Multimodal Emogram, Data Collection and Presentation

Johann Adelhardt, Carmen Frank, Elmar Nöth, Rui Ping Shi, Viktor Zeißler, Heinrich Niemann

Friedrich-Alexander Universität Erlangen-Nürnberg, Germany
{shi,adelhardt,batliner,frank,noeth,zeissler,
niemann}@informatik.uni-erlangen.de,

Summary. There are several characteristics not optimally suited for the user state classification with Wizard-of-Oz (WOZ) data like the nonuniform distribution of emotions in the utterances and the distribution of emotional utterances in speech, facial expression, and gesture. In particular, the fact that most of the data collected in the WOZ experiments are without any emotional expression gives rise to the problem of getting enough representative data for training the classifiers. Because of this problem we collected data in our own database. These data are also relevant for several demonstration sessions, where the functionality of the SMARTKOM system is shown in accordance with the defined use cases.

In the following we first describe the system environment for data collection and then the collected data. At the end we will discuss the tool to demonstrate user states detected in the different modalities.

1 Database with Acted User States

Because of the lack of training data we decided to build our own database and to collect uniformly distributed data containing emotional expression of user state in all three handled modalities — speech, gesture and facial expression (see Streit et al. (2006) and for an online demonstration refer to our website [1]). We collected data of instructed subjects, who should express four user states for recording. Because SMARTKOM is a demonstration system it is sufficient to use instructed data for the training database.

For our study we collected data from 63 naive subjects (41 male/22 female). They were instructed to act as if they had asked the SMARTKOM system for the TV program and felt content, unsatisfied, helpless or neutral with the system feedbacks. Different genres such as news, daily soap and science reports were projected onto the display for selection. The subjects were prompted with an utterance displayed on the screen and were then to indicate their internal state through voice and gesture, and at the same time, through different facial expressions.

[1] http://www5.informatik.uni-erlangen.de/SmartKom/SkBook.html

Fig. 1. Screen used for recording the user states in our local environment; several genres are shown together with a state that should be played. The *upper right corner* shows which modalities are currently recorded

1.1 Data Collection Environment

For the data collection we developed a special environment which we used directly with the local SMARTKOM demonstrator of the Institute of Pattern Recognition (LME). The screen of the graphical user interface of our experiment is shown in Fig. 1.

It shows in the center a 3×3 matrix of several genres, from which the subject had to choose one, together with a user state that should be expressed by the subject. This user state is shown in the lower left corner of the screen. At the middle bottom, the utterance, which the user had to speak at the current turn, is shown. Between the bottom of the 3×3 matrix of genres and the generated user utterance a colored bar is shown, which informs the user that the system is recording his speech.

On the right of the screen several features for handling the local data collection environment are shown. In the upper right corner five kinds of data collection can be chosen. Four of them refer directly to the modalities and to the combination of modalities used in SMARTKOM. The fifth one, UseCases, was used to collect data

Fig. 2. Sample data; from *left to right*: *hesitant*, *angry*, *joyful*, and *neutral*

related to so-called UseCases, which were used for the demonstration of the SMARTKOM system.

In the middle and the lower right part of the screen control buttons for camera and microphone and the field for the identification label of the experiment are shown. These were handled by the supervisor of the experiment with mouse and keyboard.

Data collection was done separately for all three modalities and for the combination of them. Because of the lack of annotated emotional data, the facial expression was recorded throughout, although the subjects were unaware of it. During the utterance the user had to select the genre by gesture in the centered choice matrix mentioned above.

Data collection was done in several sessions so the users could get familiar with the system and with the way to express their emotions naturally facing the camera and/or microphone in an incremental way. In the first session we recorded only the facial expression as the subject had to express the related emotion by speaking the utterance. In the second session we also recorded the speech of the subject together with facial expression. In the third session gesture and facial expression were recorded, while in the final session all three modalities were recorded as indicated with the tag *"speech_gesture"* (facial expression automatically included as mentioned above) in the upper right corner of the screen.

1.2 Collected Data

Facial expression, gesture and speech were recorded simultaneously in the experiment; this made it possible to combine all three input modalities afterwards. The user states were equally distributed. The test subjects read 20 sentences per user state. The utterances were taken in a random order from a large pool of utterances. About 40% of them were repetitions of TV genres or special expressions, which did not actual

depend on the given user state, like *"tolles Programm!"* (*"nice program!"*). In other words, we chose expressions one could produce in each of the given user states. (Note that a prima facie positive statement can be produced in a sarcastic mode and by that, turned into a negative statement.) All the other sentences were multiword expressions, where the user state could be guessed from the semantics of the sentence. The subjects should align to the given text, but minor variations were allowed.

From all collected data we chose 4848 sentences (3.6 hours of speech) with good signal quality and used them for further experiments. For the experiments with prosodic analysis, we randomly chose 4292 sentences for the training set and 556 for the test set.

For the facial analysis video, sequences of ten subjects were used. These subjects where selected because their mouth area was not covered by facial hair or the microphone. As training images, we used image sequences of these subjects without wearing the headset. In the images of the test sequences, there is a headset. Some of the training images can be seen in Fig. 2.

For gesture analysis there are, all in all, 5803 samples of all three user states (note that there are only three user states for gesture as mentioned in Shi et al. (2006)), and 2075 of them are accompanied by speech. As we are interested in the combination of all three modalities, we concentrate on this subset. Of this sample subset, 1891 are used for training and the other 184 are used for testing. Since the samples were recorded according to the user states categories in facial expression and speech, we merged the data of the corresponding user states *neutral* and *joyful* into a general user state category *determined* for gesture. The data we collected for the multimodal emogram (MMEG) are described in Zeißler et al. (2006), Frank et al. (2006) and Shi et al. (2006).

