A Multimodal Interface for Access to Content in the Home

Michael Johnston
AT&T Labs Research,
Florham Park,
New Jersey, USA
johnston@research.att.com

Luis Fernando D'Haro
Universidad Politécnica de Madrid,
Madrid, Spain
lfdharo@die.upm.es

Michelle Levine
AT&T Labs Research,
Florham Park,
New Jersey, USA
mfl@research.att.com

Bernard Renger
AT&T Labs Research,
Florham Park,
New Jersey, USA
renger@research.att.com

Abstract

In order to effectively access the rapidly increasing range of media content available in the home, new kinds of more natural interfaces are needed. In this paper, we explore the application of multimodal interface technologies to searching and browsing a database of movies. The resulting system allows users to access movies using speech, pen, remote control, and dynamic combinations of these modalities. An experimental evaluation, with more than 40 users, is presented contrasting two variants of the system: one combining speech with traditional remote control input and a second where the user has a tablet display supporting speech and pen input.

1 Introduction

As traditional entertainment channels and the internet converge through the advent of technologies such as broadband access, movies-on-demand, and streaming video, an increasingly large range of content is available to consumers in the home. However, to benefit from this new wealth of content, users need to be able to rapidly and easily find what they are actually interested in, and do so effortlessly while relaxing on the couch in their living room — a location where they typically do not have easy access to the keyboard, mouse, and close-up screen display typical of desktop web browsing.

Current interfaces to cable and satellite television services typically use direct manipulation of a graphical user interface using a remote control. In order to find content, users generally have to either navigate a complex, pre-defined, and often deeply embedded menu structure or type in titles or other key phrases using an onscreen keyboard or triple tap input on a remote control keypad. These interfaces are cumbersome and do not scale well as the range of content available increases (Berglund, 2004; Mitchell, 1999).

Figure 1 Multimodal interface on tablet

In this paper we explore the application of multimodal interface technologies (See André (2002) for an overview) to the creation of more effective systems used to search and browse for entertainment content in the home. A number of previous systems have investigated the addition of unimodal spoken search queries to a graphical electronic program guide (Ibrahim and Johansson, 2002

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 376–383,
Prague, Czech Republic, June 2007.

(NokiaTV); Goto et al., 2003; Wittenburg et al., 2006). Wittenburg et al experiment with unrestricted speech input for electronic program guide search, and use a highlighting mechanism to provide feedback to the user regarding the “relevant” terms the system understood and used to make the query. However, their usability study results show this complex output can be confusing to users and does not correspond to user expectations. Others have gone beyond unimodal speech input and added multimodal commands combining speech with pointing (Johansson, 2003; Portele et al, 2006). Johansson (2003) describes a movie recommender system MadFilm where users can use speech and pointing to accept/reject recommended movies. Portele et al (2006) describe the SmartKom-Home system which includes multimodal electronic program guide on a tablet device.

In our work we explore a broader range of interaction modalities and devices. The system provides users with the flexibility to interact using spoken commands, handwritten commands, unimodal pointing (GUI) commands, and multimodal commands combining speech with one or more pointing gestures made on a display. We compare two different interaction scenarios. The first utilizes a traditional remote control for direct manipulation and pointing, integrated with a wireless microphone for speech input. In this case, the only screen is the main TV display (far screen). In the second scenario, the user also has a second graphical display (close screen) presented on a mobile tablet which supports speech and pen input, including both pointing and handwriting (Figure 1). Our application task also differs, focusing on search and browsing of a large database of movies-on-demand and supporting queries over multiple simultaneous dimensions. This work also differs in the scope of the evaluation. Prior studies have primarily conducted qualitative evaluation with small groups of users (5 or 6). A quantitative and qualitative evaluation was conducted examining the interaction of 44 naïve users with two variants of the system. We believe this to be the first broad scale experimental evaluation of a flexible multimodal interface for searching and browsing large databases of movie content.

In Section 2, we describe the interface and illustrate the capabilities of the system. In Section 3, we describe the underlying multimodal processing architecture and how it processes and integrates user inputs. Section 4 describes our experimental evaluation and comparison of the two systems. Section 5 concludes the paper.

