The Technical Processing in SmartKom Data Collection: a Case Study

Ulrich Türk

Institut für Phonetik und Sprachliche Kommunikation (IPSK)
Schellingstr. 3/II
80799 München, Germany
tuerk@phonetik.uni-muenchen.de

Abstract

This paper dicusses the specific technical features and processing steps of the multimodal data collection in the SmartKom project [1]. It gives an overview of the goals of the project and the requirements to the multimodal corpus. The processing steps from the data recording to the final distribution of the data are detailed. We focus on the problem of recording temporal synchronous data from different sources and present our manual synchronization process based on standard software and hardware. In addition, we describe shortly the logistic system for organizing the working teams and managing the processing of the data.

1. Introduction

1.1. The SmartKom project

The goal of the SmartKom project is the development of an intelligent computer-user interface which allows almost natural interaction for the user [1]. The system can recognize natural speech as well as gestures on a flat interaction area. Additionally, facial expression is analyzed. The output of the system is presented with a GUI, which is projected on the interaction area, and with synthesized speech (see figure 1 for a schematic view of the recording setup).
In the SmartKom project our institute is responsible for the collection of multimodal data and the evaluation of the system.
In Wizard-of-Oz experiments subjects are recorded in sessions of 4.5 minutes length while they are interacting with a simulated version of the 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 framerate video (used only for labeling)
- gesture coordinates captured by the SIVIT system and the graphical tablet.

1.2. The data collection

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

[1]This research is being supported by the German Federal Ministry of Education and Research, grant no. 01 IL 905.

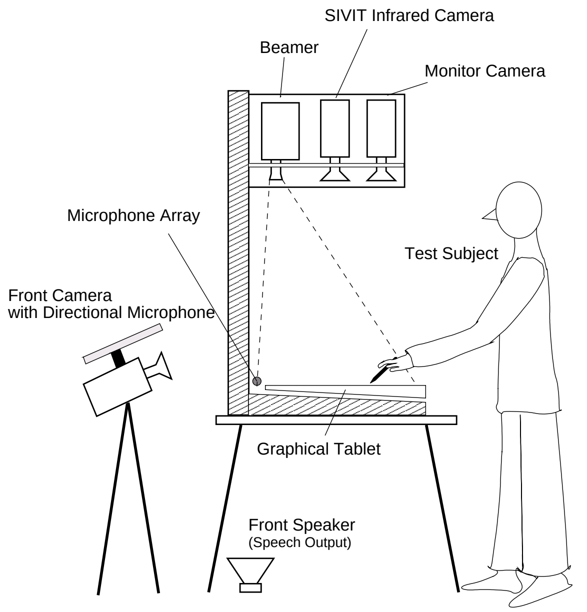

Figure 1: *Schematic illustration of the SmartKom setup used for data collection.*

2. The technical processing in detail

Figure 2 shows the processing stages together with resulting changes on the data file set. Except for the last two stages which are covered by one team each block represents a different working team.
We give at each stage description more information about how the requirements are met. We will focus especially on the capturing of synchronous multimodal data because we think that we have found an easy way to handle this task with standard equipment.

2.1. The QuickTime file format

QuickTime [2] allows the integration of several kinds of media like text, video, audio, images and vector graphics in one multimedia file format. All SmartKom files are compatible with this technology. Together with these data files we provide a socalled framework QuickTime file in the corpus. The user can playback the recorded session, select the tracks to be presented, choose the style of the presentation und navigate to individual turns in the recording. In addition, we use QuickTime's capabilities to render new data files from the existing set of files (see e.g. 2.3.1 or 2.5).

2.2. Data acquisition

During the recording of a Wizard-of-Oz experiment all 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) is recorded seperately via a FireWire bus between camera and computer and is stored as Quicktime file 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 (see 3.2 for details on storage requirements).
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. The file format is a text file containing a list of coordinates together with timestamps.
In addition to the actual recording, this stage comprises also the generation of new entry for the session in the database. Information about the speaker as well as his or her behaviour 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. The database is presented in more detail in 3.1.

2.3. Preprocessing

This stage comprises all steps that are necessary to deliver a set of synchronized data files to the following processing stages. Because the recording is done on several computers each running on its own internal clock, it is thus necessary to align all data tracks to a single master clock and to define a common start and end point.

