Bikers Accessing the Web: The SmartWeb Motorbike Corpus

Moritz Kaiser*, Hannes Mögele*, 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

Three advanced German speech corpora have been collected during the German SmartWeb project. One of them, the SmartWeb Motorbike Corpus (SMC) is described in this paper. As with all SmartWeb speech corpora (e.g. (M¨ogele et al., 2006)) SMC is designed for a dialogue system dealing with open domains. The corpus is recorded under the special circumstances of a motorbike ride and contains utterances of the driver related to information retrieval from various sources and different topics. Audio tracks show characteristic noise from the engine and surrounding traffic as well as drop outs caused by the transmission over Bluetooth and the UMTS mobile network. We discuss the problems of the technical setup and the fully automatic evocation of natural-spoken queries by means of dialogue-like sequences.

1. Introduction

The growing use of spoken human machine interfaces (HMI) results in the expansion of such applications into areas other than fixed telephone line networks or quiet acoustical surroundings such as office environments. Caused by the tremendous growth during the last decade mobile networks are necessarily at the focus of current automatic information systems. The German SmartWeb project funded by the German Ministry of Science and Education (grant number 01 IMD 01I) tries to address this issue. Ultimately information content about virtually every topic should be made retrievable from a HMI over a smart phone.

A considerable challenge is to enable motorcyclists to make use of a spoken HMI while driving. To meet the requirements of the SmartWeb project partners with regards to ASR training and dialogue design the SMC was collected in 2005. To cover the circumstances of real on-the-bike situations, recordings were done during a ride performed by professional test drivers of BMW company in real traffic environment in the City of Munich. In the following the technical setup, procedure of recording and first results as well as experiences with this difficult type of field recording will be discussed in detail.

2. Target Motorcycle Scenario

A motorcyclist will access SmartWeb or any other dialogue system in a different way than a pedestrian on a side walk, a person in a car or a building. Motorcyclists will use SmartWeb to gather information that is closely related to his or her current activity, i.e. related to traffic situations, weather, technical state of the vehicle, destination of the journey and similar matters. Complex themes like retrieval of encyclopaedic knowledge or general news will occur less likely although sites of interest might be a major topic. A motorcyclist will have to concentrate on steering and will have to watch the traffic with high attentiveness. As a consequence dialogues will often be interrupted due to traffic situations and speech may be more error prone than in a traditional hand-held situation. Driving on a motorcycle will increase the noise level from surrounding traffic, and driving at higher velocities will add a significant amount of wind noise. The increased noise will cause the speaker to raise his/her voice (Lombard effect) and additional strain may change some characteristics of the voice as well. Since the voice signal is to be transferred over Bluetooth from the helmet of the driver to the host system of the motorcycle[1] and possibly transferred over UMTS to a server, frequent transmission errors in form of glitches and disruptions will happen, caused by interferences with other Bluetooth channels, UMTS cell hand over and loss of data frames. To achieve a more widely usable speech corpus and to cover more than just the special server-based SmartWeb scenario described so far, we added two more possible subscenarios, where the dialogue system is located on the host system of the vehicle and being fed either through a Bluetooth headset or a cable connected throat microphone.

3. SMC Recording Setup

The SMC recording setup tries to adopt the above scenario as closely as possible. The scenario requires different possible types of signal flow, therefore different channels have to be recorded which cover the respective sub-scenario.

3.1. Signal sources

Motorcyclists always wear helmets and modern helmets are already prepared for communication purposes. Often helmets are equipped with speakers and microphones and in some cases with a wireless transceiver. Due to the cooperation with the Department for Science and Technique at BMW, Munich, it was possible to use a helmet with two speakers, two microphones and a built-in Bluetooth transceiver. The helmet[2] contains two microphones which

[1] A cable connection would of course be more robust but seems quite unlikely in the future, since Bluetooth connected headsets are becoming widely used these days.

