Talking and Looking: the SmartWeb Multimodal Interaction Corpus

Florian Schiel; Hannes Mögele

BAS Services; Institute of Phonetics and Speech Processing
Moltkestr. 1, 80803 München, Germany; Schellingstr. 3, 80799 München, Germany
schiel@bas-services.de; hannes@bas.uni-muenchen.de

Abstract

Nowadays portable devices such as smart phones can be used to capture the face of a user simultaneously with the voice input. Server based or even embedded dialogue system might utilize this additional information to detect whether the speaking user addresses the system or other parties or whether the listening user is focused on the display or not. Depending on these findings the dialogue system might change its strategy to interact with the user improving the overall communication between human and system.

To develop and test methods for On/Off-Focus detection a multimodal corpus of user – machine interactions was recorded within the German SmartWeb project. The corpus comprises 99 recording sessions of a triad communication between the user, the system and a human companion. The user can address/watch/listen to the system but also talk to her companion, read from the display or simply talk to herself. Facial video is captured with a standard built-in video camera of a smart phone while voice input in being recorded by a high quality close microphone as well as over a realistic transmission line via Bluetooth and WCDMA. The resulting SmartWeb Video Corpus (SVC) can be obtained from the Bavarian Archive for Speech Signals (BAS).

1. Introduction

Portable devices and so called smart phones as well as their respective network infrastructure have matured up to a point where it is possible to capture the video image of the talker's face during a dialogue and send the combined video and voice streams in real time to a server based dialogue system. In large parts of Europe this can be achieved using an effective video coding standard (3GPP) used for mobile phone video as well as the Wide-band Code Division Multiple Access (WCDMA) transmission protocol (commonly known as 'UMTS') and an infrastructure which is by now in operation from a number of telephone providers. This technique not only makes video conferencing from handheld devices possible but also allows — in principle — a speech dialogue system not only to capture the human users voice input but also her facial gestures such as eye direction, lip movement, head movement, movement of eyebrows etc. One interesting application of face video capture is the automatic detection of so called OnFocus/OffFocus (Hacker et al., 2006; Batliner et al., 2007) meaning basically the answer to two questions:

> *Is the user's focus on the display?*
> *Is she addressing the system or rather a third party or is speaking to herself?*

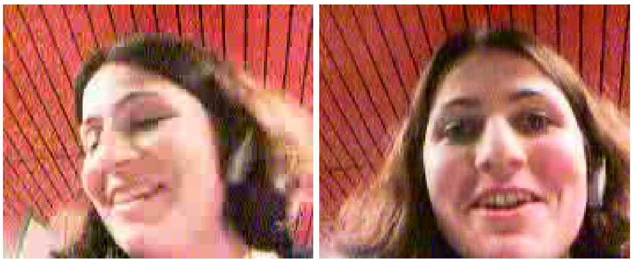

Figure 1: Left: captured frame with OffFocus; right: captured frame with OnFocus behavior (recording *i055*).

Figure 1 shows two captured frames video taken during a user — machine interaction; in the right picture the user is OnFocus, in the left picture she talks to her companion and is therefore OffFocus. Assuming a dialogue system with no push-to-talk activation, that is, the microphone is always 'open' for voice input of the user, it is of vital importance to detect utterances that are not addressed to the system to avoid misinterpreted voice input to the dialogue engine. On the other hand, a dialogue system using combined acoustic and visual output modalities might alter its output modality depending on the information whether the user is looking at the display or not: visual display output may for instance be replaced by speech synthesis output if the user's focus is definitely not on the device.

To provide realistic data for these and other questions concerning multimodal interaction by means of portable devices a part of the SmartWeb data collection (Mögele et al., 2006; Kaiser et al., 2006) was dedicated to provide a multimodal speech and video corpus of human — machine interactions.

The aim of the SmartWeb 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 (IPS) of Munich University. In the course of this data collection three new speech resources have been 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 SVC. A detailed description of SHC and SMC can be found in (Mögele et al., 2006) and (Kaiser et al., 2006) respectively.

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 annotation that covers linguistic phenomena typical for spontaneous speech.

