Development of the User-State Conventions for the Multimodal Corpus in SmartKom

Silke Steininger[†], Susen Rabold[†], Olga Dioubina[†], Florian Schiel[*]

[†]Institute of Phonetics and Speech Communication
[*]Bavarian Archive for Speech Signals (BAS)

Ludwig-Maximilians-University, Schellingstr.3, 80799 Munich, Germany
{kstein, rabold, olga, schiel}@phonetik.uni-muenchen.de

Abstract

This contribution deals with the problem of finding procedures for the labeling of a multimodal data corpus that is created within the SmartKom project. The goal of the SmartKom project is the development of an intelligent computer-user interface that allows almost natural communication with an adaptive and self-explanatory machine. The system does not only accept input in form of natural speech but also in form of gestures. Additionally the facial expression and prosody of speech is analyzed.

To train recognizers and to explore how users interact with the system, data is collected in so-called Wizard-of-Oz experiments. Speech is transliterated and gestures as well as user-states are labeled. In this contribution we will describe the development process of the User-State Labeling Conventions as an example for our strategy of functional labeling.

Key-words: multi-modal, annotation, user-states, human-machine interaction, coding conventions.

1. Introduction

The goal of the SmartKom project is the development of a multimodal dialogue system that allows the user to interact almost naturally with the computer. Among other things the emotions of the user are taken into account by the system. Since not much is known about the role emotions play in a human-machine dialogue, data is collected in Wizard-of-Oz experiments. The analysis of the interaction of the users with the simulated system can reveal which emotions occur in such a situation, in which way the emotions are expressed and in what connection. For such an analysis the data has to be labeled[1].

This contribution deals with the problem of how to define a labeling procedure for emotions, respectively. user-states[2]. We will first describe shortly how the data was collected that was used for the development of the labeling procedure. Then we describe the requirements the procedure had to meet. After that we give an overview over the steps of the development process of the procedure and some open questions.

2. Collection Of Multimodal Data

The data collection is done with the Wizard-of-Oz technique: The subjects think that they interact with an existing system but in reality the system is simulated by two humans from another room.

In each Wizard-of-Oz session spontaneous speech, facial expression and gestures of the subjects are recorded with different microphones, two digital cameras (face and sideview hip to head) and an infrared sensitive camera (from a gesture recognizer: SIVIT/Siemens) which captures the hand gestures (2-dimensional) in the plane of the graphical output. Additionally, the output to the display is logged into a slow frame video stream. Each subject is recorded in two sessions of about 4.5 minutes length each. For more information on technical details of the data collection see Türk (2001).

3. Developing the Labeling Procedures - Starting Point

3.1 Goals

The labeling of user-states in SmartKom serves two main functions:

1. The training of recognizers.
2. The gathering of information how users interact with a multimodal dialogue system and which user-states occur during such an interaction.

These two goals had to be satisfied with the labeling procedures we had to define. For practical and theoretical reasons we decided against a specific system like the "Facial Action Coding System" of Ekman (1978) where the precise morphological shape of facial expressions is coded, but used a simplified, practice-oriented system. The user-states are defined with regard to the subjective impression that a human communication partner would have, if he would be in place of the SmartKom system. This is a functional definition: Not the user-state per se is coded, but the impression the communicated emotion or state generates.

In Steininger, Lindemann & Paetzold (2002a) we already discussed this approach with regard to gestures[3]. The next paragraphs explain our approach relating to user-states.

3.2 Practical Requirements

[1] The development and structure of the gesture labeling is described in detail in Steininger, Lindemann & Paetzold (2002a). The transliteration conventions can be found in Oppermann et al. (2000). The special problem of combining the information of the different labeling steps and the transliteration is discussed in Schiel et al. (2002) at this workshop.

[2] The name “emotion labeling” was changed in “user-state labeling” because the targeted episodes in the data comprise not only emotional, but also cognitive states.

