Speaking While Driving – Preliminary Results on Spellings in the German SpeechDat-Car Database

Christoph Draxler (1), Klaus Bengler (2), Cristina Olaverri-Monreal (1)

(1) Dept. of Phonetics, University of Munich
(2) BMW Group, Munich
draxler@phonetik.uni-muenchen.de, bengler@bmw.de

Abstract

Voice-operated devices are of particular interest in mobile environments, e.g. vehicles. They promise a natural and intuitive interface to devices and services, and they offer hands-free operation, a legal prerequisite for in-car usage in many European countries.

Spelling is a common task for the operation of voice operated devices, especially under unfavorable communication conditions. This paper presents a first analysis of the error and fluency rate for 4502 utterances from the German SpeechDat-Car database. The error rate was found to be between 1.7% and 4.4% for the spelling of natural items, and between 3.6% and 7.9% for artificial letter sequences. Only 3.6% of the utterances contained hesitations. These results suggest that spelling while driving might be a suitable means of fallback interaction if specific error recovery mechanisms are implemented.

1. Introduction

Compared to the state of the art, users' expectations towards speech products are huge. They expect that the introduction of speech operation will lead to a significant reduction of distraction from the driving task and an increase of usability and ease of use. In this context at least two conditions are crucial: A broader range of functions and domains is usable via speech operation; furthermore, in-car speech recognition has to get more robust against ambient noise and wrong operation by the user finding himself – while driving – in a dual task situation [1].

In some situations spelling is used as a fallback strategy instead of command recognition, e.g. for the input of names or the correction of recognition errors. Spelling also allows the user to enter complex information which would exceed the speech recognition's vocabulary. Finally, spelling is used to stabilize human machine interaction after recognition errors or to obtain a higher reliability for ambiguous results.

SpeechDat-Car is an EU-funded project (Contract LE8334) for the recording of large speech databases in mobile environments, e.g. vehicles. These databases will be used to develop voice-operated devices and services for mobile environments [2], [3], [4].

Until Spring 2001, 9 databases have been collected (Table 1).
The SpeechDat-Car databases will be made available for product development through ELRA in 2002 [5].

Each database consists of at least 600 recording sessions from a minimum of 300 speakers. Each recording session contains approximately 130 items. The vocabulary of SpeechDat-Car extends the original SpeechDat-II vocabulary by application specific command words and phrases, language dependent items, and spontaneous speech elicitations (Table 2) [6], [7].

Language	Partner
Danish	Sonofon, University of Aalborg
English	Vocalis Ltd.
Finnish	Nokia, Technical University of Tampere
Flemish	L&H
French	L&H, Renault
German	Bosch, BMW, University of Munich
Greek	Knowledge SA, University of Patras
Italian	Alcatel, IRST
Spanish	Polytechnical University of Catalonia, SEAT
US English	Siemens, ELRA

Table 1: SpeechDat-Car databases

Count	Corpus contents
2	voice activation keywords
12	isolated digit, digit sequence, phone number, PIN code, credit card numbers
3	dates, read and spontaneous
2	word spotting phrases with application word
7	spellings, read and spontaneous
1	money amount
1	natural number
7	person, city, or company names, read and spontaneous
9	phonetically rich sentences
2	times, read and spontaneous
4	4 phonetically rich words
67	application words
2	language dependant keywords
9	spontaneous speech items

Table 2: SpeechDat-Car database contents

2. SpeechDat-Car in Germany

Robert Bosch GmbH, a leading German manufacturer of electric and electronic car equipment, was the German contractor in the SpeechDat-Car project. BMW and the Department of Phonetics and Speech Communication at Munich University (IPSK) were subcontractors to Bosch.

BMW provided a vehicle for the recordings, Bosch installed the recording platform, and IPSK performed the recordings, annotated them, and produced master DVDs for the final data distribution [8].

2.1. Recording equipment

The SpeechDat-Car recording equipment consists of a 4-channel high-bandwidth mobile recording platform in the car (PLTM), and a synchronous GSM mobile phone connection recorded on a fixed ISDN recording platform in the lab (PLTF).

For PLTM, four microphones are installed in the car: a Shure close-talk microphone, AKG mouse microphones in the A-column and midway between driver and co-driver, and a Peiker mouse microphone above the speaker to the rear of the sunvisor. The GSM phone used the AKG mouse microphone that comes with the Nokia car kit. This microphone is placed on the ceiling to the left of the center microphone (Figure 1).