2 Interacting with the system

The system described here is an advanced user interface prototype which provides multimodal access to databases of media content such as movies or television programming. The current database is harvested from publicly accessible web sources and contains over 2000 popular movie titles along with associated metadata such as cast, genre, director, plot, ratings, length, etc.

The user interacts through a graphical interface augmented with speech, pen, and remote control input modalities. The remote control can be used to move the current focus and select items. The pen can be used both for selecting items (pointing at them) and for handwritten input. The graphical user interface has three main screens. The main screen is the search screen (Figure 2). There is also a control screen used for setting system parameters and a third comparison display used for showing movie details side by side (Figure 4). The user can select among the screens using three icons in the navigation bar at the top left of the screen. The arrows provide ‘Back’ and ‘Next’ for navigation through previous searches. Directly below, there is a feedback window which indicates whether the system is listening and provides feedback on speech recognition and search. In the tablet variant, the microphone and speech recognizer are activated by tapping on ‘CLICK TO SPEAK’ with the pen. In the remote control version, the recognizer can also be activated using a button on the remote control. The main section of the search display (Figure 2) contains two panels. The right panel (results panel) presents a scrollable list of thumbnails for the movies retrieved by the current search. The left panel (details panel) provides details on the currently selected title in the results panel. These include the genre, plot summary, cast, and director.

The system supports a speech modality, a handwriting modality, pointing (unimodal GUI) modality, and composite multimodal input where the user utters a spoken command which is combined with pointing ‘gestures’ the user has made towards screen icons using the pen or the remote control.

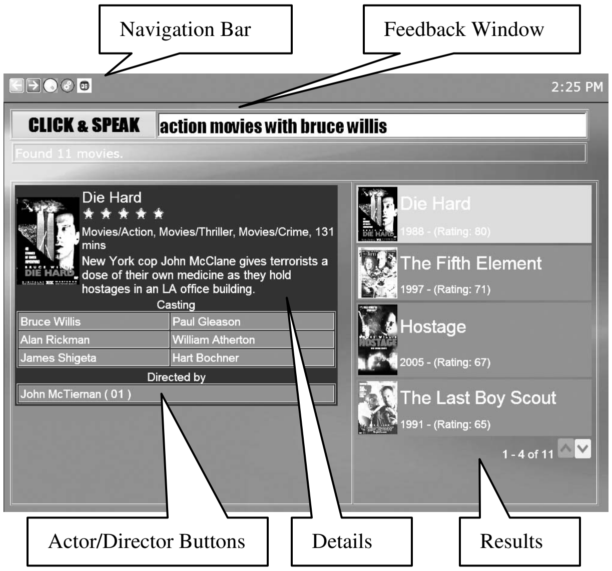

Figure 2 Graphical user interface

Speech: The system supports speech search over multiple different dimensions such as title, genre, cast, director, and year. Input can be more telegraphic with searches such as "Legally Blonde", "Romantic comedy", and "Reese Witherspoon", or more verbose natural language queries such as "I'm looking for a movie called Legally Blonde" and "Do you have romantic comedies". An important advantage of speech is that it makes it easy to combine multiple constraints over multiple dimensions within a single query (Cohen, 1992). For example, queries can indicate co-stars: "movies starring Ginger Rogers and Fred Astaire", or constrain genre and cast or director at the same time: "Meg Ryan Comedies", "show drama directed by Woody Allen" and "show comedy movies directed by Woody Allen and starring Mira Sorvino".

Handwriting: Handwritten pen input can also be used to make queries. When the user's pen approaches the feedback window, it expands allowing for freeform pen input. In the example in Figure 3, the user requests comedy movies with Bruce Willis using unimodal handwritten input. This is an important input modality as it is not impacted by ambient noise such as crosstalk from other viewers or currently playing content.

