User-State Labeling Procedures For The Multimodal Data Collection Of SmartKom

Silke Steininger[†], Florian Schiel[*], Angelika Glesner[†]

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

Ludwig-Maximilians-Universität, Schellingstr.3, 80799 Munich, Germany
{kstein, schiel, angel}@phonetik.uni-muenchen.de

Abstract

This contribution deals with the user-state labeling procedures of a multimodal data corpus that is created in 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 allow input in the form of natural speech but also in the form of gestures. Additionally, facial expressions are analyzed.

For the training of recognizers and the exploration of how users interact with the system, data is collected. The data comprises video and audio recordings from which the speech is transliterated and gestures and user-states are labeled.

This paper gives an in depth description of the different annotation procedures for user-states. Some preliminary results will be presented, particularly a description of the homogeneity of the different user-states and their most important features.

1. Introduction

Do users show emotions if they interact with a rather intelligent multimodal dialogue system? And if they do, how do the "emotions" look like? Are there any features that can be exploited for the automatic detection of the emotions? These are the main questions we want to answer with the collection and labeling of emotional data in SmartKom.

To answer these questions, data has to be collected, labeled and analyzed. This contribution deals with the second step, the labeling.

The labeling of emotions or user-states[1] 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. Why did we not use an already existing technique? One reason was: Until now, there is no settled approach to describe emotion. There is even disagreement with respect to the term "emotion" (Cowie, 2000). Another reason was that we were interested in the emotions or user states of people that interacted with a computer system. It can be assumed that the behavior changes profoundly if a human does not interact with another human, but with a machine (Jönsson & Dahlbäck, 1988). His or her emotions will change accordingly and look very differently. Therefore, 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 rather used a simplified, pragmatic 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[2] definition: Not the user-state per se is coded, but the impression the communicated emotion or state generates. The approach is in some ways similar to the technique used by Cowie (1999), where it was tried to define a "basic emotion vocabulary" that was chosen by naive subjects. In Steininger et al. (2002) we already discussed our functional approach with regard to gestures[3]. Apart from the practical reasons it has some theoretical advantages:

- A formal system incorporates assumptions about the connection between morphological shape (of parts of the face) and content (which emotion). We don't want to include these assumptions in the coding step because
- most studies of emotions concentrate on full-blown emotions or emotions played by actors. We assume that these emotions have a different appearance as the moods and subtle emotions that show up during a human-machine dialogue.

"Most pre-existing databases consisted of examples representing a few archetypical states. The rationale behind that approach is rarely spelled out, but the only obvious way to justify it is to postulate that the whole space of emotional signs can be reconstructed from information about a few cardinal types." (Douglas-Cowie, 2000). Douglas-Cowie calls this the "benign interpolation hypothesis". Like her, we don't take this hypothesis for granted, but want to find out how ecological valid user-states look like and in which different ways a certain state may be signaled. For a further discussion of this question and a detailed description of the development of the user-state procedure please refer to Steininger et al. (2002b).

In this contribution we give a detailed description of the label categories, along with major problems and some results.

[1] 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.

[2] "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.

[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. The development and structure of the gesture labeling is described in detail in Steininger, Lindemann & Paetzold (2002a).

2. The SmartKom Project

The goal of the SmartKom project is the development of an intelligent computer-user interface that allows almost natural communication between human and machine. The system does not only allow input in the form of natural speech but also in the form of gestures. Additionally the emotional state of the user is analyzed via his/her facial expression and prosody of speech. The output of the system comprises a graphic user interface and synthesized language. The graphic output is realized as a computer screen that is projected onto a graph tablet.

To explore how users interact with a machine, 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. The different functionalities of the system are developed by different partners of the project. The Institute of Phonetics and Speech Communication in Munich is responsible for the collection and annotation of the multimodal data and the evaluation of the system.

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 the labeling the video of the front camera is used (see figure 1)[4].

3. Coding Conventions for User-States

The labeling process comprises three separate procedures: Holistic labeling of the data, labeling without audio/facial expression labeling and prosodic annotation.

The labeling procedure is work in progress. The description of the categories, along with some formal criteria to help differentiate categories that can be easily mixed 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 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. Segment 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.

