SmartWeb UMTS Speech Data Collection
The SmartWeb Handheld Corpus

Hannes Mögele*, Moritz Kaiser*, Florian Schiel†

*Bavarian Archive for Speech Signals, Schellingstraße 3, 80799 München, Germany
{ariser,hannes}@bas.uni-muenchen.de
†BAS Services Schiel, Moltkestr. 1, 80803 München, Germany
schiel@bas-services.de

Abstract

In this paper we outline the German speech data collection for the SmartWeb project, which is funded by the German Ministry of Science and Education. We focus on the SmartWeb Handheld Corpus (SHC), which has been collected by the Bavarian Archive for Speech Signals (BAS) at the Phonetic Institute (IPSK) of Munich University. Signals of SHC are being recorded in real-life environments (indoor and outdoor) with real background noise as well as real transmission line errors. We developed a new elicitation method and recording technique, called *situational prompting*, which facilitates collecting realistic dialogue speech data in a cost efficient way. We can show that almost realistic speech queries to a dialogue system issued over a mobile PDA or smart phone can be collected very efficiently using an automatic speech server. We describe the technical and linguistic features of the resulting speech corpus, which will be publicly available at BAS or ELDA.

1. Introduction

SmartWeb[1] is a research project funded by the German Ministry of Science and Education (grant number 01 IMD 01I). The aim of this project is to enable the user to access semantic Web services via mobile UMTS handheld devices by means of a multi-modal user interface (Wahlster, 2004). The SmartWeb speech data collection is carried out by the BAS (Bavarian Archive for Speech Signals) at the Phonetic Institute (IPSK) of Munich University. In the course of this data collection three new speech resources will be created: The SmartWeb Handheld Corpus (SHC), the SmartWeb Motorbike Corpus (SMC) and the SmartWeb Video Corpus (SVC). These corpora form the empirical basis for the development of the man machine interface of the SmartWeb system. In this paper the main focus will be the SHC. A detailed description of SMC can be found in Kaiser et al. (2006).

Nowadays corpus creation as resource for applied research in the field of speech technology is subject to two conditions: it should be as *economical* as possible and as *realistic* as possible. The latter refers to the properties of the physical speech signal as well as to the speaking style of the speakers. In our context this means an acoustic signal transferred over UMTS and Bluetooth channel, recorded in real environment augmented by an/receive annotation that covers linguistic phenomena typical for spontaneous speech. Hence, one of the main challenges in corpus construction is how to elicit representative utterances from the participants that are as natural as possible in a real situation and in a cost efficient way.

There exist various techniques of speech data collection depending on the purpose of corpus and the intended application or target system. Of course, these techniques differ in effort and costs. Furthermore, each technique elicits speech in various speaking styles such as *read speech*, *answering speech*, *command and control speech*, *dictation speech*, *descriptive speech*, *non-prompted speech*, *spontaneous speech* and *emotional speech* (Schiel and Draxler, 2003).

Important data collection methods, which are designed to obtain natural spoken language data or multi-modal data are, for instance, *Wizard-of-Oz experiment* (WOZ) Türk (2001), *Video Task (Daily Soap Scenario)* (Peters, 2001) and *script experiments* (Rapp and Strube, 2002). *WOZ* uses a realistic simulation of all functionalities of a fully-deployed target system[2]. *Script experiments* combine features of WOZ and *prompting experiments* and were applied as one strategy within an iterative data collection approach by (Rapp and Strube, 2002).

All of these methods are able to elicit spontaneously uttered speech at different levels, but require human support and are therefore expensive and time-consuming. Goal of the SmartWeb data collection is to economically record a variety of naturally spoken requests that might be uttered by users of such a system under realistic field conditions.
the possibility to record speech data in intended speaking style in real situations.[3]

2. Methodology: Situational Prompting

2.1. Elicitation Method

The development of SitPro is based on the experience we gained in the WOZ simulation of the SmartKom system (Türk, 2001) and in a pilot study (Steininger, 2005) of human-human telephone dialogues for the SmartWeb data collection. SitPro combines *script methods* with *interview techniques* and *speaker prompting* (Gibbon et al., 1997). Speaker prompting is only used for instruction, information or feedback prompts. Repetition and question prompts are not applied.

