LingWear: A Mobile Tourist Information System

Christian Fügen [1], Martin Westphal [1,3], Mike Schneider [2], Tanja Schultz [2] and Alex Waibel [2]
fuegen@ira.uka.de, westphal@de.ibm.com, {schneider, tanja, ahw}@cs.cmu.edu

[1] Interactive Systems Laboratories
University of Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
++49 721 608 4730

[2] Interactive Systems Laboratories
Carnegie Mellon University
School of Computer Science
Pittsburgh, PA 15221, USA
++1 412 268 76

ABSTRACT

In this paper, we describe LingWear, a mobile tourist information system that allows uninformed users to find their way around in foreign cities and to ask for information about sights, accommodations, and other places of interest. The user can communicate with LingWear either by means of spontaneous speech queries or via a touch screen. LingWear automatically decides whether to respond through the integrated speech synthesis or display messages. LingWear is currently available for the cities of Heidelberg and Karlsruhe. It was designed to run on wearable computer, e.g. the Xybernaut family, and is available in both Windows and Linux versions.

Figure 1. Welcome Screen of LingWear for Karlsruhe.

1. INTRODUCTION

Due to the rapid development within the area of processors and memory modules, the performance of today's wearable computer is sufficient, to enable it to run processor and memory intensive applications. This makes it possible to develop user friendly multi modal user interfaces including speech recognition for wearables.

For this reason we have decided to develop LingWear, a mobile tourist information system that allows uninformed users to find their way around in foreign cities and to ask for information about sights, accommodations, and other places of interest. Hence we were able to make use of valuable information collected in the course of other projects like VODIS [9], C-STAR [3] and cooperations like DeepMap [5].

The following modules are integrated in LingWear:

- The **tour manager** presents some sights depending on the user's current location and on user's preferences (Figure 2). User preferences are handled through a user model. Sights that are currently open or closed are marked with special icons.
- The **navigation module** helps the user to find specified places in the city (Figure 3). It searches for the shortest path between the user's current and desired locations. The route segments can be retrieved step by step. In the near future a GPS-Module will be integrated to enhance this capability.
- The **information module** provides information about sights or other places of interests saved in the database (Figure 4). The information is presented to the user by images and short text descriptions.

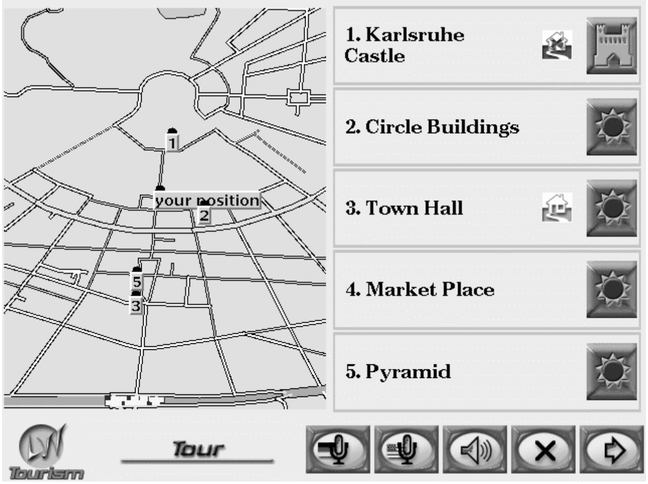
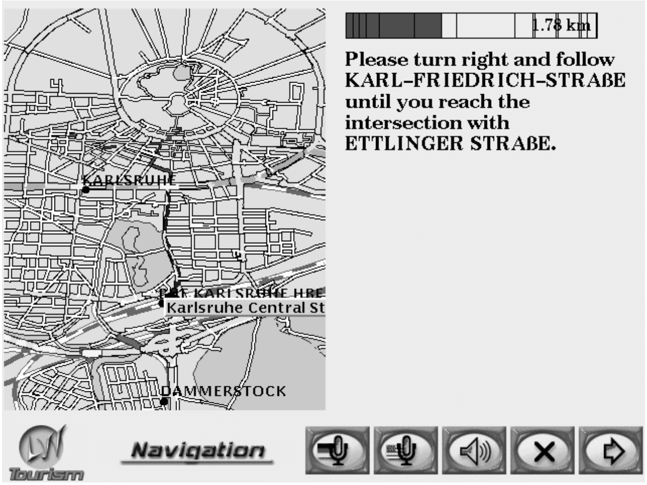