3.1 Holistic labeling of the video

The first step during user-state labeling is the so called "holistic labeling". A labeler watches the video of a session and marks each change in the state of the user. The segments that are found in this way are then assigned one of seven labels:

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

The allocation of the label is done with regard to the overall, subjective impression - not only the facial expression is taken into account but also the quality of the voice, the choice of words and the context. However, the sole usage of anger-implicating words that are uttered without any emotional expression are *not* taken as indicator for anger.

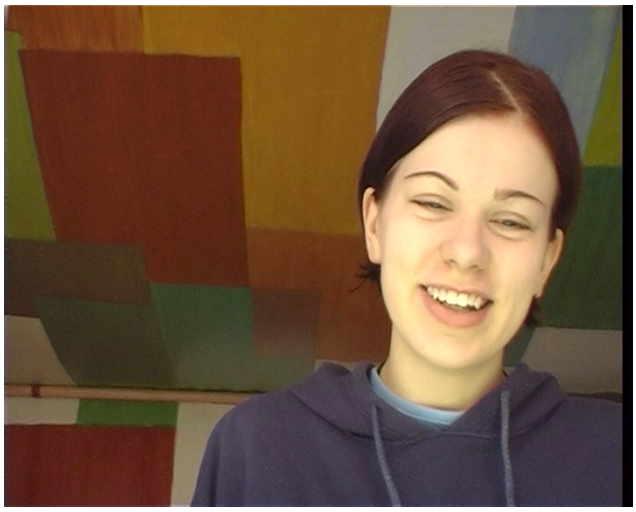

Figure 1: Example of the front view that is used for the holistic and the facial expression labeling. The picture was taken from an episode that was labeled as "joy/success" in the holistic labeling step.

Additionally the label is given a rating with regard to the intensity of the user-state (weak or strong).

Sequences during which the face is partly occluded by the hand/s of the subject are marked, as well as sequences during which the face can be seen only partly in the camera picture.

After the first labeling two correction steps follow, each done by a different, more senior labeler to find mistakes in form and content.

[4] For more information on the SmartKom project see Schiel, Steininger, & Türk (2002a) at this conference. 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. (2002b).

3.1.1 Categories

Joy/gratification (being successful): This label is given if the labeler has the impression that the user is in a positive mood, enjoys himself, is visibly content, amused or something similar. The emotion can be seen in the facial expression and/or heard in the voice. Other context information can be taken into account. However, without a smile or emotional voice, context information is not enough to warrant the label.

Amusement that (obviously or probably) stems from derision, mockery or something similar is still labeled as joy. We made this decision because sarcasm, derision etc. are exceedingly hard to detect reliably.

Formal criteria that can help: The user laughs or smiles. The corners of the mouth are curved upward. Eyes are often open. Eyebrows can be curved upwards. Teeth can be visible. The voice is often higher and/or louder. Audible laughing, friendly voice.

Problems: Most cases can be judged easily, but very faint smiles can be problematic: As a rule, smiles that look almost like neutral (are hardly detectable by a communication partner) are sorted into "neutral". If it is not clear if the user is content or if sarcasm can be suspected: The label joy is given anyhow, *as long as the user smiles or laughs*.

Homogeneity: The labeled episodes are relatively homogenous. In almost all cases the corners of the mouth are curved upward, he laughs or smiles. In the strong cases users have an open mouth and the teeth can be seen, but this is not a rule.

Anger/irritation:

This label is given if the labeler has the impression that the user is in a negative mood, is visibly not content, is irritated, annoyed, exasperated, angry, disappointed or something similar. The emotion can be seen in the facial expression and/or heard in the voice. Other context information (curses, offenses, corresponding off-talk) can be taken into account. However, without an angry facial expression or voice, context information is not enough to warrant the label.

Most cases of this category are weak: Full blown anger almost never shows up. In many cases users show their anger rather "politely" - it is obvious he or she is angry, but without strong changes in the facial expression or voice.

Formal criteria that can help: The eyebrows can be knitted, the lips can be pressed together, the user sometimes frowns and/or closes his eyes. He sometimes sighs, speaks more slowly, loudly or articulates overly clear, pauses between words. Sometimes: Deeper voice. Other indicators that can show up: Shaking of the head, moving backwards, commands that are given curtly or in a reprimanding tone.

Problems: Weak cases (which are frequent) can be problematic. Sometimes the anger can only be interpreted from the verbalization. However, an uttered curse alone is not enough to warrant the label! Voice or face have to mirror the emotion. Furrowed brows can be mixed with pondering/reflecting where they show up, too. Here, context information can help. Furrowed brows together with a forward movement are probably "pondering/ reflecting" not "anger/irritation".

