An In–Car Speech Recognition System For Disabled Drivers

Jozef Ivanecký, Stephan Mehlhase

European Media Laboratory
Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
{jozef.ivanecky, stephan.mehlhase}@eml.org

Abstract. Automatic Speech Recognition (ASR) is becoming a standard in nowadays cars. However, ASR in cars is usually restricted to activities not directly influencing the driving process. Thus, the voice-controlled functions can rather be classified as comfort functions, e. g. controlling the air condition, the navigation and entertainment system or even the mobile phone of the driver. Obviously this usage of an ASR system could be extended in two directions: On the one side, the speech recognition system could be used to control secondary functions in the car like lights, windscreen wipers or windows. On the other side, the comfort functions could be enriched by utilizing services like weather inquiries, SMS dictation or online traffic information. Compared to todays usage these extensions require a different approach than the one employed today. Controlling secondary functions in the car by voice demands the usage of a very reliable, real–time, local ASR. At the same time a large vocabulary ASR system is required for comfort functions like dictation of messages.
In this paper, we describe our efforts towards a hybrid speech recognition system to control secondary functions in the car. We also provide an extended comfort functionality to the driver. The hybrid speech recognition system contains a fast, grammar-based, embedded recognizer and a remote, server-based, LM-based, large vocabulary ASR system. We will analyze different aspects of such a design and the integration of it into a car. The main focus of the paper will be on maximizing the reliability of the embedded recognizer and designing an algorithm for switching dynamically between the embedded recognizer and the server-based ASR system.

1 Introduction

Automatic speech recognition (ASR) is becoming more and more common in todays cars [2]. The used ASR systems can be classified in two distinct classes: On the one hand there are integrated ASR systems, which control basic comfort functions like air conditioning, radio, or navigation system, e. g. to enter the address. On the other hand, todays upper class cars are utilizing speech recognition system running on a server which is accessed through the Internet. This

allows for more complex tasks, e. g. supporting inquiries for weather or traffic information.

Irrespective of the used ASR technology, in general the set of controlled in–car devices and functions does not expand to the secondary functions (e. g. lights or windscreen wipers). The driver can reach those without having to stop focusing on the driving process itself. Pressing a switch is in general, faster and more natural than to use a spoken command for such a task. However, controlling comfort functions is a more complicated process. Complex tasks like music selection require a significant amount of the driver's attention. Therefore, the driver benefits from controlling these functions by voice. In cases where the driver has to use a joystick instead of a steering wheel, e. g. due to a disability, controlling the secondary functions takes significant additional effort. Therefore, it makes sense to expand the voice control to include the secondary functions as well. The requirements for controlling secondary and comfort functions differ: On one hand a reliable, real–time speech recognition system with a safety model for incorrectly recognized commands is required for secondary functionalities. On the other hand controlling the comfort functions by voice, does not require real–time speech recognition. Also, a mis-recognized comfort function does not directly influence safety.

In this paper we describe our effort towards the implementation of a hybrid ASR system. A real–time, grammar–based embedded recognizer is used to recognize secondary functions commands directly in the car. A large vocabulary, LM–based recognizer connected via the Internet is used for advanced comfort functionality. We investigate different methods for dynamically switch between those recognizers, which is an important step towards reaching the aforementioned goals.

Remaining parts of the paper are organized as follows: In Section 2 we define secondary and comfort functions of a car. In Section 3 we describe the design of the two different ASR systems used for in–car speech recognition. Section 4 describes experiments used to evaluate the in–car speech recognition. In Section 5 a brief summary is provided.

2 Secondary and Comfort Functions

We define 3 classes of functions available in a car. They differ in terms of availability, simplicity of usage and required promptness of the reaction.

1. *Secondary functions:* Obligatory functionality of each car which does not belong to the primary functions (accelerator, breaks, steering wheel, ...). Examples are the different kind of lights, car horn or windscreen wipers. They are easily accessible and intuitively to operate. The reaction time of all these devices is instant and reliability is very high.
2. *Basic comfort functions:* Optional equipment of a car related to driving comfort, e. g. air conditioning or radio. They are usually easily accessible but not always intuitively to operate. As before, the reaction time is instant. Malfunctioning is not significantly influencing car usability.