Figure 3 Handwritten query

Pointing/GUI: In addition to the recognition-based modalities, speech and handwriting, the interface also supports more traditional graphical user interface (GUI) commands. In the details panel, the actors and directors are presented as buttons. Pointing at (i.e., clicking on) these buttons results in a search for all of the movies with that particular actor or director, allowing users to quickly navigate from an actor or director in a specific title to other material they may be interested in. The buttons in the results panel can be pointed at (clicked on) in order to view the details in the left panel for that particular title.

Figure 4 Comparison screen

Composite multimodal input: The system also supports true composite multimodality when spoken or handwritten commands are integrated with pointing gestures made using the pen (in the tablet version) or by selecting items (in the remote control version). This allows users to quickly execute more complex commands by combining the ease of reference of pointing with the expressiveness of spoken constraints. While by unimodally pointing at an actor button you can search for all of the actor's movies, by adding speech you can narrow the search to, for example, all of their comedies by saying: "show comedy movies with THIS actor". Multimodal commands with multiple pointing gestures are also supported, allowing the user to 'glue' together references to multiple actors or directors in order to constrain the search. For example, they can say "movies with THIS actor and THIS director" and point at the 'Alan Rickman' button and then the 'John McTiernan' button in turn (Figure 2). Comparison commands can also be multimodal; for example, if the user says "compare THIS movie and THIS movie" and clicks on the two buttons on the right display for 'Die Hard' and the 'The Fifth Element' (Figure 2), the resulting display shows the two movies side-by-side in the comparison screen (Figure 4).

3 Underlying multimodal architecture

The system consists of a series of components which communicate through a facilitator component (Figure 5). This develops and extends upon the multimodal architecture underlying the MATCH system (Johnston et al., 2002).

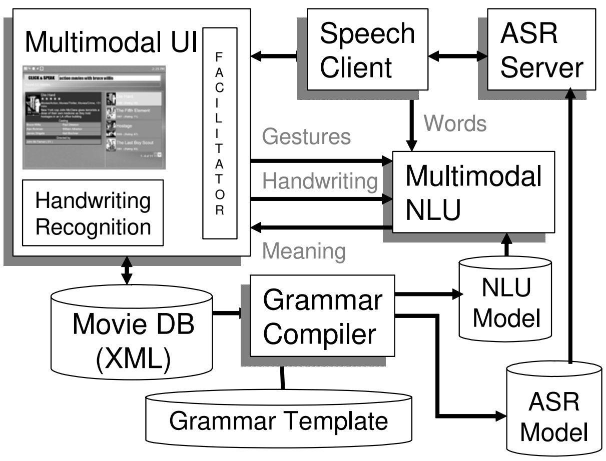

Figure 5 System architecture

The underlying database of movie information is stored in XML format. When a new database is available, a Grammar Compiler component extracts and normalizes the relevant fields from the database. These are used in conjunction with a predefined multimodal grammar template and any available corpus training data to build a multimodal understanding model and speech recognition language model.

The user interacts with the multimodal user interface client (Multimodal UI), which provides the graphical display. When the user presses 'CLICK TO SPEAK' a message is sent to the Speech Client, which activates the microphone and ships audio to a speech recognition server. Handwritten inputs are processed by a handwriting recognizer embedded within the multimodal user interface client. Speech recognition results, pointing gestures made on the display, and handwritten inputs, are all passed to a multimodal understanding server which uses finite-state multimodal language processing techniques (Johnston and Bangalore, 2005) to interpret and integrate the speech and gesture. This model combines alignment of multimodal inputs, multimodal integration, and language understanding within a single mechanism. The resulting combined meaning representation (represented in XML) is passed back to the multimodal user interface client, which translates the understanding results into an XPATH query and runs it against the movie database to determine the new series of results. The graphical display is then updated to represent the latest query.

The system first attempts to find an exact match in the database for all of the search terms in the user's query. If this returns no results, a back off and query relaxation strategy is employed. First the system tries a search for movies that have all of the search terms, except stop words, independent of the order (an AND query). If this fails, then it backs off further to an OR query of the search terms and uses an edit machine, using Levenshtein distance, to retrieve the most similar item to the one requested by the user.