[1] http://www.smartweb-projekt.de/

[2] For further informations of WOZ especially in the SmartKom project cp. (T¨urk, 2001)

[3] Note that the same technique will be used for all SmartWeb speech data collection.

prompt category	mode of topic
standard	given
individualized	open
script	open, given

Table 1: SHC prompt category and corresponding mode of topic

topic	example
soccer	team, group
navigation	public transport, pedestrians
community	restaurant
information	tourist information, points of interest

Table 2: Examples for SmartWeb topics

This resulted in three different prompt categories, called *standard prompts*, *individualized prompts* and *script prompts*.
In a *standard prompt* the subject is told a topic (cp. table 2) to which she/he is supposed to pose a query (see example table 4).
An *individualized prompt* is a prompt for which the subject provides his/her own topic (see example in table 6).
A *scripted prompt* simulates a three-turn conversation as frequently found in dialogue between human and machine (see example in table 5).
The automatic prompting system of SHC simulates two interlocutors. The *instructor* (female voice) gives directions about the situation and the topics, while the *operator* (male voice) answers the subjects' questions or gives feedback like the SmartWeb system.

2.2. Prompts Preparation

2.2.1. Text-prompts provide a basis

A *prompt unit* consists of a system prompt (pro), a variable silence interval, followed by a recording (rec) and a possible system answer (opr). Six prompt units are bundled into a thematic *action unit*. The silence interval at the end of a system prompt allows us to adjust the time between end of prompt playing and start of the recording. The length of this pause depends, on the one hand, on task complexity, script composition and individual prompt structure and, on the other hand, on recording situation and subjects' mental workload. A precise adjustment of these factors is necessary to minimize artefacts in the recordings resulting from subjects' mental overload.

2.2.2. Audio-prompts Generation

The audio-prompts are generated from the text-prompts using the text-to-speech system of AT&T[4]. We decided against pre-recorded prompts for two reasons: first the situation is more realistic using synthesized voices, second the usage of pre-recorded speech would have exceeded our budget. We selected the German AT&T male voice *Reiner* for the operator and the German female voice *Klara* to inpersonate the instructor. Unfortunately, it turned out that the spelling of the text prompts has to be adjusted manually so that the synthesized speech can be understood well over the UMTS phone line. This applies mainly for proper names and foreign words; some examples in German are given in table 3. We collected the adapted spellings into a synthesis lexicon that can be used for future semi automatic adjustments of text-prompts.

[4]http://www.naturalvoices.att.com

correct orthography	adapted orthography
kurvigste Route	kur wiggste Route
empfangene Messages	empfangene Messädsches
Britney Spears	Brithnie Sspiers

Table 3: Examples for manually adapted spelling for synthesis

flow	text prompt	pause
pro-010	Please think of the soccer WM. You want to get information about results and games of several teams. First you want to know how the last game of Germany against Costa Rica ended.	4000
rec-010	*What was the result of the match Germany against Costa Rica?*	
pro-010	Now you want to get information about who else plays in the group of England and the Netherlands.	4000
rec-020	*What other teams are in the group of Britain and Holland?*	
pro-031	Next you are interested in the time of the next game of Mexico against Ukraine.	4000
rec-030	*When is the match Mexico against Ukraine*	
...		

Table 4: Example of SmartWeb recording dialogue using a *standard prompt scheme*. *pro* denotes the (female) instructor's voice; *opr* denotes the (male) operator's voice; *rec* indicates a recording of the users's elicited query.

2.3. Recording Procedure

A few days berfore the recording the subject receives preliminary information with a first short explanation about the recording procedure and the request to prepare some topics he would like to ask the Smartweb information system. These notes are necessary for the recording session to manage the *individualized prompts* that are filled by the subject with topics and names of their individual background. This elicitation method has been tested in a pilot study (Steininger, 2005). During this pilot study we noticed that it was very easy for subjects accommodate to these individualized prompts.
The recordings take place in real environments and with real equipment. This means that the *external context* of the subject is realistic. The subjects are induced to visualize a situation in which they might use an information system

flow	text prompt	pause
pro-010	In the following block the topic is general knowledge. Please try to find out who painted the ceiling fresco of the Sistine Chapel.	3000
rec-010	*Who did the ceiling painting in the Sistine Chapel?*	
opr-010	The fresco was done by Michelangelo.	1500
pro-020	You were not able to properly understand the name. Please ask for a repetition	2000
rec-020	*Could you say that again?*	
...		

