TOWARDS SPONTANEOUS SPEECH RECOGNITION FOR ON-BOARD CAR NAVIGATION AND INFORMATION SYSTEMS

Martin Westphal and Alex Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
{westphal,waibel}@ira.uka.de
http://isl.ira.uka.de

ABSTRACT

Speech recognition is seen to be of great benefit in on-board car navigation systems and assistance. The command word approach will be used for applications in the near future since the small active vocabulary and the hierarchical structure is much easier to cope with, from the developers' side. An alternative approach, using spontaneous speech input, is far more complex but provides the user with an interface that is very intuitive and has fewer restrictions. The user can rely upon his or her experience in inter-human communication and utter spontaneous queries. In this paper, we describe the requirements and the collection of a continuous car speech data base and show first recognition results obtained under different environmental conditions in the car.

Keywords: spontaneous speech recognition, speech based car navigation interface

1. INTRODUCTION

Speech recognition technology in the car has a number of convincing advantages and first products have already appeared on the market. However, it is clear that speech recognition in the car is a difficult task due to the noisy environment. In the European project VODIS [1] two different approaches, namely the *command word approach* and the *spontaneous speech approach*, were investigated. VODIS aims to control not only the car phone and audio components like radio, but also the navigation system. Depending on the task one or the other approach is appropriate [2]. A navigation demonstrator system that allows spontaneously uttered queries was developed by the Interactive Systems Laboratories in Karlsruhe, Germany and Pittsburgh, USA. In this section we review the two approaches mentioned above and show why it was necessary to collect a continuous speech database in the car to provide our demonstrator system with a recognizer capable of processing speech recorded in the real car environment.

1.1. Limitations of the Command Word Approach

Using spoken digits to dial a phone number or selecting from a personal phone directory by just uttering the name, is an appreciable help and very much increases the safety in the car. The first speech based approach to include other control functions like selecting the radio station, controlling the CD/cassette player, or entering destinations in a car navigation system, is made by means of command words. Compared to continuous speech, the recognition process is simpler and requires fewer technical resources. A set of commands can be defined that matches the desired functionality. With an increasing number, one would typically arrange the commands with a similar context, that is for example controlling the same device, within a hierarchical structure. This does not only reduce the size of the active vocabulary but can also give the user guidance in form of an active command word list in a small display. These hints are very important since one can not expect the user to memorize all the commands nor to know which hierarchical level is currently active.

Let us consider the following example: The driver is hungry and is looking for a place to eat. Each time he utters a command the small display will provide a new list from which he can choose. He steps through the menus of his navigation device by using the following command words:

"*NAVIGATION*"
"*ENTER DESTINATION*"
"*OTHER DESTINATIONS*"
"*RESTAURANT*"
"*NATIONALITY*"
"*ITALIAN*"
"*RESTRICTION*"
"*CLOSEST*"
"*SELECT DESTINATION*"

Unless he is not very familiar with the system and this sort of query he might look at the possible choices on the display each time before uttering one of the commands. Note that he would never be able to enter that kind of information while driving without speech and that he does not need his hands for a tactile interface. His focus however is not on the road for quite a while. For queries with such a degree of complexity, the command word approach reaches its limit. Besides being awkward, it is questionable whether this use of speech technology guarantees safety in the car. Nevertheless, for the near future this speech driven approach can provide a basic functionality and opens up new possibilities for on-board navigation and information systems.

1.2. The Spontaneous Speech Approach

In general, a short human conversation in the car is not considered a dangerous distraction from traffic. Allowing a variety of familiar expressions for human machine interaction, the user can easily access a wide range of functionality without being distracted too much. Compared to the command word approach, spontaneous speech allows a far more user friendly and faster input:

"*Take me to the nearest Italian restaurant!*"

However, for the machine, it is much harder to recognize continuous speech. In the car, we even have to deal with *spontaneous* speech since the user still concentrates on the traffic which might result in false starts, hesitations, or ungrammatical sentences. Also, for the interpretation of the query a natural language understanding component, that can process such input, is necessary. The following example illustrates the usefulness of such an interface:

User: "*Where .. uh ..How far is the nearest post office?*"

System: The nearest post office is about 2 miles from here!