Homogeneity: The labeled episodes are very inhomogenous. Almost everyone shows his or her irritation in a different way. Relatively consistent is only a loud voice.

Figure 2: Example of an episode that was labeled as "anger/irritation" in the holistic labeling step.

Helplessness:

This label is given if the labeler has the impression that the user is helpless, confused or interrogative. The strong cases comprise helplessness and distress: The user does not know what to do and/or has no idea how to go on. The weak cases comprise episodes in the dialogue where the user is puzzled, a bit confused, wants to know something, has no concrete plan how to go on. Uttering a question is not enough for the label, there has to be the impression of at least a bit confusion.

The emotion can be seen in the facial expression and/or heard in the voice. Other context information can be taken into account. However, without a corresponding facial expression or voice, context information is not enough to warrant the label.

Formal criteria that can help: The eyebrows are curved upwards (symmetrically or asymmetrically), on the forehead can be seen horizontal lines. The eyes tend to be open. The mouth can be open. Hesitations are an indicator, as well as stuttering. Backward movements. Head shaking. Erratic gestures. Inquiring intonation in the voice.

Problems: "Helplessness" can be mixed with "pondering/reflecting". A good discrimination criterion is the question if the users seems to feel "in control" or "out of control"[5]. The first is an indicator for "pondering/reflecting", the last an indicator for "helplessness". Additionally, "helplessness" more often shows itself in the upper part of the face (eyebrows, forehead), "pondering/reflecting" more in the lower part (mouth).

Homogeneity: An upward movement of the eyebrows and horizontal lines of the forehead show up very consistently. A hesitant voice is a relatively consistent indicator too.

[5] The "control" dimension is often used in structural models of emotion, for an overview see Scherer (1999).

Pondering/reflecting:

This label is given if the labeler has the impression that the user is thinking hard. The task of the subject leads to the fact that for the most part of the session he watches the display concentrated, reads or searches the display. This behavior is *not* labeled as "pondering/reflecting". There have to be visible or audible indicators, like biting on the lips.

"Pondering/reflecting" episodes mostly show up during decision making - after new information was presented or during a phase of planning the way to proceed.

The emotion can be seen in the facial expression and/or heard in the voice. Other context information can be taken into account. However, without a corresponding facial expression or voice, context information is not enough to warrant the label.

Formal criteria that can help: The eyebrows are knitted, the user frowns (and indicators for "anger/irritation" are missing!), the user chews at his lips, the mouth is partly open, whetting of lips, looking to the ceiling. Inhaling and holding of breath, low muttering. Hesitations are an indicator, as well as stuttering. Foreward movements.

Problems: See "helplessness". Movements of the mouth can have different reasons than "pondering/reflecting", for example to arrange lipstick or to whet dry and/or itching lips. If "pondering/reflecting" is not the clear reason than the label "unidentifiable episode" is given for such cases.

Homogeneity: Chewing on the lip and other movements of the mouth show up very often. "Pondering/reflecting" is one of the most homogenous labels at least with respect to the facial expression.

Figure 3: Example of an episode that was labeled as "pondering/reflecting" in the holistic labeling step.

Surprise:

This label is given if the labeler has the impression that the user is surprised. There is a fast movement in the face (in most cases of the eyebrows) or an abrupt vocal reaction in reaction to an external stimulus.

The emotion can be seen in the facial expression and/or heard in the voice. Other context information can be taken into account. However, without a corresponding facial expression or voice, context information is not enough to warrant the label.

Formal criteria that can help: The eyebrows make a fast upward movement. The eyes and/or mouth are opened. The head is sometimes moved backwards. The voice is sometimes louder and/or higher. Exclamations are uttered.

Problems: Almost none, because most episodes are very obvious. A fast movement or a sudden change are the most important indicators. If they are missing, the episode is probably not surprise.

Homogeneity: The fast eyebrow movement shows up very consistently. Vocal indicators are not as homogenous.

Neutral:

This label is given if no emotional or cognitive state can be detected by the labeler in the face or the voice. It is also given if an emotional or cognitive state is so faint that the assignment of a label seems inappropriate.

Formal criteria that can help: Relaxed face. Calm voice, no distinctive prosodic features.

