The SmartKom Multimodal Corpus at BAS

Florian Schiel*, Silke Steininger†, Ulrich Türk†

*Bavarian Archive for Speech Signals (BAS)
†Institut für Phonetik und Sprachliche Kommunikation

University of Munich, Schellingstr. 3, 80799 München, Germany
{schiel,kstein,tuerk}@phonetik.uni-muenchen.de

Abstract

In this contribution we announce and describe in detail the new multimodal corpus evolving from the publicly funded German SmartKom project. The first release of the corpus (BAS SK-P 1.0) has been finished end of 2001 and will be ready for distribution to the scientific community in July 2002. The SmartKom corpus will be the first of a new generation of Language Resources (LR) designed for a more or less complete data gathering of human-machine communication combining acoustic, visual and tactile input and output modalities. Since the funding of about EU 2 Mio for this LR is 100% public, the corpus will be available without royalties via the Bavarian Archive for Speech Signals (BAS) at the University of Munich.

1. Introduction

This paper gives a detailed specification of the upcoming distribution series BAS SK at the Bavarian Archive for Speech Signals (BAS) located at the University of Munich. Aside from the specs of the first release we also give some background information that might be useful for the prospective user of the resource. Note that the SK Biometric Corpus is not reported in this paper.

The paper is organized as follows: In the second section we give a brief description of the project frame work of SmartKom (SK). The SK corpora are not produced as a pure infrastructural initiative (like for instance the CGN corpus or the BNC); therefore it might be interesting for the user of the corpus to know under which motivation these data were produced and for which purposes. The third section describes the basic recording technique using Wizard-of-Oz experiments (WOZ) and naive users. Section 4 gives a very brief overview how the gathered data are processed and annotated. Readers who are only interested in the basic facts of the resource might skip the first three sections and go right to section five which sums up all specifications for the distributed resource. Note that although these are in some cases specific to the first release, most of the specs will hold for all SK corpora released within the next two years by BAS. Section six gives the terms of availability and the intended release plan of the SK corpora, while the last section is dedicated to a brief discussions of the weaknesses of the corpus and prospective uses.

2. SmartKom

2.1. The Project

The goal of the SmartKom project (SK) is the development of an intelligent computer-user interface which allows almost natural interaction for the user[1]. The system recognizes natural speech as well as gestures above a flat interaction area. Additionally, facial expression is analyzed. The output of the system is synthesized speech and a GUI, which is either projected on the interaction area or shown on a portable Web Pad or PDA (see figure 1 for a schematic view of the recording setup). The focus of the SK project

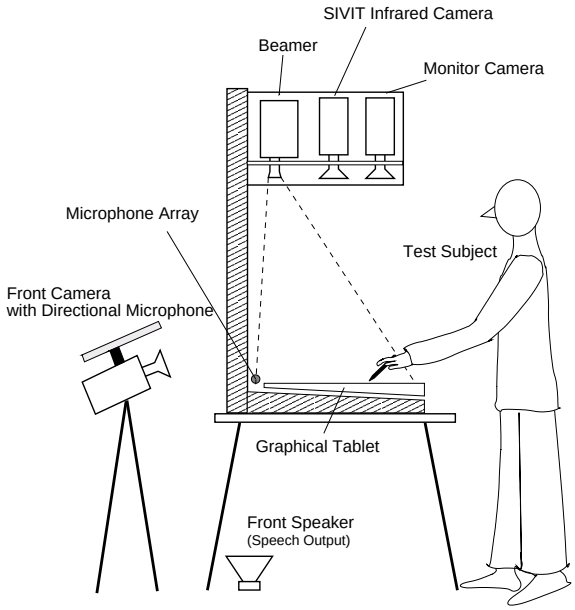

Figure 1: *SmartKom Public setup used for data collection.*

does not aim at the development of application products but to investigate the new possibilities in multimodal communication, in this case by combining natural speech input with 2D gestures and mimic expressions. To demonstrate the possibilities of multimodal communication SK is implemented in three different technical example scenarios:

- SK Public : a publicly accessible information interface
- SK Home : intelligent communication assistant at home
- SK Mobil : portable communication assistant

Within of each technical scenario the SK core system covers several task domains. For example: hotel/restaurant in-

[1] http://smartkom.dfki.de/

formation, cinema, weather forecast, MP3-Jukebox, VCR programming, TV program (EPG), E-Mail/Fax assistant, scheduler, navigation etc. Furthermore, the SK system should be capable to verify identities and/or recognize users by biometric measures (voice, signature and hand contour).