3. *Advanced comfort functions:* Optional equipment of a car related to driving comfort, e. g. navigation system or traffic information systems. In general, they are rather complex to operate and the reaction time is not instant. Some of these functions require Internet access. Malfunctioning affects only the comfort of the driver.

Secondary functions are easily accessible in any car and there is seemingly no need to use voice control. However, the situation is fundamentally different in cars modified to be used by disabled driver. Depending on the level of disability, controlling secondary functions with ordinary control levers may vary from easy to impossible. In the latter situation, speech recognition might be a more natural way to control the secondary functions of a car.

3 Hybrid Speech Recognition

Because of the different requirements for the aforementioned in–car functions, it is difficult to use a single ASR system. For the secondary and basic comfort functions it is necessary to use a real–time local ASR system with very high recognition accuracy. This is achieved by a small vocabulary grammar–based system directly integrated into the car. The advanced comfort functions often require a large vocabulary, but do not require as high accuracy and low latency as ASR for the secondary functions. We are using a LM–based recognition server accessed through the Internet to provide this functionality. Finally, we designed a system which dynamically switches between the two recognition systems to provide a uniform interface to the user.

In the literature the term *Hybrid Speech Recognition* is used to describe a combination of HMM and ANN-based recognizers. In this paper however, we use it to refer to the combination of a grammar–based, real–time recognizer with a server–based, large vocabulary recognizer.

3.1 ASR for Secondary and Basic Comfort Functions

Embedded recognizers were originally designed to run on significantly slower hardware than available today. Therefore, in case of a small grammar the real–time requirement is easily satisfied. The main challenge for such a system is to meet the very low error rate requirements. An incorrect recognition can trigger an unwanted action, which, in a certain ill–timed moment, can lead to dangerous situations, e. g. switching off the lights during the night or switching on the opposite turning signal. Therefore, a robust safety model in case of an incorrect recognition is needed.

We are using commercially available embedded recognizers (Loquendo and SVOX). To run the recognizer we used the same platform as in [1]. We were focusing mainly on grammar and application design to achieve maximal accuracy and reliability. Usually if the grammar offers a big variety of commands the error rate of the recognition increases. Therefore, we tried to minimize the grammar

size and avoid acoustic similarities between the commands. As there are many ways to toggle specific devices, we focused on the most common short and long forms. For instance, for turning on the high beams the short form is "*Fernlicht an*" whereas the long form is "*Das Fernlicht einschalten*" [1]. The vocabulary size of the resulting grammars is only around 30 words.

The system is operating in *Push-to-Talk* (P2T) mode, which means that the system is only listening while a button is pressed. The *Push-to-Activate* (P2A) mode, in which the user only pushes the button once to indicate the start of the utterance, could be easier to use. However, we decided for the P2T system for accuracy reasons. Especially at high speeds the automatic end-pointing needed in the P2A system poses a problem due to the environmental noise.

Irrespective of the activation mode, the button used is serving also safety purposes. If the user presses the button again shortly after the recognition finished, he cancels the initiated action. Such a behavior should avoid unwanted situations caused by incorrect speech recognition and consecutive actions.

3.2 ASR for Advanced Comfort Functions

In order to provide the user with the comfort functions as defined in Section 2, the speech recognition system must be able to deal with a large vocabulary. Therefore, it is no longer feasible to use a grammar–based recognition system. We decided to use a server–based, large vocabulary speech recognition system. It is located in a computing center and consequently requires an in–car Internet connection to be available. Some cars already come with support for mobile network connectivity, it is possible to place UMTS routers in the car. If that is not possible the tethering capabilities of smart-phones could be used.

Regarding the recognition time, there are two considerations to take into account: On one hand, in case of accessing the advanced comfort functions it is no longer necessary to provide the user with recognition results in real–time. On the other hand, it is also important that processing is not taking too long as the driver gets distracted from driving when the system is not working as he expects to, i. e. not reacting to his voice input promptly. Given that the audio data needs to be transferred to the server which in turn sends back the recognition result using a possibly slow and unreliable mobile Internet connection, it was necessary to build a robust system which can handle outages in a non-disruptive way.