4 Evaluation

After designing and implementing our initial prototype system, we conducted an extensive multimodal data collection and usability study with the two different interaction scenarios: tablet versus remote control. Our main goals for the data collection and statistical analysis were three-fold: collect a large corpus of natural multimodal dialogue for this media selection task, investigate whether future systems should be paired with a remote control or tablet-like device, and determine which types of search and input modalities are more or less desirable.

4.1 Experimental set up

The system evaluation took place in a conference room set up to resemble a living room (Figure 6). The system was projected on a large screen across the room from a couch.

An adjacent conference room was used for data collection (Figure 7). Data was collected in sound files, videotapes, and text logs. Each subject's spoken utterances were recorded by three microphones: wireless, array and stand alone. The wireless microphone was connected to the system while the array and stand alone microphones were around 10 feet away.[1] Test sessions were recorded with two video cameras – one captured the system's screen using a scan converter while the other recorded the user and couch area. Lastly, the user's interactions and the state of the system were captured by the system's logger. The logger is an additional agent added to the system architecture for the purposes of the evaluation. It receives log messages from different system components as interaction unfolds and stores them in a detailed XML log file. For the specific purposes of this evaluation, each log file contains: general information about the system's components, a description and timestamp for each system event and user event, names and timestamps for the system-recorded sound files, and timestamps for the start and end of each scenario.

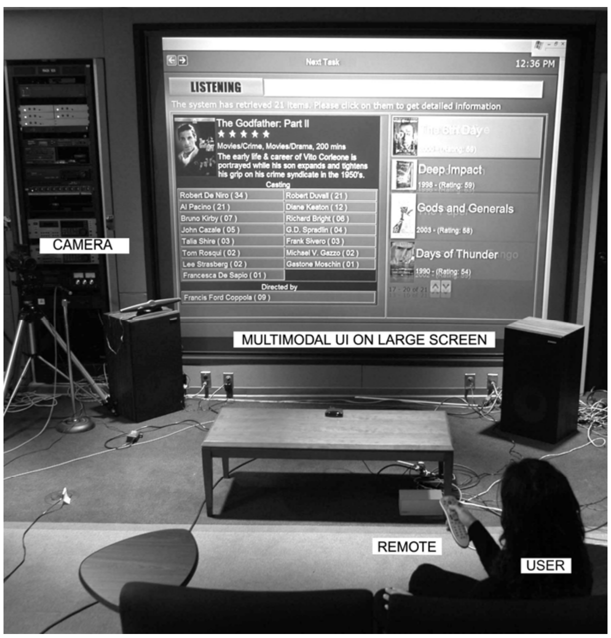

Figure 6 Data collection environment

Forty-four subjects volunteered to participate in this evaluation. There were 33 males and 11 females, ranging from 20 to 66 years of age. Each user interacted with both the remote control and tablet variants of the system, completing the same two sets of scenarios and then freely interacting with each system. For counterbalancing purposes, half of the subjects used the tablet and then the remote control and the other half used the remote control and then the tablet. The scenario set assigned to each version was also counterbalanced.

Figure 7 Data collection room

Each set of scenarios consisted of seven defined tasks, four user-specialized tasks and five open-ended tasks. Defined tasks were presented in chart form and had an exact answer, such as the movie title that two specified actors/actresses starred in. For example, users had to find the movie in the database with Matthew Broderick and Denzel Washington. User-specialized tasks relied on the specific user's preferences, such as "What type of movie do you like to watch on a Sunday evening? Find an example from that genre and write down the title". Open-ended tasks prompted users to search for any type of information with any input modality. The tasks in the two sets paralleled each other. For example, if one set of tasks asked the user to find the highest ranked comedy movie with Reese Witherspoon, the other set of tasks asked the user to find the highest ranked comedy movie with Will Smith. Within each task set, the defined tasks appeared first, then the user-specialized tasks and lastly the open-ended tasks. However, for each participant, the order of defined tasks was randomized, as well as the order of user-specialized tasks.

At the beginning of the session, users read a short tutorial about the system's GUI, the experiment, and available input modalities. Before interacting with each version, users were given a manual on operating the tablet/remote control. To minimize bias, the manuals gave only a general overview with few examples and during the experiment users were alone in the room.