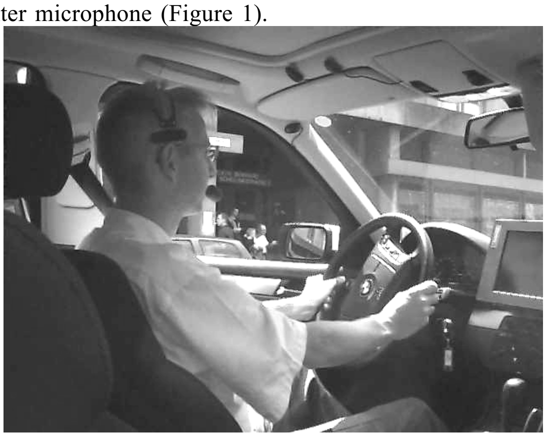

Figure 1: VEHIC1DE speaker position

In the boot of the car, an industry PC was installed. The microphones were connected to the PC via a DataTranslation four channel digital audio card. The PLTM recordings are have a sample rate of 16 KHz and 16 bit quantization, the PLTF recordings have a sample rate of 8 KHz, 8 bit alaw compression.

2.2. Recordings

In the German SpeechDat-Car recordings (VEHIC1DE), the speaker is actively driving. He or she reads prompts or answers questions presented on a display mounted near the center of the dashboard. Readability of the prompts was tested for all subjects. The experimenter on the co-driver seat controls the recording progress via an IR-keyboard.

Before a recording session starts, the speaker is briefed about the operation of the car and the recording procedure. During the first few recordings, the experimenter gives instructions on how to speak the current items. When the speaker is familiar with the task, no more instructions are given. In case the speaker makes some serious mistake, e.g. reading instead of spelling, the experimenter can repeat the recording of the current item.

VEHIC1DE recordings started in May 1999 and ended in Nov. 2000. Each recording session took about 40 to 55 minutes, and speakers were encouraged to do two sessions in sequence with a short recreational break in between. For the second session, either new traffic conditions were chosen, e.g. from town traffic to highway traffic, or some care settings were changed, e.g. windows or roof opened.

VEHIC1DE contains a total of 646 recording sessions by 338 speakers (179 male and 159 female). There are 179 speakers in the age class 18-30 years, 79 in the class 31-45 years, and 80 over 45 years. The spelling items are distributed evenly over all recording sessions. The traffic situations are biased towards city traffic:

Traffic	Environment	Count
stopped	engine running	84
city	no noise	83
	noise	137
low speed	no noise	83
	noise	94
highway	{no} noise	82
	audio	83

Table 3: Environment conditions

4502 spelling items are included in the final database.

2.3. Annotation

The VEHIC1DE recordings were annotated using the close-talk channel. Of the other channels, a subset of about 10% was checked acoustically.

The SpeechDat-Car annotation is an orthographic annotation with a set of markers for speech, signal, and noise phenomena (Table 4).

Type	Annotation	Comment
Speech	*word	mispronunciation or word fragment
	**	incomprehensible speech
Signal	~word | word~	signal truncation at begin or end
Noise	[int] | [sta]	intermediate or stationary non-articulatory noise
	[spk]	articulatory noise
	[fil]	filled pause or hesitation
	[dit]	dial tone

Table 4: SpeechDat-Car noise markers

Note that in VEHIC1DE annotations, [dit] was used to mark the prompt beep in the PLTM signal.

The annotations were performed using the WWWTranscribe software. This software supports the annotator with editing buttons for frequent editing tasks and performs a syntactic consistency check for annotation texts [8].

3. Analyses

The VEHIC1DE vocabulary specifies 7 spelling items which can be divided into the three classes SN, RN, and RA (Table 5):

Type	Task	Content
SN	spontaneous	natural item: first name
RN	read	natural item: person, company, or geographical name
RA	read	artificial letter sequence

Table 5: Spelling classes

Two parameters were measured:

- *Error rate*: percentage of incorrect spellings, and
- *Fluency rate*: percentage of utterances containing hesitations.

Both error rate and fluency rate were computed for the three classes of spelling items (SN, RN, RA). The results are shown in dependency of gender, age class, and traffic conditions.