2.3.1. Generation of "visible" coordinate logfiles

In a first step the coordinate data files must be made accessible for the editor in order to perform the temporal alignment. Our idea was to use the sprite track feature of Quicktime where small bitmaps (called sprites) can be blended over another video track. The parameter of the sprites like for example the position or orientation can be changed dynamically.

Figure 3 shows an overlay of three video tracks:

- the infrared video from the SIVIT camera, showing a greyscale picture of subject's gestures
- the video track of the graphical system output

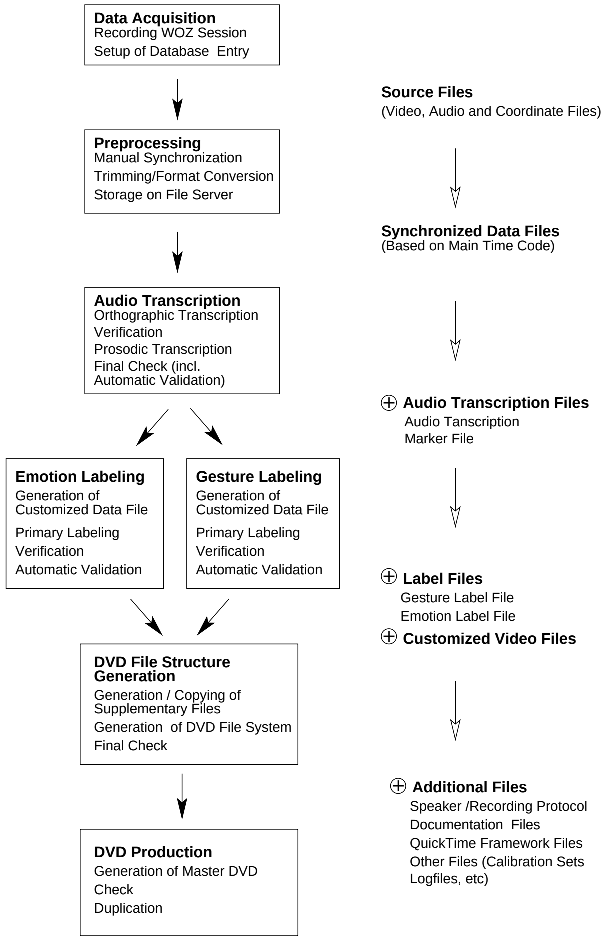

Figure 2: *Processing stages in SmartKom data collection (left) and corresponding changes in the data file set (right).*

- the visualization of the SIVIT coordinates.

The coordinate data is displayed by a moving dot with a comet-like tail which gives not only a graphical representation of the current position but also of the history of coordinates at previous time steps.

2.3.2. Manual synchronizing

The actual synchronization process is done manually on a video editing workstation running Adobe Premiere [3].
The process is done in two steps:
First the audio files which are recorded seperately on the audio capture computer are aligned to the two video camera tracks. This is done by aligning visually the waveforms of the audio tracks to the waveforms of the audio tracks recorded by the cameras.
Second the remaining video tracks are aligned by matching characteristic changes in the video tracks e.g. a change of graphical system output or a stroke of a hand gesture. The circumstance that the test subject starts each session with a finger tap on the web persona helps with this task.
At this point we have aligned all data tracks at the beginning of the recording. As we will present in 2.3.3 the data tracks keep an acceptable alignment over the length of the recording (4.5 minutes).

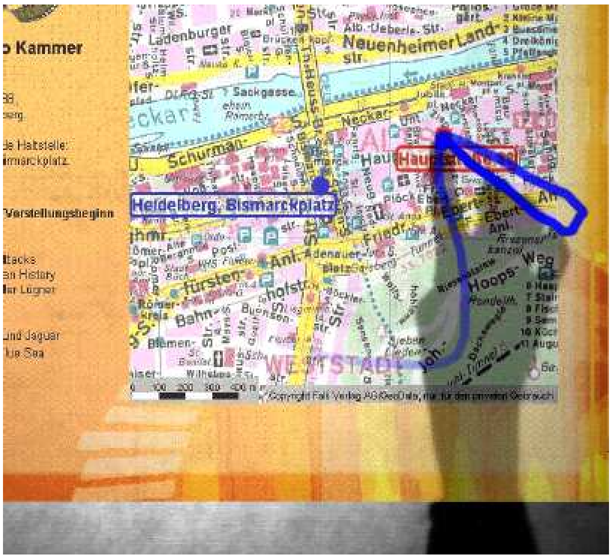

Figure 3: *Coordinate logfile of the SIVIT gesture recognizer visualized by comet-like data plot.*

In a last step the start and end point of the recording is defined and all tracks are cut to equal length. At the same time some file format conversions are done like e.g. downsampling the audio data files to 16kHz. The coordinate files are updated: the temporal shift and trimming which was performed on the visual coordinate tracks must also be considered in the raw data files.

