Modality Fusion in a Route Navigation System

Topi Hurtig
topi.hurtig@helsinki.fi

Kristiina Jokinen
kristiina.jokinen@helsinki.fi

University of Helsinki
PL9 (Siltavuorenpenger 20A)
FIN-00014 University of Helsinki

ABSTRACT

In this paper we present the MUMS Multimodal Route Navigation System which combines speech, pen, and graphics into a PDA-based multimodal system. We focus especially on the three-level modality fusion component which we believe provides an accurate and more flexible input fusion than the usual two-level approaches. The modular architecture of the system supports flexible component management, and the interface is designed to enable natural interaction with the user, with an emphasis on adaptation and users with special needs.

Categories and Subject Descriptors

H5.2 [**Information Systems**]: Information Interfaces and Presentation – *user interfaces.*

General Terms

Design, Experimentation, Human Factors.

Keywords

dialogue processing, human computer interaction, multimedia, user interfaces, cognitive modeling

1. INTRODUCTION

In recent years, multimodal interactive systems have become more feasible from the technology point of view, and they also seem to provide a reasonable and user-friendly alternative for various interactive applications that require natural human-computer interaction. Although speech is the natural mode of interaction for human users, speech recognition is not yet robust enough to allow fully natural language input, and human-computer interaction suffers from the lack of naturalness: the user is forced to follow a strictly predetermined course of actions in order to get a simplest task done. Moreover, natural human-human interaction does not only include verbal communication: much of the information content is conveyed by non-verbal signs, gestures, facial expressions, etc., and, in many cases, like giving instructions of how to get to a particular place, verbal explanations may not be the appropriate and most effective way of exchanging information. Thus, in order to develop next generation human-computer interfaces, it is necessary to work on technologies that allow multimodal natural interaction: it is important to investigate coordination of natural input modes (speech, pen, touch, hand gestures, eye movement, head and body movements) as well as multimodal system output (speech, sound, images, graphics), ultimately aiming at intelligent interfaces that are aware of the context and user needs, and can utilise appropriate modalities to provide information tailored to a wide variety of users. Furthermore, the traditional PC environment as a paradigm for human-computer interaction is changing: small handheld devices need adaptive interfaces that can take care of situated services, mobile users, and users with special needs. The assumption behind natural interaction is that the various users in different situations need not browse manuals in order to learn to use the digital device; instead the users could exploit the strategies they have learnt in human-human communication and thus interaction and task completion with the device would be as fluent and easy as possible.

In this paper we describe our work on a PDA-based multimodal public transportation route navigation system that has been developed in the National Technology project PUMS. The main research points in the project have been to set to:

- modality fusion on the semantic level
- natural and flexible interaction model
- presentation of the information
- usability of the system
- architecture and technical integration.

We focus on the interaction model and the modality fusion component that takes care of the unification and interpretation of the data flow from the speech recognizer and the tactile input device. The fusion component works on three levels instead of the conventional two. We also discuss the architecture of the system. The paper is organized as follows. In Section 2 we first discuss the related research and set the context of the current research. Section 3 presents the MUMS multimodal navigation system, especially its interface and architecture. Section 4 takes a look at the modality fusion component, and section 5 finally draws conclusions and points to future research.

2. RELATED RESEARCH

An overview of multimodal interfaces can be found e.g. in [9]. Most multimodal systems concentrate on speech and graphics or tactile input information, and the use of speech and pen in a multimodal user interface has been extensively studied. For instance, Oviatt and her colleagues [11; 12] studied the speech and pen system Quickset, and found out that the multimodal system can indeed help disambiguating input signals, which improves the system's robustness and performance stability. Gibbon et al. [3] list several advantages of multimodal interfaces. E.g. the use of different modalities offers different benefits and also the freedom of choice: it is easier to point to an object than talk about it by speaking, and the users may also have personal preferences of one modality over another. Jokinen and Raike [7] point out that multimodal interfaces have obvious benefits for users with special needs who cannot use all the communication modes.