2.2. The data collection

Within the SK project BAS is responsible for the collection of multimodal data and the evaluation of the system prototypes (not reported here).
In Wizard-of-Oz experiments (see section 2) subjects are recorded in sessions of 4.5 minutes length while they are interacting with a simulated version of the SK system. During these sessions all capture devices of the system are used for collecting data:

- audio is captured using a directional microphone, a microphone array with 4 channels and (alternating) a headset or a clip microphone.
- video is recorded by to two standard DV cameras (one for the facial expression, one for the side view of the subject) and by an infrared camera which is part of the gesture recognizer SIVIT (Siemens).
- the graphical output is recorded in a low frame rate video (used only for labeling).
- gesture coordinates captured by the SIVIT system and the graphical tablet.

The data collection in SmartKom serves two distinct purposes. First, it is used as training and test material in the developing process of the different recognizers (speech, gesture, facial expression). Second, it provides insight into the human-machine interaction. The information obtained here is used during the concept phase in SmartKom, especially when modeling the interaction strategies.

2.3. Partners

The SK consortium consists of 7 industrial and three academic partners, namely Daimler/Chrysler, DFKI, European Media Lab (EML), Philips, Media Interface, Siemens, SONY, Universities of Erlangen, Munich, Stuttgart. Furthermore, two sub-contractors, Sympalog and ICSI, are involved. For further information please consult smartkom.dfki.de. Partners will have access to the data right after their edition; other parties have access after one year of blocking period exclusively via the BAS.

3. Recording with WOZ

To explore how users interact with the SK system, data is collected in so-called Wizard-of-Oz experiments: The subjects have to solve certain tasks with the help of the system (e.g. planning a trip to the cinema). They are made believe that the system they interact with is already fully functional. Actually, many functions are only simulated by two ”wizards”, who control the system from a separate room. Each subject is recorded in two sessions of about 4.5 minutes length each.
The basic recording parameters of each session are:

- Technical scenario (Public, Home, Mobil)
- Primary task (eg. programming the VCR)
- Secondary task (eg. look for TV shows tonight in EPG)
- User profile (gender, age, education, technical background, etc.)
- Emotions evoked yes/no
- Hand gesture or pen gesture used
- background noise (on/off, type, level)
- background of front camera (different patterns)
- results of user questionnaire

Most of the above data (and more) are stored directly into the project database and exported later into the user profile file (SPR) and the recording protocol (RPR). Technical scenarios are defined by blue prints to ensure consistent technical recording conditions. These also include the simulated GUI (see (Beringer, 2001) for details). Task domains are defined by so called task flow charts which define the possibilities of the simulated system (figure 2). During the

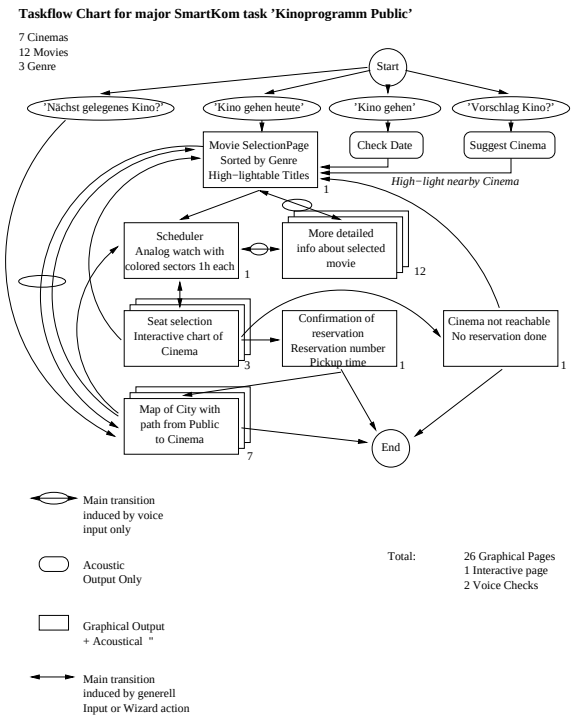

Figure 2: *SmartKom task flow chart for cinema information*

session recording all captured data are stored on a set of currently five Windows NT workstations. Each computer is dedicated to record a single data stream because continuous capturing of audio and, especially, of video demands a great amount of computing power.
The data from two video streams (front view and side view) are recorded separately via a FireWire bus between camera and computer and are stored as Quicktime files encoded in DV. The video signal of the infrared camera is digitized by