2.3.3. Discussion on the synchronization process

A big challenge in the SmartKom data collection was the ability to record different data sources and to combine them to one temporal consistent data entity. Due to a limited budget we had to use standard components for the recording studio. While audio hardware with synchronizing features is available, video systems which allow capturing of multiple streams can only be found in the high end (and also high price) market. The problem how the other data sources like the gesture coordinate data in our project can be integrated in such a setup still remains.
Another aproach would have been to synchronize the timebase of the recording computers via network using e.g. the NTP protocol [4]. This would have required the design of new capturing tools in order to profit from the distributed master clock. Beside the fact that this would have caused a great delay for the start of the SmartKom data collection, this aproach failed because of missing APIs for the video hardware.
Our current synchronization process is based on the assumption that the timing deviations of the different capturing processes stay in an acceptable range. We try to achieve this by a setup which is as homogenous as possible. The three video capture computers e.g. are equipped and configured identically.
We made test recordings in order to specify the timing deviations in our setup. These test recordings were limited to a length of 9 minutes as the size of the video files would otherwise hit the 2GB limit for the file size given by the operation system. All data tracks were recorded and manually aligned as described in the previous section. We found an increasing timing deviation between the video tracks and the remaining tracks which reached the maximum of one full video frame (40 ms in PAL) after 9 minutes. If we assume a linear increasing drift we end with a timing deviation of 20ms in a typical SmartKom session of 4.5 minutes. Deviations in this range are below the range human beings normally can perceive. Dialogue systems like SmartKom operate with much higher timing windows when combining events from different modalities. We consider therefore the timing deviation as acceptable for this multimodal corpus.

2.3.4. Data storage on server

Finally the aligned data files are stored on a Linux file server. The working teams of the following process stages access these files over the network in order to prevent repeated copying of large files. The actual path to the data is noted in the database sheet.

2.4. Audio transcription

Audio transcription is a time consuming manual process. It includes orthographic transcription (proper names, hesitations, noise and pronuncation variants are also tagged), a verification step and prosodic transcription (primary and secondary stress, intonational movement at phrase bounderies).
During the orthographic transcription a marker file is generated which describes the start and end point of each turn of the human-machine-interaction. This timing information is for example used later to display the transcription turn-wise in the video stream in a subtitle-style.
The final transcription file (called TRL file) is checked manually and validated automatically against the transcription conventions. More information on the audio transcription process can be found in [5] .

2.5. Emotion and Gesture Labeling

These worksteps require customized video files which are rendered from the source data files on the file server. Figure 4 shows a frame of this video track used for gesture labeling. It contains four different views on the video data in one picture (front camera, side camera, graphical system output, infrared camera merged with the system output) and an audio track. In addition, the audio transcription of the current turn is inserted on the top. The file format is AVI with Cinepak encoding as this the only video format which can be used together with the labeling tool Interact [6]. For the emotion labeling a simplified version (only the front camera together with the audio TRL and an audio track) is used.
Both customized video tracks are created by a specially programmed Java application using the Quicktime API for Java [8]. The label files generated in these stages are stored in ASCII text files. They are later integrated in the framework Quicktime file in the same manner as the audio TRL file. For more information on gesture labeling see [7].

2.6. Final data preparation

In this stage all source files and annotation resp. label files of a session are collected, checked finally and prepared for distribution on DVD-R media. A specially designed tool provides help with the required working steps.
Supplementary files for the sessions are generated: for example a speaker protocol file, which gives information about the characteristics of the subject or a recording protocol file describing parameters of a recording. Final checks of the source files are done. They include semi-automated checks of cross-references in the data files (e.g. same speaker id used in speaker protocol