On the other hand, there are also disadvantages of multimodal interfaces: coordination and combination of modalities requires special attention on the system as well on the interpretation level, and from the point of view of usability, there is a danger that the users are exposed to cognitive overload by the stimulation of too many media. Especially in route navigation tasks, the system should guide the user accurately and quickly and provide necessary assistance in cases which are likely to be complicated and confusing (e.g. in our case provide information about the number of bus or tram changes the user needs in order to get to her destination), and also allow several levels of details to be included in the guidance depending on the user's needs (Section 3.3).

The system described in this paper is based on Interact-system [6] which aimed at studying methods and techniques for rich dialogue modelling and natural language interaction in situations where the interaction had not been functional or robust enough. The application dealt with public transportation in a city and the system provided information about bus routes and timetables. The system also showed some basic multimodal aspects in that an interactive map was demonstrated together with the mainly speech-based interface. In the follow-up project PUMS, the main goal of research is to integrate a PDA-based graphical point-and-click interface with the user's speech input, and to allow the system to output in speech as well as drawing on the map. Besides the technical challenges, an important goal is also to investigate possibilities for natural interaction in a route navigation task where the system is to give helpful information about the route and public transportation.

3. MULTIMODAL NAVIGATION

3.1 Interface

A sample dialogue is shown in Example 1, and the corresponding map input in Figure 1.

Example dialogue 1: The user first receives information about the transportation possibilities, and then asks the system to instruct how to get to the desired destination.

U: Uh, how do I get from the Railway station ... uhh

S: Where was it you wanted to go?

U: Well, there! + <map gesture>

S: Tram 3B leaves Railway Station at 14:40, there is one change. Arrival time at Brahe Street 7 is 14:57.

U: When does the next one go?

S: Bus 23 leaves Railway Station at 14:43, there are no changes. Arrival time at Brahe Street 7 is 15:02.

U: Ok. Navigate.

S: Take bus 23 at the Railway Station at 14:43.

U: Navigate more.

S: Get off the bus at 14:49 at the Brahe Street stop.

U: Navigate more.

S: Walk 200 meters in the direction of the bus route. You are at Brahe Street 7.

U: How long does it take?

S: It takes 19 minutes.

U: I see, ok.

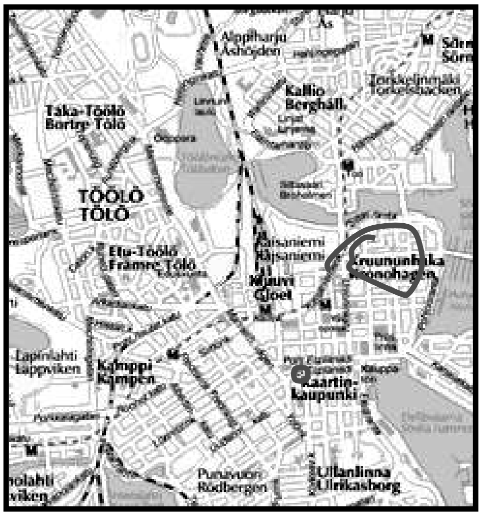


Figure 1. Sample tactile input.

Because of the rather limited functionality and task-specific nature of the system, the user is limited to a handful of ways of forming a spoken route enquiry. This will reduce the load of the speech recognizer resulting in a more robust recognition process. The touch-screen map interprets all tactile input as location data, so a tap on the screen denotes a pinpoint coordinate location, whereas a circled area will be interpreted as a number of possible locations. The map can also be freely scrolled and zoomed in real time. The inputs are recorded simultaneously and time stamped for later modality fusion phase processing.

Users are not restricted to any specific timing or form of input. The dialogue is initiated by the user, who can e.g. make corrections to already submitted input at any time, as shown in example dialogue 2.

Example dialogue 2: The user corrects the departure location as the system has made a recognition error.

