USER MODELLING IN A DIALOG SYSTEM

Ralf Kompe, Martin Emele, Silke Goronzy, Robert Mencl, Sunna Torge
Sony International Europe (GmbH) Sony Corporate Labs Europe
Advanced Software Lab
Hedelfinger Str. 61
D-70327 Stuttgart
{kompe, emele, goronzy, mencl, torge}@sony.de

ABSTRACT

In people's home the amount of on-line available digital content such as video, music, or pictures is rapidly increasing. First home server products are available. Efficient selection of content will be a problem for users. We developed a flexible multi-modal dialog system to overcome this. A key-component is the user model, which learns very fast from user behaviour. All information like key-words are determined automatically and scored according to their relevance. No manual setting is required.

Keywords

User model, recommendation, multi-modal dialog

1. INTRODUCTION

The amount of digital content rapidly increases. Electronic program guides (EPG) become more and more popular. Home servers with large storage space become available. People will have large amounts of videos, music titles, and pictures on-line. It will be a problem to select content efficiently, in particular if the user does not have a specific item in mind.

Therefore a system is needed which knows the user preferences very well and which can be operated in an intuitive and flexible manner, without the necessity of reading manuals or memorising a set of commands or key-words. The user should be free how to formulate a request and which I/O modality to use: Certain wishes are easier to express by speech others easier by gestures, e.g. pointing or crossing out or even a combination of both.

Easy access of content as well requires meta-data to be associated with the content, which can be retrieved from broadcasted digital content, via the Internet or generated by the user. In any case meta-data generation needs to be supported by automatic methods like speech recognition; we will not discuss this in this paper. Rather we want to focus on the user modelling aspect and give a description of our dialog system for content selection as far as necessary.

Proceedings of the
1st annual GC-UPA Track
Stuttgart, September 2003

2. DIALOG SYSTEMS

2.1 Overview

Current dialog systems range from pure graphical interaction to pure speech driven interaction. Concerning spoken dialog, systems are available which provide a simple menu-driven question/answer type of dialog. Sometimes speech input is even restricted to simple commands rather than spontaneous speech. Others allow flexible dialogs where the user can take the initiative and guide the dialog. The quality of speech recognition technology has undergone remarkable progress over the past years, however, it is still far behind human capabilities in particular in the presence of noise, room reverberation, or limited channel bandwidth. To cope with this, usually vocabulary and grammar are restricted and often close-talking microphones are necessary [6]. Prosodic information is sometimes used to support the interpretation process [8], which usually is conducted with a parser based on some type of grammar.

Recently research shifted its interest towards multi-modal dialog, where simultaneous I/O through different modalities is possible [7,11,14]. Multi-modal interaction increases robustness, efficiency, expressiveness, and the ability to correct errors [3].

2.2 Qbit

We developed a multi-modal dialog system (Qbit: **Q**uery **b**y **i**nteraction **t**echnologies). Figure 1 depicts the components of Qbit. They were all developed within Sony apart from the dialog manager, which was obtained from SemanticEdge [12]. The application specific dialog model defining the flow of dialog was designed by Sony in cooperation with Semantic Edge.

Qbit provides several simultaneous input channels. Currently, continuous speech recognition and pointing gestures on a touch screen are

implemented. Speech interpretation relies on a robust coarse concept spotting rather than on interpretation of the exact meaning. On the output side simultaneous output by graphics or speech are possible. Research on haptics I/O (force-feedback) is still at a rather initial stage, however, it will play an important role in the future [9].

Typically the dialog system would e.g. upon a request for a movie for tonight display (or output by speechCore of the system are interaction management and underlying knowledge bases. On a semantic level the analysis hypotheses of the different input modalities are merged and scored. Cross-modal references are resolved, i.e., an utterance *play these* with *these* referring to a selection on the screen is interpreted correctly [5]. Based on the best scored hypothesis the dialog manager decides about the next action to be taken. This can be an output to the user or some actions on device-side. We use a mixed-initiative dialog system, where the user can take the initiative and guide the direction of the dialog. That means (s)he does not necessarily have to provide an exact answer to the questions of the system. On the other hand in order to reduce complexity of speech recognition the user is not free to arbitrarily change the topic of conversation. In case of output to the user the media fission [4] decides depending on a set of rules and the current context how to present the information to the user.

The interaction management relies on several knowledge-bases. The application is defined by discourse and domain model. The context model allows situation-specific decisions. Personalised interaction is established by a user model which will be described in detail in the next section.

depending on the context) to the user a list of 5-10 alternatives, leaving him/her some possibility to select manually without overloading him/her with irrelevant information. This list is computed with the user model described below. In case of too many well or similar matching suggestions the dialog system needs to ask the user for further constraints. In case of no good matches the dialog system as well would clarify with the user possible alternatives.