[3] Our gesture coding system also defines hand gestures functionally (not morphologically). A labeled unit is coded with regard to the intention of the user, i.e. with regard to his (assumed) discrete goal.

To satisfy the two goals of the labeling process mentioned above the following requirements had to be met. They apply to transliteration, gesture and user-state labeling.

1. The labels should refer to the functional level[4], not the morphological level. For theoretical reasons we want to use a functional coding system (see below). However, the decision is also made for practical reasons since the structural coding of e.g. facial expressions is exceedingly time consuming.

2. The labels should be selective. Functional codes (as indirect measurements) are not as exact as direct methods, therefore exceptional care has to be taken to find labels that are well-defined, easy to observe and unproblematic to discriminate by means of objective (communicable) criteria. This is even more true for user-states than for gestures because communicable criteria for the discrimination of functional user-state categories are hard to find.

3. The coding system should be fast and easy to use.

4. The resulting label file should facilitate automatic processing (a consistent file structure, consistent coding, non-ambiguous symbols, ASCII, parsability) and preferably should be easy to read.[5]

5. The main categories and most of the modifiers should be realized as codes and not as annotations, in order to heighten consistency. Annotations (free comments and descriptions that don't follow a strict rule) are more flexible, but codes (predefined labels from a fixed set) increase the conformity between labelers.

4. Definition of the User-State Coding System

The questions that have to be solved to detect user-states automatically are: Which features of the face and of the voice contribute to an emotional impression - and in which degree does each feature contribute to the impression? Which of these features can be detected automatically?

If we already knew the answers it would make sense to define coding conventions that mark these features in the data. But since we are far from answering these questions conclusively we decided to use another strategy: The labelers mark beginning and end of a user-state sequence and sort it into one of several subjective categories.

A human in a conversation with another human is able to judge which emotion or user-state his or her communication partner shows. Therefore he or she should be able to discriminate relevant user-states in a video. Of course the labeler does not know which emotion is truly present in his communication partner/a human in a video and he or she will make mistakes. But he or she should be good enough to use his emotion-detection capability to keep the conversation smooth. This goal is the same for the system - it should be able to detect which user-state is present in its communication partner to keep the conversation smooth.

This consideration we used for the definition of the user-state coding system.

4.1 First Step: Pretest - Labeling with some defined subjective categories

First we decided to look for several categories that were deemed interesting for user-state recognition: "anger/irritation", "boredom/lack of interest", "joy/gratification (being successful)", "surprise/amazement", "neutral/anything else". A few sessions were labeled with these categories. Beginning and end were defined by an observable change in the emotional state of the user. It was marked if the user-state seemed "weak" or "strong".

In the first step each session was labeled by at least two different labelers. After the labeling the categories were discussed. "Boredom/lack of interest" was excluded because it could not be distinguished from "neutral". "Neutral" and "anything else" were separated into two different categories because many sequences were found where the users definitely did not show a neutral expression but no meaningful label could be given. Two new categories were included to describe user-states that occurred quite often in the data and are important in the context of human-computer interaction: helplessness and pondering/reflecting.

The label "anything else" comprises three cases:

1. Grimaces with no emotional content, for example playing with the tongue in the cheek, twitching muscles etc. (about 65%).

2. Emotional sequences that have no label in our system, for example disgust (about 5%).

3. States that seem to have an emotional or cognitive meaning, but cannot be decided upon by the labelers (about 30%).

The three cases were put together into one category because they all comprise sequences that are not suited as training material.

Cases like number 2 (disgust etc.) are very uncommon in our context and because of this an extra category was not deemed worthwhile. Cases like number 1 (grimaces for physiological reasons) sometimes look very similar to user-states, but have a different meaning - therefore they have to be distinguished from neutral.

Cases like number 3 would be interesting to analyze further because the comprise complex or difficult to understand user-states. They are sorted into the "anything else" category simply for practical reasons: The other labels should be selective, therefore any label that cannot be categorized for certain has to be sorted into "anything else".

