CONVERSATIONAL SPEECH SYSTEMS FOR ON-BOARD CAR NAVIGATION AND ASSISTANCE

Petra Geutner, Matthias Denecke, Uwe Meier, Martin Westphal and Alex Waibel

pgeutner@ira.uka.de
Interactive Systems Laboratories
University of Karlsruhe (Germany)
Carnegie Mellon University (USA)

ABSTRACT

This paper describes our latest efforts in building a speech recognizer for operating a navigation system through speech instead of typed input. Compared to conventional speech recognition for navigation systems, where the input is usually restricted to a fixed set of keywords and keyword phrases, complete spontaneous sentences are allowed as speech input. We will present the interaction of speech input, parsing and the necessary reactions to the requested queries. Our system has been trained on German spontaneous speech data and has been adapted to navigation queries using MLLR. As the system is not restricted to command word input, a parser is necessary to further process the recognized utterance. We show that within a lab environment our system is able to handle arbitrary spontaneous sentences as input to a navigation system successfully. The performance of the recognizer measured in word error rate gives a result of 18%. The parser has also been evaluated and yields an error rate of 20%.

1. INTRODUCTION

The technique of speech recognition has been used in more and more applications over the last years. Speech recognition in the car [2] is one of the applications that will be of high interest in the future. Whereas applications in this domain so far mostly have been restricted to hand free operation of the telephone, the demand for other functionality, like e.g. controlling radio and cassette or using car navigation systems with speech input, is steadily growing. Especially for navigation systems the necessity of typed input for a query can be tedious and in some situations even dangerous. Even though speech input seems an adequate and convenient way to overcome this problem, recognizing navigation queries still suffers from many problems:

1. the noisy car environment which automatically leads to performance decreases compared to lab environments,
2. the large number of confusable street and city names that have to be recognized by the system, where a high confusability will also lead to performance degradations.

Research on speech input for navigation systems so far has focused on input of certain keywords, city and street names. The system presented here adds one more level of complexity to the problem: our research concentrates on permitting arbitrary spontaneous sentences as navigation queries. As a consequence

1. the vocabulary of the system has to include more than just keywords and destinations,
2. the system has to be able to handle effects inherent in spontaneous speech (false starts, hesitations, ungrammatical sentences etc.).

2. SPONTANEOUS SPEECH VERSUS KEYWORDS

So far, speech research for navigation systems has concentrated on being able to recognize single keywords or fixed keyword phrases as accurate as possible. Legal input for a state-of-the-art system would therefore be a single street name “Main Street” or a street crossing “Peachtree at the corner of Second Avenue“. Our goal was to permit a user to phrase his query without any restrictions, thereby allowing “Main Street” as well as “What is the shortest path to Main Street?”, “I’d like to go to Main Street.” or even “I am in a real hurry, give me directions to Main Street”. Beside the need of robustness with respect to noisy environments and the confusability of street names, this approach of course increases the complexity for the underlying speech recognition engine of the navigation system.

When allowing spontaneous navigation queries, the vocabulary used as input to the system will go far beyond a simple list of street names and other destinations. It will also have to include words extracted from typical queries directed to the system. Secondly a parser is needed that is able to parse the recognition output properly and identify what the requested input to the navigation system actually is.

Even though the vocabulary of our system includes more than the conventional keywords, there will always be some words unknown to the recognizer as an unlimited recognition dictionary is impossible. But as we will show later task completion (meaning: giving the right directions to a certain destination) is possible even though not every single word was recognized correctly. To achieve this, the correct recognition of a street name and the identification of the actual destination within the recognizer output is enough. To find the actual destination of a query a semantic case-frame parser with an