Qbit functions as a central user interface, which allows to operate an arbitrary and dynamically changing number of devices through a planning module (plug-and-play). This translates complex user wishes into sequences of actions of possibly several devices or services based on their abstract function models. It also infers new functionalities arising from the combination of devices through a home network [13].

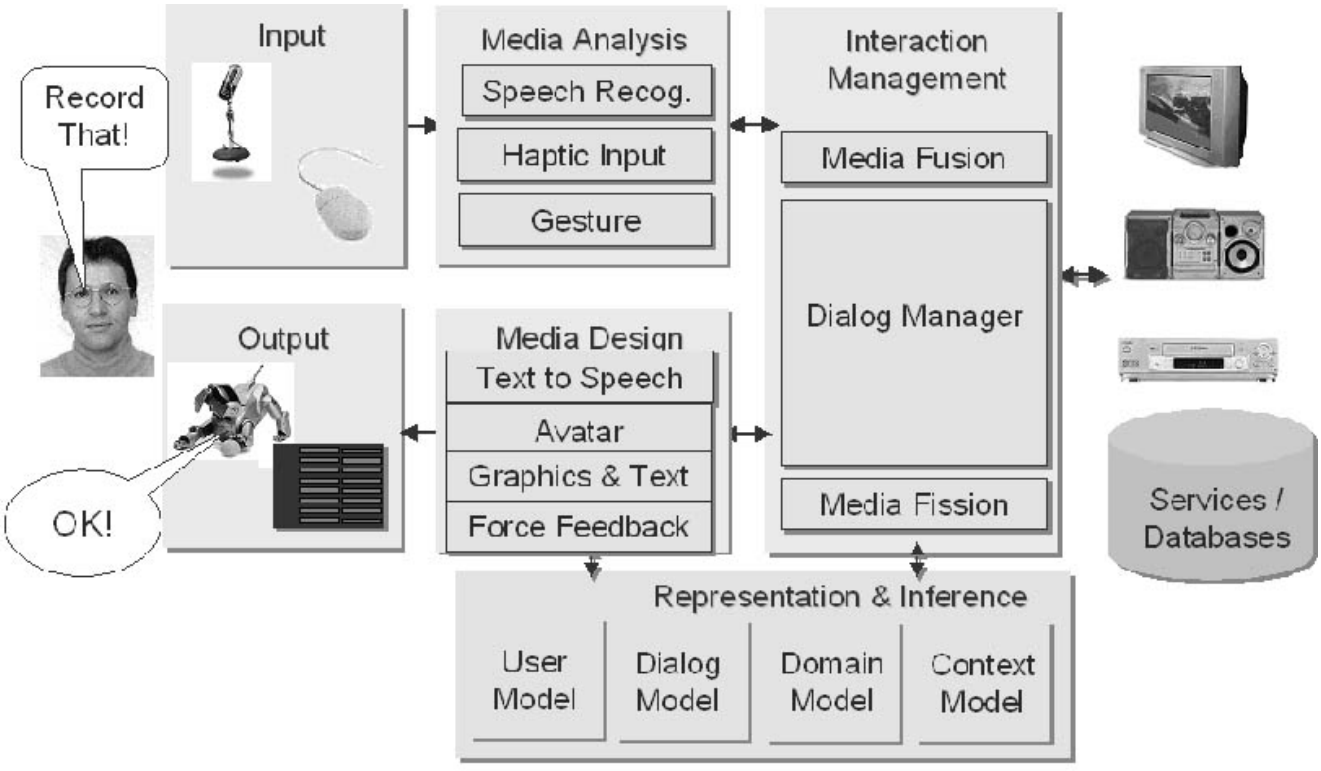

Figure 1: Structure of the dialog system Qbit

3. USER MODELING

3.1 Why personalisation?

Users want to have efficient access to the increasing amount of digital content available on-line in homes. Often, underspecified and vague requests like

- Some happy music,
- The pictures of the vacation in Venice,
- A movie on Saturday night,

need to be fulfilled, such that the set of matching items has to be restricted according to the user's preferences, or the user needs to be asked for further constraints. A genre hierarchy as e.g. used by Amazon for music is tiring to use and often it is not clear how to categorise a title [10]. In general, using a genre hierarchy is in most cases not the way people want to access their music at home.

3.2 Key features

We originally developed our user model for TV program selection, so that the following is focused on this. However, meanwhile we have successfully adapted it to music content and it can be easily applied to still picture selection, news filtering, etc.