an external analog-DV-converter and then recorded in the same way. Video data encoded with DV offer a high picture quality and show very few artifacts which could be otherwise problems for image processing algorithms. However, the advantage in quality comes together with large amounts of data. For capturing the audio data we use a ten track audio card. The recorded audio files are stored in Windows WAVE format, using a resolution of 16 bit and a sampling frequency of 48 kHz.
Capturing of the graphical system output is done via a screen capturing tool. It allows recording to a video file in AVI format at low frame rates; in our setup capturing at 4 fps is sufficient because the interaction speed is low.
The remaining data tracks, the coordinate files of the SIVIT gesture recognizer and of the graphical tablet are recorded with specially designed tools. In addition to the actual recording, this stage comprises also the generation of a new entry for the session in the database. Information about the user as well as his or her behavior during the recording or details about the setup are stored here. In the following stages gradually more and more information is added e.g. about the current state of the session in the processing pipeline.

4. Processing and Annotation

After the session recording several non-synchronized data streams have to be integrated into a common QuickTime[2] frame and then annotated in severals steps (figure 3). To control this rather complicated process we use a project database where each member of the SK staff (currently about 26 persons) may monitor and control the different stages of manual and automatic processing. The scope of this paper does not allow us to present all processing and labeling steps in detail. Therefore we give here the latest publications to these topics:

- Post-processing, synchronization, DVD production: (Tuerk, 2001)
- Transliteration and labeling of natural speech, prosodic annotation: (Oppermann et al., 2000)
- Annotation of gestures: (Steininger et al., 2001)
- Annotation of user states: (Steininger et al., 2002b), (Steininger et al., 2002a)

See the next section, part 'Annotations' for a detailed description of the annotations contained in the final SK corpora.

5. Specifications

5.1. Distribution Format

The data of each recording session (max. 4.5min) are stored on one DVD-5 (max. 4.7GByte). The file system is UDF and should therefore be readable by all platforms. All data streams are synchronized to the absolute time scale. Individual streams may differ at the end of the recording by about 20 msec due to different time bases in the th recording devices. All data streams are incorporated into the main

[2]http://developer.apple.com/techpubs/quicktime/quicktime.html

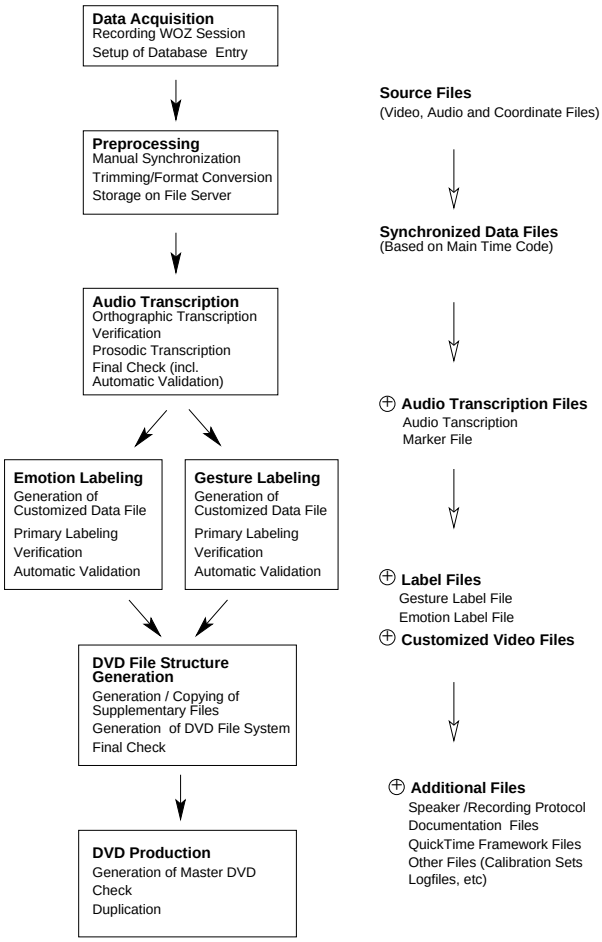

Figure 3: *Processing stages in SmartKom data collection (left) and corresponding changes in the data file set (right).*

QuickTime frame that covers the total recording. However, the single streams are accessible in their respective file formats as well, since the QuickTime format simply refer to these streams. the following listing shows the root directory of a SK DVD:

```
Data/         # QT signal files
Doc/          # General Documentation
Quicktime/    # QT software for Win, Mac
Readme.67.0   # sessions on this DVD
Readme        # structure of this DVD
mar/          # Turn cut files
rpr/          # Recording protocols
spr/          # User profiles
Annotation/   # All annotation files
```

All annotations are stored in the original distribution format as well as in the BAS Partitur Format (BPF) in the subdir "Annotation" (see below).