Problems: Sometimes it is difficult to judge if an episode with a faint state should be sorted into "neutral" or into the respective category. Most frequently this is the case for "joy/success" or "pondering/reflecting". See also "joy/success".

Unidentifiable episodes:

This label is given if the user is neither neutral, nor any of the other labels can be assigned to the episode.

Three cases can be discriminated:

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".

4.4 Labeling of the video without audio information

This step is done mainly to get training material for a recognizer. A second labeler group (that was not involved in the holistic labeling) watches the same video without the audio information. To speed up the process the labeler is informed about the segment borders and the occurrence of "neutral"segments which he or she can ignore. For the other segments he or she assigns new labels from the facial

impression.

4.4.1 Categories

Joy/gratification (being successful): This label is given if the facial expression of the user imparts one of the following impressions to the labeler: A a positive mood, user enjoys himself, is visibly content, amused or something similar.

Formal criteria that can help: The user laughs or smiles. The corners of the mouth are curved upward. Eyes are often open. Eyebrows can be curved upwards. Teeth can be visible.

Anger/irritation:

The facial expression gives the impression of a negative mood, being not content, irritation, being annoyed, exasperation, anger, disappointment or something similar.

Formal criteria that can help: The eyebrows can be knitted, the lips can be pressed together, the user sometimes frowns and/or closes his eyes. He sometimes sighs. Other indicators that can show up: Shaking of the head, moving backwards.

Helplessness:

The facial expression gives the impression of helplessness or an interrogative state.

Formal criteria that can help: The eyebrows are curved upwards (symmetrically or asymmetrically), on the forehead can be seen horizontal lines. The eyes tend to be open. The mouth can be open. Backward movements. Head shaking. Erratic gestures.

Pondering/reflecting:

The facial expression gives the impression that the user is thinking hard. There have to be visible indicators, it is not enough if the user watches the display concentrated or searches the display with his eyes.

Formal criteria that can help: The eyebrows are knitted, the user frowns (and indicators for "anger/irritation" are missing!), the user chews at his lips, the mouth is partly open, whetting of lips, looking to the ceiling. Inhaling and holding of breath, moving lips. Foreward movements.

Surprise:

The facial expression gives the impression that the user is surprised. There is a fast movement in the face (in most cases of the eyebrows).

Formal criteria that can help: The eyebrows make a fast upward movement. The eyes and/or mouth are opened. The head is sometimes moved backwards.

Neutral:

Labeled neutral episodes are ignored in this labeling step. However, it can be necessary to change an episode into neutral in this step that was assigned one of the other labels in the holistic step. This is the case if without the audio information the indicators for a certain category are missing (mostly the voice).

Unidentifiable episodes:

As in the holistic labeling.

4.5 Prosodic annotation of the audio stream with formal criterions

This step captures the information that is contained in the voice: A labeler listens to the audio file and marks if any of the labels below occur, together with their time of occurrence.

The labels are adapted from Fischer (1999), who used the same conventions in the Verbmobil project. For more information on the usage of prosodic features as indicators of emotional speech please refer to Batliner et al. (2000).

A word can have more than one prosodic marker. The labels are given with respect to the "normal" speaking habits of a subject: If someone habitually articulates clearly or speaks emphatically only the very clearly articulated words or the words with strong emphasis are marked.

4.5.1 Categories:

PAUSE_PHRASE: Irregular pause on a phrasal level/between units of meaning. Pauses between sentences or between main clause and subordinate clauses are not meant, except if the pause is very long.

PAUSE_WORD: Irregular pauses between words.

PAUSE_SYLL: Irregular pauses between syllables of a word.

LENGTH_SYLL: Lengthening of a syllable. Can occur at any syllable of a word.

EMPHASIS: Emphatic accentuation/strong emphasis on a word or syllable.

STRONG_EMPH: Very strong emphatic accentuation/very strong emphasis on a word or syllable.

CLEAR_ART: Clearly articulated speech. Clear articulation can be seen as a weak version of hyperarticulated speech. The speaker uses (tries to use) less colloquial speech and less dialect, the speech emulates that of newsreader.

HYPER_ART: Hyper-articulated speech. Very strong increase of the clear articulation.

LAUGHTER: Speech overlapped by laughter or sighing.

5. Frequency of the Labels

The following tables show the frequency of the different categories in the holistic and the facial expression labeling step. Please note that the episodes vary greatly with respect to the duration, therefore we gave the percentage of the number of the episode (third column) and the percentage of the duration (fourth column). "Neutral" episodes for example are very long on average, whereas "surprise" episodes are relatively short.

