RECENT ADVANCES IN LINGWEAR: A WEARABLE LINGUISTIC ASSISTANT FOR TOURISTS

Christian Fügen[1], Tanja Schultz[2], Jia-Cheng Hu[2], Alex Waibel[1,2]

[1]Interactive Systems Labs
University of Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
{fuegen,waibel}@ira.uka.de

[2]Interactive Systems Labs
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA
{tanja,jiacheng,ahw}@cs.cmu.edu

ABSTRACT

In this paper we describe our recent advances in LingWear, a wearable linguistic assistant for tourists. LingWear allows uninformed users to find their way in foreign cities or to ask for information about sightseeing, accommodations, and other places of interest. Moreover, the system allows the user to communicate with local residents through integrated speech-to-speech translation.

Furthermore, the graphical user interface (GUI) of LingWear runs also on small hand-held devices (e.g. Compaq's iPAQ). In this client-server solution the main components of the system are running on a wireless connected server. The user can query LingWear either by means of spontaneous speech or via touch screen and receive the system's responds either by the integrated speech synthesis or by display messages.

1. INTRODUCTION

Due to the rapid development in the area of hand-held devices, we expect the performance of such devices to be sufficient in the near future, in order to run processor and memory intensive applications. Therefore, it is our believe that the development of user friendly multimodal user interfaces including speech recognition and translation are within the reach for small wearable devices.

Driven by this expectation we developed LingWear [1], a mobile tourist information system that allows uninformed users to find their way in foreign cities as well as to ask for information about sightseeing, accommodations, and other places of interest. Moreover, the system allows the user to communicate with local residents through integrated speech-to-speech translation. However, due to the lack of current computing power and memory storage of small hand-held devices a client-server model with a wireless communication to a LingWear server is adopted. This gives us first access to the newly developed platform and furthermore allows us to step-by-step migrate all other modules into the hand-held device.

The next section gives a short overview of LingWear's architecture and describes the variety of available modes. In section 3, we present our latest achievements in speech and language processing. Section 4 presents the translation module of LingWear. We describe some results of our experiments in domain portability by extending semantic grammars by hand or by automatic learning for the new medical domain. In section 5, we deal with our client-server approach for LingWear. Section 6 concludes the paper and gives an outlook on future work.

2. ARCHITECTURE OF LINGWEAR

The implementation of LingWear followed the standard design principles of light interfaces, which allow high flexibility and makes it easy to add new modules. Following this concept LingWear is based on a central communication server (ComServer). Although all messages are forced to go through the communication server, this central communication has several advantages over a distributed communication or communication via bus:

- Since all modules which are connected to the ComServer are known by it, an error message can be returned, if a module is not accessible.
- As a result of the direct communication between the modules, messages are solely sent to the individual module given by an ID. Grouping of IDs allows message broadcasting to a group of modules.
- The direct communication reduces the processor load, since the rest of the modules do not have to analyze messages.

2.1. Modes of LingWear

For a clear arrangement, we have divided LingWear into several modes, whereby each of the modes is represented by a special topic. The following modes are integrated in LingWear:

- The **tour mode** displayed in Figure 1 presents information about sightseeing. The selection depends on

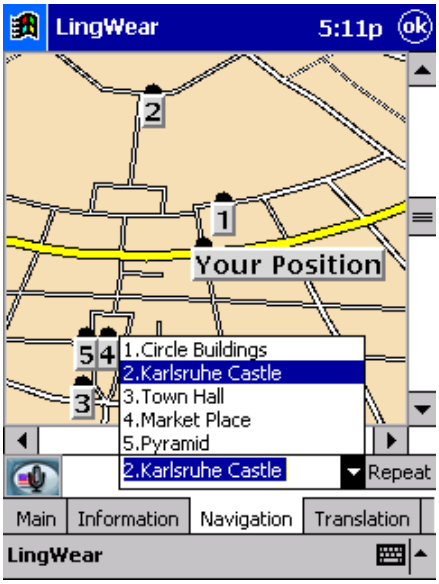

Figure 1. Tour mode.