5.2. Recorded Signals

In Smartkom a typical session file contains the following tracks:

- Video of the face, frontal, DV format.
  The DV camera is in a fixed position so that the head of the user is in the the inner 9th square of the frame. The

user may move his/her head outside of this frame. No additional lightning was used. The background colors may change according to the recording specs, but not within one session; no moving objects in the frame except the user.

- Video of upper body, from left, DV format.
  Basically the same conditions as the front camera. White background color throughout the SK Public recording.

- Video of infrared camera directed on display to capture hand gestures, from top, DV format.
  The camera is part of the SIVIT gesture analyzer by Siemens. The camera is a standard PAL camera which delivers an analog video signal in b/w. The background is a special infrared reflective surface. Only objects on this surface appear in the picture as black shadows (background appears in white). Any graphics projected on the background do not appear in the video. The video signal is digitized into a DV compatible stream and then captured by a standard FireWire card.

- Audio in 10 channels (microphone array (4), directed mic, headset (2), background noise (2), system output) captured by a 10-channel audio card with 48 kHz. Data are filtered to 8 kHz and down-sampled to 16 kHz before included into the QuickTime frame.

- Graphical system output captured by a screen capture application at 4fps, AVI format.

- combined video frame with face, upper body, system output and infrared, AVI format (see figure 4).

- Coordinate log files: output of either the gesture recognition system (finger tip) or the output of the graphic tableau (pen tip)

For performance reasons all streams are captured on different computers. Coordinate log files are transformed into a sprite track to make coordinates visible in the video signals. Then all raw signals are synchronized, cut and integrated into a QT frame.

Figure 4 shows four data streams of a SmartKom recording within a single flattened video frame. In the upper left quadrant the video signal of the face camera is shown; in the upper right quadrant the video signal of the body from the left; in the lower left quadrant the displayed output of the system, in the lower right quadrant the output of the system and as an overlay the video signal of the infrared camera that captures the user's gestures. The shown frame is actually from a video stream that was calculated from the original QT frame; the QT Player Pro is principally capable to show many video streams simultaneously, however the performance on a standard Intel platform is still unsatisfying.

5.3. Annotations

The BAS SK corpora contain the following annotations:

- SmartKom Transliteration of audio channels
- Turn segmentation
- Segmentation and labeling of gestures in the 2D plane
- Segmentation and labeling of user state (facial and speech)
- Segmentation and labeling of user state from facial expression only
- Segmentation and labeling of complex prosodic features to recognize 'emotions'