The resulting SVC corpus is part of the total SmartWeb corpus package and is freely available to the scientific community via the Bavarian Archive for Speech Signals (BAS) as well as the European Language Resources Distribution Agency (ELDA)[1]. All descriptions in this paper are based on version 2.10 of the SmartWeb Corpus distribution of BAS.

In this paper we describe the recording methodology to yield almost real-life recordings of a human – machine interaction, the recording technique, the post-processing and segmentation, give details about the annotation, meta information and the overall properties of the resulting SVC corpus as can be obtained from BAS.

2. Recording Methodology

2.1. Environment

Recordings took place in five different real–life environments, some indoors (office, lobby, public cafe) and some outdoors (courtyard, park) with varying acoustic and lighting conditions, changing sources of background noise and visual background (resulting for example from different weather conditions: sunny with blue sky or cloudy). These conditions were not controlled for the experiment but have been documented in the recording protocol (see below).

2.2. Facial Video Capture

The user was not being instructed to keep the mobile phone in a certain position. However, a small control display shows the recorded video and therefore provides the user with some intuitive guidance to keep the face somewhere within the video capture area of the camera. We found that this sufficiently influences the user to keep the smart phone in a position where her face was visible most of the time when the user is looking at the display. In case that the user looks elsewhere the face sometimes vanishes totally or partly from the video capture area. Since the user communicates with the system via a Bluetooth microphone, the microphone is visible in the video recording; the second microphone is attached to the collar of the user and does not interfere with the facial video.

2.3. Communication Scenario

To provoke head movements and facial expressions combined with speech directed to the system and to third parties we choose a triad communication scenario: the system and two human subjects (the user and a companion). The user interacts directly with the system while her companion tries to interfere the communication with remarks and additional requests for information. Thus the recorded user is often distracted from her task (to gather information about a certain topic), looks to and from the display and addresses her companion to answer questions. The companion is not allowed to watch the display's output. Therefore the user often paraphrases or reads given system output to her companion.

2.4. Situational Prompting

Since the real SmartWeb system was not operational yet at the time of recording, the dialogue between system and user is controlled using the so called situational prompting (SitPro) elicitation method as used in (Mögele et al., 2006). SitPro combines script methods with interview techniques and speaker prompting. Here, speaker prompting is only used for instruction, information or feedback prompts; hence the user should not simply repeat a sentence prompted by the system.

The automatic prompting system used for the SVC data collection simulates two interlocutors. The *instructor* (female voice) gives directions about the situation and the topics, while *smartweb* (male voice) answers the subjects' questions or gives feedback like the real SmartWeb system. This results in three different prompt categories, called standard prompts, individualized prompts and script prompts.

In a *standard prompt* the subject is told a topic to which she is supposed to pose a query:

> *instructor:* Imagine you just arrived at the station Berlin Zoo. Ask SmartWeb for instructions on how to get to the soccer stadium. (beep)
> *user:* How do I get to the soccer stadium from here?

An *individualized prompt* is a prompt for which the subject provides her own topic. For that purpose the user has been asked to bring a pre-fabricated list of 6 topics of general interest with her to the recording. The system then refers to these topics during the recording such as:

> *instructor:* Please refer now to your list of interesting topics. At first I'd like you to pose a general question about the topic number 6 on your list. (beep)
> *user:* Where is the Neue Pinakothek in Munich?
> *instructor:* Very good! Now please pose a more detailed question about the same topic. (beep)
> *user:* Um ... who is the architect of the Neue Pinakothek?

A *scripted prompt* simulates a three-turn conversation as frequently found in dialogues between human and machine:

> *instructor:* Ask SmartWeb about the weather in Berlin on Sunday. (beep)
> *user:* Can you give me the weather forecast for Berlin on Sunday?
> *smartweb:* Partly cloudy with 35% chance of rain. Temperatures between 20 and 23 degrees Celsius.

Aside from these prompt types the instructor asks the user to relay certain information to her companion:

> *smartweb:* The soccer game tonight is being sold out!
> *instructor:* Please inform your partner that the

[1] Distribution fees apply.

> game is sold out. (beep)
> *user to her companion:* Hey, listen: the game is already sold out. What do we do now?

or the instructor asks the user to read information from the display:

> *instructor:* Please read the information shown on the display to your partner. (beep)
> *user to her companion:* ...