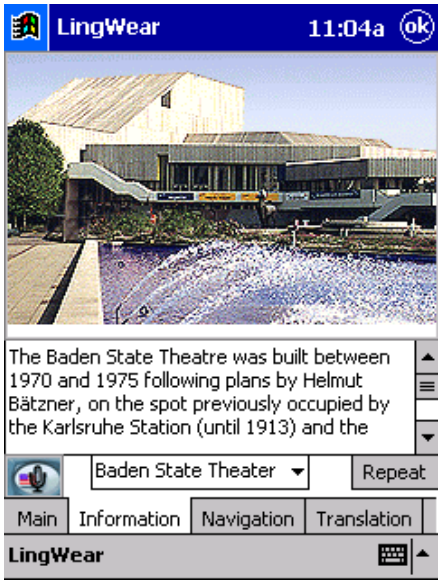

Figure 2. Information mode.

Figure 3. Translation mode.

the user's current location and preference. User preferences are handled through a user model. It is possible to attach individual icons to the sightseeing places to identify whether the event is open or closed.

- The **navigation mode** supports the user in finding the shortest route to specified places in the city. The route can be retrieved step by step, and additional information about sights can be presented along the way. Currently we are investigating the usefulness of a GPS-augmented navigation.
- The **information mode** as displayed in Figure 2 provides information about sightseeing or other places of interests as stored in a database. The information is presented to the user via images and short text descriptions.
- The **translation mode** as shown in Figure 3 enables non-native visitors to communicate with local residents, a necessary function in situations like making a hotel reservation, or visiting a physician. .

In addition to the presentation of the information on the screen, speech output is synthesized. For English, German, and Arabic we are currently using the speech synthesis system Festival [2], for Japanese the Fujitsu VoiceSeries provided by Animo Ltd.

3. SPEECH AND LANGUAGE PROCESSING

The speech recognizer used in LingWear was built using the Janus Recognition Toolkit, JRTk [3]. In the current LingWear system we are applying IBIS [4], our recently developed one pass decoder, which is part of JRTk. Besides several other advantages compared to our old Janus three-pass search, like smaller memory usage and higher recognition speed, IBIS allows us to decode along context free grammars beside the classical statistical n-gram language models (LM).

3.1. Speech Recognition

The typical speech recognizer in LingWear consist of a fully continuous system using approx. 2,000 context-dependent acoustic models with 16 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 calculated. A Linear Discriminant Analysis is applied, followed by a speaker-based maximum likelihood signal adaptation. In order to reduce time delays for the user, the recognizer works in run-on mode. The update of the VTLN-factors and the adaptation matrices are done while the system waits for new user input. To further speed up the run-time behavior without suffering from severe accuracy loss, we are applying fast score computation methods (BBI) and phone lookaheads. The overall vocabulary size of the recognizers is usually about 5000 words.

Due to our modular architecture we can easily add recognizers in several languages to our system. Our default system works with an English and a German recognizer. The default language for the navigation, tour, and information mode is English, while German is used in the dialogue-mode during translation. We are currently working on the integration of a multilingual recognizer, which significantly reduces the costs of system's maintenance. Furthermore it allows to switch to another language by simply switching the linguistic knowledge sources on the fly.

	LM	CFG
WA	76.12%	76.26%
correct sentences	40.57%	51.23%
RTF on PIII, 1 GHz	0.20	0.16
memory requirements	35 MB	35 MB
vocabulary size	2035	2035

Table 1. Comparison between a 3-gram LM and a CFG on the navigation domain measured on ~250 queries.

3.2. Parsing

As can be seen from Table 1 context free grammars (CFGs) outperform statistical n-gram LMs in both, the recognition speed and the number of fully correct recognized sentences achieving a significant relative gain of 20%. Furthermore, due to the effect that all hypothesis produced by a CFG are parsable, the hypotheses usually are either correct or incorrect. This makes it easier for a confidence based dialogue management to decide whether more clarification is needed.

