star Wars here please” [points at the To-See-List]

Table 5: Multimodal interaction and task examples in the MADFILM domain. User utterances are within quotation marks, user actions within brackets.

Table 6 shows part of one of the use scenarios exemplifying MADFILM interaction capabilities.

Utterance	GUI Action
U1: Do you have any good comedies starring Jim Carrey?	-
S1: Yes, there are these 5 matching movies.	Displays a list of titles on the main panel
U2: Who directed this?	Points at one of the 5 titles in the list
S2: Frank Darabont.	Textual display of the name Frank Darabont
U3: Is there a picture of him?	-
S3: I found the following image of Frank Darabont.	Displays a photograph of F. Darabont.
U4: Put these on my to-see-list please.	Selects three of the titles with the cursor
S4: Certainly.	The selected titles are moved to the to-see-list panel

Table 6: Excerpt from one of the use-case scenarios derived from the MADFILM dialog corpus illustrating multimodal interaction in terms of the search and organization problems. U = user, S = system.

3 The MADFILM System

The unimodal ADFILM system is a dialog system that gives movie recommendations to users who state their movie preferences using NL (Johansson, 2003). ADFILM is a prototype focusing on the new-user cold-start problem and uses typed input. In the multimodal application – MADFILM – the interaction is enhanced by allowing users to organize retrieved recommendations using (mouse) gestures in combination with NL utterances as outlined in Section 2. This feature puts additional demands on the input interpretation, and amounts to the system architecture. MADFILM exists in a Swedish version and is built in Java.

3.1 System Architecture

MADFILM accommodates coordinated and simultaneous multimodal input by maintaining one separate thread for each modality, and base the integration on timestamps, much like the SPICE system (Kellner & Portele, 2002). Figure 1 shows a conceptual view of the MADFILM information flow and its relation to system modules.

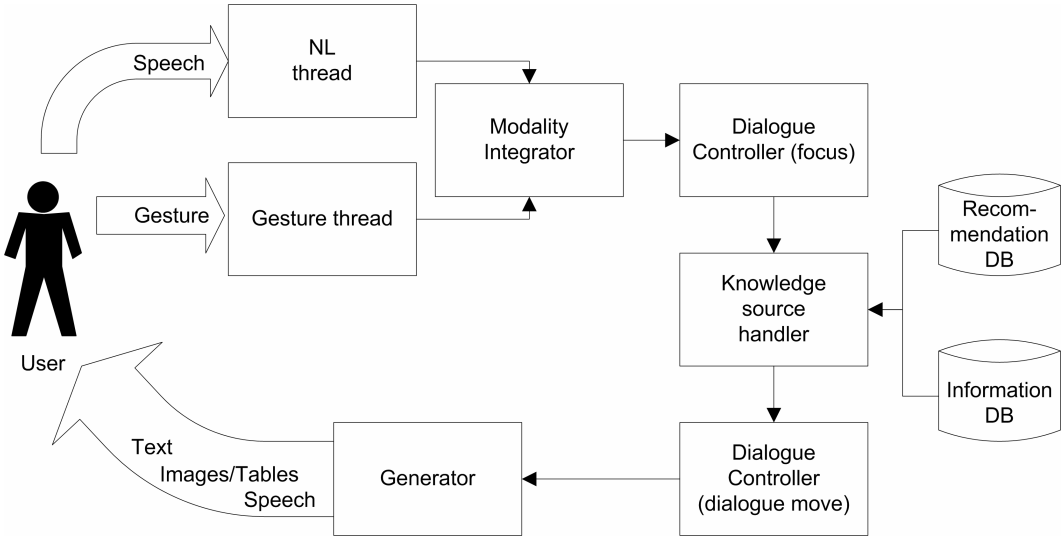

Figure 1: Conceptual overview of the MADFILM information flow.

The *NL Thread* understanding component is responsible for creating a feature structure representation of the semantic content of the user utterance. It utilizes the Nuance speech recognizer and a semantic component that creates the semantic representation. Speech grammars are constructed to handle the language identified in section 2.2.

The *Gesture Thread* interpretation component monitors the GUI's event model which is time stamped by the *Integrator*. In the integration phase, the feature structure from the NL thread is enriched with the mouse information from the gesture thread. Gesture information management consists of keeping track of selected items in the various lists in the GUI, and mouse pointer coordinate tracking. This is important since users do not have to click to select e.g. a list in the GUI, but may leave the pointer hovering or circling over a desired area such as the to-see-list (see Figure 2). Previous research suggests that a suitable time lag between speech and an accompanying indexical gesture is four seconds (Kellner & Portele, 2002; Oviatt, 1999), and currently this is the strategy adopted. It is the Integrator's responsibility to keep track of the timestamps. When the Integrator has merged the interpreted gesture and the NL utterance, the modified feature structure is ready to be passed to the first dialog control phase.