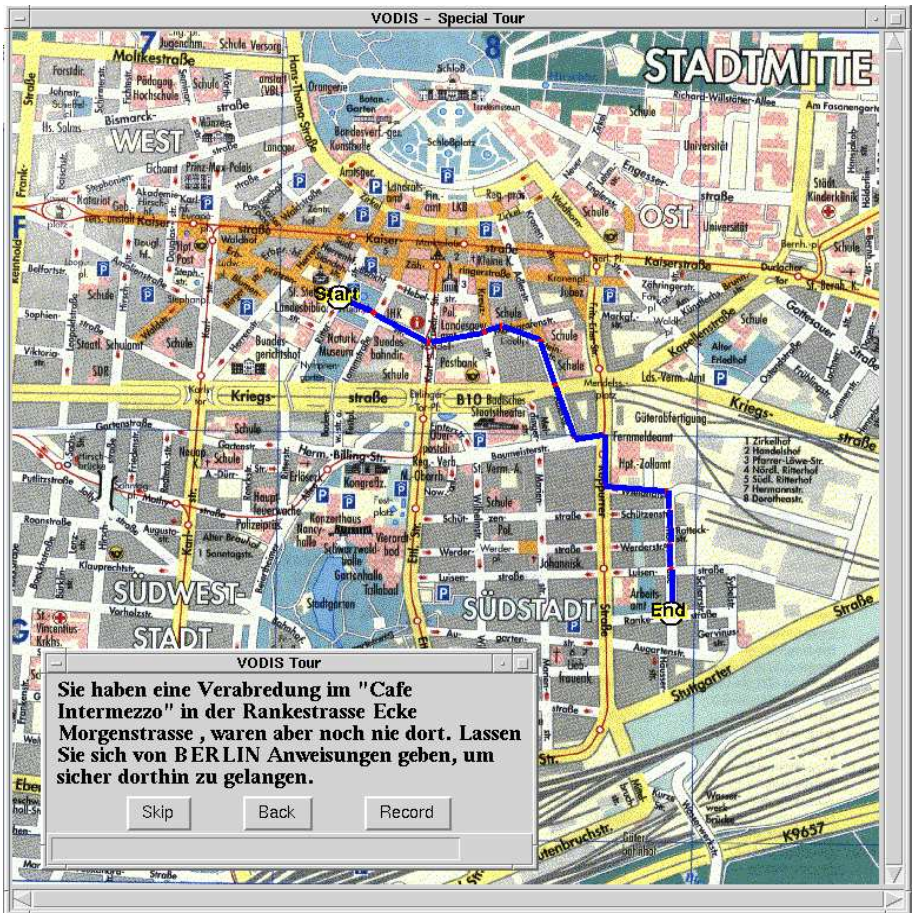

Figure 1: Example prompt of the **Data Collection Tool** used to simulate a navigation system

underlying semantic grammar is used. Both speech recognition system and parser were also trained to be able to handle spontaneous speech effects as coughing, laughter, lip smacks, breathing and so on.

3. DATA COLLECTION

For the purpose of training and and testing our speech recognition system especially designed for spontaneous navigation queries, both speech and text data had to be collected at the Interactive Systems Laboratories in Karlsruhe. Therefore a data collection tool has been developed at the University of Karlsruhe that simulates the use of a car navigation system.

The tool simulates a car trip by prompting the user with certain situations he has to react to. Figure 1 shows one of the prompts the system will come up with while collecting data. The prompt here reads as follows:

"You have an appointment at the Cafe Intermezzo in Rankestrasse at the corner of Morgenstrasse, but you have never been there before. Ask the system to give you directions to get there"

About 50 speakers were asked to participate in the data collection process and on average about 10 sentences were recorded by each of these speakers. In addition to these data that were transcribed afterwards, a text corpus of about 1700 sentences was collected as training material for language modeling training. Finally five spontaneous sentences were recorded from 100 speakers during a data collection actually taking place in a running car under different environment conditions (fan off/on, city/highway traffic, radio on/off, rain yes/no, windows open/closed etc.). This data will be used to train a robust speech recognition engine that allows the input of navigation queries not only in lab but also in noisy car environments.