The results in Table 1 are based on a LM built from a text corpus with ~261K words, which has achieved a perplexity of 12.14. The navigation CFG consists of 138 rules with 900 nodes and 1053 arcs. We do not compile one large final state graph out of the grammar, but we are using several rule based final state graphs, which are linked together by their non terminal symbols. This gives us the advantage of having small systems. Spontaneous nonverbal speech events as well as nonhuman noises are also supported, by using filler words in the decoder. This results in fewer restrictions to the user and needs no special treatment when writing grammars.

Another advantage when working with CFGs instead with n-gram LMs in IBIS is, that an extra parser becomes superfluous, because the parsing will already be done during decoding.

We are using modular semantic grammars to model system knowledge. Semantic grammars are known to be more robust against ungrammaticalities in spontaneous speech and recognition errors [5]. However, they are usually hard to expand in order to cover new domains. Therefore, we are using modular semantic grammars. Each subgrammar covers the dialogue acts required for one subdomain. An additional grammar provides cross-domain dialogue acts such as common openings and closings. Also location-dependent proper names are located in a separate grammar file. This makes extensions to new locations straightforward. The assignment of domain tags to different sub-grammars allows us to switch easily between navigation, global translation, and task specific (e.g. medical) translation mode in one speech recognizer.

3.3. Dialogue Management

We are using ARIADNE [6] as dialogue manager in Ling-Wear. Due to the fact, that this system is very new, the integration is not finalized yet. The usage of ARIADNE has the following advantages:

- We are able to specify all the linguistic knowledge at one location and to share it with all speech an language processing modules. Furthermore, by initiating a bi-directional communication between ARIADNE and Janus, we are able to support the above mentioned confidence based dialogue management or a weighting of special CFG rules in different dialogue states.
- Due to a clear separation of generic dialogue processing algorithms from domain and language specific knowledge sources, ARIADNE supports Rapid Prototyping.
- The above mentioned modular semantic grammar formalism is extended by an object oriented technique in grammar specification through vectorized context free grammars [7].

4. TRANSLATION

The translation is based on Interlingua as an interchange format. It makes additional use of the modular semantic grammars mentioned above. Interlingua was initially developed for a travel planning and hotel reservation domain in the context of C-STAR.

The analysis component of our Interlingua-based MT module takes a sentence as input and produces an Interlingua representation as output. For the medical domain we used a task-oriented Interlingua based on domain actions. Examples of domain actions are giving information about the onset of a symptom (e.g. *I have a headache*) or asking a patient to perform some action (e.g. *wiggle your fingers*).

In order to investigate on the portability of our speech-to-speech translation system to the medical domain, we compared the extension of a seed grammar by hand with one done by automatic learning. The seed grammar covered the domain actions but did not cover very many ways to phrase each domain action. Both the human and the machine-learned extension show improved performance over the seed grammar. However, the human extended grammar tended to outperform the automatically learned grammar in precision, whereas the automatically learned grammar tended to outperform the human extended grammar in recall. This result indicates that humans are capable of formulating correct rules, but may not have time to analyze the amount of data that a machine can analyze. Currently only a small set of translations are possible in the medical domain, but the grammars are constantly being extended [8].

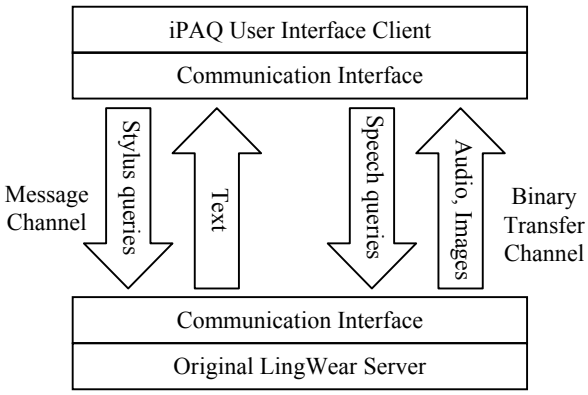

Figure 4. Client-Server communication.

