Towards Multimodal Interaction with an Intelligent Room

Petra Gieselmann, Matthias Denecke

Interactive Systems Lab
Universität Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
petra@ira.uka.de

Interactive Systems Lab
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, 15213
denecke@cs.cmu.edu

Abstract

There is a great potential for combining speech and gestures to improve human computer interaction because this kind of communication resembles more and more the natural communication humans use every day with each other. Therefore, this paper is about the multimodal interaction consisting of speech and gestures in an intelligent room. The advantages of using multimodal systems are explained and we present the gesture recognizer and the dialogue system we use. We explain how the information from the different modalities is parsed and integrated in one semantic representation.

1. Introduction

Nowadays, there is an increasing demand for a human centered system architecture with which humans can naturally interact so that they do no longer have to adapt to the computers, but vice versa [1]. Therefore, it is important that the user can interact with the system in the same way as with other humans - via speech and gestures. This kind of multimodal human-machine interaction facilitates the communication for the user of course, whereas it is quite challenging from the system's point of view. For example, we have to cope with spontaneous speech and gestures, bad acoustical and visual conditions, different dialects and different light conditions in a room and even ungrammatical or elliptical utterances which still have to be understood correctly by the system. Therefore, we need a multimodal interaction where missing or wrongly recognized information could be resolved by adding information from other knowledge sources.

Within the European Union funded project FAME (Facilitating Agent for Multicultural Exchange), we are developing an intelligent meeting room which serves as an information butler; it assists the user by providing information, switching on or off different devices, playing media files, retrieving information in the internet, giving some information on the available media which could be played, displaying some pictures or video streams via a beamer, making tea or coffee, informing the user about the general abilities of the system, etc. Meetings and lectures can take place in this room and whenever the users explicitely or implicitely require additional information from the room or tasks to be done by the room such as changing the state of a device in the room, this information butler comes into consideration.

This paper is about multimodal dialogue management with integrated speech and gestures for a successful human machine communication in intelligent rooms. The following section gives an overview of the advantages of using speech and gesture in multimodal interaction and the reasons why multimodal interaction is more efficient for the user than unimodal one are explained in detail. The third section deals with our dialogue system ARIADNE and its resources for dialogue processing. Then the gesture recognizer is shortly explained. In the fifth section, we describe how multimodal parsing works and how the different modalities are integrated. The sixth chapter contains a conclusion and an outlook of further activities in this area.

2. Advantages of multimodal interaction

A classical example for multimodal man machine interaction with speech and gestures is Bolt's ”Put that there” [2]. Since that time, lots of research has been done in the area of speech recognition and dialogue management (For more details see [3], [4], [5], [6]) so that we are now in the position to integrate continuous speech and to have a more natural interaction. Although the technology was much worse in these times, the vision was very similar: to build an integrated multimodal architecture which fulfills the human needs.

The two modes can complement each other easily so that ambiguities can be resolved by sensor fusion. This complementarity has already been evaluated by different researchers such as [6], [7], etc. and the results showed that users are able to work with a multimodal system in a more robust and stable way than with a unimodal one. Both modes could therefore serve for mutual disambiguation. For example, gestures can easily complement to the pure speech input for anaphora resolution. From a purely linguistic point of view, it is quite difficult to resolve pronouns because there are often different candidates for the antecedent of the pronoun in the discourse. This task is of course a lot easier with complementing pointing gestures which determine about which object the user is talking.

Another reason for multimodal interaction, as explained in

[8], is the fact that in some cases the verbal description of a specific object is too long or too complicated compared to the corresponding gesture and in these cases humans tend to prefer deictic gestures than spoken words. On the other hand, there are also some cases where deictic gestures are not used because the object in question is too small, it is too far away from the user, it belongs to a group of objects, etc.; here, also the principles of Gestalt theory have to be taken into account which determine whether somebody pointed to a single object or to a group of objects eg. (for more details on Gestalt theory see [9]).

Besides, there is also empirical evidence for the fact that the user performance is better in multimodal systems than in unimodal ones, as explained in [6], [10], [5] etc. Of course, it is not per se better to have a multimodal system than a unimodal one, but it depends on the type of action being performed by the user. As already mentioned by [11], gesture input is advantageous, whenever spatial tasks have to be done. Although there are no actual spatial tasks in our case, there are some situations where the verbal description is much more difficult than a gesture and in these cases, users prefer gestures.

Speech seems to be the more important modality which is supported by gestures as in natural human human communication [12]. This means that the spoken utterance guides the interpretation of the gesture; for example, the use of demonstrative pronouns indicates the possible appearance of a gesture. Furthermore, different studies can be found which prove that speech and gestures are coexpressive (see for example [13] or [14]) which means that they present the same semantic idea, although different modalities are used. Therefore, multimodal fusion should be done on semantic level and not on the level of the modes as it is done in [2] for example.