At the end of each session, users completed a user-satisfaction/preference questionnaire and then a qualitative interview. The questionnaire consisted

[1] Here we report results for the wireless microphone only. Analysis of the other microphone conditions is ongoing.

of 25 statements about the system in general, the two variants of the system, input modality options and search options. For example, statements ranged from "If I had [the system], I would use the tablet with it" to "If my spoken request was misunderstood, I would want to try again with speaking". Users responded to each statement with a 5-point Likert scale, where 1 = 'I strongly agree', 2 = 'I mostly agree', 3 = 'I can't say one way or the other', 4 = 'I mostly do not agree' and 5 = 'I do not agree at all'. The qualitative interview allowed for more open-ended responses, where users could discuss reasons for their preferences and their likes and dislikes regarding the system.

4.2 Results

Data was collected from all 44 participants. Due to technical problems, five participants' logs or sound files were not recorded in parts of the experiment. All collected data was used for the overall statistics but these five participants had to be excluded from analyses comparing remote control to tablet.

Spoken utterances: After removing empty sound files, the full speech corpus consists of 3280 spoken utterances. Excluding the five participants subject to technical problems, the total is 3116 utterances (1770 with the remote control and 1346 with the tablet).

The set of 3280 utterances averages 3.09 words per utterance. There was not a significant difference in utterance length between the remote control and tablet conditions. Users' averaged 2.97 words per utterance with the remote control and 3.16 words per utterance with the tablet, paired t (38) = 1.182, p = n.s. However, users spoke significantly more often with the remote control. On average, users spoke 34.51 times with the tablet and 45.38 times with the remote control, paired t (38) = -3.921, $p < .01$.

ASR performance: Over the full corpus of 3280 speech inputs, word accuracy was 44% and sentence accuracy 38%. In the tablet condition, word accuracy averaged 46% and sentence accuracy 41%. In the remote control condition, word accuracy averaged 41% and sentence accuracy 38%. The difference across conditions was only significant for word accuracy, paired t (38) = 2.469, $p < .02$. In considering the ASR performance, it is important to note that 55% of the 3280 speech inputs were out of grammar, and perhaps more importantly 34% were out of the functionality of the system entirely. On within functionality inputs, word accuracy is 62% and sentence accuracy 57%. On the in grammar inputs, word accuracy is 86% and sentence accuracy 83%. The vocabulary size was 3851 for this task. In the corpus, there are a total of 356 out-of-vocabulary words.

Handwriting recognition: Performance was determined by manual inspection of screen capture video recordings.[2] There were a total of 384 handwritten requests with overall 66% sentence accuracy and 76% word accuracy.

Task completion: Since participants had to record the task answers on a paper form, task completion was calculated by whether participants wrote down the correct answer. Overall, users had little difficulty completing the tasks. On average, participants completed 11.08 out of the 14 defined tasks and 7.37 out of the 8 user-specialized tasks. The number of tasks completed did not differ across system variants.[3] For the seven defined tasks within each condition, users averaged 5.69 with the remote control and 5.40 with the tablet, paired t (34) = -1.203, p = n.s. For the four user-specialized task within each condition, users averaged 3.74 on the remote control and 3.54 on the tablet, paired t (34) = -1.268, p = n.s.

Input modality preference: During the interview, 55% of users reported preferring the pointing (GUI) input modality over speech and multimodal input. When asked about handwriting, most users were hesitant to place it on the list. They also discussed how speech was extremely important, and given a system with a low error speech recognizer, using speech for input probably would be their first choice. In the questionnaire, the majority of users (93%) 'strongly agree' or 'mostly agree' with the importance of making a pointing request. The importance of making a request by speaking had the next highest average, where 57% 'strongly agree' or 'mostly agree' with the statement. The importance of multimodal and handwriting requests had the lowest averages, where 39% agreed with the former and 25% for the latter. However, in the open-ended interview, users mentioned handwriting as an important back-up input choice for cases when the speech recognizer fails.

[2] One of the 44 participants videotape did not record and so is not included in the statistics.
[3] Four participants did not properly record their task answers and had to be eliminated from the 39 participants being used in the remote control versus tablet statistics.