Table 5: Example of SmartWeb recording dialogue using a *scripted prompt scheme*. *pro* denotes the (female) instructor's voice; *opr* denotes the (male) operator's voice; *rec* indicates a recording of the users's elicited query.

flow	text prompt	pause
pro-010	Please refer now to the notes with your prepared topics. These topics should be the focus of the following six prompts. Please start by asking a general question about your topic number six.	6000
rec-010	*Where is the new Pinakothek in Munich?*	
pro-020	Now please ask a second, more specific question about this sixth topic.	5000
rec-020	*Uhm ... who was the architect?*	
pro-030	Very good! Now please ask a question about topic number one.	5000
rec-030	*Which country has the largest fresh water consumtion?*	
...		

Table 6: Example of SmartWeb recording dialogue using a *individualized prompt scheme*. *pro* denotes the (female) instructor's voice; *opr* denotes the (male) operator's voice; *rec* indicates a recording of the users's elicited query.

like SmartWeb. By the use of subjects' imagination of such possible real situations the *intrinsic context* is close to reality. We found that the external and internal context of the participants establish a sufficient communicational context to motivate natural statements.

2.4. Situational Prompting at a Glance

- preparation and open instructions
- detailed instructions of the test subjects
- external context: real environments
- internal context: imagination of possible real situations
- prompts bundled in thematic action units
- only acoustic prompting
- synthesised prompt speech
- two interlocutors on system side
- transferable to other domains and scenarios
- mobile recording equipment

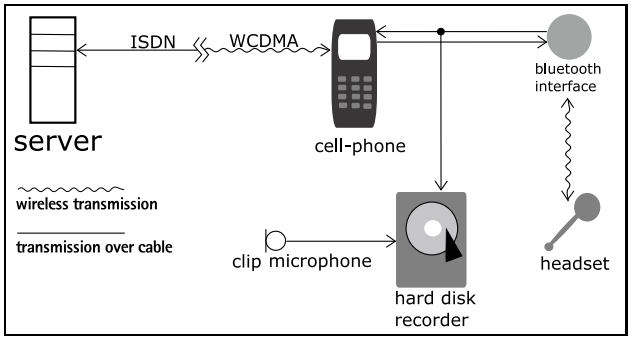

Figure 1: Signal flow of SHC recording.

3. Corpus Collection

3.1. Recording Technique

The speaker uses a standard UMTS cellular phone[5] in combination with either a wireless[6] or cable connected headset (for the various used headset types see table 8). A second microphone is attached to the collar of the speaker. Both microphone signals, the collar microphone and the output of the Bluetooth microphone after the Bluetooth transmission, are recorded by a portable hard disc recorder throughout the recording session, which lasts approximately 30 minutes. The Bluetooth signal is transmitted via WCDMA (UMTS) and the telephone network of the German Telecom to an ISDN speech server at the BAS. Figure 1 depict an overview of the signal flow.
The speech server records the UMTS speech signal for the total length of the session as well as the individual queries starting after the prompt beep and ending either with the maximum recording time of 12 seconds or when the silence detection signals the end of the speaker's turn. This results in four types of recorded speech signals:

- high quality[7] recording of Bluetooth mike
- high quality recording of collar mic
- UMTS quality[8] recording of total session
- UMTS quality recording of individual queries

The same recording technique was also used successfully in the speech and video data collection for SMC and SVC.

[5] Siemens U15
[6] Bluetooth 1.1
[7] 16bit; 44,1kHz; PCM
[8] 8bit; 8kHz; ALAW

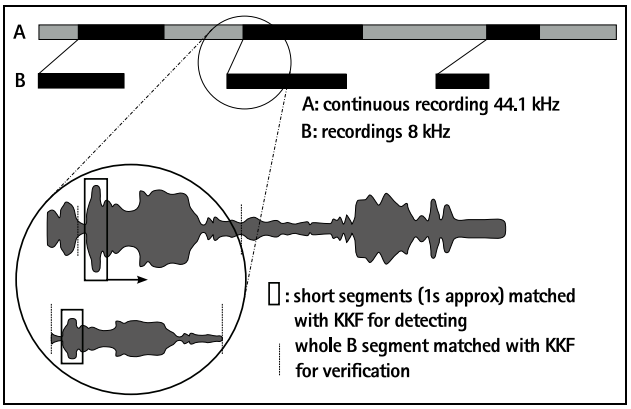

Figure 2: Automatic segmentation of harddisk recordings using a cross-correlation technique.