Different classifications of gestures can be found in the literature (cf. [13], [14], etc.). Most of them order the gestures according to their function from pure gesticulation via semaphoric gestures to real sign language. For the time being, we want to concentrate on deictic gestures as a good starting point for the multimodal integration. But since these gestures only represent a small part of the whole human computer interaction [11], the system will be extended to handle all the different types of gestures and also other information given by cameras as well.

The advantage of using multimodal systems compared to unimodal ones lies not only in the fact that there are more efficient, but also the number of task-critical errors can be reduced, the systems are more flexible and most of the users prefer multimodal systems [11]. Furthermore, as we have already seen, the combination of both modalities depends very much on the semantics. It has also already been evaluated that multimodal input seems to be complementary in content, not redundant [11] which means that the combination of both modalities has to take into account both modes at the same time.

3. The dialogue manager ARIADNE

We use the language and domain independent dialogue manager ARIADNE [15]. This dialogue manager is specifically tailored for rapid prototyping because only the domain and language dependent components have to be implemented for a new application, whereas the general concepts are already available and can be reused. Therefore, vectorized context-free grammars and inheritance mechanisms are used which are explained below. Besides, possibilities to evaluate the dialogue state and general input and output mechanisms are already implemented which can then be applied in the actual application.

Furthermore, multidimensional feature structures are used [16] which means that not only semantic information can be saved at the nodes of the tree but also information on the input modality and for example also confidence measures of this input. In this way, it is possible to ask the user specifically for the words which could only be recognized with a very low confidence measure for example.

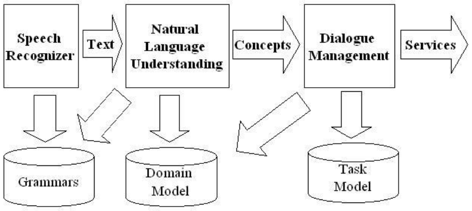

Figure 1: The Dialogue Manager and its Resources

The dialogue manager uses different kinds of task and domain dependent resources (see figure 1): An ontology, a specification of the dialogue goals, data base rules, a grammar and generation templates. Besides, the dialogue strategy decides how new information could be interpreted and integrated.

3.1. Dialogue grammar and domain model

First of all, the input of the user is parsed by means of a context-free grammar which is enhanced with information from the domain model. This domain model contains all the different concepts the system knows for understanding the users' utterances; it is build up as an ontology with objects, actions and properties which could inherite from each other. Therefore, it is also possible to access the domain independent general ontology which consists of concepts such as different speech acts and general goals, objects and properties from which specific objects, actions and properties could then inherit in the domain dependent part. This combination of grammar and domain model is possible by means of the vectorized context-free grammars which consist of non-terminals of n-dimensional vectors of partially organized elements [17].

Furthermore, it is in this way also possible to separate syntactic from semantic information in the grammar. Because of this separation, the re-use of syntactic information is possible. Therefore, the construction of complex noun phrases for example can be found in a general part of the grammar whereas the actual semantic instantiations of the objects are in the domain dependent part.

This grammar used by the dialogue manager can be converted in a non-vectorized context-free grammar and used in this way by the speech recognizer so that both components use the same linguistic knowledge base.

3.2. Task model

The task model specifies dialogue goals which can be seen as the description of a form which is filled by means of the dialogue between human and machine [15]. This means that the dialogue goals are specified by the information the user gives in the discourse and that they consist of objects, actions and properties which are defined in the ontology. Therefore, the dialogue goals are the connection between the domain model and the services the dialogue manager can execute.

If a dialogue goal is recognized, the dialogue manager searches for the corresponding parameters in the discourse, such as objects, properties and actions. If the feature structure is still underspecified, a clarification dialogue is initiated.

3.3. Generation templates

All the information from the room to the user is given in natural language. Therefore, so called generation templates are used in which the utterance of the dialogue manager depending on the dialogue state and the current situation is defined. This dialogue state is defined by the information in the dialogue goals [15].

Therefore, in these generation templates it is determined on one hand what the room asks the user in which situation and on the other hand, it is specified here what the dialogue manager expects as an answer and how this new information is integrated in discourse.

3.4. Databases

The database contains objects which are in the room and their properties. In this way, the dialogue manager can search for different instances of objects and their places in such a database. This is of special importance for the integration of gesture information in the dialogue manager, as you will see below.