Further support for input modality preference was gathered from the log files, which showed that participants mostly searched using unimodal speech commands and GUI buttons. Out of a total of 6082 user inputs to the systems, 48% were unimodal speech and 39% were unimodal GUI (pointing and clicking). Participants requested information with composite multimodal commands 7% of the time and with handwriting 6% of the time.

Search preference: Users most strongly agreed with movie title being the most important way to search. For searching by title, more than half the users chose 'strongly agree' and 91% of users chose 'strongly agree' or 'mostly agree'. Slightly more than half chose 'strongly agree' with searching by actor/actress and slightly less than half chose 'strongly agree' with the importance of searching by genre. During the open ended interview, most users reported title as the most important means for searching.

Variant preference: Results from the qualitative interview indicate that 67% of users preferred the remote control over the tablet variant of the system. The most common reported reasons were familiarity, physical comfort and ease of use. Remote control preference is further supported from the user-preference questionnaire, where 68% of participants 'mostly agree' or 'strongly agree' with wanting to use the remote control variant of the system, compared to 30% of participants choosing 'mostly agree' or 'strongly agree' with wanting to use the tablet version of the system.

5 Conclusion

With the range of entertainment content available to consumers in their homes rapidly expanding, the current access paradigm of direct manipulation of complex graphical menus and onscreen keyboards, and remote controls with way too many buttons is increasingly ineffective and cumbersome. In order to address this problem, we have developed a highly flexible multimodal interface that allows users to search for content using speech, handwriting, pointing (using pen or remote control), and dynamic multimodal combinations of input modes. Results are presented in a straightforward graphical interface similar to those found in current systems but with the addition of icons for actors and directors that can be used both for unimodal GUI and multimodal commands. The system allows users to search for movies over multiple different dimensions of classification (title, genre, cast, director, year) using the mode or modes of their choice. We have presented the initial results of an extensive multimodal data collection and usability study with the system.

Users in the study were able to successfully use speech in order to conduct searches. Almost half of their inputs were unimodal speech (48%) and the majority of users strongly agreed with the importance of using speech as an input modality for this task. However, as also reported in previous work (Wittenburg et al 2006), recognition accuracy remains a serious problem. To understand the performance of speech recognition here, detailed error analysis is important. The overall word accuracy was 44% but the majority of errors resulted from requests from users that lay outside the functionality of the underlying system, involving capabilities the system did not have or titles/cast absent from the database (34% of the 3280 spoken and multimodal inputs). No amount of speech and language processing can resolve these problems. This highlights the importance of providing more detailed help and tutorial mechanisms in order to appropriately ground users' understanding of system capabilities. Of the remaining 66% of inputs (2166) which were within the functionality of the system, 68% were in grammar. On the within functionality portion of the data, the word accuracy was 62%, and on in grammar inputs it is 86%. Since this was our initial data collection, an un-weighted finite-state recognition model was used. The performance will be improved by training stochastic language models as data become available and employing robust understanding techniques. One interesting issue in this domain concerns recognition of items that lie outside of the current database. Ideally the system would have a far larger vocabulary than the current database so that it would be able to recognize items that are outside the database. This would allow feedback to the user to differentiate between lack of results due to recognition or understanding problems versus lack of items in the database. This has to be balanced against degradation in accuracy resulting from increasing the vocabulary.

In practice we found that users, while acknowledging the value of handwriting as a back-up mode, generally preferred the more relaxed and familiar style of interaction with the remote control. However, several factors may be at play here.

The tablet used in the study was the size of a small laptop and because of cabling had a fixed location on one end of the couch. In future, we would like to explore the use of a smaller, more mobile, tablet that would be less obtrusive and more conducive to leaning back on the couch. Another factor is that the in-lab data collection environment is somewhat unrealistic since it lacks the noise and disruptions of many living rooms. It remains to be seen whether in a more realistic environment we might see more use of handwritten input. Another factor here is familiarity. It may be that users have more familiarity with the concept of speech input than handwriting. Familiarity also appears to play a role in user preferences for remote control versus tablet. While the tablet has additional capabilities such handwriting and easier use of multimodal commands, the remote control is more familiar to users and allows for a more relaxed interaction since they can lean back on the couch. Also many users are concerned about the quality of their handwriting and may avoid this input mode for that reason.