Prompt types and topics are randomly generated individually for each recording session; only the percentage of prompt types is kept constant.
SitPro has been shown to result in almost realistic user queries to a information system (Mögele et al., 2006); the additional instructions for relaying information to a third party ensure frequent head movements and change of speech style.

2.5. User Instruction

The user is instructed to interact with the automated prompting system in order to request information from the system on topics of interest, to read off information of the simulated display, to relay information to the companion, and to answer the interposed companion's questions. The task of the companion is to distract the caller from her task. Thus, a restricted triadic communication scenario is being established in which the companion only observes the caller but not the system's speech and display output.

3. Recording Technique

For the recordings we use a Nokia 6680 mobile phone with a built-in face capture camera in combination with a Bluetooth headset and parallel thereof a collar microphone connected to a hard disc recorder. The Bluetooth microphone (either Plantronics M 3500 or Samsung WEP 150 MBE) is connected to the mobile phone and its signal is transmitted to the prompting server while the signal of the collar microphone is connected to a portable hard disc recorder (iRiver). The mobile phone is set into a simple video recording mode with audio capture turned off to avoid the periodic warning beep that signals the other party that a recording is taking place. After the video recording has started, the user connects over a standard WCDMA line to the automated prompt server which then controls the complete dialogue following a randomly created situational prompting script (VXML). The VXML script defines which instructions are given to the user (female voice), when and for which length the request of the user is being recorded and a possible (simulated) system responses (male voice). The prompt server records the total dialogue as well as the prompted user requests as defined in the VXML script. After the recording session the video file (3GPP without Audio) and the harddisc recording (WAV) are downloaded to the server.

4. Post-processing and Segmentation

The original video file (3GPP) uses the codec H.263 with size 176x144, 24bpp and 15fps, no audio. It is transformed into an MPEG1 standard video stream and manually synchronized to the audio track recorded by the server. Both video streams are then included into the final corpus.
The individual prompt recordings of the server are automatically synchronized to the harddisc recording using a cross correlation technique thus yielding in a segmentation of the harddisc recording into synchronized prompt recordings. Both recordings are part of the final corpus. They reflect a realistic scenario on what speech input the SmartWeb system might receive from a triad communication situation. Since the user is often being distracted, many user inputs do not fit within the recording windows of the VXML script. Furthermore, all OffFocus related speech is naturally outside the server recording windows as well. Therefore the total server recording was manually segmented into speech chunks independently from the VXML script. This second segmentation of the session is also incorporated into the annotation of the user's speech (see next section) so that the transcripts are closely aligned to this segmentation.

5. Annotation

All recorded user output has been annotated using a subset of the annotation scheme used in the SmartKom project (Rabold & Biersack, 2002). The speech output of the prompts server as well as the audible speech of the companion were not transcribed.
In the SmartWeb transcripts prosodic markers and markers for superimposed speech have been omitted. On the other hand the annotation was augmented by a set of word tags for different types of OffTalk, time markers for the second segmentation of the recording mentioned in section 4 and phonemic transcripts of all dialectal variants.
The OffTalk markers used in the SVC annotation are:

<SOT> spontaneous Off-Talk
<POT> paraphrased Off-Talk
<ROT> read Off-Talk
<OOT> other Off-Talk

where *spontaneous Off-Talk* denotes speech directed to the companion, *paraphrased Off-Talk* is a special case of spontaneous Off-Talk where the user is relaying information provided by the system to her companion. *Read Off-Talk* denotes reading (literally) from the display while *other Off-Talk* builds the garbage class (e.g. muttering to himself).
OffTalk always implies OffFocus, while OnTalk (everything that is not tagged as OffTalk) always implies OnFocus. Unfortunately, this does not mean that Talk and Focus are equivalent, since the user might be OffFocus while being silent (e.g. listening to her companion). There exists no OffFocus labeling of the SVC video stream, although this would be desirable.
Figure 2 illustrates two different instances of OffTalk behavior. The top left picture shows OffTalk without or barely perceptible head movement; only the eye focus changes. The top right picture shows a clearly visible case of OffTalk: the user moves her head in direction to her companion. For reference a typical OnFocus situation is shown below.
Annotations were produced in a three-level scheme where the annotator of the first level produced a basis transcript, the second level added all tags and the third level verified the work of the the first two levels. Transcripts are stored