4. System Overview

As can be seen in figure 2 our system consists of a speech recognition system that is able to handle German navigation queries. Here the user of the system utters the following request: "How do I uhm get to the theatre?". The hypothesized output of the recognizer is then fed into a semantic case-frame parser. The output of the parser is piped into a general manager. Within this general manager a dialogue manager decides if the parsed output is specific enough to be given as input to the navigation. If the parsed sentence does not

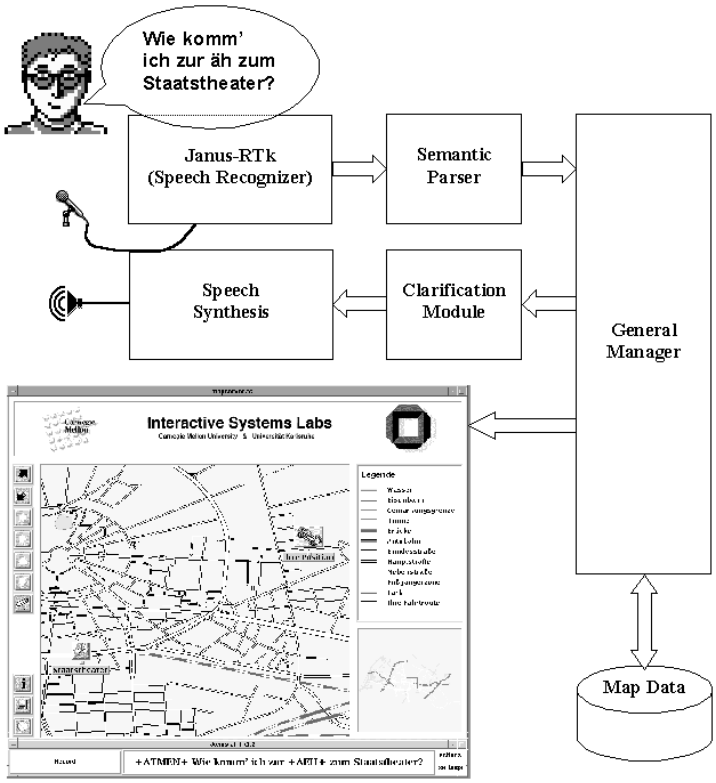

Figure 2: System Overview

include a specific destination, the general manager initiates a clarification dialogue with the user to further narrow down the actual destination. An example clarification dialogue looks as follows:

```
User:   "I want to go to the next restaurant."

System: "Do you want to go to a Chinese,
         Italian or American-style
         restaurant?"

User:   "Italian."
```

As soon as the destination of the query is fully specified, the general manager retrieves the necessary map coordinates from the map database and passes this information to the navigation system. Here the route is calculated and directions are given to the user.

5. THE SPEECH RECOGNITION ENGINE

The speech recognition engine presented in this paper is trained with the Janus Recognition Toolkit (JRTk) developed at the University of Karlsruhe (Germany) and the Carnegie Mellon University (USA)[1]. It is initialized through an existing German system, trained on more than 30 hours of speech from about 1500 German speakers and adapted to the navigation speech data collected at the Interactive Systems Laboratories.

The speech signal is first sampled with 16 kHz and then preprocessed. After a short time Fourier transform every 10 ms and a window size of 16 ms, we apply a Melscale filter bank and calculate 13 cepstral coefficients and their first and second order derivatives. To enhance speaker and channel robustness a speech based cepstral mean subtraction is done before a final LDA transform reduces the feature space to a 32 dimensional feature vector. The acoustic model built on that feature space has 2500 clustered polyphone classes each modeled as a mixture of 32 Gaussian with diagonal covariance matrices. For speed reasons we use a global BBI tree and phoneme look-ahead scores provided by a small context independent system trained with the same speech data. We also use a single search pass which results in a small increase in error rate but also in shorter turnaround times.

To train the language model 1700 utterances from our navigation

speech data collection (see section 3) and about 200 utterances added by hand were used. Classes were defined for towns, streets, numbers, neighbourhoods, points of interests (POI), names and places like shops, hotels and so forth. By mapping those words to a class symbol we reduced the dependency on a certain town or environment where the recognition system is to be used. Additionally we omitted unfrequent words and ended up with a total of about 800 words required for the navigation task. Finally the language model was calculated using the filtered text which means that all elements of a certain class will have the same language model probability. To be able to recognize all street names within the city of Karlsruhe, these streets were added as pronunciation variants of the class symbol into the dictionary. The current speech recognition system for the city of Karlsruhe consists of about about 1700 street names and 30 neighbourhoods.