3.5. Dialogue strategy

Finally, the dialogue strategy defines how the different kinds of information are evaluated in a specified dialogue state. It consists of different interaction patterns which define how information can be added respectively deleted in the discourse.

4. The gesture recognizer

The gesture recognizer detects pointing gestures by means of a stereo camera in the room. Therefore, a user model is created first which consists of the positions of the head and the hands of the user. Pointing gestures can then be recognized in real time by means of Hidden Markov Models which has been trained on example trajectories. The determination of the pointing direction is done by the line from head to hand.

The gesture recognizer sends the point of the user's hand and the pointing direction to the dialogue manager which searches in its database of all the objects in the environment whether it is pointed to a specific object it knows or not. If no object can be found, at which the user might have been pointed, the gesture event is ignored and it is assumed that it was an accidental gesture; therefore, the error of the ideal line pointing to this special object is calculated and if the error is more than $45\,^{\circ}$, it is assumed that the user did not want to point to this object. If an object is found at which the user might have been pointed, but the confidence is quite low, the dialog manager asks the user whether he really meant this object. And if the error is of course low, it is assumed that the user wants to point to this object and the semantic representation of this object is put in the discourse of the dialogue manager and this information can then be unified with the information from the speech recognizer.

5. Multimodal integration

First of all, a short example should illustrate the advantages of multimodal interaction in such an intelligent room: When the user enters the room, the system asks him what it can do for him. The user can then give some explicite commands, such as switching on the beamer he points to, or he can ask for help when he does not know how to connect his laptop to the beamer for example. Besides, implicite commands are also recognized by the system; this means that when the user complains for example that it is too dark in the room, the system will ask him, whether it should switch on the lights. Furthermore, the user can point to the VCR for example and asks what the system can do with this device, and the system answers that videos can be played, recorded, etc. This means that the user can interact with the room in the same way as with a human assistant helping him with the different technical devices.

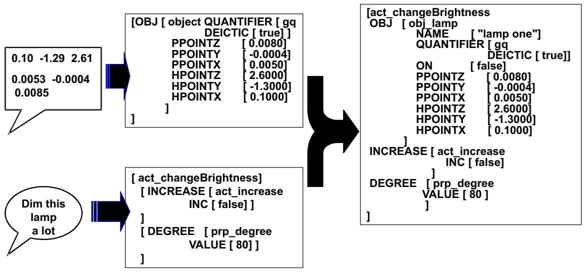

Figure 2: Multimodal Integration of Speech and Gestures

There are different possibilities how speech and gesture are combined and how they interfere in multimodal interaction:

- Speech only: The user just says something without using his hands at all.
- Gesture only: The user is pointing to a specific object; he might also say something, but this does not carry any additional semantic information. This is above all the case when the user answers to a clarification question from the dialogue manager. For example, imagine that the room understood that the user wanted to switch on something, but did not understand which object should be switched on; then the dialogue manager asks the user "What do you want me to switch on?" and the user points to the object and says "this one". In this case, the gesture carries all the information and the dialogue manager can even ignore the speech input.
- Speech and Gesture: The user says something and points to a specific object. In this case, both input streams have to be merged, as you can also see on figure 2, where the user says "dim this lamp a lot" and points to the lamp at the same time. In this case, the information from the speech input is the action to be executed, whereas the object which should be manipulated is given by the gesture.

Of course, the first two possibilities are actually unimodal, the only problem here is that additional information from the other modality has to be ignored. But the actual multimodal interaction lies in the third possibility. In this case, both information have to be merged semantically, regardless of their modality.

More specifically, this means that both recognizers, the speech recognizer and the gesture recognizer, send their input to the dialogue manager and then the input from the speech recognizer is analysed by means of the context-free grammar and the domain model and a semantic representation is constructed and compared against the dialogue goals. At the same time, the information from the gesture recognizer is processed by the dialogue manager and it is checked whether the user might have been pointed to an object in the room. If so, the semantic representation of this object is also put in the discourse and the information is unified.

Furthermore, there are special dialogue goals which compare whether the gesture and the spoken object matches. In this way, the semantics of the different modalities can be taken into account.

The difference compared to pure spoken dialogue management lies in the fact that now two input streams have to be considered at the same time - or nearly at the same time taking into account that there is empirical evidence that gestures normally precede the corresponding spoken utterance [11]. Therefore, we are able to process two input streams at one turn which are then interpreted jointly for the sake of the multimodal interaction.

6. Conclusions

The system presented here is a first step towards a complete multimodal interaction with speech and gestures in intelligent rooms. We started with a combination of speech and pointing gestures and showed in this scenario how both modalities can be successfully integrated by taking into account the semantics of the information from the different modalities.