[2] "BMW Systemhelm" with WCS-2 communication controller

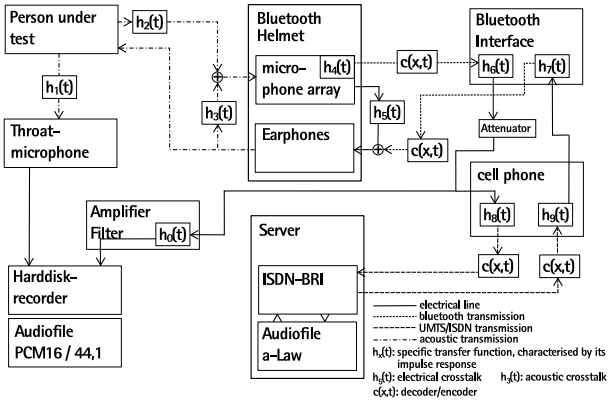

Figure 1: signal flow diagram

are located on top of the forehead, embedded into the EPS padding with a distance of approx. 7cm. The two signals are processed by a DSP to achieve a high directivity towards the mouth of the driver.

For most of the recordings a throat microphone (Type KEP 33-S) was used in parallel. It consists of a plastic bracket which can be fitted closely around the neck. On both ends it carries microphone capsule with 3cm in diameter. Those capsules are placed left and right of the larynx.

Both the throat microphone and the helmet microphones use standard electret capsules (thus avoiding magnetic field noise).

3.2. Signal Transmission

The intra-helmet speech signal is transmitted by the built-in Bluetooth circuit to an external Bluetooth transceiver which delivers the signal to an attenuator. The attenuator splits the signal to one of the line inputs of the hard disk recorder as well as to the microphone input of the UMTS mobile phone. The mobile phone is configured to use WCDMA protocol (UMTS) exclusively.

The throat microphone is directly connected to the other line input of the hard disk recorder.

The signal coming from the helmet is subject to several influences roughly described in figure 1. $h_2(t)$ and $h_3(t)$ describe the transfer functions of the acoustic path, while $h_3(t)$ models the acoustic crosstalk between the headphones and the microphones. All other $h_x(t)$ refer to electrical transfer functions, e.g. low pass filters and partially to electrical crosstalk. All functions of the form $c_y(x,t)$ stand for compression and expansion algorithms applied to the signal on digital transmissions. Those occur between the helmet and the Bluetooth interface and between the mobile phone and the voice server. Noise sources are not included in figure 1 but can be considered to influence nearly every part of the transmission except between the UMTS-Network and the server.

3.2.1. Bluetooth transceiver

The Bluetooth transceiver is an off-the-shelf device usually used for the extension of phones which do not support Bluetooth. It cannot be configured by any means and has a simple 2.5 mm stereo jack. It expects a 2 Volts bias on the microphone line to start. This voltage is usually supplied by every mobile phone to operate external electret microphones. The model used in the SMC recordings is the *B-Speech Bluetooth hands-free* adapter.

3.2.2. Amplifier

The amplifier is a self-made circuit with adjustable gain and an approx. cut-off frequency of 16 kHz. For the motorcycle recordings the gain was set to 14,5dB. The main task of the amplifier was to provide a signal to the hard disk recorder input with an amplitude of the same order of magnitude as the signal from the throat microphone. The second task was to separate the recorder circuit from the 2V bias provided by the mobile phone, because this bias seemed to carry some high frequency noise.

3.2.3. UMTS Mobile Phone

For all recording sessions a *Siemens U15* mobile phone, which is similar to the *A835* from Motorola, was used to connect to the server. This UMTS phone allows the user to select the type of mobile network[3]. To maintain reproducible conditions throughout the recording the phone was configured to connect only to the UMTS network of the *T-Mobile Corporation*. However, at the time of the recordings[4] it turned out that not all areas in Munich were suitable covered by the UMTS network. Therefore a special route through the city was selected where a full coverage with UMTS base stations could be guaranteed. Leaving this route caused the connection to be severed and the recording was interupted.

3.3. Signal Recording

3.3.1. Hard Disk Recorder

For all recordings the iRiver iHP120 was used. This device features one analog stereo input and output, one stereo headphone output and one S/PDIF input and output. Input sensitivity can be adjusted over a wide range and the analog input features a 2 V supply voltage for electret microphones.

For SMC the hard disk recorder was configured to record at 16 bit, 44.1 kHz on the analog input; no compression applied. This results in theoretical capacity of roughly 50h.