A smooth handling of the various forms of clarifications is essential. The *Dialog Controller* handles focus management, and checks for focus switches and inconsistencies. If no inconsistencies in the focus management phase are found a database query can be constructed based on the resolved focus and incoming user utterance. Dialogs in MADFILM are represented as a network of generic states. As suggested by the corpus analysis, the system drives the dialog as a recommendation dialog trying to extract quantification of preferences (e.g. preferences over all movies in a specific genre, or movies starring a particular actor), and atomic preferences (i.e. rating of a particular movie). User initiatives are often interleaved information requests in the recommendation dialog, as in Table 3.

The *Knowledge Source Handler* queries the information database – or the recommendation engine depending on the user's request – and retrieves a result set. The dialog controller now checks for results that might require sub-dialog (e.g. too many or no hits), and generates an appropriate dialog move.

The dialog move is realized by the *Generator*. It consists of a speech generation component that plays pre-recorded system utterances. It can also produce a textual representation in the system feedback panel if the user so wishes (see Figure 2). Pre-recorded speech and template-based text generation are suitable techniques for the restricted domain, and fast to develop. This is in line with the iterative development method employed in this project. The Generator also consists of a component that creates a graphical representation of the retrieved information which is presented in the main result panel of the GUI as images, and text (see Figure 2).

3.2 Graphical User Interface

The GUI consists of several panels, and follows a suggested interaction model for multimodal TV information systems in (Ibrahim & Johansson, 2002), with the addition of implications from the object manipulation category from the dialog analysis. These include the to-see-list and the trashcan list. Figure 2 shows a screen shot of the system.

The largest panel is the main information panel where the title or person under discussion is displayed. The recognized user utterance is displayed in the user panel below the main panel. To the left, the corresponding system utterance is displayed as text, should the user choose to turn system speech off. The right-most part of the screen consists of two panels that are interesting from a gesture point of view. The top right panel is the current user's to-see-list, which is a list of titles that the user has been recommended and has expressed a wish to see (either by telling the system in the dialog, or by moving a title there from e.g. the main information panel). The panel below – the trashcan – is where movies disliked by the user end up. The main panel, the to-see-list, and the trashcan are the main interaction areas from a gesture point-of-view since the user can drag and drop movie items on these panels. The user can also address these manipulations by speech, or a combination of speech and gesture as Tables 5 and 6 suggest.

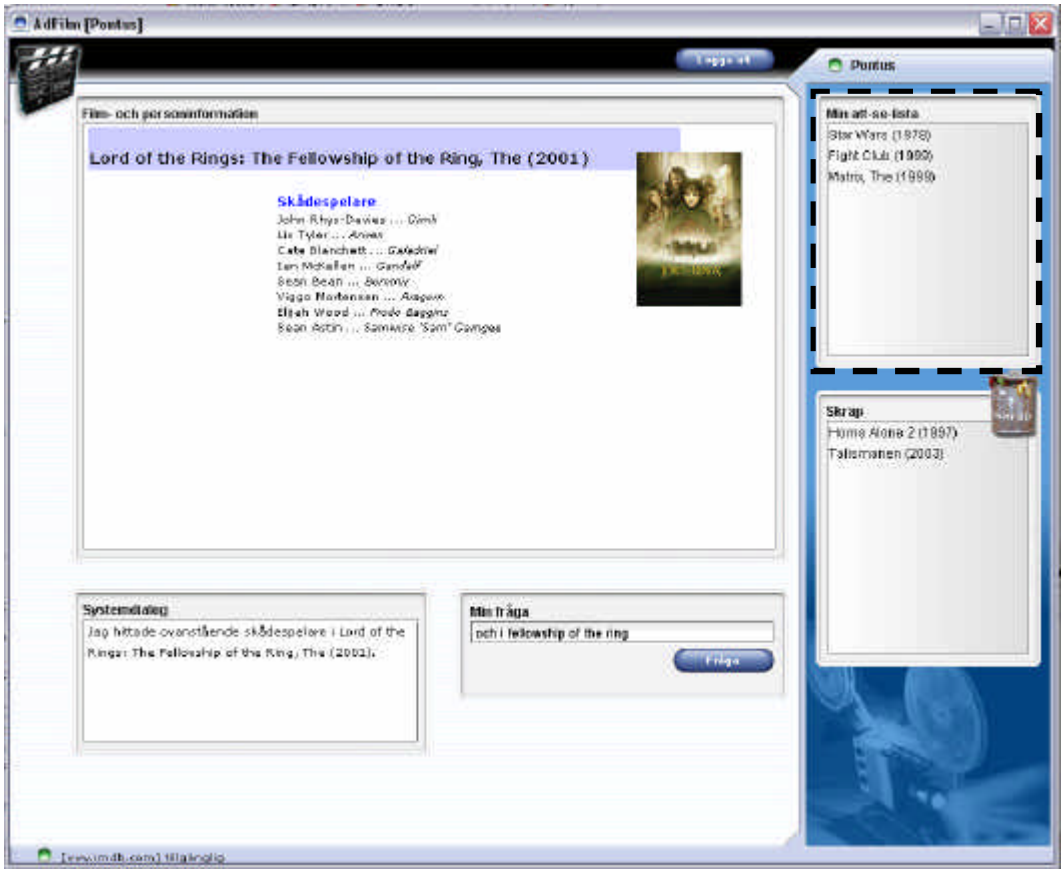

Figure 2: MADFILM graphical user interface screen shot. Actor information for the movie "The Fellowship of the Ring" is currently discussed. The dashed black box surrounding the To-See-List is not visible in the actual application, but shows one of the "pointer-sensitive" areas, whose coordinates are used by the gesture component in order to help interpret referring expressions such as the word 'here' in the utterance: "Put Lord of the Rings here".