User: "*Okay, take me there*"

To understand the last utterance, context information is also needed. Due to the high complexity and the problems with speech recognition under noisy conditions, it will take several years until such systems are available on the market.

1.3. Towards an On-board Navigation Demonstrator for Spontaneous Speech

In [3] we described a first laboratory demonstrator for the spontaneous speech approach allowing to enter spontaneous navigation queries that are recognized, parsed and then replied using a map display. This demonstrator is using a recognizer for clean spontaneous speech based on the Janus Recognition Toolkit. To run such a system in the car, we need a continuous speech recognizer that can cope with the adverse conditions.

Since most studies are based on the command word approach, huge effort was made to collect command words, proper names, connected digits and letter sequences in the car environment. In the MoTiV data collection [4] also a limited number of spontaneous queries were recorded but the main goal was to provide a database for command word recognizers (see for example [5]) and small vocabulary continuous speech recognition, like digits and letters.

2. DATA COLLECTION

Our aim is to develop and evaluate a continuous car speech recognizer and to study the effects arising in this environment. Due to the very limited amount of available continuous speech data recorded in a real car environment, we performed our own data collection. In this section, we describe the requirements and the collection of a car speech data base.

2.1. Requirements

Although we can expect that a specific user sitting in the car will speak a number of utterances so that the recognizer can adapt, we have to provide a speaker independent system in the first place. For the training of such a system, we need many different speakers covering both genders, all ages, dialects and so forth. It is also necessary to cover different car types since they have a large influence on the recorded speech. A complete coverage would have surely extended the scope of the project so we collected 43 speakers between ages 18 and 64 in three different cars.

The content of the utterances was oriented on the requirements of possible applications. One part consists of spontaneous navigation queries. The amount is relatively small since such queries have to be transcribed manually. Furthermore, the vocabulary and the resulting phonetic context (polyphones) very much depend on the navigation scenario. Therefore, the largest portion consists of read newspaper articles which are easily available and do not need manual transcriptions. The vocabulary is significantly larger and, as a consequence, also the polyphone coverage. In order to allow the recognition of proper names, we also collected spoken as well as spelled city and street names.

For the study of environmental effects in the car, it was very important to log the recording conditions as accurate as possible. A laptop allowed to verify the quality of the speech recording and to record the environmental conditions. After each recording the conditions were determined according to table 1 and in special situations (e.g. "indicator"), a predefined comment was selected.

For the audio recordings we used the same microphones as for part of the MoTiV collection [4]. The room microphone AKG C400 was installed at the car ceiling just above the windscreen. Simultaneously, we recorded the speech with a close-talking microphone Sennheiser HMD 410. The latter one is less affected by noise and was also used for earlier laboratory speech data collections.

road type	speed [km/h]	road condition	fan	window	weather
engine off	0	normal	off	closed	dry
city	0-30	tunnel	low	open	drizzle
federal or secondary road	30-60	gravel	medium		rain (wiper)
highway	60-90	cobble	high		wet road
	90-120				
	>120				

Table 1: recording conditions

2.2. Database Statistics

According to the requirements, a first set of 10393 utterances from 43 speakers was recorded. It was collected in 3 different cars with two microphones at a sampling rate of 16 kHz. With an average duration of 4.3 seconds, the entire database amounts to a total of 12½ hours per channel. Most of the utterances contain continuous speech and a smaller portion covers isolated or spelled names. Table 2 gives the exact number of utterances for different partitions and table 3 gives statistics on utterance, word, and vocabulary counts for different categories.

	utterances	*percentage*
Utterances total	**10393**	**100 %**
Set:		
Test	1385	13,3 %
Training	9008	86,7 %
Gender:		
female	2384	22,9 %
male	8009	77,1 %
Car type:		
BMW 3	4654	44,8 %
Ford Escort	2564	24,7 %
Honda Civic	3175	30,5 %