There exist two different types of annotations: Script-like annotations for the linguistic (TRS) and prosodic labeling (TRP) and the BAS Partitur File (BPF) that summarize all annotations of a dialog partner. Therefore, you will find two BPF files in each BAS SK session: one regarding the (human) user of the system, the other regarding the system itself (linguistics only).
The following example shows an extract from a SmartKom BPF. For better readability the file is abbreviated to the first 12 words of the dialogue and the header block is omitted.

```
TRS:    0       <"ah> [NA] [B2]
TRS:    1       hallo [PA] [B3 fall] . <A> <P>
TRS:    2       kennst [NA]
TRS:    3       du
TRS:    4       den [B2]
TRS:    5       Wetterbericht [PA]
TRS:    6       f"ur
TRS:    7       heute
TRS:    8       abend [B3 fall] ? <P>
TRS:    9       <:<#> na:> [NA] [B2] ,
TRS:    10      vergi"s [PA]
TRS:    11      es [B3 fall] . <#>
...
SUP:    42,43   w104_mt_SMA.par @1m"ochtest @1du
SUP:    55      w104_mt_SMA.par Pl"atze . <P>2@>
SUP:    56      w104_mt_SMA.par <:<#> hier3@:>
SUP:    61      w104_mt_SMA.par bitte . <P>4@>
ORT:    0       <"ah>
ORT:    1       hallo
ORT:    2       kennst
ORT:    3       du
ORT:    4       den
ORT:    5       Wetterbericht
ORT:    6       f"ur
ORT:    7       heute
ORT:    8       abend
ORT:    9       na
ORT:    10      vergi"s
ORT:    11      es
...
KAN:    0       QE:
KAN:    1       hal'o:
KAN:    2       k'Enst
KAN:    3       d'u:+
KAN:    4       d'e:n+
KAN:    5       v'Et6#b@r"ICt
KAN:    6       f'y:6+
KAN:    7       h'OYt@
KAN:    8       Q'a:b@nt
KAN:    9       n'a+
KAN:    10      f6g'Is
KAN:    11      Q'Es+
...
TRN:    66560   197888  0,1,2,3,4,5,6,7,8,9,10,11 002
TRN:    377984  43776   12,13,14,15 004
...
NOI:    1;2     <A>
NOI:    9       <#>
NOI:    11;12   <#>
...
USH:    0       244480  Neutral
USH:    244480  519040  "Uberlegen/Nachdenken
USH:    517760  25600   Hand im Gesicht
...
USM:    0       515840  Neutral
USM:    515840  216960  "Uberlegen/Nachdenken
USM:    517760  25600   Hand im Gesicht
...
USP:    1364144 3936    27      CLEAR_ART
USP:    1377776 3536    30      CLEAR_ART
```