By means of such an information butler in intelligent rooms, we want to cut the cord that tied users of a multimodal system to classical input devices such as mouse, pens or other pointing devices and let them use just their hands which leads to a more natural and easier communication of course. Future work will be done in the area of integrating more cameras to be able to analyse also other gestures and focus of attention. Besides, the possibility to use two or even more gestures during one utterance will be integrated.

An interesting research topic for the future is also the assertion that multimodal speech is different from unimodal one which means that within multimodal interaction language seems to be briefer and syntactically simpler [11].

Furthermore, there is some empirical evidence that there are large individual differences between users in timing of speech and gestures: Some tend to show a simultaneous behaviour whereas others are interacting more in a sequential way [11]. But since this study was done with pen input and not with real gesture input which is of course more natural, it might be interesting to see whether these individual differences could also be noticed in our system. Then user models could be used to help here.

7. Acknowledgements

Part of this work was carried out within the FAME project (Facilitating Agent for Multicultural Exchange) and has been funded by the European Union as IST-2000-28323.

8. References

[1] Coen, M.H.,“The Future of Human-Computer-Interaction or How I learned to stop worrying and love my Intelligent Room”, In: IEEE Intelligent Systems, March/April 1999.

[2] Bolt, R.A., “Put that there: Voice and Gesture at the Graphics Interface”, In: Computer Graphics, 14(3), pp. 262-270, 1980.

[3] . Johnston, M., Cohen, P., McGee, D., Oviatt, S., Pittman, J., Smith, I., “Unification-based Multimodal Integration”, In: Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the European Chapter of the ACL, pp.281-288, Somerset, New Jersey, 1997.

[4] Johnston, M. “Unification-based multimodal parsing”, In: Proceedings of the International Joint Conference of the ACL and the International Committee on Computational Linguistics, ACL Press, pp. 624-630, Montreal, Canada, August 1998.

[5] Cohen, P., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L. and Clow, J., “QuickSet: multimodal interaction for distributed applications”, In: Proceedings of the 5th ACM international conference on Multimedia, p.31-40, Seattle, Washington, United States, November 09-13, 1997.

[6] Oviatt, S., “Mutual Disambiguation of Recognition Errors in a Multimodal Architecture”, In: Proceedings of the Conference on Human Factors in Computing Systems (CHI’99), ACM Press: New York, pp. 576-583, 1999.

[7] Nakagawa, S. and Zhang, J.X., “An input interface with speech and touch screen”, In: Transactions of the Institute of Electronic Engineers Japan C, 114-C(10), pp. 1009-1017, 1994.

[8] Landragin, F., “The Role of Gesture in Multimodale Referring Actions”, In: Proceedings of the 4th International Conference on Multimodal Interfaces, 2002

[9] Wertheimer, M., “Untersuchungen zur Lehre von der Gestalt II”, In: Psychologische Forschung 4, pp. 301-350, 1923.

[10] Ando, H., Kitahara, Y. and Hataoka, N., “Evaluation of multimodal interface using spoken language and pointing gesture on interior design system”, In: Proceedings of the International Conference on Spoken Language Processing (ICSLP’94), Vol. 2, pp. 567-570, Yokohama, Japan, 1994.

[11] Oviatt, S., “Ten Myths of Multimodal Interaction”, In: Communications of the ACM, 42(11), pp.74-81, Nov. 1999.

[12] Corradini, A., Wesson, R., Cohen, P., “A Map-based System Using Speech and 3D Gestures for Pervasive Computing”, In: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI’02), pp. 191-196, October 14-16, Pittsburg (PA, USA), 2002.

[13] Quek, F., McNeill, D., Bryll, B., Duncan, S., Ma, X., Kirbas, C., McCullough, K-E, and Ansari, R., ”Multimodal Human Discourse: Gesture and Speech”, In: ACM Transactions on Computer-Human Interaction (TOCHI), 9(3), September 2002.

[14] McNeill, D., Duncan, S., “Growth points in thinking-for-speaking”, In: D. McNeill (Ed.), “Language and Gesture”, Cambridge: Cambridge University Press, 2000.

[15] Denecke, M.,”Generische Interaktionsmuster für aufgabenorientierte Dialogsysteme”, PhD Thesis, University of Karlsruhe, 2002.

[16] Denecke, M. and Yang, J., “Partial Information in Multimodal Dialogue”, In: Proceedings of the International Conference on Multimodal Interfaces, 2000.

[17] Denecke, M., “Object-oriented Techniques in Grammar and Ontology Specification”, In: Proceedings of the Workshop on Multilingual Speech Communication, 2000.