3.3.2. Server

The main server is a PC with *AMD K7 500MHz* and a total amount of RAM of 2^{28} Byte (approx. 250 MB) running *SuSE Linux 9.0* with kernel 2.4.21. The connectivity to the fixed telephone line network is provided via an *AVM Fritz!PCI v.2.0 (rev 2)* ISDN controller, using the standard kernel driver for CAPI[5]. To record audio signals from the phone line, a special interpreter was developed at our site in Java using JNI. The interpreter is capable of sending and receiving A-law encoded signals to and from a telephone line following simple scripts. The scripts are formatted in a subset of Voice-XML (QXML) to allow simple sequences of prompts and recordings. The properties of the server are:

- parses QXML (subset of Voice-XML, i.e. no grammar, no ASR)

[3] which is *not* possible with most other UMTS phones.

[4] Aug-Dec 2005

[5] The Common ISDN Application Programming Interface

- playback of arbitrary audio files
- recording of prompted audio signals into raw A-law files
- detection of DTMF signals on a form
- detection of DTMF signals during recording/playback for barge-in
- ability to fetch new XML documents through HTTP from arbitrary hosts
- variable duration of recording determined by silence detection and/or hard limitations.
- saving and transmitting session related information like calling line identification to arbitrary hosts.
- record the total length of the recording in parallel to the prompted recordings

The file server was configured to record at 8 bit, 8 kHz and A-law compression. Two set of files were generated by the server. One continuous file covering the whole session, and several short segments containing chunks of utterances.

3.4. Session control

As mentioned earlier the recordings took place in normal urban traffic in the city of Munich with all possible problems for a motorcyclist performing two tasks simultaneously. Therefore the SMC scenario requires an almost hands-free operation during the speech recording for obvious reasons: the driver cannot handle the mobile phone during the ride. Furthermore, the driver should be able to interrupt the recording session any time to avoid potentially dangerous distractions in certain traffic situations. This results in two technical problems: how can the server identify the speaker and how can the speaker interrupt the recording session.

We mounted a small device on top of the fuel tank of the motorcycle with three buttons named “CALL”, “START”, “STOP”: The Call button controls a relay breaking the microphone line to the mobile. Since the mobile phone monitors the microphone impedance and interprets interruptions of the microphone circuits as a kind of default action, pressing the button twice results in redialing and pressing it for a longer period results in hang up.

The other two buttons issue DTMF signals. The server was modified to detect these signals and to put the recording process on hold or to continue the process respectively. Prior to any recording the recording supervisor calls the server on a different mobile phone. The server recognises the supervisor by means of the calling line identification and starts a special script to activate the forthcoming recording session identified by the session number typed in. After that a simple call of the recording mobile phone to the server will initiate the recording.

Thus, with one button the driver is able to establish a connection to the voice server while the other two enable the driver to interrupt the session as long as is required. .

3.5. Test Vehicle

After some test drives BMW supplied us with a *R 1200 RT* motorcycle with a 110bhp air cooled engine. The control device was mounted in a water-proof box on top of the tank, the mobile phone was tugged in a pocket of the driver and the remaining electronics stored in a small knappsack. There was no galvanic connection to the vehicle’s electric system.

4. Situational Prompting

A special technique was used to ensure that the utterances of the test person were as close as possible to a real HMI dialogue. Since Wizard-of-Oz recordings were considered not feasible for economic reasons, the *situative prompting* scheme as described in (Mögele et al., 2006) was chosen to improve the naturalness of queries uttered by the test person.

In a nutshell, the test person was instructed by a female voice to imagine a certain situation connected with a task to solve while a male voice simulated the responses of the Smartweb system. During each thematic block of 6 queries the test person was prompted to ask questions concerning the situation resulting in a small dialogue between test person and recording system. This method yielded natural spoken utterances with a higher variety in phrasing than by means of simple text prompting.

5. Speaker recruitment

Recruitment of speakers was performed by BMW research facilities among a selected list of BMW employees. Public recruitment was not viable due to limitations in insurance policies. To reduce the risk of accidents only experienced drivers were selected who already had performed several driving tests in the past. It should be mentioned here that although the percentage of female motocyclists is rather high in Germany, it turned out to be very difficult to recruit female drivers for the SMC.

6. Recording procedure

The recording procedure is similar to the SmartWeb UMTS Handheld corpus which is described in (Mögele et al., 2006), albeit the SMC scenario was more difficult to handle because no supervisor could be present during the recording. In the following we give a short overview with focus on the special measures taken for the SMC recordings.