Table 2: Utterance Statistic

category	*utterances*	*words*	*vocabulary*
Dictation	6562	65511	12891
Navigation	582	4102	582
spoken names	1330	1330	1077
spelled names	1327	9350	33
Commands	592	592	15

Table 3: Word Statistic

For each utterance we also logged the recording conditions such as road type, road condition, speed, fan, window and weather. No restrictions were given concerning these items. As an example figure 1 and 2 give an idea of the road types and the speed distribution of the collection. Contrary to most other conditions, the speed can be measured (is measured anyway in the car) and thus could be used as an additional input feature for the speech recognition process.

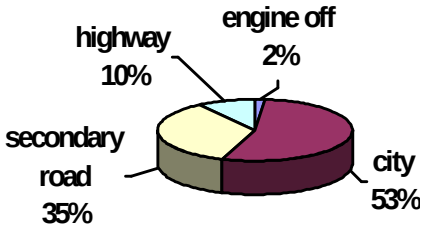

Figure 1: road types

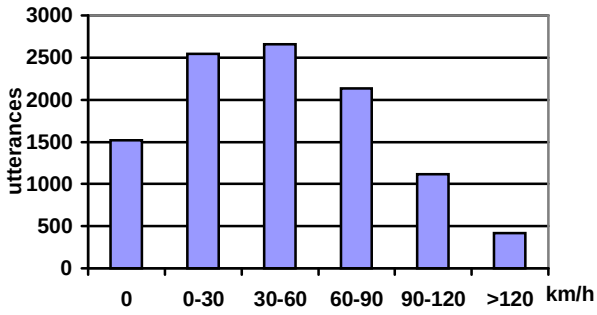

Figure 2: speed distribution

Additionally to the first set, we collected a second set with the 10 test speakers uttering each of 30 navigation queries under 12 different controlled conditions giving a total of 3600 utterances. The 12 conditions cover different speed ranges, fan settings and special cases such as indicator, open window, and acceleration (see table 4). The queries cover not only questions about any of 1715 streets from the German city of Karlsruhe but also street numbers, neighborhoods, points of interests, and less specific questions like

"*Where can I drink a coffee here*?".

3. RECOGNITION EXPERIMENTS

The collected database was used to train and evaluate a speech model for car speech (CarTrain). No data from one of the 10 test speakers was used for training. Our clean speech recognizer (LabTrain) trained on 30 hours of speech recorded in a quiet office environment was also tested for the different conditions to determine which aspects cause a performance degradation. Figure 3 shows word error rates over all conditions for the two recognizers. Note that for our application some errors are tolerable as long as they do not change the semantics of the input and lead to the desired response. The two most frequent substitutions are "*den*" vs. "*dem*" (both meaning "*the*" in English) and "*zur*" vs. "*zum*" (both meaning "*to the*" in English). These cases do not degrade the overall performance of the navigation system.

Training and testing was done based on the JANUS-3 speech recognition toolkit. After sampling the audio signal at 16 kHz 13 mel-frequency cepstral coefficients and their first and second order derivatives are computed. A speech based cepstral mean subtraction helps to enhance channel robustness. Finally this input vector is reduced by linear discriminant analysis (LDA) into a 32 dimensional feature vector. The acoustic model uses fully continuous mixture Gaussian densities based on 2500 decision-tree clustered context-dependent sub-phones. Both systems use the same language model and a 3k dictionary including all streets of Karlsruhe.

Using the clean speech recognizer together with the close talking microphone results in a word error rate of 13.2% for condition 01. This condition is very similar to the office environment since the engine and fan are turned off. Note that the same microphone type was also used to record the training data. This setup turned out to be robust for most conditions.

No.	*condition:* speed [km/h]	fan	special
01	0	**off**	**engine off**
02	0	low	**indicator**
03	25	low	**open window**
04	0	**high**	
05	0	**medium**	
06	**0**	**low**	
07	**25**	low	
08	**50**	low	
09	**75**	low	
10	**100**	low	
11	**125**	low	
12	**acceleration**	low	

Table 4: Recording conditions of set 2.