Another finding is that it is important not to underestimate the importance of GUI input. 39% of user commands were unimodal GUI (pointing) commands and 55% of users reported a preference for GUI over speech and handwriting for input. Clearly, the way forward for work in this area is to determine the optimal way to combine more traditional graphical interaction techniques with the more conversational style of spoken interaction.

Most users employed the composite multimodal commands, but they make up a relatively small proportion of the overall number of user inputs in the study data (7%). Several users commented that they did not know enough about the multimodal commands and that they might have made more use of them if they had understood them better. This, along with the large number of inputs that were out of functionality, emphasizes the need for more detailed tutorial and online help facilities. The fact that all users were novices with the system may also be a factor. In future, we hope to conduct a longer term study with repeat users to see how previous experience influences use of newer kinds of inputs such as multimodal and handwriting.

Acknowledgements Thanks to Keith Bauer, Simon Byers, Harry Chang, Rich Cox, David Gibbon, Mazin Gilbert, Stephan Kanthak, Zhu Liu, Antonio Moreno, and Behzad Shahraray for their help and support. Thanks also to the Dirección General de Universidades e Investigación - Consejería de Educación - Comunidad de Madrid, España for sponsoring D'Haro's visit to AT&T.

References

Elisabeth André. 2002. Natural Language in Multimodal and Multimedia systems. In Ruslan Mitkov (ed.) *Oxford Handbook of Computational Linguistics*. Oxford University Press.

Aseel Berglund. 2004. *Augmenting the Remote Control: Studies in Complex Information Navigation for Digital TV*. Linköping Studies in Science and Technology, Dissertation no. 872. Linköping University.

Philip R. Cohen. 1992. The Role of Natural Language in a Multimodal Interface. In *Proceedings of ACM UIST Symposium on User Interface Software and Technology*. pp. 143-149.

Jun Goto, Kazuteru Komine, Yuen-Bae Kim and Noriyoshi Uratan. 2003. A Television Control System based on Spoken Natural Language Dialogue. In *Proceedings of 9th International Conference on Human-Computer Interaction.* pp. 765-768.

Aseel Ibrahim and Pontus Johansson. 2002. Multimodal Dialogue Systems for Interactive TV Applications. In *Proceedings of 4th IEEE International Conference on Multimodal Interfaces*. pp. 117-222.

Pontus Johansson. 2003. MadFilm - a Multimodal Approach to Handle Search and Organization in a Movie Recommendation System. In *Proceedings of the 1st Nordic Symposium on Multimodal Communication.* Helsingör, Denmark. pp. 53-65.

Michael Johnston, Srinivas Bangalore, Guna Vasireddy, Amanda Stent, Patrick Ehlen, Marilyn Walker, Steve Whittaker, Preetam Maloor. 2002. MATCH: An Architecture for Multimodal Dialogue Systems. In *Proceedings of the 40th ACL.* pp. 376-383.

Michael Johnston and Srinivas Bangalore. 2005. Finite-state Multimodal Integration and Understanding. *Journal of Natural Language Engineering 11.2.* Cambridge University Press. pp. 159-187.

Russ Mitchell. 1999. TV's Next Episode. *U.S. News and World Report*. 5/10/99.

Thomas Portele, Silke Goronzy, Martin Emele, Andreas Kellner, Sunna Torge, and Jüergen te Vrugt. 2006. SmartKom–Home: The Interface to Home Entertainment. In Wolfgang Wahlster (ed.) *SmartKom: Foundations of Multimodal Dialogue Systems.* Springer. pp. 493-503.

Kent Wittenburg, Tom Lanning, Derek Schwenke, Hal Shubin and Anthony Vetro. 2006. The Prospects for Unrestricted Speech Input for TV Content Search. In *Proceedings of AVI'06*. pp. 352-359.