Our system differentiates from others in the following ways [1,2]:

- It learns the user preferences very fast and fully automatically. The user does not have to specify any preferences or key-words manually.
- It automatically generalises from past observations of the user behaviour without the need for any manually generated knowledge-base.
- User's are not classified into categories of similar behaviour. Rather real individual profiles are built. We do not use something like *people who watched the same show as you also watched X*.
- The suggestions usually fit very well the taste of the user, however (s)he is free to adjust the number of suggestions.
- The user is free about the constraints he specifies. For example, all of the following requests are possible: *I want to watch TV, … watch a report, … a report about X.*

Our system is efficient in terms of CPU and memory usage and can thus be implemented on the client side. Therefore, no information on the user's preferences has to be stored on any server.

3.3 Basic Algorithm

Before describing our algorithm we have to define the usage of a few terms:

- *Article (a):* contains for one broadcasted or recorded TV show (music title) all available meta-data as usually distributed with in an EPG like title, actors, artists, directors, genre, abstract of TV show, description of music, date and time broadcasted/ recorded/viewed, play-time.
- *User history (h):* stores everything the user did and all the articles of shows the user watched. It contains all relevant information like article, minutes viewed, zapping behaviour, switched on just for this show, etc.
- *User input:* Everything the user does such as turning on the TV or requesting a movie to be recorded.
- *User profile:* a weighted list of key-words/-phrases based on a (filtered) user history. Key-phrases are taken from all text in the articles including title, genre. By this genre becomes just one parameter of many.
- *User model:* comprises user profile, matching algorithm, and inference engine providing generalisation.

The core of the algorithm consists of the computation of the weights (w) of words/phrases (p) given a specific article:

$$w(p|a) = I_1(p|a) * I_2(p|h) / I_3(p|d) \qquad (1)$$

where d denotes the whole database of articles, and I_i denotes a measure of importance. The term $I_2(p|h)/ I_3(p|d)$ defines the weights contained in the user profile, i.e., the importance of a phrase in the user history normalised by the importance of the phrase in the whole database. Multiplying this with $I1$, the importance of the phrase in the given article, implies a matching of a specific article with user profile. Phrases with neglectable weights are discarded from the user profile for efficiency. The functions I_i are based on the commonly used inverse document frequency [1], which basically gives words a high score given a certain document or set of documents if they occur frequently in this and just rarely in others.

We score each article of the future TV program a_l by the following algorithm:

1. filter the user history depending on the request, e.g., the day of the week specified by the user, if applicable.
2. compile the user profile for the filtered history based on the weights $I_2(p|h) / I_3(p|d)$
3. generalisation: extend the user profile by related words or phrases, where these get a slightly lower weight than the corresponding word actually being in the user profile
4. for each phrase p_k in the history compute $w_{kl}(p_k|a_l)$ according to equation (1)
5. the score w_l for this article a_l is given by the normalised sum $\Sigma_k w_{kl}(p_k|a_l)$

Then the $N(w_l)$ TV shows corresponding to the articles with the greatest score w_l are suggested to the user. The number N depends on the weights,

such that only few or even none suggestions are provided if none matches well. *N* can also be influenced by the user, depending whether (s)he wants few or many suggestions.

The related words or phrases mentioned in step 3 are taken from a thesaurus. We compute the thesaurus automatically from the whole database of articles on the basis of co-occurrences of words. Thereby an application dependent thesaurus is computed which achieves better results than a generic thesaurus.

3.4 Evaluation

This basic algorithm has been evaluated by a number of users. People were asked to mark in the past TV program what they watched. Then they got a suggestion for tonight's TV program and had to rate the quality of the first, second, and third suggestion as being excellent, good, neutral, not too bad, or bad. Figure 2 shows the average rating for 20 users having marked 3-4 TV shows from the past program, i.e. a user history of 3-4 items. The suggestions are rather good even after such a short learning period: 80% of the first suggestion was rated as either excellent or good. Note that in each of these tests the user was not asked to compare the suggestions with the full program. So (s)he is not aware of what (s)he missed. The rating was done in rather general terms.

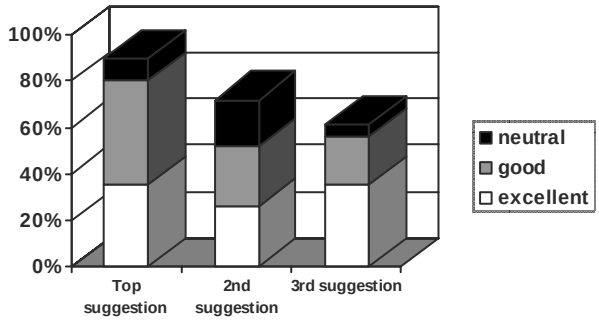

Figure 2: Quality of suggestions (history of 3-4 items)

3.5 Extensions

Meanwhile we have extended our algorithm in order to better model different kinds of user behaviour. These will be described informally in the following.

We can now provide the user with suggestions which depend on the time of day and whether the day is on a weekend or not.