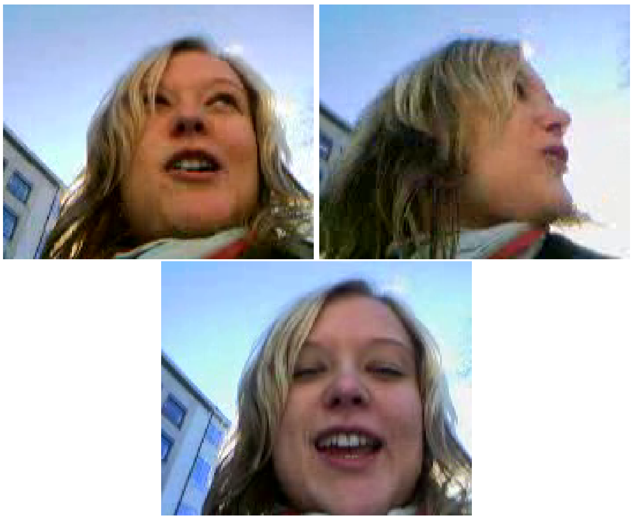

Figure 2: Top: two different OffFocus situations; below: OnFocus behavior (recording *i077*).

in text files with extension 'trl' that conform the Verbmobil and SmartKom transcription format.

Example (translated):

i065_man-0000rec-010: <ZA 197.260> could we go to the stadium by feet ? <P> and how about <!1'bout#baUt> the weather tomorrow ? <ZE 202.965>

i065_man-0000rec-020: <ZA 214.133> that<POT> will<POT> be<POT> at<POT> least<POT> #five<POT> kilometers<POT> to<POT> the<POT> soccer<POT> stadium<POT> . <ZE 221.961>

i065_man-0000rec-030: <ZA 226.065> any sights nearby ? <ZE 228.954>

i065_man-0000rec-040: <ZA 237.314> where's the next taxi stand ? <P> but<SOT> that's<SOT> probably<SOT> fucking<SOT> expensive<SOT> , huh<SOT> ? <ZE 24 3.318>

This short example shows four recorded user inputs. The first is a query directed to the system; it contains a tagged dialectal variant of 'about'; please note the segmentation tags ZA and ZA labelling the beginning and end of the speech chunk in seconds from begin of recording. The second input consists entirely of paraphrased OffTalk since the user only relays information that she received from SmartWeb to her companion. The third input is again an user query elicited by the SitPro method. The forth input partly shows spontaneous OffTalk directed to the companion.

In an ideal working dialogue system the first and third and – partly – forth input would be processed as valid queries adressing the system, while the third input as well as the second part of the forth would be ignored.

Transcripts are also provided in the BAS Partitur Format (BPF) (Schiel et al., 1998) as well as an ATLAS compatible Annotation Graph (Bird et al., 2000).

6. Meta Data

Extensive speaker information is stored in XML formatted *speaker protocols* for each recorded user: sex, age, handedness, profession, mother tongue, mother tongue of both parents, experience with dialogue systems and search engines, glasses, beard, baldness, smoker/non-smoker, piercing, other props.

For each recording session an XML formatted *recording protocol* is provided. It contains the technical setup, experimenter, user and companion IDs, recording date and time, location, recorded audio and video tracks, available annotation as well as environmental conditions during the recording.

The original *recording script* of each recording is provided in two forms: a simple text file and the original VXML script that was used during the experiment.

7. Summary Corpus Properties

Table 1 shows the most prominent figures of the SVC corpus. About half of the uttered words are classified as OffTalk which indicates that the triad communication scheme of SVC works fine. As it can be expected spontaneous OffTalk is the most frequently occurring form of OffTalk. The high percentage of paraphrasing and read OffTalk is caused by the explicite instruction to the speakers; in a natural setting we would expect a much lower frequency for these types of OffTalk. Finally, the percentage of other OffTalk (muttering to himself etc.) is with 3.1% of the OffTalk very low, probably caused by the (possibly embarrassing) presence of the companion. In other studies (Oppermann et al., 2001) where no third party was present this percentage was significantly higher.

number of recordings / speakers	99
female / male	63 / 36
age	15 - 64
indoor recordings	72
lobby	63
office	2
public cafe	7
outdoor recordings	27
court yard	17
public park	10
total words	25151
total words spontaneous OffTalk	5177
total words paraphrasing OffTalk	4385
total words read OffTalk	2917
total words other OffTalk	400
percentage OffTalk	51.2%
total size	17.2GB
media	5 DVD-R

Table 1: The SVC in numbers.