In order to decrease the recognition time, the service uses a custom network protocol to transfer the audio data in small chunks. The protocol allows the server to send back partial results as soon as they are available. With this protocol it was possible to create a service already starting to process the audio data while sending. Optimizing the server-side processing of the received audio signal allows to further decrease the perceived decoding time. Using this technique, we were able to reduce the perceived recognition time factor from around 3 down to around 1. The *perceived recognition time* specifies the time the user perceives as waiting time from finishing to speak until the system reacts to his input. The

[1] German terms to switch on the high beams.

actual recognition time can differ, mainly due to the time needed to transfer the data to the server.

The recognition system we are using is working with a language model with a vocabulary size of over 1 million words, specifically tailored for mobile search and dictation applications. The server–based system is designed to be highly scalable and can serve many clients at the same time without performance degradation.

3.3 Which one to use?

The audio signal is always processed by the in–car recognition system. A control application has to decide if the command was aimed at the secondary or basic comfort functionality or whether it is part of the advanced comfort functions. We evaluated 3 different approaches on how to distinguish between them:

1. *Confidence score:* Only the confidence score of the recognized utterance is taken into account. If the score is below a certain threshold, the audio signal is sent to the server–based recognizer.
2. *Out of grammar model:* If the recognition result is tagged as out of grammar (OOG), the audio signal is sent to the server–based recognizer. The confidence score is not taken into account.
3. *OOG model with trigger word:* As the previous method, but a special key word has to precede the "out of grammar" part.

If the decision algorithm decides that the utterance has to be sent to the server–based recognizer, the application informs the user about it and waits for a reply from the server. This kind of functionality assumes a working Internet connection as explained in Section 3.

4 Evaluation

The evaluation is split into two major aspects. The first aspect is to examine the speech recognition accuracy for different grammars and noise levels. The second aspect is evaluating the switching between the local and the remote recognizer. In order to evaluate our system we recorded a test set. For data collection the P2T mode was used and the microphone was at a distance of $20-30$ cm to the speaker. The recorded data consists of 10 speakers (4 female and 6 male voices) of which 2 were non-native German speakers. For each we recorded 2×30 commands, containing

- 10 long commands for controlling secondary functions (*den Blinker links ausschalten, die Lichthupe einschalten, ...*),
- 10 short commands for controlling secondary functions (*Blinker links an, Lichthupe, ...*),
- 5 commands controlling comfort functions with a trigger word (*Komfortfunktion: Wettervorhersage für Heidelberg, Komfortfunktion: Radio: SWR3 wählen, ...*), and

	SER	SA	AER	AA	ASCF
			Quiet environment		
Long form - reduced grammar	2 %	84 %	2 %	94 %	84.51 %
Long form - full grammar	15 %	80 %	1 %	94 %	84.56 %
Short form - full grammar	9 %	91 %	3 %	97 %	84.01 %
			Noisy environment		
Long form - reduced grammar	0 %	94 %	0 %	84 %	81.88 %
Long form - full grammar	13 %	86 %	1 %	98 %	81.38 %
Short form - full grammar	11 %	88 %	6 %	93 %	77.36 %

Table 1. Speech recognition and action accuracy (SER – Sentence Error Rate, SA – Sentence Accuracy, AER – Action Error Rate, AA – Action Accuracy, ASCF – Average Sentence Confidence Score).

– 5 commands controlling comfort functions without a trigger word (*Wettervorhersage für Heidelberg, Radio: SWR3 wählen, ...*).

The recording took place in 2 different environments: A quiet office environment and a noisy environment with in–car noise up to 80 dB, responsible for low SNR and the Lombard effect during the recording.

4.1 Speech Recognition

For the recognition accuracy test we created two different grammars. The first grammar is covering only the long forms of the commands and was designed to be used only with the first 10 test sentences recorded by each speaker. The second grammar is covering all commands for the secondary functions. The second one was used for all recorded commands to examine whether the error rate is getting worse with bigger command variety in the recognition grammar as expected. However, more important than the speech recognition accuracy is the accuracy of the actions triggered by the voice command. Even an incorrectly recognized command can trigger the correct action. Therefore we examined action accuracy as well as recognition accuracy.

In Table 1 the results for the sentence accuracy and the action accuracy obtained on the test set are shown. From the speech recognition point of view the most important results are the sentence accuracy (SA) and sentence error rate (SER). It is difficult to decide which combination of grammar and set of commands to use based on these results alone. In the quiet environment the short form commands with the full grammar give the best accuracy, whereas in the noisy environments the long forms with the reduced grammars give the best results.