4.2 Second Step: Holistic labeling with the conventions

In a second step the sessions were labeled with the following fixed set of categories:

- joy/gratification (being successful)
- anger/irritation
- helplessness
- pondering/reflecting
- surprise
- neutral

[4] "Functional code" or "functional unit" is sometimes defined differently by different authors. We use the term in accordance with Faßnacht (1979) for a unit that is defined with regard to its effect or its context.

[5] Many of the practical criteria were adopted from the transliteration conventions for speech in SmartKom, see Oppermann et al. (2000).

- unidentifiable episodes

Consistency was achieved by two correction steps. Final correction was done by the same corrector for every session. Difficult episodes were discussed.

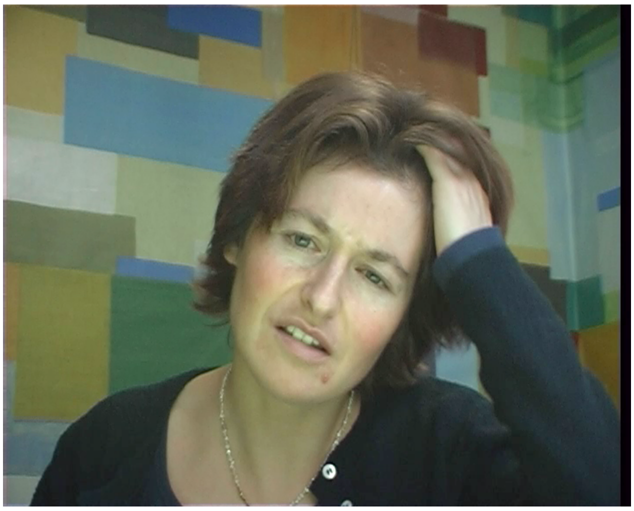

Figure 1: Example of the front view that is used for the holistic and the facial expression labeling. The picture was taken from an episode that was labeled as "anger/irritation" in the holistic labeling step.

4.3 Third step: Finding features

The categories are assigned according to the subjective impression of the labelers. Nevertheless the goal is to find detectable features. Additionally the categories have to be describable with observable criteria - otherwise no one else apart from the labelers will be able to understand the content of the labels.

Therefore, for each category some characteristic features were listed. A feature was included in the list if it occurred regularly or if it seemed very distinctive of a category for some subjects.

This step of the development process is still in progress. At the moment the features are simply an aid for labeling. However, the feature list could be studied with objective methods to judge which features are good candidates to be "indicators" for a category.

4.4 Fourth Step: Overcoming some limitations

With the holistic labeling system we were relatively sure to catch all relevant user-state episodes and to sort them into selective categories. However, a serious problem had to be solved: For the recognition of facial expressions the coding system was not well suited. Because of the holistic approach the labels included not only information from the facial expression, but also from the voice and from the context. This is a problem because a facial expression recognizer derives information only from the facial expressions and a prosody recognizer derives information only from the voice.

First, we tried to solve the problem with a special marker of the source for a category: voice or face. But it turned out that it was very difficult to make the judgment with regard to the source. Additionally, only very few episodes with the source "voice" could be found.

We abandoned the source marker and included two different labeling steps: Labeling of the facial expression without audio and prosodic labeling.

For the facial expression labeling a different labeler-group watched the videos without audio. The labelers started with a pre-segmented file (from the holistic labeling) to avoid missing subtle episodes that are hard to perceive without audio and context information. This pre-segmentation was derived from the holistic labeling - the names of the categories (apart form "neutral") were deleted, the borders were retained.

Since it seemed to be difficult to use the functional approach with regard to the voice, we adopted a formal coding system that was used in Verbmobil (Fischer, 1999) and changed it to suit our needs in SmartKom.