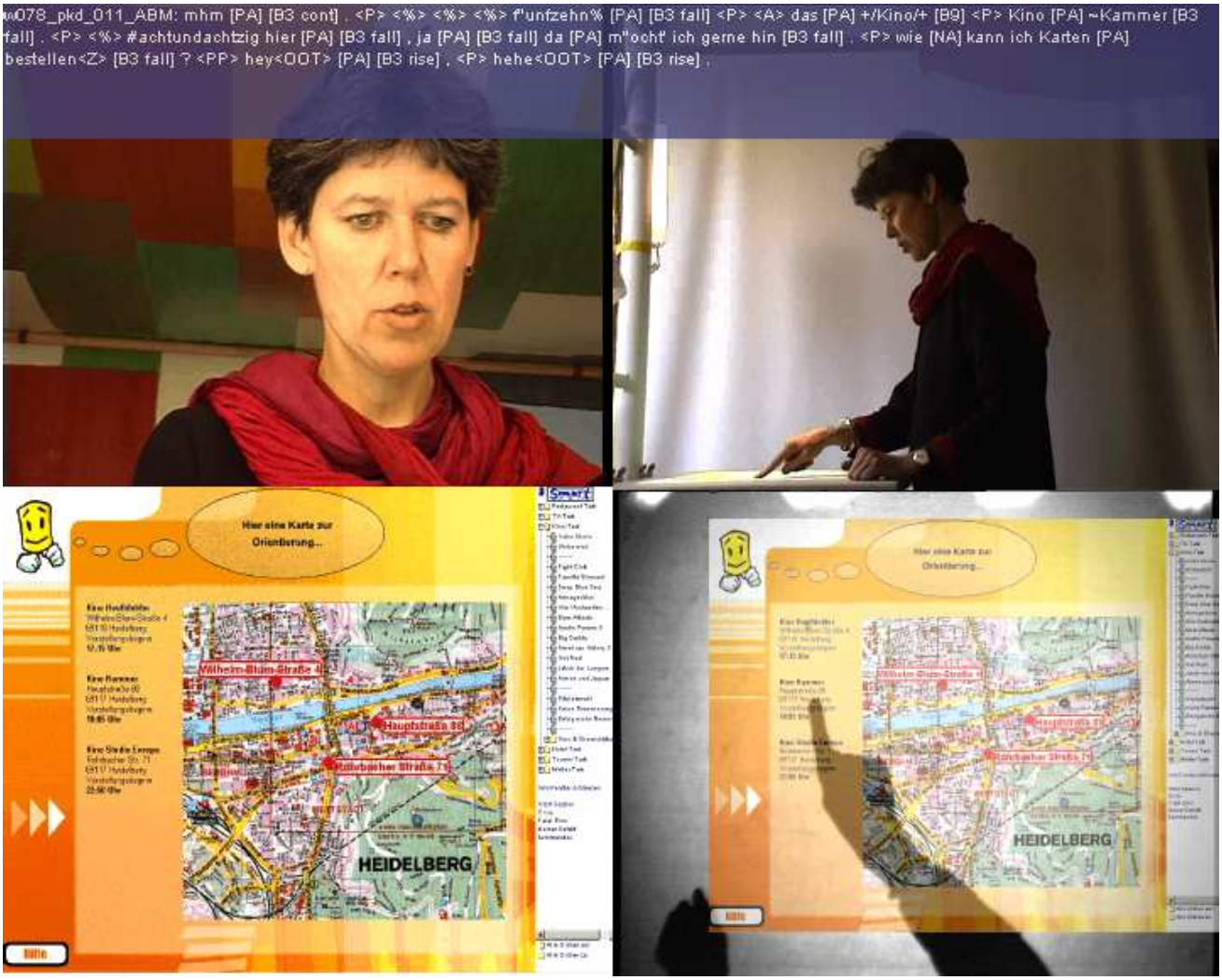

Figure 4: Four synchronized video streams extracted from a SmartKom QT file (see text)

```
USP:    3437728 5856    63      EMPHASIS
USP:    3983392 14992   73      PAUSE_SYLL
...
GES:    265600  32000   U-Geste U - "uberleg - \
        p re Stift nicht erkennbar     640
GES:    376320  30080   I-Geste I - tipp +     \
        re Stift nicht erkennbar
GES:    515200  29440   R-Geste R - emot -     \
        re Hand 393600  8320 "Uberlegung/Nachdenken
...
```

In this example the following tier blocks are contained (see references for details about labeling systems and conventions):

- TRS : SmartKom transliteration (Oppermann et al., 2000)
- SUP : Labeling of cross talk between user and system
- ORT : Lexical entity
- KAN : Citation form in SAM-PA
- TRN : Turn segmentation
- NOI : Noise labeling
- USH : User state labeling using video and audio (Steininger et al., 2002b)
- USM : User state labeling using video only (Steininger et al., 2002a)
- USP : Prosodic labeling of features for user state detection
- GES : Labeling of 2D gestures (Steininger et al., 2001)

For a discussion of the BPF with regard to multimodal data annotation please refer to (Schiel et al., 2002). Detailed descriptions of the individual annotations are stored in the documentation section of each BAS SK DVD or available on the Web[3].

5.4. User Profile

The profile of each user is stored in a file named `UID.spr` where `UID` denotes the unique user id. The user profile is structured in XML-like tag regions and contains the following entries:

- speaker-id
- sex

[3] www.bas.uni-muenchen.de/Bas/BasFormatseng.html

- date of birth
- height
- weight
- right/left handed
- school degree
- school in which state
- profession
- mother tongue
- mother tongue of mother
- mother tongue of father
- dialect
- bilingual
- foreign languages
- bilingual languages
- cultural environment
- speech/singing training
- experience with computers
- experience with dialog systems
- glasses
- smoker
- beard
- piercing
- jewels
- free comments

Only the last field may contain free text; all other fields are parsable. Languages are defined according to ISO 639.2; German according to the RVG1 system[4].

5.5. Recording Protocol

Each recording session is described in detail in the recording protocol file `w057_pk.rpr` where `w057_pk` is the unique session identifier. Here `w057_pk` denotes a WOZ recording in SK Public and the task domain Cinema. The RPR file contains the following blocks of information:

- user: UID
- session parameters: session_id, recorded_domains, atmosphere, background_pattern, pen_mode, evoke_emotions, content_variation, recording_date, recording_location, recording_setup, experimenter, wizard_speech_output, wizard_navigation, session_sequence_no.
- data tracks:
  - video: defines technical setting for all video tracks.
  - gesture: defines coordinates of SIVIT or graphic tableau.
  - audio: defines audio channels and mics.
  - environment: defines distortion of wizard voice, background noise back and front.
  - annotations: defines used annotation formats.
  - QuickTime: defines QT formats for tracks.
- behavior: describes classes of behavior of user during the test.
- free comment on behavior
- other comments

Apart from the last two entries all other fields are parsable.

5.6. Numbers of BAS SK-P 1.0

The first release BAS SK-P 1.0 will contain a minimum of 90 session recordings of 45 users.

- Gender: 25 female, 20 male
- Computer experience: 42
- German nationality: 36
- Age 16 – 24 : 18
- Age 25 – 45 : 21
- Age above 45 : 6

The approx. size of the first release will be 400 GByte.
Please note that the corpus is not thoroughly homogeneous: Not all modalities are recorded in each session. Prospective users of the LR should download the recording protocols (RPR) beforehand and then choose the appropriate session for their respective search topic.