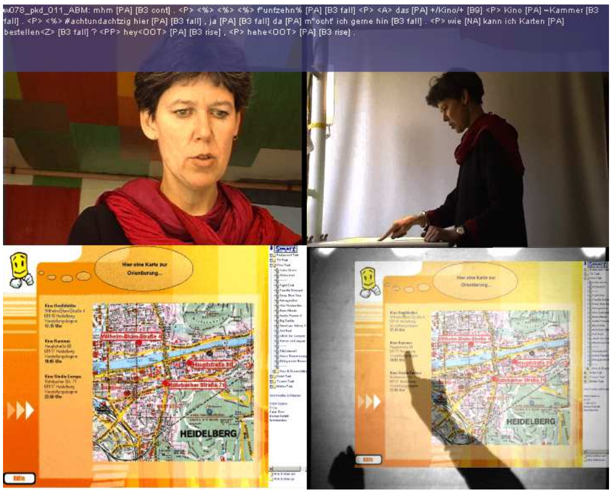

Figure 4: *Video stream used for gesture labeling.*

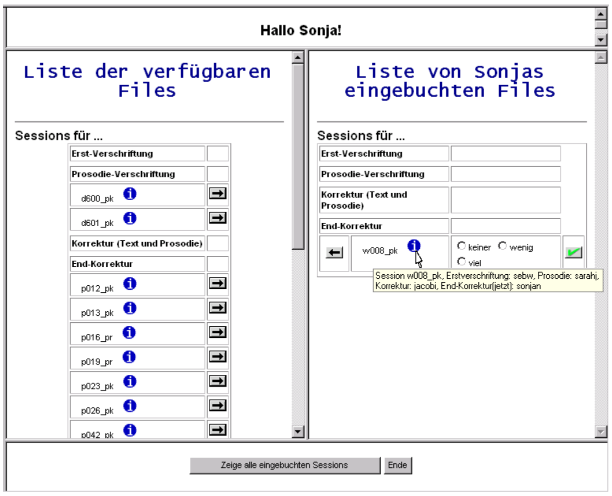

Figure 5: *An example of a web interface for access to the SmartKom database.*

and recording protocol) as well as manual consistency checks. The framework Quicktime file is generated in the the next step. It offers the possibility to choose among different views on the data files and use these specialized views for playback of the recorded session. In the last step the DVD file structure is generated and additional files like for example technical documents, or useful tools for viewing the data are copied.

2.7. DVD Production

Finally, all data is moved to the DVD-R station, an Apple Macintosh running DVD Toast by Adaptec. A master DVD is burnt and verified; it serves then as source for duplication. As soon as the DVD is verified sucessfully, all data is removed from the server.

3. Database and data logistics

3.1. The database

The database used in SmartKom data collection stores two kinds of information:

- information on the subjects, the recordings and the recorded data itself
- information on the current process state of each data set.

The database is designed with FilemakerPro and can be accessed over several web interfaces. Customized interfaces are supplied for each working team. Figure 5 shows the interface for the audio transcription group. On the left side is a list of recordings waiting for the next editing step, on the right side is a list of reserved sessions for a particular user. These interfaces allow a reservation of a recording for an editing step and releasing the reservation after an editing step is done. In addition they support the entry of comments in the database. For queries the user can list the previous editors who worked on the data set (see the small tooltip window in figure 5).

3.2. Data logistics

The database mentioned is a key element in organising the recordings and the further processing of the data. Due to the fact that each recording makes up a data amount of 4 GB (temporarily even up to 6.5 GB) and because up to 30 recordings are processed in parallel, the whole processing must be well planned. Currently our file server offers 200 GB disc space and the processing time needed from the recording to burning the DVD is about 5 weeks. These facts restrict the number of test subjects to 3 per week.

4. Further improvements and Outlook

The current recording setup will be further improved in order to make it easier to handle for the wizards. Other improvements are required in the field of organizing the working teams. We are developing a tool to monitor the processing of sessions on the file server and to support planning new recordings.
Starting in Summer 2001 we will evaluate already existing SmartKom modules by integrating them in our setup and logging their output. This task will require the design of new working stages and tools to handle this data.

5. References

[1] <http://smartkom.dfki.de/index.html>

[2] <http://www.apple.com/quicktime/>

[3] <http://www.adobe.com/products/premiere/main.html>

[4] <http://www.eecis.udel.edu/ ntp/>

[5] Beringer, N., Burger, S. and Oppermann, D., “Lexikon der Transliterationen”, SmartKom Technisches Dokument 02-00, 2000

[6] <http://www.mangold.de/>

[7] Steininger, S., “Labeling of Gestures in SmartKom — Concept of the Coding System” SmartKom Report No. 2, to appear in the proceedings of the Gesture Workshop 2001

[8] Maremaa, T., Stewart, W., “QuickTime for Java: A Developer Reference”, 1999, Morgan Kaufmann Publishers