The weights in the user profile incorporate whether the user has watched a TV show from the beginning to the end and if he turned the TV on at the beginning of the show and turned it off afterwards. This indicates that (s)he really intended to watch this show. If the TV is just on for a longer period the shows of this period are being taken into account with a lower weight.

A forgetting factor ensures that "older" items in the history influence the user profile less than newer items. This, allows for a change of the user's taste.

Note that our algorithm implicitly handles the problem of favourite shows being broadcasted only seldom. Relevant for our algorithm is only the number of times the user actually watched the show relative to the number of times it was broadcasted.

4. CONCLUSIONS AND DISCUSSION

The concern most frequently raised is that items important for the user might be missed. This can never be avoided but just minimised, however, fact is that people will need sufficient support by systems for contents selection. There should always be the option for manual selections and to adjust the degree of filtering done by the system. An explanation function should help users to understand certain decisions and to correct these.

There are many details in daily operation of such a system which need to be taken into account. E.g., switching on a certain TV show accidentally can cause problems. As another example, user requests for tomorrow 10 pm have to be interpreted carefully: Everything shown between now and tomorrow 10 pm should be taken into consideration, because it could be recorded. As a third example, it might be obvious from past user behaviour that (s)he does not watch a show lasting for two hours or more after 10 pm.

Our system is a good step towards a flexible personalised user interface for content access, and we have overcome some of above mentioned problems, however, there is still a lot of research necessary to make user modelling fool-proof and user interfaces really flexible.

Our next step will be that the user himself can specify within the dialog with the system for a current TV show or music title if he considers it to be funny, happy, sad, exciting, … These attributes should be automatically incorporated into the scoring function.

Also, we need to capture with our algorithm the following questions: Was the user unable to watch a show although it was interesting? Did the user zap because there was nothing interesting or because there were two interesting shows at the same time? Did the user watch a potentially interesting show deliberately not, because there was another one broadcasted on a different channel at the same time, or because it was just a repetition?

5. REFERENCES

[1] P. Baudisch, Dynamic Information Filtering. Ph.D. Thesis. *GMD Research Series 2001, No. 16. GMD Forschungszentrum Informationstechnik*, St. Augustin

[2] M.M Beaulieu, M. Gatford, X. HUang, S.E. Robertson, S. Waler, P. Williams, Okapi at TREC-5, 5th Text retrieval conference, NUIST, Gaithersburg (1997)

[3] P. Cohen, Multimodal Interaction: a new focal area for AI, Int. Joint Conf. on Artificial Intelligence IJCAI (2001)

[4] Ch. Elting, G. Möhler, Modelling output in the EMBASSI multi-modal dialog system, *Int. Conf. On Multimodal Interfaces* ICMI (2002)

[5] Ch. Elting, S. Rapp, G. Möhler, S. Strube, Multimodal input processing and output generation in EMBASSI System, *Conf. Multimodal Interfaces* ICMI (2003)

[6] S. Goronzy, Robust Adaptation to Non-Native Accents in Automatic Speech Recognition, Springer (2002)

[7] T. Herfet, Multimodal Assistance for Infotainment & Service Infrastructures. *Proc. Human Computer Interaction Conf.,* Berlin (2003)

[8] R. Kompe: Prosody in Speech Understanding Systems, *Lecture Notes in Artificial Intelligence*, Springer, Heidelberg (1997)

[9] G. Michelitsch, A. Ruf, H. van Veen, J. van Erp, Multi-finger haptic interaction within the MIAMM project, *Conf. EuroHaptics*, Edinburgh (2002)

[10] F. Pachet, D. Cazaly, A taxonomy of musical genres, *Conference RIAO Computer-assisted information searching on Internet* (2000)

[11] N. Reithinger, C. Lauer, L. Romary, MIAMM-multidimensional information access using multiple modalities, Multimodal Dialogue Systems, Copenhagen (2002)

[12] SemanticEdge: www.semanticedge.de

[13] S. Torge, S. Rapp, R. Kompe, Serving complex user wishes with an enhanced spoken dialog system, Int. Conf. On Spoken Language ICSLP (2002)

[14] W. Wahlster, SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. *Proc. Human Computer Interaction Conf.,* Berlin (2003)

Referent

Ralf Kompe obtained the diploma degree in computer science and the Ph.D. at the University of Erlangen in 1989 and 1996, respectively. From 1989 to 1990 he was research assistant at McGill University, Montreal working on speech recognition. From 1991 to 1996 he was a member of the research staff of the Institute for Pattern Recognition, Univ. of Erlangen working on the recognition of prosodic information and its use in parsing, speech understanding, and dialog control. In 1997 he joined the Sony Stuttgart Technology Centre. He is currently as a general manager responsible for user interface activities. Current research interests are in multi-modal interaction, user modelling, information retrieval, and usability. He is author or co-author of more than 70 publications including 7 journal articles and one monograph.