3.1. Error rate

The *spelling target* is either the prompt text displayed by the recording platform in the car, or the correct spelling of the first name of the speaker. The spelling transcript is the part of the transcription text for the entire utterance that corresponds to the actual spelling. Note that for the analysis of spelling errors, all noise markers except [dit] for the system beep are removed from the transcription text.

A spelling is considered *correct* if

- the spelling transcript and the spelling target match exactly, or
- the spelling transcript includes the spelling target, i.e. it is surrounded by other text

There are a number of variants for spelling the German special letters Ä, Ö, Ü and ß. The umlauts can be spelled either by their corresponding phonemes /E:/, /2:/, and /y:/ (German SAM-PA), or the word “Umlaut” preceding or following the corresponding base letter.

“SZ” is the standard spelling for ß, but “scharf{es} S”, “doppel S”, or “dreier{les} S” (which is a regional variant for Svabian) may also be used. Note that “SZ” and "doppel S” are ambiguous: the letter sequence “S Z” can be acoustically indistinguishable from “SZ”, and “doppel S” can also be used for the letter sequence “S S”.

The spelling target contained non-letter items and non-native diacritics as well: apostrophes, double quotes, hyphens, periods, and the French accent diacritics ´, `, ^. If such items were not spelled, or if only some generic word was used for the accent diacritics, the spelling transcript was considered a variant and hence could be correct.

Furthermore, lower and upper case can be indicated by using key words, e.g. “groß”, “klein”, etc. Although they were not necessary for SpeechDat-Car prompts, some speakers used them nevertheless. These spellings were also considered as variants that could be correct.

A spelling is considered *incorrect* if

- the spelling transcript differs from the spelling target in at least one letter,
- the spelling transcript contains non-spelling words, or
- the spelling transcript begins before the recording beep.

Note that if a letter was transcribed as being mispronounced it was considered different from the target letter.

A total of 168 (=3,73%) incorrect spellings was found (Table 6, Table 7, Table 8).

	F	M
SN	2,63%	2,92%
RN	2,51%	4,47%
RA	4,95%	6,18%

Table 6: Spelling errors vs. gender

	18-30	31-45	46-99
SN	3,35%	2,24%	1,95%
RN	3,42%	3,29%	4,07%
RA	4,75%	6,77%	6,58%

Table 7: Spelling errors vs. age class

	Stopped	Town	Low speed	Highway
SN	1,19%	4,09%	1,69%	3,03%
RN	2,15%	4,01%	2,86%	4,37%
RA	3,57%	5,48%	4,57%	7,88%

Table 8: Spelling errors vs. environment conditions

3.2. Fluency rate

For the fluency rate analysis, the occurrences of the noise marker [fil] were counted. The basis assumption is that speakers produce more hesitations with increasing workload. Workload may be caused by external factors, e.g. traffic conditions, unfamiliar recording environment, etc., or internal influences, e.g. self-correction and re-reading of prompts, etc.

Note that for the analysis of the fluency rate, only the transcription text was used, not the speech signal. The markers [fil] and [spk] can be used functionally or based on the acoustic impression. Inter-transcriber variability may thus result in inconsistent use of these markers. Furthermore, some words may also be interpreted as hesitation indicators: “ach”, “ah”, “ja”, “oh”, “ne” and similar. Whether or not such a word could be substituted by a [fil] marker cannot be determined from the transcription alone.

160 (= 3.55%) spelling items were transcribed with the [fil] marker (Table 9, Table 10, Table 11).

	F	M
SN	2,30%	2,34%
RN	3,90%	3,94%
RA	2,64%	3,24%

Table 9: Fluency rate vs. gender

	18-30	31-45	46-99
SN	1,96%	1,49%	3,90%
RN	3,31%	4,63%	4,72%
RA	2,23%	4,51%	3,29%

Table 10: Fluency rate vs. age class

	Stopped	Town	Country	Highway
SN	1,19%	2,27%	1,69%	3,64%
RN	1,67%	4,47%	3,66%	4,62%
RA	2,38%	1,83%	3,43%	4,24%

Table 11: Fluency rate vs. environment conditions

4. CONCLUSION

The overall error rate is low. This is especially true for conditions that are comparable to realistic use cases (i.e. SN and RN).