3.2. Speaker Recruitment

There are only a few restrictions on participating in the SmartWeb data collection. Speakers should speak German fluently and be familiar with the use of a cellular phone. Dialect speakers and foreign language speakers with accent are recorded, but not specifically recruited. Non-students and persons over 35 years are systematically recruited, to form a counter-balance to the students that typically volunteer for such recordings close to a university. Participants are equally distributed in gender. Up to now we have recorded 156 speakers meeting the mentioned criteria. There are 86 female and 79 male subjects aged 13 to 70 years.

3.3. Post-processing

3.3.1. Automatic Query Segmentation

The individual query recordings of the speech server are automatically aligned to the high quality hard disc recordings using an un-supervised cross correlation scheme (Kaiser and Schiel, 2005) yielding the identical turn segmentations as in the server recordings (see figure 2). The result of this segmentation is stored in a marker file for each session. The three-column table in the marker file lists the recording number which corresponds to the file name of the server recording together with the first and last sample of the segment within the hard disc recording. Using these values it is possible to cut out corresponding segments from the hard disc recording automatically (for instance for ASR training).

3.3.2. Segmentation with MAUS

All SHC recordings have been segmented into phonemic segments (extended German SAMPA[11]) using the MAUS method (Schiel (1999), Schiel (2004)). To avoid the strong disturbances and drop outs of the Bluetooth and UMTS transmission lines the harddisk recording containing the collar microphone was used as the base signal for segmentation. But even with this relatively clean signal we found that the signal preceeding and tailing the actual speech input contains too much crosstalk and background noise for statistical modeling to work properly. The effect of this is a spreading of the initial and final phonemic segments into non-speech parts of the signal. The reason for this lies probably in the low S/N ratio of the field recordings. Therefore we applied a simple speech tracking algorithm that detects the actual begin and end of the user's query to approx. 250msec accurarcy and then restricted the MAUS search to this interval. This method yielded satisfying results. The speech detection algorithm will be incorporated as an option into the public domain software package MAUS[9] in the next release.

rec ID	first sample	last sample
oip-0150rec-010	14202434	14504226
oip-0150rec-020	14970211	15384045
oip-0150rec-030	15929514	16305813
...	...	...
npm-0120rec-020	62301339	62736771
npm-0120rec-030	63294210	63643046
npm-0120rec-040	64062461	64566634
npm-0120rec-050	64943552	65441512
npm-0120rec-060	65981664	66376733

Table 7: marker file entries

3.4. Transliteration

For annotating the speech data we use the client-server application *WebTranscribe*. WebTranscribe (Draxler, 2005) is a modern platform-independent web-based annotation framework with a client/server architecture for the transliteration of speech signals. The annotation client for SmartWeb is JavaWebStart application and consists of a signal display, two separate editors and a set of editing and label buttons which add pre-defined marker symbols into the text. The recorded speech is transcribed in a two-stage process. For the basic transcription the first editor shows the prompt in text form as context information for the transcriber. The second editor is used for the basic orthographic transcription. On the second level of the transcription process the first editor displays the basic transcript from the first stage; the transcriber then corrects/augments this transcript to yield the final form of transcript. All transcripts are stored in a PostgreSQL database. As annotation scheme we use a reduced Verbmobil/SmartKom transliteration set[10] to tag items like foreign words, acronyms, spelling, lengthening and so on. In addition, we judge the signal quality for each recording. All final transliterations are passed through a quality assurance stage where one of two experienced transcribers reviews all data before distribution.

4. SHC Data

4.1. Contents

The current release (1.3) of SHC contains 155 recording sessions with 11.930 recordings and approx. 147.932 token words. 4.815 different word forms were encountered; a large number of these are proper names.