5. LINGWEAR ON IPAQ

Due to the lack of computing power and memory storage of current hand-held devices, a client-server based approach in combination with a wireless connection to the LingWear server is adopted. For this purpose the given architecture is extended by an extra communication interface (see Figure 4).

The communication interface acts as the server to process user queries and to generate feedback to the iPAQ. This design allows to integrate all input modalities which are supported by the iPAQ. As a result the user can still communicate with LingWear by stylus or by speech queries. Furthermore, it is possible to connect several hand-held devices to the LingWear server by just creating new instances of the communication interface for each device.

In the current implementation, the user's speech queries are recorded on the iPAQ and transmitted to the LingWear server to be recognized and processed. The result in form of images, text files or synthesized speech will be displayed on the iPAQ screen or played back via it's speakers. For the transmission process, we are using two channels:

- A message channel, which forwards the standard communication like pen based queries, mode switching commands, server feedback or recognition and translation texts to or from the LingWear server.
- A binary transfer channel, for transmitting binary files, like recorded speech, synthesized audio and images.

Compared to the monolithic version of LingWear running on a PC, the client-server based solution comes to the cost of time delays due to the audio file transmission.

6. CONCLUSION AND FUTURE WORK

In this paper we presented recent advances in LingWear, which is prepared to run on small hand-held devices. The current client-server based solution enables us to do user studies on this platform in order to further improve the interface.

By using the new dialogue manager ARIADNE, it will be easier to move to extend the scenario to new cities, because only the city specific databases have to be updated. In this context we are thinking about automatically generating the underlying databases of sights and their descriptions, by e.g. connecting to a city's web server.

Especially for such a system like LingWear it is very important to have the ability to incorporate unknown words into the running speech and language processing modules. We will address this issue in the future by adding multiple input modalities like handwriting and gestures.

7. ACKNOWLEDGEMENT

We would like to thank Céline Morel for designing the layout of LingWear. Our thanks also to Donna Gates, Chad Langley, Alon Lavie, Lori Levin, Kay Peterson, Alicia Tribble, and Dorcas Wallace for writing and integrating the translation modules and grammars.

8. REFERENCES

[1] C. Fügen, M. Westphal, M. Schneider, T. Schultz, A. Waibel: *LingWear: A Mobile Tourist Information System*. In Proc. of the Human Language Technology Conference, HLT-2001, San Diego, March 2001.

[2] A. W. Black, P. Taylor: *The Festival Speech Synthesis System: system documentation*. Technical Report HCR/TR-83, Human Communication Research Center, University of Edinburgh, Scotland, UK, 1997.

[3] 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.

[4] H. Soltau, F. Metze, C. Fügen and A. Waibel: *A One pass-Decoder based on Polymorphic Linguistic Context Assignment*. In Proc. of the Automatic Speech Recognition and Understanding Workshop, ASRU-2001, Madonna di Campiglio, Trento, Italy, December 2001.

[5] 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.

[6] M. Denecke: *Rapid Prototyping for Spoken Dialogue Systems*. In Proc. 19th International Conference on Computational Linguistics, COLING-2002, Teipei, Taiwan, August 2002.

[7] M. Denecke: *Object-Oriented Techniques in Grammar and Ontology Specification*. In Proc. of the Workshop on Multilingual Speech Communication, MSC-2000, Kyoto, Japan, 2000.

[8] A. Lavie, L. Levin, T. Schultz, C. Langley, B. Han, A. Tribble, D. Gates, D. Wallace, and K. Peterson: *Domain Portability in Speech-to-Speech Translation*. In Proc. of the Human Language Technology Conference, HLT-2001, San Diego, March 2001.