Experiments on data collected in a lab environment resulted in a 18% word error rate for spontaneous navigation queries. Although this means that 18% of all words were not recognized correctly, not all requests directed to the system containing recognition errors must necessarily fail. When measuring the performance of the speech recognizer in terms of task completion instead of word error rate, more than 82% of all queries are "correct". Here correctness means that the destination the user wants to get instructions to is recognized correctly. When considering task completion as sole performance criterion, in only about 5% of all queries the destination of the request was misrecognized.

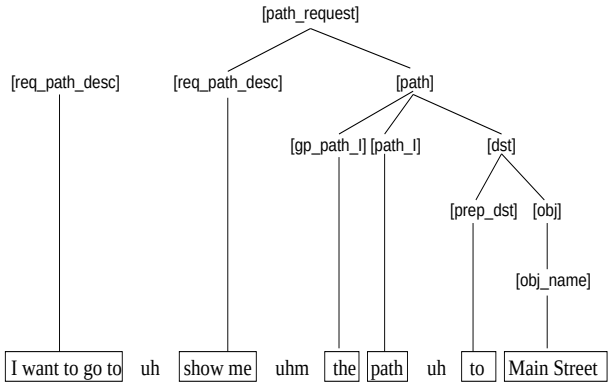

Figure 3: Example Parse Tree.

6. PARSER

The parser used to further process the hypothesis from the speech recognition engine is a semantic case-frame parser. It uses a semantic grammar to extract a representation of an utterance. The meaning of the utterance is then represented in case frames. In this sense it does not matter if the input is possibly ill-formed and does not adhere to grammatical rules. The parser ignores non-matching fragments and focuses on important keyword phrases, thereby being extremely well-suited for ungrammatical input. Example for such an ill-formed request that is typical for spontaneous speech is the following request: "I want to go to uh show me uhm the path uh to Main Street?". Both speech recognizer and parser are able to handle spontaneous sentences like this successfully.

Our parser uses a context-free semantic grammar. The parse algorithm uses a parse chart and beam search algorithm which yields parse times below one second for one utterance. Figure 3 gives an example parse tree for an ungrammatical sentence.

Currently our semantic grammar consists of approximately 200 rules and is able to handle about 700 words, not considering street names and other map-specific data. Development of the grammar has been done based on part of the 1900 spontaneous utterances available for language model training. The following is an example rule of the grammar:

```
[req_path_desc]
     ( where is )
     ( show me )
     ( i need to find )
     ( i want to go to )
     ( what is the shortest path to)
     ( i'd like to go )
```

Evaluation of the parser yields a 20% error rate, which means that 80% of all sentences passed from the recognizer to the parser are not parsed completely. A high percentage of errors is due to out-of-vocabulary words that are not included in the grammar.

7. CONCLUSIONS

This paper describes our latest achievements in developing a first German prototype system that allows spontaneous speech input for on-board car navigation and assistance. We show that with a 18% word error rate only 5% of the navigation queries do not contain the requested destination. We presented the interaction of speech input, parsing and the necessary reaction to spontaneous navigation queries.

8. ACKNOWLEDGEMENTS

This research is part of the VODIS project, and was partly supported by the Language Engineering Program of the European Community. We greatfully acknowledge their support. The views and conclusions contained in this document are those of the authors.

9. REFERENCES

1. M. Finke, J. Fritsch, P. Geutner, K. Ries, and A. Waibel. *The JanusRTk Switchboard/Callhome 1997 Evaluation System*. Proceedings of the LVCSR Hub5-e Workshop, May 1997. Baltimore, Maryland.
2. X. Pouteau and L. Arevalo. *Robust Spoken Dialogue Systems for Consumer Products: a Concrete Application*. Proceedings of the 1998 International Conference on Spoken Language Processing (ICSLP), 1998. Sydney, Australia.