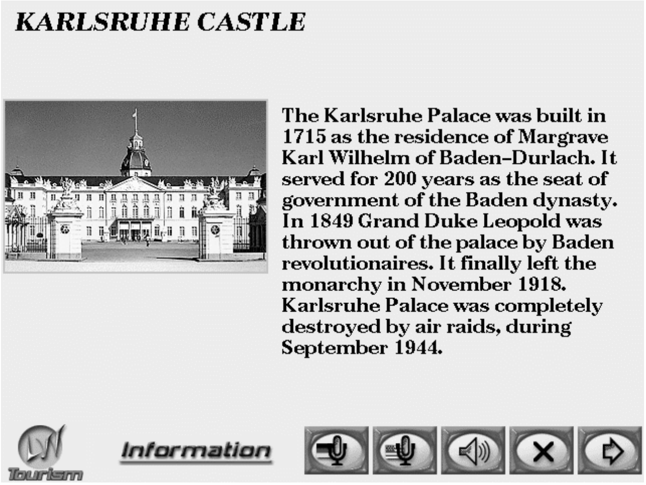
Figure 2. Screen shot of the tour mode.

[3] Now with: European Speech Research – IBM Deutschland Entwicklung GmbH



Please turn right and follow KARL-FRIEDRICH-STRAßE until you reach the intersection with ETTLINGER STRAßE.

Figure 3. Screen shot of the navigation mode.



KARLSRUHE CASTLE

The Karlsruhe Palace was built in 1715 as the residence of Margrave Karl Wilhelm of Baden-Durlach. It served for 200 years as the seat of government of the Baden dynasty. In 1849 Grand Duke Leopold was thrown out of the palace by Baden revolutionaires. It finally left the monarchy in November 1918. Karlsruhe Palace was completely destroyed by air raids, during September 1944.

Figure 4. Screen shot of the information mode.

- The **translation module** helps foreign visitors to communicate with local residents, as required when making hotel reservations, physician-visits etc. (Figure 6 and Figure 7). It accepts user queries in either English or German, and can produce translations in any of the target languages, English, German and Japanese. The translation output is both displayed and spoken.

We have currently modified our system to also support medical queries (Figure 7). Such a capability allows for example an English-speaking patient on vacation in Germany to describe his symptoms to a German-speaking doctor, and receive in return instructions and other medical advice.

2. MODULES OF LINGWEAR

The following figure shows the architecture of LingWear and the dependencies of the component modules:

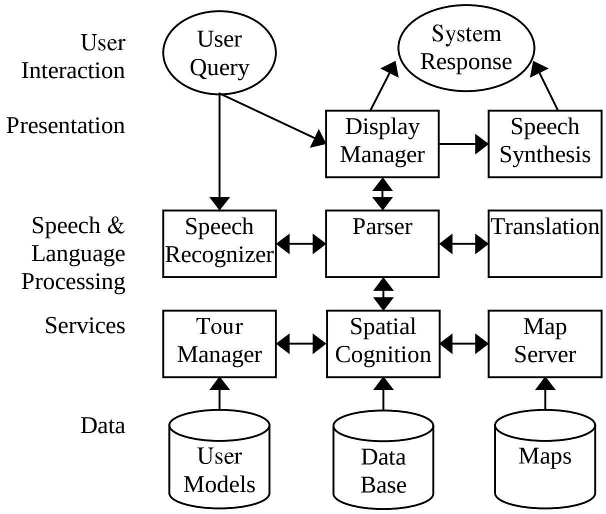


Figure 5. Module dependencies of LingWear.

2.1 User Interaction and Presentation

User queries, given by keyboard or via touch sensitive display, are processed by the display manager. Speech user queries are processed directly by the speech recognizer. The results of both components are passed to the parser for further analysis.

The Display Manager is responsible for presenting the current status of the system to the user in suitable form. In addition it arranges pictures, texts, and icons with given layouts and depending on their size on the display. The output can take place either in HTML or directly in Tcl/Tk.

Text information is also synthesized via an integrated speech synthesis system. For English and German we are currently using the speech synthesis system Festival [1], and for Japanese the Fujitsu VoiceSeries provided by Animo Ltd.