For the prosodic labeling the transliteration files are filtered: Only the orthographic transcript remains so that the transliteration labels don't divert the prosodic labelers. For the labeling prosodic features like pauses, irregular length of syllables and other prosodic features which could reveal the emotional state of the particular user are marked. There are nine categories for the prosodic labeling:

1. Pauses between phrases
2. Pauses between words
3. Pauses between syllables
4. Irregular length of syllables
5. Emphasized words
6. Strongly emphasized words
7. Clearly articulated words
8. Hyperarticulated words
9. Words overlapped by laughing

The labels were chosen according to the requirements for the User-State recognition group in SmartKom and are thought to represent prosodic features that are indicative of emotional speech. Hyperarticulated words for example, can be indicative of anger. However, it is still not known very well which prosodic features occur during which emotional states. Nevertheless, by the comparison between the holistic labeling and the prosodic labeling it should be possible to detect relevant user-states in speech. For more information on the usage of prosodic features as indicators of emotional speech please refer to Batliner et al. (2000).

For a detailed description of the labels and concrete examples for the labeling procedure please refer to our paper at the main conference (Steininger, Schiel & Glesner, 2002b).

4.5 Open Questions

We have to state clearly that the user-state labeling procedure is work in progress. The description of the categories, along with some formal criteria to help differentiate categories that can be mixed easily is not complete. After it's completion, the intercoder agreement has to be measured. At the moment, we can only use the extent of corrections that are done in each correction step as a rough indicator how reliable the labeling procedure probably is:

Holistic labeling: About 20% of all labels are changed with regard to content. About 10% of the segment borders are changed. This is the case for correction step 1 as well as 2.

Facial Expression labeling: Only one correction step exists. Segments borders have to be corrected almost never. Changes of labels with regard to content occur in about 20% of the cases.

Prosodic labeling: Only one correction step exists. Changes of labels with regard to content occur in about 20% of the cases. Changes of time markers occur in about 50% of the cases.

One other problem that remains are mixed emotions. Since there is no category for mixed emotions, all such cases have to be sorted into "anything else". However, the problem is not as big as it seems: Since we use categories that are defined mainly by subjective impression not mainly by formal criteria, it is rare that a labeler has the impression of a mixed emotion[6]. As already mentioned, the labeler take the viewpoint of a communication partner and try to discern which state his opponent is in. On this level, there almost always is an integrated impression of only one emotion at a time. Many emotional states are mixed of course if one analyses them closely. With a formal system like FACS (Ekman, 1978), mixed emotions correspond to mixed expressions: The face may show anger (for example with a frown) and surprise (for example with an open mouth). In a functional system like ours the viewpoint is taken that it is not known if a frown always means anger and an open mouth always means surprise. If the frown and the open mouth leave the observer (labeler) with the impression of reflecting then this label is given. That is to say that a mixed state on the formal level can lead to a new (holistic) impression on the functional level. Actually this is quite often the case. In most instances there is a clear message for a communication partner. We label only this "clear message", not the subtle undercurrents.

Of course the overall impression can also be of a mixed state. In this case the label “anything else” is given since only very few mixed states were found. Since for the voice a formal system is used and in one labeling step the facial expression is judged without the audio information mixed states for speech and facial expression can occur. In some cases they will be real mixed states but in some cases they will occur because of labeling mistakes.

In our view, formal and functional systems can complement each other, but cannot replace each other because they refer to different levels.

A third important open question is the “anything else” category. For practical reasons some of the most interesting cases “disappear” into this category, namely the episodes that cannot be categorized neatly. Of course it would be of great interest to analyze these difficult episodes further. How could this be done? It is no option to ask the subjects what they felt in the case of an unidentifiable user-state, because with the functional approach the emotions are labeled that are transmitted to a communication partner. Introspective evaluation of the emotion by the user will give a different picture because of effects of social conventions (among other things). To include recordings of other modalities could be helpful: Hesitant movements for example could give hints about the user-state “helplessness”. However, we decided against using additional visual context information because we wanted to focus the labelers on the face and on changes in the voice accepting that some episodes remain unidentifiable. Adding such information later can change the impression (which is highly context dependent), therefore the whole labeling process has to be done again. An interesting option would be to have the unidentifiable episodes judged by a group of naive, untrained labelers (without giving them predefined categories). In this way it could be analyzed if the unidentifiable episodes are episodes that are difficult to understand by a communication partner or if at least some of them form a user state not yet identified as important.