U: How do I get from the Opera House to the Central Bus Station?

S: Bus 43 leaves Orton School at 11:40, there are no changes. Arrival time at the Central Bus Station is 11:54.

U: No, I wanted to leave from the Opera House!

S: Tram 6 leaves the Opera House at 11:41, there are no changes. Arrival time at the Central Bus Station is 11:51.

S: Ok, great.

In addition to requesting route guidance, the user can, as shown in Example 1, also present questions about route details: travel times, distances, the stop count, etc

3.2 System Description

The system consists of a PDA client device and a remote system server. The system server handles all processing of the user-provided information, and, apart from a light-weight speech synthesizer, the PDA can be considered only a simple user interface. The system is built on the Jaspis architecture [14], which is a flexible distributed and modular platform designed originally for speech systems. However, due to its configurability, it has been modified for the use of multiple modalities.

The system is connected to an external routing system and database, which returns, for each complete query, a detailed set of route information in XML format. This information is stored in a local database and is used for creating route summaries and providing the user with detailed route information. A high-level diagram of the system architecture is shown in Figure 2.

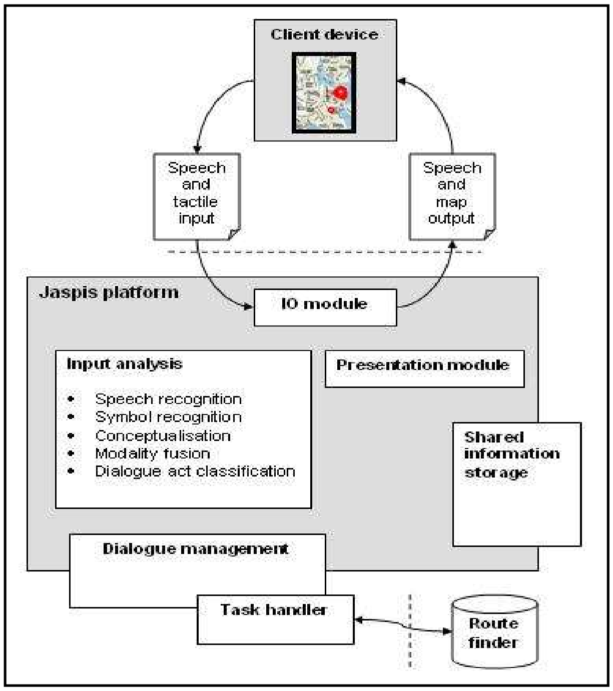


Figure 2. System architecture

The processing of received user input begins with the recognition of each modality and the attaching of high-level task-relevant concepts, e.g. "explicit_speech_location" to input units. The next phase, the fusion of modalities results in an N-best list of user input candidates. In the final phase of input processing, a dialogue act type is attached to each of the fused inputs. The process then advances to the dialogue management module, which, while being able to access the dialogue history, attempts to determine user intentions, chooses the input candidate that best fits the situation and task at hand, and carries out the corresponding task. These processes are explained in detail in Section 4. Depending on the content of the user input and the state of the dialogue, the dialogue management module forms a generic response, which is then accessed by the presentation module. The presentation module formats the response according to the set user preferences and the client hardware in use, after which the information is ready to be sent and presented in the client device.

3.3 Route Navigation

The system can perform two basic tasks: provide timetable information for public transportation and provide navigation instructions for the user to get from a departure place to a destination. In order for the system to be able to retrieve route information, at least the departure and arrival locations must be provided by the user. If the user does not provide all necessary information to execute a full database query, the system prompts the user for the missing information. As shown in Examples 1 and 2, the user can provide information either by voice or a map gesture, and the user can also correct or change the parameter values. When all necessary information has been collected, the route details will be fetched from the route database.