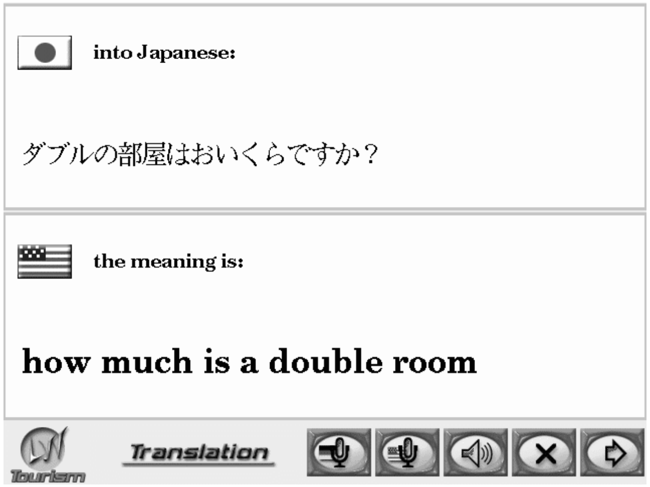
2.2 Speech and Language Processing

2.2.1 Speech Recognizer

The speech recognition engine used in LingWear is a part of the Janus Recognition Toolkit [6]. The full continuous English and German system uses approx. 2,500 context-dependent acoustic models with 32 Gaussians per model. Cepstral Mean Normalization is used to compensate for channel variations. In addition to the mean-subtracted mel-cepstral coefficients, the first and second order derivatives are also calculated. Linear Discriminant Analysis is applied to reduce feature dimensionality to 24, followed by a speaker-based maximum likelihood signal adaptation. With all optimisations the recognizer runs in nearly real time. To reduce the waiting periods for the user, the recognizer works in run-on mode.

We are using a class-based trigram language model to cover a vocabulary of approx. 5,000 words. Classes are introduced for places, streets, hotels, and all other location-dependent and some location-independent types like numbers. This allows us to easily switch between different cities by only providing location-dependent lists for the language model classes. More work has to be done for providing dictionary entries for these lists, especially for English pronunciations of proper names, like German street names. For this purpose we are using Festival together with some post processing scripts to remove misspellings at word composita boundaries.

To support also medical user queries, a domain specific, class-based language model was built. In addition to the medical domain, information for switching between different modes, e.g.

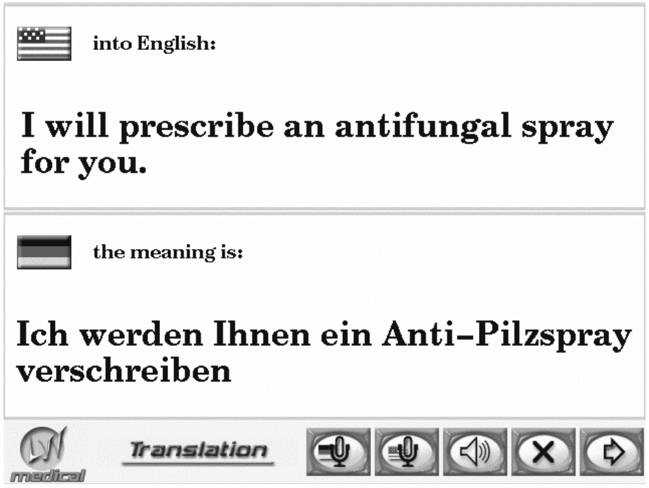

into Japanese:

ダブルの部屋はおいくらですか？

the meaning is:

how much is a double room

Figure 6. Screen shot of the translation mode.


into English:

I will prescribe an antifungal spray for you.

the meaning is:

Ich werden Ihnen ein Anti–Pilzspray verschreiben

Figure 7. Screen shot of the medical translation mode.

“Switch to Navigation”, is included in the language model. Wherever a model switch occurs, the appropriate search object is selected.

2.2.2 Parser

The parser analyses the hypothesis received from the speech recognizer and, according to the content of the hypothesis, decides which further components are needed to produce the system response. In addition, clarification dialogues are generated for underspecified user queries.

We are using SOUP as a parser, which was developed at the Carnegie Mellon University. It is a stochastic, chart-based, top-down parser, which was designed for real-time analysis of spoken language with very large, multi-domain semantic grammars [8]. SOUP achieves flexibility by encoding context-free grammars, specified for example in the Java Speech Grammar Format, as probabilistic recursive transition networks. Robustness is achieved by allowing skipping of input words at any position and producing ranked interpretations that may consists of multiple parse trees.

In LingWear, modular semantic grammars are used to model system knowledge. Semantic grammars are known to be robust against ungrammaticalities in spontaneous speech and recognition errors. However, they are usually hard to expand to cover new domains. For this reason we are using modular semantic grammars. Each sub-grammar covers the dialogue acts required for one sub-domain. An additional grammar provides cross-domain dialogue acts such as common openings and closings. All grammars share one library with common concepts, such as time expressions. Also location-dependent proper names are located in a separate grammar file. This makes extensions to new locations straightforward. Each grammar is associated with a special tag that reflects the domain of that grammar. Currently we are using tags for the domains navigation (NAV), travel planning (TPL), hotel reservation (HTL), medical (MED) and mode switching (SWI).