This corroberates the belief that the quality of speech production is influenced by the primary driving task to an acceptable degree. The degree of interference can be estimated by having a look at error and fluency rates in situations where the car is stopped. This shows that there is a remaining subset of problems caused by the inherent structure of the spelling task itself ranging from 1.19% to 3.57% error rate reflecting the task difficulty.

As expected no significant differences could be found between male and female subjects. The differences between the age groups have to be investigated further but might be caused by other factors than cognitive capabilities.

The comparison of error rates in different traffic situations reveals that there is a remarkable influence of different situations, where the low-speed country road condition seems to have the lowest impact on speech production.

The results show that spelling can be used as a fallback strategy in speech driven human machine interaction. But one has to take into account that besides other problems like different spelling strategies and probably reduced recognition rates a given rate of user mistakes has to be handled. In most cases the value of spelling might be reduced regarding the influence of different traffic situations. The effects that can be observed under higher workload are increased rate of mis-spellings and hesitation phenomena. Therefore we have to consider how existing recognition technology is working if speech fluency is reduced by workload.

5. OUTLOOK

The given results show that spelling could be used as a fallback strategy in speech driven human machine interaction. But one has to take into account that besides other problems like different spelling strategies and probably reduced recognition rates a given rate of user mistakes has to be handled. One has to keep in mind that the value of spelling might be reduced regarding the influence of traffic situations.

As this study was focused on effects appearing during spelling, it would be of interest to get more knowledge about the dependencies between continuous speech and driving activities. These analyses can be performed on SpeechDat-Car material as well.

While the SpeechDat-Car project standardized the gathering of speech data, future investigations should also put more emphasis onto the standardization and description of the actual traffic situation. Possible classification schemes are described in [10]. In the case of small error rates a differentiated situation taxonomy could help to get a better understanding of the situative influence.

A more detailed study of speaker performance not only for spelling, but also for reading and spontaneous speech is currently being carried out by Cristina Olaverri in her MA thesis (expected end date: Nov. 2001).

6. ACKNOWLEDGEMENT

Part of the SpeechDat-Car project was funded by the Commission of the European Communities, Telematics Applications Programme, Language Engineering, Contract LE8334.

7. REFERENCES

[1] Bengler, K. (2000) Automotive Speech-Recognition - Success Conditions Beyond Recognition Rate. *Proc. LREC 2000, Athens, pp. 1357-1360*

[2] SpeechDat Family: http://www.speechdat.org/

[3] Heuvel, H., Bonafonte, A., Boudy, J., Dufour, S., Lockwood, Ph., Moreno, A., Richard, G. (1999) SpeechDat-Car: Towards a Collection of Speech Databases for Automotive Environments, *Proc. Nokia-COST249 Workshop, Tampere*

[4] Van den Heuvel, H., Boudy, J. Comeyne, R., Euler, S. , Moreno, A. Richard, G. (1999) *The SpeechDat-Car multiligual speech databases for in-car applications: some first validation results. Proc. Eurospeech 99, Budapest, pp. 2279-2282*

[5] ELRA/ELDA: http://www.icp.grenet.fr/ELRA/

[6] Dufour, S. (1998) Specification of the car speech database (definition of corpus, scripts and standard), Car environments and speaker coverage, SpeechDat-Car report SD1.12, 1998

[7] Draxler, Chr. (1999) Specification of Database Interchange Format. SpeechDat-Car Report LE-8334-D1.3.3

[8] Draxler, Chr., Grudszus, R., Euler, St., Bengler, K. (1999) First experiences of the German SpeechDat-Car Database Collection in Mobile Environments. *Proc. Eurospeech 99, Budapest, pp. 919-922*

[9] Draxler, Chr. (1997) WWWTranscribe - A Modular Transcription System Based on the World Wide Web. *Proc. Eurospeech 97, Rhodes*

[10] Fastenmeier, W. (1995): Die Verkehrssituation als Analyseeinheit im Verkehrssystem. In: *W. Fastenmeier (Hrsg.)(1995): Autofahrer und Verkehrssituation. Neue Wege zur Bewertung von Sicherheit und Zuverlässigkeit moderner Straßenverkehrssysteme. Reihe Mensch - Fahrzeug - Umwelt, Band 33. Köln: Verlag TÜV Rheinland*; Bonn: Deutscher Psychologen Verlag. S. 27-78.