As pointed out by [2], one important aspect in route navigation is to be able to give the user information that is suitably chunked. The initial response the system produces for a valid route query is a route summary, based on which the user is able to accept or decline the fetched route. The spoken summary contains the time of departure, the form of transportation, the line number (where applicable), the number of changes from a vehicle to another, the name or address of the final destination, and the time of arrival. The route suggestion is also displayed on the map as shown in an example in Figure 3.

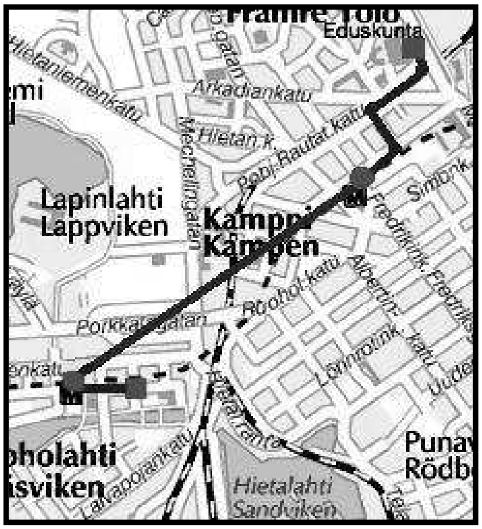


Figure 3. Sample graphical representation of a route (verbal route description in example dialogues 3 and 4).

The summary is kept as brief and informative as possible, since it also functions as an implicit confirmation for the route information. The user can also traverse the list of possible routes with simple commands, e.g. "*next*" and "*previous*", as shown in Example 1.

If the user is content with a provided route suggestion, navigation can be initiated by the command "*navigate*". Depending on the set user preferences, a suitable amount of navigation information is then presented in the desired level of detail. Sample dialogues of the navigation levels *default* and *detailed* are shown in Examples 3 and 4.

Example dialogue 3: default navigation level.

U: Navigate.
S: Take the subway from the Kamppi stop at 07:43.
U: Navigate more.
S: Get off the subway at 07:46 at the Ruoholahti stop. Santa Street 5 is located 50 meters to the East.
U: Ok, thanks.

Example dialogue 4: detailed navigation level.

U: Navigate.
S: Walk 50 meters south and take a right turn.
U: Navigate more.
S: Walk 100 meters west and take a left turn.
U: Navigate more.
S: Walk 100 meters south and take a right turn. The Kamppi subway station is 50 meters in front of you.
U: Navigate more.
S: Take the subway from the Kamppi station at 07:43.
U: Navigate more.
S: Get off the subway at the first stop, Ruoholahti, at 07:46.
U: Navigate more.
S: Santa Street 5 is located 50 meters to the East.
U: Ok.

The default navigation level is intended for users familiar with the system and comfortable with traveling in the area, whereas the detailed navigation level provides valuable details for novice users and for user with special needs (e.g. visually impaired users). The default navigation is also preferable in route planning stage when the user is more interested in getting the useful information in advance than using the system for online guiding. Although in these cases the user can set the navigation level herself, we can also envisage that it would be possible for the system to adapt itself, by its knowledge of the particular situation and learning through interaction with the user, when to switch to a more detailed navigation mode.

As pointed out by [8], using natural language to give route descriptions is a challenging task due to the dynamic nature of the navigation itself: online navigation requires that the system must not focus only on the most relevant facts, but on the facts which are most salient for the user in a given situation. Of course, it is not possible to use knowledge of salient landmarks in MUMS, as it is impossible to determine exactly what is visible for the user. However, as we mentioned earlier, from the usability point of view it is important that the information through different media in multimodal systems is unambiguous and coordinated in a way that the user finds satisfying, and especially that verbal descriptions take into account those important elements that are available and "visible" in the environment. Cognitive aspects of dynamic route descriptions can thus be exploited in the MUMS system so as to design system output that is clear and transparent. For instance, route instructions are generated with respect to the landmarks and their relative position in regard to the user ("on your right", "in front of you", "first stop"), and in accordance with the changes in the user's current state ("Walk 50 meters south and take a right turn"). Route descriptions are supported by the back-end database (kindly provided by the Helsinki City Transportation Authority), and it contains information about the landmarks such as the main sightseeing points, buildings, hotels and shops. The database also contains distances, and although the meter-wise accurate walking instructions may not be realistic, they can be used in the application, since the users are already familiar with this type of information through the popular web-based interface.