The output of the parser is converted to typed feature structures as defined in [2], which simplifies processing and interpretation by other components of the system. The notion of a type in a feature structure refers to the fact that every feature structure is assigned to a type from a type hierarchy. Moreover, for every type, a set of appropriate features is specified so that type inference is possible. Naturally, feature structures are well-suited for representing partial information. They do not adequately represent ambiguity, however. For this reason, we are using underspecified feature structures as introduced in [4]. In addition to feature structures, they are able to leave disjunctions unresolved.

Since underspecified feature structures represent unresolved disjunctions, they provide a good point of departure for generating clarification dialogues.

2.2.3 Translation

The Translation Module was developed mainly during the C-STAR-II-Project and integrated into LingWear. Within C-STAR-II, we have been developing a translation system for the broad domain of travel planning, including hotel reservation.

We have extended our system by grammars for covering also translations in the medical domain. Currently only a small set of translations are possible, but the grammars are constantly being extended.

Both translations are based on Interlingua as an interchange format, and make use of modular semantic grammars, which allows us to easily expand our system to new languages. Such grammars have also been shown to be effective in providing accurate translation for limited domains [10]. This gives us the possibility to use once more the SOUP parser. The assignment of domain tags to different sub-grammars allows us to switch easily between navigation, global translation, and medical translation mode.

2.3 Services

The spatial cognition, together with the map server and the tour manager are used for all navigation and information queries.

The tour manager presents some sights depending on the user’s location and on his preferences. When starting LingWear, the user is asked for his name, and when using LingWear for the first time, for his preferences as well. Preferences may include sites like churches, museums, bars and restaurants. Each object in the database is also assigned an importance factor. When suggesting sites, both importance and distance from the user's current location are taken into account.

The spatial cognition is responsible for all navigation queries. It is the interface to the database, which includes all objects like sights and restaurants, together with optional short descriptions, pictures, and hours of operation. The location of all objects is also saved in the database. The system's current mode determines what information is returned to the parser module.

The map server is responsible for drawing the desired portions of the map. Desired route segments within the map are shown highlighted. It is also possible to include additional text information and icons in the map, for an easy identification of specified objects by the user.

3. COMMUNICATION

As shown in Figure 8, a communication server (ComServer) is integrated into the system for the communication between the different modules. Modules can connect to this ComServer via socket with a given ID. After connection they have to specify an entry procedure, which is called every time a message is sent to a module specified by ID.

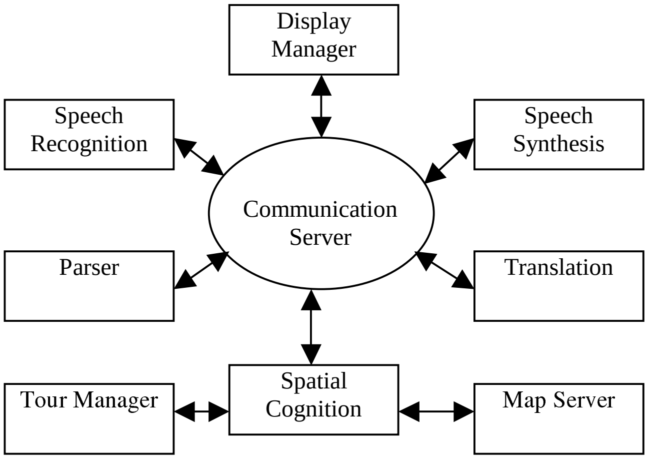


Figure 8. Communication between the modules.

Although all messages must go through the communication server, our central communication has several advantages over a distributed communication or communication via a bus:

- All connected modules are known by the ComServer, so error messages can be returned, if a module is not reachable.
- The communication between the modules is direct. So that the message is only sent to the module given by the ID. Grouping of IDs to virtual IDs is also allowed, for broadcasting one message to a group of modules. This is necessary for example when switching between different modes.
- The direct communications reduces processor load, because unaddressed modules do not have to analyze the message.

The databases and the map server are connected directly to the spatial cognition.

The communication format between the different modules is FIPA (Foundation of Intelligent Physical Agents) [7].

3.1 Communication Flow

An incoming English or German speech user query is processed by the corresponding recognizer and then transferred via the ComServer to the parser.

The parser analyses the hypothesis by producing a parse tree and, depending on the parse tree, the typed feature structure. Each user query belongs to a specific use case, which means, that two features are used to specify a user query: the grammar ID tag and the use case. In each mode besides the mode switching grammar, only the grammars belonging to the mode are active.

Depending on the grammar ID and the use case, the modules involved in producing the system response are defined.