[9] http://www.phonetik.uni-muenchen.de/Bas/software

[10] http://www.is.cs.cmu.edu/trl_conventions/projects/verbmobil_entrance.html contains a detailed description of the Verbmobil tag set.

type of headset	out	in	SUM
[b]Samsung WEP 150 MBE	22	4	26
[b]PlantronicsM 3500	4	3	7
[b]Siemens HHB 505	17	9	26
[b]Logitech	9	36	45
[c]GN netcom GN 2100 flex boom 82	40	10	50
SUM	92	62	154

Table 8: Microphone types and number of indoor and outdoor recordings ([b]Bluetooth, [c]cable)

Table 9 shows the distribution of action units (6 prompts each) over the SmartWeb domain topics while table 10 shows the number of action units per prompting category. The situational prompting scheme seems to work fine: The same prompts elicit a various range of subject queries. Examples are given in the tables 12, 13 and 14 which depict a number of recorded queries after the same single prompt (on top). The recorded data are spoken with natural prosody and contain various features of spontaneously uttered speech like *disfluencies*, *pauses*, *hesitations*, *false starts*, *ungrammatical sentences*, *interruptions* etc. Examples of some typical linguistic phenomena are given in table 11 (in German).

4.2. Speech Signals

Since we use different combinations of recording equipment — the cellular phone is combined with five cable and Bluetooth headsets (see table 8) — in different noisy environments (indoors and outdoors), the signal quality varies tremendously. The acoustical environment ranges from quiet office to very noisy train station (with S/N of less than 6dB). Permanent and transient background noise occurs quite frequently as well as strong crosstalk from single or groups of speakers.

4.3. Annotations

The corpus provides one transliteration file for each recording session as well as BAS Partitur Format (BPF) files for each individual recording. A marker file contains the alignment between server and harddisk recordings.

4.4. Meta Data and Lexicon

The parameters of each recording session are described in a XML formatted recording protocol. Speaker charcteristics are listed in a XML formatted speaker profile.
Based on the spoken words of the total corpus a list of manually checked canonical pronunciations was added to the corpus. The dictionary contains 4.815 word forms. Pronuciations are coded in German extended SAMPA[11].

4.5. Distribution Scheme

The SHC will be distributed on DVD-R via the BAS[12] and the ELDA[13]. Since the corpus was 100% funded by the German government, SHC will be distributed without royalties

[11]http://www.bas.uni-muenchen.de/Bas/BasSAMPA
[12]http://www.bas.uni-muenchen.de/bas
[13]http://www.elda.org

topic	SUM
community	82
soccer	563
navigation	318
open topic	430
tourist information	295
public transport	129

Table 9: Number of action units counted by topics

category	SUM
prompted	847
individualized	617
scripted	353

Table 10: Number of action units counted by categories

or other restrictions. A first publicly available release is to expected in June 2006.

5. Conclusion

A speech corpus SHC with realistic speech queries spoken to a mobile hand-held HMI device has been collected and will soon be available via BAS and ELDA. Signals undergo a complex chain of speech transmissions including Bluetooth and UMTS and are prone to real-life background noise. To save time and costs we developed a new fully automatic prompting scheme called *situational prompting* that enables us to collect natural spoken queries for real-life applications without using the costly Wizard-of-Oz method. All 11.930 recordings of the SHC will be transliterated in a reduced Verbmobil tagging scheme and phonemically segmented using the MAUS method adapted

linguistic phenomenon	example
disfluency, pause	wer h= ha= hat sonst noch in diesem <P> Film mitgespielt ?
false start, hesitation	<hm> wann ist die <”ah> wann f”angt die Fu”sballmeisterschaft an ?
interruption, hesitation	gibt es aktuelle Nach_ <P> _richten von Herrn Moshammer ?
repetition	zeigen Sie mir bitte eine Karte von von der Umgebung .
ungrammatical, hesitation	<”ah> wie viele $U-Bahn-Linien gibt in ˜Paris ?
self-corrections, hesitation	wie komme ich zum Hauptbahnhof <”ah> vom Hauptbahnhof zu einem chinesischen Restaurant ?