4. INFORMATION FUSION

4.1 Modality Fusion

One of the central tasks in a multimodal system is carried out by the modality fusion component. Challenges are faced not only in the low-level integration of signals but rather in the construction of the user's intended meaning from the meanings contributed by parallel input modes. The classic example of coordinated multimodal input is Put-That-There –system by [1], which combined spoken commands with hand gestures so as to enable to user to manipulate block world items. In CUBRICON [10], the user could coordinate speech and gestures in a military planning task. In the QuickSet architecture [4; 5], speech and pen input are integrated in a unification-based model where multimodal information is represented as typed feature structures. Feature structures support partial meaning representation, and unification takes care of the combination of compatible structures, thus facilitating consistent structure building from underspecified meanings. In the MUMS-system, however, the semantic information from the input modalities is not represented as typed feature structures but as a rather modest semantic representation, and thus the modality integration is developed to be a three stage process where each level transforms and manipulates the incoming information so as to provide the combined meaning representation for the user input.

In technological projects that have focused on building large multimodal systems, such as the SmartKom project [15], modality integration takes place in the backbone of the system and is divided on different sub-levels due to practical reasons. In practise it does not seem possible to work in a sequential way by unifying more and more consistent information so as to reach the appropriate interpretation of the user intentions, but integration seems to take place on different levels depending on the information content. For instance, it may be possible to integrate utterances like "*From here* + <point to a map place>" by rather low-level fusion of information streams, but it may not be possible to interpret *"I want to get there from here* + <point>" in a similar way, without having access to discourse level information that confirms that the first location reference "*there*" refers to the location given earlier in the dialogue as the destination, and is thus not a possible mapping for the pointing gesture. Blackboard architectures, like Open Agent Architecture, thus seem to provide a more useful platform for multimodal systems which need asynchronous and distributed processing.

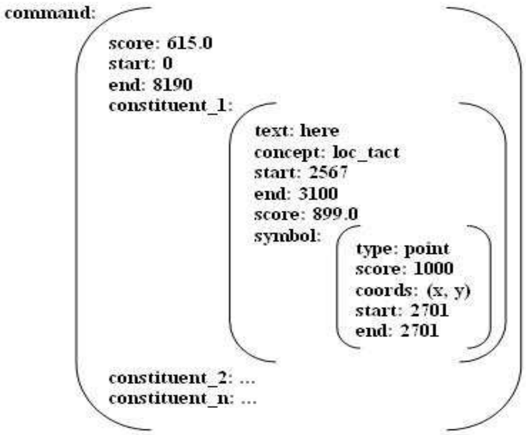
4.2 Three Level Fusion

Approaches to multimodal information fusion can be divided into three groups, depending on the level of processing involved: signal level fusion, feature-level fusion, and semantic fusion. In semantic fusion, the concepts and meanings extracted from the data from different modalities are combined so as to produce a single meaning representation for the user's action. Usually semantic fusion takes place in two levels [3]: first multimodal inputs are combined and those events that belong to the predefined multimodal input events are picked up, and then the input is handed over to the higher-level interpreter which uses its knowledge about the user intentions and context to finalize and disambiguate the input.

We introduce a three-level modality fusion component which consists of a temporal fusion phase, a statistically motivated weighting phase, and a discourse level phase. These phases correspond to the following levels of operation:

- production of legal combinations
- weighting of possible combinations
- selection of the best candidate.