6. Availability and Release Plan

The first release of the BAS SK corpora will contain data collected in the SK Public technical scenario. The release date is scheduled for July 2002. Distribution format will be DVD-5 (100 volumes). Inquiries can be send to bas@bas.uni-muenchen.de. Note that although the corpus is license-free this does not wave the basic distribution costs of EU 255 per volume.
Two other releases are scheduled for 2003: BAS SK-H 1.0 and BAS SK-M 1.0. Please refer to the BAS Web documentation for details end of 2002.
Following the BAS policies the summarized annotation and other meta files will be made available for free on the BAS FTP server[5].

[4] www.bas.uni-muenchen.de/Bas/BasRVG1eng.html

[5] ftp://ftp.bas.uni-muenchen.de/pub/BAS

7. Discussion

The BAS SK corpus is certainly a task-oriented corpus. This means that data were gathered and annotated with certain aims and have therefore a limited re-usability for other purposes. However, following our experiences with the Verbmobil Speech Corpus we hope that the free distribution of this corpus will encourage other groups in the scientific community to work with the data and add further annotations to the corpus. Aside from that we took some effort to produce not solely the required data for the SK project aims but to enrich the basic corpus with other well-tried data. For instance the SK transliteration is largely conform with the Verbmobil transliteration, additional audio channels were recorded (microphone array) and the recorded situation is not exclusively determined by the performance of the SK prototype system.

From our experience so far, the BAS SK corpus is a "difficult" corpus meaning that the data contain a lot of "real world" conditions like

- moving faces (sometime even outside of the frame)
- very differing illumination
- differing background light and/or noises
- users of all levels of computer competence
- users with only second language knowledge
- users with beards, glasses, jewelery etc.
- complex task dialogs; no fixed schemes to solve a task
- un-supervised and naive users; user were basically told that they could do anything to solve the task; some were very experimental and tried to push the system to its edges

There are many thinkable uses for this "LR":

- non-prompted speech recognition with different microphone setups
- face tracking
- eye focus tracking
- user state recognition
- lip reading
- behavioral studies
- gesture recognition
- dialog modeling
- multimodal discourse analysis

We assume that most parties interested in this corpus will not order the full set of recordings (at least not in the near future as long as our distribution medium is restricted to 5 GByte) but rather download only the annotation files and analyze those.

8. Acknowledgments

This research is being supported by the German Federal Ministry of Education and Research, grant no. 01 IL 905. We thank all SK project partners that contributed to the BAS SK corpus.

9. References

N. Beringer. 2001. Evoking gestures in smartkom - design of the graphical user interface. *Springer "Gesture Workshop 2001", London*, page to appear.

D. Oppermann, S. Burger, S. Rabold, and N. Beringer. 2000. Transliteration spontansprachlicher Daten - Lexikon der Transliterationskonventionen. TechDok 02-V4, The SmartKom Project.

F. Schiel, S. Steininger, N. Beringer, U. Tuerk, and S. Rabold. 2002. Integration of multi-modal data and annotations into a simple extendable form: the extension of the bas partitur format. *LREC Workshop on "Multimodal Resources", Las Palmas, Spain*, page to appear.

S. Steininger, B. Lindemann, and T. Paetzold. 2001. Labeling of gestures in SmartKom - The coding system. *Springer "Gesture Workshop 2001", London*, page to appear.

S. Steininger, S. Rabold, O. Dioubina, and F. Schiel. 2002a. Development of the user-state conventions for the multimodal corpus in SmartKom. *LREC Workshop on "Multimodal Resources", Las Palmas, Spain*, page to appear.

S. Steininger, F. Schiel, and A. Glesner. 2002b. Labeling procedures for the multimodal data collection of SmartKom. *Proceedings of the 3rd Int. Conf. on Language Resources and Evaluation, Las Palmas, Spain*, page to appear.

U. Tuerk. 2001. The technical processing in the SmartKom data collection: A case study. *Proceedings of EUROSPEECH Scandinavia*, pages 1541–1544.

K. Weilhammer, F. Schiel, and U. Reichel. 2002. Multitier annotations in the Verbmobil corpus. *Proc. of the 3rd Int. Conf. on Language Resources and Evaluation, Las Palmas, Spain*, page to appear.