6.1. Preparation

First, each test person was carefully instructed by a staff member of the BAS. Instructions were given about the usage of the buttons to control the session, the prompting, the route to take, the kind of questions and instructions to be expected from the system and what to do in case of system failure or critical traffic situations. Since the driver was not able to consult any written documents during the test, no *individualised promptings* ((Mögele et al., 2006)) were used except basic route planning tasks. Although in fact every test person took the same route through the city they had to imagine a certain biker route or journey they were on or were planning to do.

Then the driver was equipped with the recording devices consisting of the amplifier, the hard disk recorder and the mobile phone. All devices had to be connected properly to ensure correct operation. The recording supervisor did a number of checks to ensure proper operation.

6.2. Recording

To minimise the risk of accidents, the test person first drove around the block several times to adapt to the motorbike and to prepare for the recording. Then he/she was instructed to go on the pre-defined route and to start the recording by pressing the Call button twice as soon as the traffic situation permitted.

7. Recorded SMC Signals

7.1. Numbers

46 sessions have been recorded as described above and carefully examined. 8 sessions had to be discarded because of technical problems. Each session consisted of a maximum of 12 query blocks of 6 queries each. A relatively large number of prompts turned out to be useless because the driver did not respond to the prompt, probably being distracted by trafic. In total the recordings resulted in 2835 recorded queries with approximately 31.900 running words.

These numbers seem to be moderate compared to other speech corpora productions. However, considering the limited budget and the effort to recruit test drivers a larger number of speakers would have exceeded the scope of this project.

7.2. Data of each Recording

The queries are transliterated according to a reduced *Verbmobil Transliteration Scheme*[6] in a two-stage process using WebTranscribe ((Draxler, 2005), (Draxler, 1997)); the transliteration will be completed mid of 2006.

The high quality signals recorded to hard disk were time aligned to the server recordings using a cross correlation technique which yields a proper segmentation of the hard disk recording into query chunks. Thus each recording session consists of the following data:

- individual prompted queries (max. 72) (A-law, 8kHz)
- complete server recording throughout the session (A-law, 8kHz)
- hard disk recording of Bluetooth helmet microphone (PCM 16bit, 44,1kHz)
- hard disk recording of throat microphone (PCM 16bit, 44,1kHz)
- transliteration
- alignment from hard disk recording to individual prompted recordings

[6] http://www.is.cs.cmu.edu/trl_conventions/projects/verbmobil_entrance.html

7.3. Noise on recordings

The prominent noise source is of course the engine of the motorcycle. Also, transient sounds from gear-changing were picked up. In some cases the engine noise reached the level of the speech signal, particularly when the test person drove with the visor open. The second prominent noise source was traffic noise but with notably lower levels than the engine noise even when running idle. Line and quantisation noise occurred, too, but not at high levels. Wind noise was surprisingly low. Only in tests on the Autobahn was wind noise a significant problem.[7]

7.4. Level of speech

Since the signal processor of the helmet applies an automatic gain control on the signal, the level of recorded speech fluctuates. In most cases the level rises rapidly at the beginning of the utterance to decline immediately after to a normalised level. The time between attack to the normalised level usually is about one second.

7.5. Loss of connection

After determining a route with proper UMTS support, total connection losses occurred rarely. Sometimes glitches in the recordings occurred. Most of the glitches resulted from dropped packets during the wireless transmission from the mobile phone to the base station. Dropouts caused by the bluetooth transmission were rare.

7.6. Throat Microphone

The throat microphone registered no significant noise from engine and traffic. Some noise from friction with nearby clothing could be heard. Recordings contain spectral components below 3500 Hz and in this range the speech quality is very high; distortions could not be detected.

8. Availability

The SMC will be distributed on DVD-R via the BAS[8] and the ELDA[9]. A first publicly available release is to be expected in Aug 2006.

9. References

Christoph Draxler. 1997. WWWTranscribe – a modular transcription system based on the world wide web. In *Proceedings of the Eurospeech 1997*, Rhodes, Greece.

Christoph Draxler. 2005. WebTranscribe – an extensible web-based speech annotation framework. In *Proceedings of TSD 2005*, Karlsbad, Czech Republic.

Hannes Mögele, Moritz Kaiser, and Florian Schiel. 2006. Smartweb UMTS speech data collection: The SmartWeb Handheld Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation 2006*, page to appear, Genova, Italy. ELRA.

[7] Autobahn recordings are not contained in the SMC.

[8] www.bas.uni-muenchen.de/bas

[9] www.elda.org