In our implementation, the first two levels of fusion take place in the input analysis phase, and the third level fusion takes place in the dialogue manager. After recognition and conceptualization each input type contains the recognition score and a timestamp for each concept. The first level of fusion consists of using a rule-based algorithm for finding out all ways of legally combining the information (concepts) in the two input modalities (voice and pointing gesture), creating an often large (> 20) set of input candidates. The only restriction is that in a single modality the temporal order of events must be preserved. The formalism is now based on location-related data only, but can be easily configured to support variable types of information. An example of a single candidate (command) is shown in Figure 4.



```
command:
    score: 615.0
    start: 0
    end: 8190
    constituent_1:
        text: here
        concept: loc_tact
        start: 2567
        end: 3100
        score: 899.0
        symbol:
            type: point
            score: 1000
            coords: (x, y)
            start: 2701
            end: 2701
    constituent_2: ...
    constituent_n: ...
```

Figure 4. The structure of a user command.

In the second level of fusion, all the input candidates created in level 1 undergo a weighting procedure based on statistical data. Three kinds of weighting types are at the moment in use, each of which contains multiple parameters:

- overlap
- proximity
- concept type

Overlap and proximity have to do with the temporal qualities of the fused constituents. As often suggested, for example by [13], temporal proximity is the single most important factor in combining constituents in speech and tactile data. Examples of concept types are implicitly named locations, e.g. *"Brahe Street 7"* and location gestures. The weighting is carried out for each fused pair in each input candidate, based on which the candidate is then assigned a final score. An N-best list of these candidates is then passed on to the third and final level.

In the third level, dialogue management attempts to fit the best ranked final candidates to the current state of dialogue. If a good fit, a candidate is chosen and the system will form a response. If not, the next candidate in the list will be evaluated. Only when none of the candidates can be used, the user will be asked to rephrase or repeat his/her question. A more detailed description of the fusion component can be found in [].

5. DISCUSSION AND CONCLUSIONS

We have presented the MUMS system, which provides the user with a helpful and natural route navigation service. We have also presented the system's interaction model and its three-level modality fusion component. The fusion component consists of a temporal fusion phase, a statistically motivated second phase, and a third discourse level phase. We believe that the fusion component provides accurate and more flexible input fusion, and that the component architecture is general enough to be used in other similar multimodal applications as well.

We aim at studying the integration and synchronisation of information in multimodal dialogues further. The system will be extended to handle more complex pen gestures, such as areas, lines and arrows. As the complexity of input increases, so does the task of disambiguation of gestures with speech. Temporal disambiguation has also been shown to be problematic; even though most of the time speech precedes the related gesture, sometimes this is not the case. Taking all these situations into account might result in doubling of modality combinatorics.

Since multimodal systems depend on natural interaction patterns, it is also important to study human interaction modes and gain more knowledge of what it means to interact naturally: what are the users' preferences and what are appropriate modality types for specific tasks. Although multimodality seems to improve system performance, the enhancement seems to apply only on spatial domains, and it remains to be seen what kind of multimodal systems would assist in other, more information-based domains.

We have completed the usability testing of the system as a whole. The targeted user group for the MUMS system is mobile users who quickly wish to find their way around. The tests were conducted with 20 participants who were asked to use the system in scenario-based route finding tasks. The test users were divided into two groups, and the first one was instructed to use a speech-based system with multimodal capabilities, while the other one was told to use a multimodal system which one can talk to. The tasks were also divided into those that were expected to be preferable for one or the other input modes, and those that we considered neutral with respect to the input mode so as to assess the users' preferences and the effect of the users' expectations on the modalities. The results show that the system itself worked fine, although sometimes the connection to the server was slow or unstable. Speech recognition errors also caused problems and the users were puzzled at the repeated questions. There was a preference for the tactile input although we had expected the users to resort to the tactile mode more often in case of verbal communication breakdowns. On the other hand, tactile input was also considered a new and exciting input mode, and this newness aspect may have had influenced the users' evaluation. In general, all users considered the system very interesting and fairly easy to use. The detailed analysis of the evaluation tests will be reported in the project technical report.