4 Conclusion

The collected dialog corpus in the MADFILM project suggests that quantifications and clarification sub-dialogs are an important part of movie recommendation dialogs. While this can be accommodated in unimodal NL dialog systems, other features in the user study (especially visual memory aids for managing global focus shifts, and combined dialog and direct manipulation for organizing and referring to lists of movie objects) suggest that a movie recommendation system should cater for simultaneous and coordinated integration of NL and direct manipulation.

Direct manipulation in the form of mouse pointer interaction has been integrated as a new input modality in the MADFILM recommender dialog system.

The concurrent processing of NL dialog and gesture threads in MADFILM allows for coordinated and integrated multimodal interaction for searching and organizing information in the movie domain. This interaction model is derived from a distilled human-human dialog collection, and based on previous research on multimodal interaction on an electronic program guide application. Immediate future steps include evaluating the MADFILM prototype on a set of end-users to derive new – and modify existing – use scenarios and to extend the system's capabilities to handle phenomena not covered by the first iteration's use scenarios.

The unimodal ADFILM system is focused on movie recommendations, where the user model is mainly used as a recommendation basis. Future research also includes evaluating what implications multimodal capabilities have on the user model acquisition process in a recommender dialog system.

5 Acknowledgements

Many thanks to Jenny Isberg and Sophie Öhrn for their work on collecting the dialog corpus. This research is financed by Vinnova, Swedish Agency for Innovation Systems, and Ceniit, Center for Industrial Information Technology.

6 References

Androutsopoulos, I., Ritchie, G., Thanisch, P. (1995). "Natural language interfaces to databases–an introduction," Journal of Language Engineering, vol. 1, no. 1, pp. 29–81.

Bolt, R. A. (1980). Put-that-there: Voice and gesture at graphical interfaces. Computer Graphics 14, pp. 262–270.

Burke, R. D., Hammond, K. J., Young, B. C. (1997). "The FindMe approach to assisted browsing," IEEE Expert, vol. 12, no. 4, pp. 32–40.

Cohen, P. R. (1991). The role of natural language in a multimodal interface. Technical note 514, Computer dialog Laboratory, SRI International.

Cohen, P. R., Johnston, M., McGee, D., Oviatt, S. L., Clow, J., Smith, I. (1998). "The efficiency of multimodal interaction: A case study". In Proceedings of International Conference on Spoken Language Processing, Australia.

Degerstedt, L., Jönsson, A. (2001). "Iterative Implementation of Dialogue System Modules". In Proceedings of Eurospeech 2001, Aalborg, Denmark.

Ibrahim, A., Johansson, P. (2002). "Multimodal dialog systems for interactive TV applications". In Proceedings of 4th IEEE International Conference on Multimodal Interfaces, Pittsburgh, USA, pp. 117–222.

Johansson, P. (2003). "Natural language interaction in personalized EPGs". In Proceedings of the UM'03 3rd Workshop on Personalization in Future TV, Pittsburgh, USA.

Johansson, P., Degerstedt, L., Jönsson, A. (2002). "Iterative development of an information-providing dialog system". In Proceedings of 7th ERCIM Workshop, Chantilly, France.

Jönsson, A. (1997). "A model for habitable and efficient dialogue management for natural language interaction". Natural Language Engineering 3(2/3), pp. 103-122, Cambridge University Press.

Kellner, A., Portele, T. (2002). “SPICE - a multimodal conversational user interface to an electronic program guide”. In ISCA Tutorial and Research Workshop on Multi-Modal Dialogs in Mobile Environments, Kloster Irsee, Germany.

Larsson, S., Santamarta, L., Jönsson, A. (2000). “Using the process of distilling dialogs to understand dialog systems”. In Proceedings of 6th International Conference on Spoken Language Processing (ICSLP2000/INTERSPEECH2000), Beijing, China.

Oviatt, S. L. (1999). “Mutual disambiguation of recognition errors in a multimodal architecture”. In Proceedings of CHI 1999, pp. 576–583.

Streit, M. (1999). “Interaction of speech, deixis and graphical interface”. In Proceedings of the workshop on Deixis, Demonstration and Deictic Belief, Utrecht, The Netherlands, pp. 69–75.

W3C (2000). “Multimodal requirements for voice markup languages: W3C working draft 10 July 2000”. [Online]. Available: http://www.w3.org/TR/2000/WD-multimodal-reqs-20000710.