Table 2 summarizes all data types provided for each recording session.

8. Availability

The SVC corpus will be made available for unrestricted scientific and commercial usage latest in Sept 2008 (one year after project closure). Interested parties might obtain preliminary copies of the corpus at BAS.

Type	Format
Original face video, 176x144, 24bpp, 15fps	3GPP
Video with synchronized audio track	MPG1
Server recording, total dialogue via Bluetooth microphone and WCDMA	8bit,alaw 8kHz
ditto for individual prompts	*ditto*
ditto manually segmented into chunks	*ditto*
Harddisc recording, total dialogue, collar microphone	16bit,pcm 44,1kHz
Server segmentation	text
Manual segmentation	text
Transcript SmartKom standard	text,TRL
BAS Partitur Format	text,SAM
Annotation graph ATLAS	XML
Speaker protocol	XML
Recording protocol	XML
Recording script	VXML

Table 2: Data provided for each recording session.

Please contact Florian Schiel *schiel@bas.uni-muenchen.de* or refer directly to the BAS catalogue at: *www.bas.uni-muenchen.de/Bas*.

9. Acknowledgments

The work presented in this paper as well as the resulting multimodal corpus SVC was funded by the German Ministry for Education and Science (BMBF) under grant number 01 IM D01 I as well as by the Bavarian Archive for Speech Signals (BAS). The authors thank all colleagues within the SmartWeb project for their help and support.

10. References

H. Mögele, M. Kaiser, F. Schiel. 2006. SmartWeb UMTS Speech Data Collection: The SmartWeb Handheld Corpus. *Proc. of the LREC 2006, Genova, Italy*.

M. Kaiser, H. Mögele, F. Schiel. 2006. Bikers Accessing the Web: The SmartWeb Motorbike Corpus. *Proc. of the LREC 2006, Genova, Italy*.

Chr. Draxler, K. Jänsch. 2004. SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. *Proc. of the IV. International Conference on Language Resources and Evaluation, Lisbon, Portugal*.

A. Batliner, Chr. Hacker, M. Kaiser, H. Mögele, E. Nöth. 2007. Taking into Account the User's Focus of Attention with the Help of Audio-Visual Information: Towards less Artificial Human-Machine-Communication. In: *Krahmer, Emiel; Swerts, Marc; Vroomen, Jean (Eds.) AVSP 2007 (International Conference on Auditory-Visual Speech Processing 2007*, 51-56.

Chr. Hacker, A. Batliner, E. Nöth. 2006. Are You Looking at Me, are You Talking with Me – Multimodal Classification of the Focus of Attention. In: *Sojka, P.; Kopecek, I.; Pala, K. (Eds.) Text, Speech and Dialogue. 9th International Conference, TSD 2006, Brno, Czech Republic, Springer Berlin Heidelberg*, 581-588.

F. Schiel, S. Burger, A. Geumann, K. Weilhammer. 1998. The Partitur Formatat BAS. *Proc. of the 1st Int. Conf. on Language Resources and Evaluation 1998, pp. 1295-1301, Granada, Spain*.

D. Oppermann, F. Schiel, S. Steininger, N. Beringer. 2001. Off-Talk - A Problem for Human-Machine-Interaction. *Proc. of the EUROSPEECH 2001, Scandinavia, Aalborg, Danmark*.

St. Bird, D. Day, J. Garofolo, J. Henderson, Chr. Laprun, M. Libermann. 2000. ATLAS: A flexible and extensible architecture for linguistic annotation. *Proceedings of LREC 2000*.

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

S. Rabold, S. Biersack. 2002. SmartKom Transliteration Manual. *http://www.phonetik.uni-muenchen.de/Forschung/SmartKom/Konengl/engltrans/engltrans.html*.