Finally, another important user group for the whole project is the visually impaired, whose everyday life would greatly benefit from an intelligent route navigation system. The work is in fact conducted in close collaboration with visually impaired users, and we believe that the Design-for-all principles will result in building better interfaces for "normal" users, too, especially considering verbal presentation of the navigation information and naturalness of the dialogue interaction.

6. ACKNOWLEDGMENTS

The research described in this paper has been carried out in the national cooperation project PUMS (New Methods and Applications for Speech Recognition). We would like to thank all the project partners for their collaboration and discussions.

7. REFERENCES

[1] R.A. Bolt. Put-that-there: Voice and gesture at the graphic interface. *Computer Graphics*, 14(3):262-270, 1980.

[2] Cheng, H., Cavedon, L. and Dale, R. Generating Navigation Information Based on the Driver's Route Knowledge. In B. Gambäck and K. Jokinen (eds.) *Procs of the DUMAS Final Workshop Robust and Adaptive Information Processing for Mobile Speech Interfaces*, COLING Satellite Workshop, Geneva, Switzerland. pp. 31-38, 2004.

[3] Gibbon, D., Mertins,I. and Moore, R (eds.). *Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology, and Product Evaluation*. Kluwer, Dordrecht, 2000.

[4] Johnston, M., Cohen, P.R., McGee, D., Oviatt, S., Pittman, J. and Smith, I. Unification-based multimodal integration. *Procs of the 8th conference on European chapter of the Association for Computational Linguistics*, 281-288, Madrid, Spain, 1997.

[5] Johnston, M. Unification-based multimodal parsing. *Procs of the 36th conference on Association for Computational Linguistics*, 624-630, Montreal, Canada, 1998.

[6] Jokinen, K., Kerminen, A., Kaipainen, M., Jauhiainen, T., Wilcock, G., Turunen, M., Hakulinen, J., Kuusisto, J., Lagus, K. Adaptive Dialogue Systems - Interaction with Interact. The 3rd SIGdial Workshop on Discourse and Dialogue, Philadelphia, U.S., 2002.

[7] Jokinen, K. and Raike, A. Multimodality – technology, visions and demands for the future. *Proceedings of the 1st Nordic Symposium on Multimodal Interfaces*, Copenhagen, September 2003.

[8] Maass, W. From Visual Perception to Multimodal Communication: Incremental Route Descriptions. In Mc Kevitt, P. (ed.), *Integration of Natural Language and Vision Processing: Computational Models and Systems*, Volume 1, pp. 68-82. Kluwer, Dordrecht, 1995.

[9] Maybury, M. and Wahlster, W. Readings in Intelligent User Interfaces. Morgan Kaufmann, Los Altos, California, 1998.

[10] Neal, J.G. and Shapiro, S.C. Intelligent Multi-media Interface Technology. In J.W. Sullivan and S.W. Tyler (eds.) *Intelligent User Interfaces*, Frontier Series, ACM Press, New York. pp. 11-43, 1991.

[11] Oviatt, S. *Advances in Robust Processing of Multimodal Speech and Pen Systems.* In Yuen, P.C. and Yan, T.Y. (eds.) Multimodal Interfaces for Human Machine Communication. World Scientific Publisher, London, UK, 2001.

[12] Oviatt, S., Cohen, P.R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J. and Ferro, D. Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions. *Human Computer Interaction*, 15(4): 263-322, 2000.

[13] Oviatt, S., Coulston, R., Lunsford, R. When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. *Proceedings of the Sixth International Conference on Multimodal Interfaces* (ICMI 2004), Pennsylvania, USA, October 14-15, 2004.

[14] Markku Turunen. A Spoken Dialogue Architecture and its Applications. PhD Dissertation, University of Tampere, Department of Computer Science A-2004-2, 2004.

[15] Wahlster, W., Reithinger, N. and Blocher, A. SmartKom: Multimodal Communication with a Life-Like Character. In Proceedings of Eurospeech2001, Aalborg, Denmark, 2001.