5. Conclusion

With the example of the user-state labeling we show a way to handle the problem of finding a labeling system that is consistent, fast and catches the most important episodes in a human-machine dialogue. Since as yet there is not known enough about good indicators for user-state recognition we decided against a formal/morphological system. Instead we define the labels after practical experience with the data, in this way circumventing the danger of missing important aspects by making assumptions about indicators for automatic detection that cannot be justified very well yet.

Additionally, by combining holistic labeling, labeling of the facial expression and a formal system for the speech we can make up for the disadvantages a purely holistic, functional coding system would have. Through comparing the different label files it is possible to analyze and process the data from many different points of view, looking at the whole or at parts at will.

It is also possible to combine the user-state labels with the gesture labels or the speech transliterations. It could be interesting to analyze which kinds of gestures occur during which kinds of user-states. During helplessness there should be less interactional gestures and more searching gestures, for example. The comparison between the gesture labels and the transliterations is especially interesting with regard to reference words that are possibly uttered. A combination of all three modalities could be useful to analyze the question if there are more hesitations and aborts in the speech and gestures during angry and/or helpless episodes.

With the traditional way of annotating input modalities separately such comparisons are not possible. The labeling of data of multimodal systems allows new ways of studying human-machine interaction. However, this will be successful only if the coding conventions allow the combination of the labeling of the different modalities with ease.

6. References

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E., 2000. Desperately Seeking Emotions Or: Actors, Wizards, and Human Beings. In R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.): *Proc. of the ISCA Workshop on Speech And Emotion*. Belfast: Textflow.

Ekman, P., & Friesen, W. V., Facial Action Coding System (FACS), 1978. *A technique for the*

[6] With the expeption of “sarcasm”: Cases where the user is smiling and laughing, but it can be suspected that he is also scornful are labeled as “joy/gratification”. Sarcasm is hard to detect reliably, therefore we decided againgst a special label.

measurement of facial action. Palo Alto, Ca.: Consulting Psychologists Press.

Faßnacht, G., 1979. *Systematische Verhaltensbeobachtung*. München: Reinhardt.

Fischer, K., 1999. Annotating Emotional Language Data. *Verbmobil Report 236*.

Oppermann, D., Burger, S., Rabold, S., & Beringer, N., 2000. Transliteration spontanprachlicher Daten-Lexikon der Transliterationskonventionen-SmartKom. *SmartKom Technisches Dokument Nr. 2*.

Schiel, F., Steininger, S., Beringer, N., Türk, U., & Rabold, S., 2002. Integration of multi-modal data and annotations into a simple extendable form: the extension of the BAS Partitur Format. To appear in the *Proc. of the 3rd Int. conf. on Language Resources and Evaluation, Workshop On Multimodal Resources And Multomodal Systems Evaluation,* Las Palmas, Spain.

Steininger, S., Lindemann, B., and Paetzold, T., 2002a. Labeling of Gestures in SmartKom - The Coding System. To appear in *Proc. of the Gesture Workshop*, London: Springer.

Steininger, S., Schiel, F., & Glesner, A., 2002b. User-State Labeling Procedures For The Multimodal Data Collection Of SmartKom. To appear in the *Proc. of the 3rd Int. conf. on Language Resources and Evaluation,* Las Palmas, Spain.

Türk, U., 2001. The technical processing in the SmartKom data collection: A case study. *Proc. of Eurospeech*, Scandinavia, p. 1541-1544.