Table 11: Examples for encountered linguistic phenomena

Instructor: **Please check again with SmartWeb whether the information was correct.**
Are you sure that public transportation continues that late?
What time did you say?
... catch you right? Did I catch you right?
Excuse me, until when will there be public transportation tonight?
Until which time will there be public transportation in Leipzig?
You're sure that the underground runs until two o'clock at night?

Table 12: Examples for queries elicited by *standard prompts* in the thematic context of *public transportation*: After being prompted with the text in the top line, the speakers uttered various responses. (translated to English for better readability)

Instructor: **There is a bus stop near Hegelstrasse. Would you like to learn more about the schedule?**
Yes.
Yes, I'd love to.
When does the next bus leave towards the train station at this bus stop?
Yes, I'd like to know when the next bus leaves.
Yes. Yes . I wanna catch the last bus that leaves from this bus stop.
Yes, please . What time the buses are leaving there? What number would that be and where does it go?
Nah, just where is that next bus stop? I'll be waiting there for the next bus.
No, thanks.

Table 13: Examples for queries elicited by *scripted prompts* in the thematic context of *navigation*: After being prompted with the text in the top line, the speakers uttered various responses. (translated to English for better readability)

to noisy fi eld recordings. The public release of SHC will contain three different quality speech signals, a segmentation in queries, transliteration, phonemic segmentation and a manually checked pronunciation dictionary and will be available mid of 2006.

6. References

Christoph Draxler. 2005. WebTranscribe - An Extensible Web-based Speech Annotation Framework. In *Proceedings of 8th International Conference, TSD 2005*, Karlovy Vary, Czech Rebublic, September. Springer Verlag.

Dafydd Gibbon, Roger Moore, and Richard Winski, editors. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Walter de Gruyter.

Moritz Kaiser and Florian Schiel. 2005. Techniques of Speech Data Collection. SmartWeb Technical Document No. 09 Version 1.0, University of Munich, Institute of Phonetics.

Instructor: **Imagine, you've got the wrong answer. Tell Smartweb about it.**
That's not correct.
I beg your pardon?
Well, these are not the right bulletin boards to meet other people.
No, no, I'd like to see other bulletin boards.
Please correct. Please correct.
Nah, that can't be right.
That answer must be wrong.

Table 14: Examples for queries elicited by *individualized prompts* in the thematic context of *community*: After being prompted with the text in the top line, the speakers uttered various responses. (translated to English for better readability)

Moritz Kaiser, Hannes Mögele, and Florian Schiel. 2006. Bikers Accessing the Web: The SmartWeb Motorbike Corpus. In *Proceedings of the LREC 2006*, page to appear, Genova, Italy, May. ELRA.

Benno Peters. 2001. 'Video Task' oder 'Daily Soap Scenario'. Ein neues Verfahren zur kontrollierten Elizitation von Spontansprache. http://www.ipds.uni-kiel.de/pub_exx/bp2001_1/Linda21.html.

Stefan Rapp and Michael Strube. 2002. An Iterative Data Collection Approach for Multimodal Dialogue Systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 661–665, Las Palmas, Canary Islands, Spain, May.

Florian Schiel and Christoph Draxler. 2003. *Production and Validation of Speech Corpora*. Bastard Verlag, München.

Florian Schiel. 1999. Automatic Phonetic Transcription of Non-Prompted Speech. In *Proceedings of the International Conference of Phonetic Sciences 1999*, pages 607–610, San Francisco, USA.

Florian Schiel. 2004. MAUS Goes Iterative. In *Proceedings of the IV. International Conference on Language Resources and Evaluation*, pages 1015–1018, Lisbon, Portugal.

Silke Steininger. 2005. Data collection pilot study - human-human telephone dialgoues. SmartKom Technical Document No. 01 Version 1.0, University of Munich, Institute of Phonetics.

Ulrich Türk. 2001. The Technical Processing in SmartKom Data Collection: a Case Study. In *Proceedings of Eurospeech 2001 Scandianavia*, pages 1541–1544, Aalborg, Denmark, September.

Wolfgang Wahlster. 2004. Smartweb: Mobile applications of the semantic web. http://smartweb.dfki.de/Vortraege/SmartWeb-Wahlster-KI-2004-LNAI.pdf.