- For navigation queries the spatial cognition is asked for the shortest way between two desired locations on the map. The path is highlighted and transferred, together with a generated description of the way, to the display manager.
- For information queries the spatial cognition is asked for information about the specified objects. All available information, e.g. text, images, etc., is collected and given to the display manager.
- If the user wants to know what there is to see around here, then the tour manager is asked for some objects depending on the user's current location and preferences defined in the user model. The objects are ordered by descendent ratings, whereby already-visited objects are left out.
- The translation becomes accessible when switching to the translation mode. Each user query, apart from queries to switch between different modes, will be treated as a query for the translation and sent to the translation module. The translation module produces the results for all available languages in parallel, which are sent directly to the display manager.

4. CONCLUSION AND FUTURE WORK

We have shown a system that allows uninformed users to find their way around in foreign cities and to ask for information about sights, accommodations, and other places of interest. The medical translation expedites a visit to a German doctor in the case of illness or injury.

The system is now in a state, where we can pass it to users, who are unfamiliar with the system, for real user studies. For this purpose two additional modules are already integrated into LingWear:

- A data collection module stores all recordings with the speaker information and the hypothesis produced by the speech recognizer.
- A history module logs all messages, which are sent through the system into one file. An integrated simulation component allows us to rerun a complete log file for error analysis and to make improvements of the system visible.

The system is continually updated and improved. New approaches in speech recognition, such as integrating new words into a running system and decoding along context free grammars will be integrated as soon as they have proven effective. We would expect to get a large gain in speed and accuracy by decoding along context free grammars, because of the more limited search space and because no additional parser is needed.

Unsupervised adaptation of the systems acoustic component in order to adapt to the current user will be integrated soon. As well as archiving the adaptation parameters together with the user preferences in the user model.

The display manger will be extended by a multimodal component, which detects user input via the touch screen; e.g., handwriting and other gestures. This would make it possible to specify new objects in the map and also allows the addition of user comments to objects, such as "This is my favourite restaurant".

For automatical output of the route segments while the user is walking, a GPS module will be integrated. This also allows us to give additional information about sights and other interesting objects along the route.

One major disadvantage of the system are the databases with the location-dependent information, because this information must be collected manually. Therefore a module should be integrated, which builds automatically the location-dependent databases, including images, sight descriptions, and new dictionary and language model entries for streets and other proper names. This could be done directly via Internet by connecting to a city's web server.

5. ACKNOWLEDGMENTS

We would like to thank Céline Morel for the graphic design of LingWear and Florian Metze and Thomas Schaaf for training the English and German recognizers. Our thanks also to Donna Gates, Chad Langley, Alon Lavie, Lori Levin, Kay Peterson, Alicia Tribble, and Dorcas Wallace for writing and integrating the medical translation modules and grammars.

6. REFERENCES

[1] A. W. Black, P. Taylor: *The Festival Speech Synthesis System: system documentation*, Technical Report HCR/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, 1997. Available at http://www.cstr.ed.ac.uk/projects/festival.html.

[2] B. Carpenter: *The Logic of Typed Feature Structures*, Cambridge University Press, 1992.

[3] C-STAR: Consortium for Speech Translation Advanced Research. Homepage: http://www.c-star.org.

[4] M. Denecke: *A Programmable Multi-Blackboard Architecture for Dialogue Processing System*, in Proc. of the Workshop on Spoken Dialogue Systems, ACL/EACL-1997.

[5] DeepMap: cooperation with the European Media Lab (EML), Heidelberg. More information at http://www.eml.org/english/research/deepmap/deepmap.html

[6] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal: *The Karlsruhe-Verbmobil Speech Recognition Engine*, in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97, Munich, Germany, 1997.

[7] FIPA (Foundation of Intelligent Physical Agents) Homepage: http://www.fipa.org

[8] M. Gavaldà: *SOUP: A Parser for Real-World Spontaneous Speech*, in Proc. of the 6th International Workshop on Parsing Technologies, IWPT-2000, Trento, Italy, February 2000.

[9] P. Geutner, M. Denecke, U. Meier, M. Westphal and A. Waibel: *Conversational Speech System For On Board Car Navigation And Assistance*, in Proc. Of ICSLP '98, Adelaide, Australia, 1998.

[10] M. Woszczyna, M. Broadhead, D. Gates, M. Gavaldà, A. Lavie, L. Levin, A. Waibel: *A Modular Approach to Spoken Language Translation for Large Domains*, in Proc. of AMTA-1998.

[11] J. Yang, W. Yang, M. Denecke, A. Waibel: *Smart Sight: A Tourist Assistant System*, 3rd Inter-national Symposium on Wearable Computers, ISWC-1999, San Francisco, California, October 1999.