Taking the action accuracy (AA) and more importantly the action error rate (AER) into account, Table 1 gives a better indication which is the safest grammar and commands combination. The smallest AER and biggest AA are

	With trigger word	Without trigger word
	Quiet environment	
Reduced grammar	36 %	59 %
Full grammar	42 %	60 %
	Noisy environment	
Reduced grammar	42 %	46 %
Full grammar	42 %	61 %

Table 2. Maximal sentence confidence score for the comfort function commands with the secondary function grammars.

always achieving using the long form of commands. Whether the grammar should also contain the short forms is subject to practical testing we plan to do in the future.

The table shows also the average sentence confidence scores[2]. We did not take into account the confidence score during the evaluation. However, using also such an information is an option how to further eliminate incorrect actions caused by an incorrect recognition result. On the other side the result rejection based on the confidence score will decreased the action accuracy. The number of commands from the recognition test with confidence score below 50 % was 5. In 4 of these 5 cases the recognition was incorrect. Therefore, if we used a minimum sentence confidence score for the secondary functions of 50 %, it would further reduce SER or AER but AA as well.

4.2 Speech Recognizer Selection

The recognizer selection tests included all three approaches described in Section 3.3. For the confidence score approach we re–used the grammars used for the tests in Section 4.1. With those grammars we tried to recognize the recorded commands aimed at the comfort functions. Of course the recognizer produced a recognition result containing a sentence from the grammar. But now the sentence confidence score is taken into account as well. Therefore, we examined the maximum score a sentence for a comfort function would gain, which are listed in Table 2. Comparing these values with the sentence confidence scores reported in Table 1, in all cases we observe a satisfactory difference. The lowest confidence scores were achieved for commands containing a trigger word. The best result was achieved with the combination of using such a trigger word and the grammar containing only the long forms.

For the garbage–based experiments, we modified the recognition grammar to include also an out of grammar (OOG) model. In the experiment with garbage preceded by a trigger word a command "<Trigger word> OOG;" was added. In the other experiment just the command "OOG;" was added. We were observing

[2] Confidence score of a particular recognizer was scaled into to the range 0 to 100.

	Quiet env.	Noisy env.
OOG w/o trigger word	76 %	84 %
OOG with trigger word	0 %	0 %

Table 3. Out of grammar (OOG) recognized for secondary function commands.

how many times the result "OOG" appeared among the recognized commands for secondary functions and how many times "OOG" did not appear among the comfort functions commands.

Table 3 shows how often "OOG" was returned when feeding secondary function commands into the speech recognition engine. We did the experiment with and without the trigger word "Komfortfunktion" which is not part of the remaining grammar. The results indicate, that for a reliable separation of secondary and comfort functions, the usage of some kind of trigger word is necessary. In the following experiment we used the grammar containing the trigger word and used the comfort function commands as input for the recognizer. In nearly all cases (98 % in quiet, 100 % in noisy environment) the recognizer returned the "OOG" indicator.

In case of the comfort functions the error rate, i. e. cases in which the output should be "OOG" but was not, is more important than the accuracy. A comfort function command which is accepted by a secondary function grammar could trigger an unwanted action on the secondary functionality in the car. The error rate measured in quiet and noisy environment was 0 %. Consequently, the results are confirming the previous indication, that the usage of an adequate trigger word is a reliable way to determine which recognizer to use.

5 Summary

In this paper we described various aspects of the development of an in–car speech recognition application for disabled drivers. We analyzed the usage options for disabled drivers and identified possible risks that need to be minimized. First experiments indicate the feasibility of our approach, but also unveiled the need for further work and massive testing in a real–life environment in order to maximize safety of the driver. The speech recognition accuracy in this case is only of secondary importance.

References

1. J. Ivanecký, S. Mehlhase, M. Mieskes. An Intelligent House Control Using Speech Recognition with Integrated Localization. In: *Proc. of Ambient Assisted Living – 4th AAL congress*, Berlin, 2011
2. P. Heisterkamp. Linguatronic product-level speech system for Mercedes-Benz cars. In: *Proc. of the first international conference on Human language technology research*, San Diego, 2001