2 Multimodal Emogram (MMEG)

During the system development there always exists a problem if the input and output happen to take place simultanenously. It is also unwise to registrate any kind of output at the end of the processing queue without knowing the steps between input and output. In particular, if there are several steps for processing with statistical methods, as is the case in SMARTKOM with user states, it is absolutely necessary to know the intermediate results. The *multimodal emogram*[2] is a tool to show the results of user state processed in three modules: prosody, facial expression and gesture analysis.

The advantages of the evaluation tool are the following:

- quick presentation of recognizer results in the running environment
- systematic evaluation
- demonstration of functionality without the full system environment
- possibility to show the user the difficulty to act "as if" he were in the corresponding user state

[2] http://www5.informatik.uni-erlangen.de/SmartKom/SkBook.html

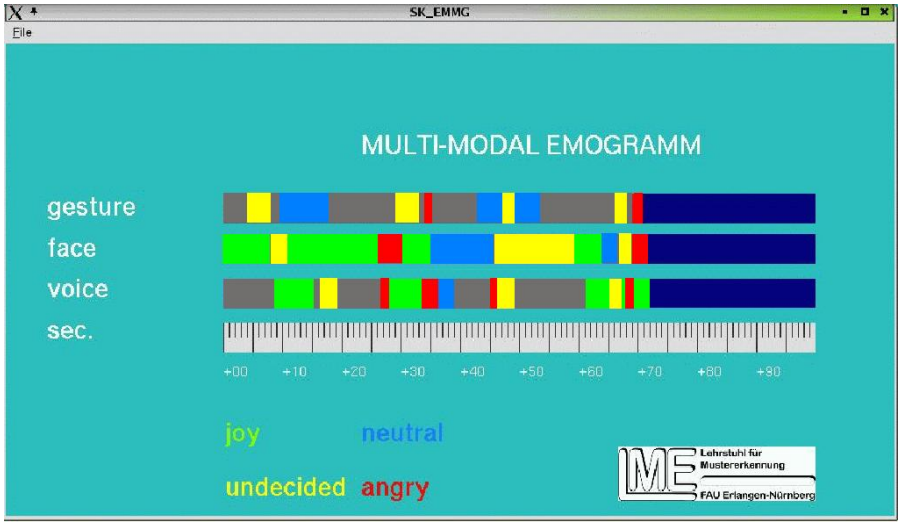

Fig. 3. Presentation of the results of the user state analysis in three modalities: speech, gesture and facial expression

For evaluating the system it is absolutely necessary to know the final result, but it is also important to know the results of the stages between input and output. In the case of error during the early development stage the developers are able to localize system components that produce errors.

In our case this is the presentation of the user states as recognized by the three modules. Their output is taken by the tool and is shown in a presentation where the user states are aligned parallel to the modalities. The presentation shows in this manner the detected user state of each module/modality at a certain time. So the recognition result of each single modality can be analyzed also for itself and at runtime.

The MMEG provides a compressed and transformed presentation of the user utterance in form of user states in quasi–real time. Just as the sonagram shows the spectral energy and the formants of the speech, the multimodal emogram shows user states as they occur in speech, facial expression and gesture. At the moment there are some problems in presentation concerning time alignment, synchronization and real-time behavior. But the idea of presentation generation in real time is realized, and it relies only on technical aspects such as the dependency on the word lattice or the power of the CPU. Another positive side effect of the evaluation tool is the possibility to show the functionality of the components processing the user state without running the full system.

For the robust classification of the user state it is necessary to analyze the input of three different modalities. Therefore, these single results have to be combined in a fusion component, which is performed by the *media fusion module* in the SMART-

KOM environment. By considering the result of the fusion, it is possible to know the contribution of each modality to the final recognized user state. By knowing the results of each system component, it is possible to give a clue to the current user state and thus to increase the system performance in several tuning steps.

Beyond evaluation the multimodal emogram delivers very valuable information of the analysis in the system. So it is possible to adapt it to several conditions:

- A user state is not recognized: the tool shows the results of the different modalities; with the help of these results it is possible to find out at run time why the user state is not recognized,
- Testing of new side conditions like new light conditions or new microphones.
- Training of system conformation behaviour for presentation tasks.

For the tuning task it is helpful to show the contribution and result of each modality. Like the sonagram shows the different frequencies in the speech signal, the multimodal emogram shows the results of user state expressed in speech, facial expression and gesture.

The user states are presented with the help of colored bars as shown in Fig. 3. For each modality there is one bar. We have four colors for the presentation:

- red: angry/anger
- green: joyful/joy
- yellow: hesitant/undecided
- blue: neutral

For the recognition of user states in gesture there is the restriction that only three user states are defined: angry, hesitant, and joyful together with neutral as the third class.

References

C. Frank, J. Adelhardt, A. Batliner, E. Nöth, R.P. Shi, V. Zeißler, and H. Niemann. The Facial Expression Module, 2006. In this volume.

R.P. Shi, J. Adelhardt, A. Batliner, C. Frank, E. Nöth, V. Zeißler, and H. Niemann. The Gesture Interpretation Module, 2006. In this volume.

M. Streit, A. Batliner, and T. Portele. Emotion Analysis and Emotion Handling Subdialogs, 2006. In this volume.

V. Zeißler, J. Adelhardt, A. Batliner, C. Frank, E. Nöth, R.P. Shi, and H. Niemann. The Prosody Module, 2006. In this volume.