Only for speeds over 75 km/h the error rates increased to values over 20%. For these conditions, we also observed a greater loudness of the uttered speech which indicates that the Lombard effect also plays a role here.

For an on-board navigation system it is highly desirable to have a built in microphone that is mounted in the car cabin. The database we collected and described above provides us with simultaneously recorded utterances over the head-mounted close talking microphone and a car-mounted room microphone. This way, we can directly compare recognition results on the two channels. From figure 3 one can see that using the room microphone (car mic) with the clean speech recognizer leads to severe degradations of the recognition performance. For the clean speech condition 01 we find an error rate of 17.4% due to the microphone mismatch. For all other conditions the performance losses are higher, especially for a high fan setting and high speeds.

The car speech recognizer was trained with data recorded with the room microphone under real driving conditions. This training helped to improve the recognition results for most conditions. For some of the conditions they even became comparable with the close talking results. For others, like for a high fan setting (condition 04), the error rate is still at a level of almost 30%. The indicator (condition 02) seems to affect the recording of the room microphone and was also not well compensated by the car speech training. The clean condition 01 gives the worst results with 30.1%. This is a very rare case in our car speech data base and thus a mismatch between training and test environment.

In the future, we aim at improving the car speech and the clean speech recognizer by noise reduction and adaptation methods. We already observed fundamental differences between the effectiveness of such methods for continuous versus single word recognition.

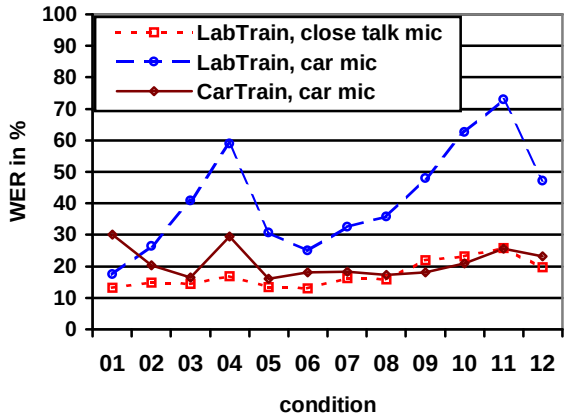

Figure 3: Word error rates of the two recognizers. The clean speech recognizer (LabTrain) was tested with a close talking microphone (close talk mic) and a room microphone mounted in the car cabin (car mic). The car speech recognizer (CarTrain) was only tested with the car-mounted microphone.

4. SUMMARY

By providing a continuous speech database for car speech, we could build a spontaneous speech recognizer for a navigation task in the car. The results are comparable to a clean speech recognizer trained on a very large database of spontaneous speech and tested with the same type of close talking microphone that was used to record the training data.

REFERENCES

[1] VODIS: "Advanced Speech Technologies for Voice Operated Driver Information Systems", EC Language Engineering Project LE 1-2277.
VODIS-URL: http://isl.ira.uka.de/VODIS

[2] D. Van Compernolle: "*SPEECH RECOGNITION IN THE CAR – From Phone Dialing to Car Navigation*", Eurospeech '97, pp 2431-2334, Rhodes, Greece, 1997

[3] P. Geutner, M. Denecke, U. Meier, M. Westphal and A. Waibel: "*Conversational Speech Systems For On-Board Car Navigation And Assistance*", ICSLP '98, Adelaide, Australia, 1998

[4] D. Langmann, H. Pfitzinger, T. Schneider, R. Grudszus, A. Fischer, M. Westphal, T. Crull, U. Jekosch: "*CSDC – The MoTiV Car Speech Data Collection*", ICLRE '98

[5] A. Fischer and V. Stahl: "*Subword Unit Based Speech Recognition In Car Environments*", ICASSP '98, pp 257-260, Seattle, USA, 1998