The number of labeled episodes in the facial expression step is less than in the holistic labeling step. This is the case because in the facial expression labeling the episodes that were coded earlier because of the voice or other information are changed to neutral.

User-State	N	% Number	% Duration
Neutral	1253	43,7%	71,6%
Pondering/Reflecting	689	24,0%	14,0%
Joy/Success	370	12,9%	7,2%
Anger/Irritation	205	7,1%	2,8%
Helplessness	182	6,3%	3,3%
Surprise	99	3,4%	0,6%
Unidentifiable Episodes	72	2,5%	0,6%

Table 1: Frequency of the different labels within holistic labeling. Number of labeled sessions: 97. Number of labeled episodes: 2870.

User-State	N	% Number	% Duration
Neutral	830	45,3%	74,3%
Pondering/Reflecting	447	24,4%	14,3%
Joy/Success	194	10,6%	4,8%
Anger/Irritation	75	4,1%	1,3%
Helplessness	194	10,6%	4,4%
Surprise	47	2,6%	0,4%
Unidentifiable Episodes	46	2,5%	0,6%

Table 2: Frequency of the different labels within facial expression labeling. Number of labeled sessions: 97. Number of labeled episodes: 1833.

The ratio of "pondering/reflecting" is similar in both labeling steps. "Joy/gratification" drops a bit in the facial expression step. "Anger/irritation" and "helplessness" show the biggest differences. "Anger/irritation" are less and "helplessness" episodes are more frequent in the facial expression step than in the holistic step. We observed that many "anger/irritation" sequences are identified because of vocal indicators (for example loud voice or reprimanding tone), this could explain why their number drops during facial expression labeling. Perhaps some of the "anger/irritation" sequences look like "helplessness" without context information. This is only an ad hoc explanation and has to be analyzed further to be sure.

By far the most frequent of the categories is "pondering/reflecting" which is probably the case because of the context of searching for information with the help of a computer assistant. "Joy/Success" is frequent also, which is not surprising: The task was judged as fun and interesting by the subjects. The low frequency of "anger/irritation" is unfortunate, because it would be desirably to have training material to train recognizers to detect this category. To get subjects (naturally) angry, however, is very difficult - if they agree to participate in a test, they tend to be in a friendly and cooperative mood.

Taking together these first results with respect to the frequency of user-states in the context of a human-machine dialogue and the experience with respect to the possible (formal) indicators for the user-states, a few assumptions can be made:

- "Anger/irritation" (or similar user-states) data (in a natural setting) for the training of recognizers is hard to come by. A solution could be data collections that concentrate on anger and try hard to evoke it. However, the episodes that we collected show a great variation in form. We fear that this variation will get only slightly better with more data. The goal to detect "anger" in a human-machine dialogue will remain a difficult one.

- "Helplessness" seems to have relatively consistent indicators. It is not very frequent, but in combination with "anger/irritation" (with which it perhaps shares similarities) it is perhaps a better candidate for automatic recognition than anger alone.

- "Pondering/reflecting" and "joy/success" seem to be more promising candidates for automatic recognition. They can be collected easily, because they seem less difficult to evoke than for example anger. Additionally, there are some formal criteria that show up consistently (corners of the mouth, chewing on the lip).

- "Surprise" seems to have relatively consistent indicators, too. However, it's frequency is very low and therefore it's probably not worthwhile to try to detect it.

4. Summary

The challenge we faced at the start of the project was to develop labeling systems for multimodal data that were tailored to the task, fast and suited for the training of recognizers. We decided not to label on the morphological level but to define the labels with regard to the intent of the user respectively his obvious communication goal. Since we do not yet know which features of the gestures or the user-states carry the vital information (and which are perhaps automatically recognizable), we decided against a (pure) formal system.

We hope that with our labels we catch the most interesting (non verbal) episodes of the dialogue with all the relevant information. Finding indicators usable for automatic recognition through analyzing these episodes will be the next challenge.

Acknowledgments

This research is being supported by the German Federal Ministry of Education and Research, grant no. 01 IL 905. We give our thanks to the SmartKom group of the Institute of Phonetics in Munich that provided the Wizard-of-Oz data. Many thanks to Alexander Borkowski for the help with analyzing the data and Bernd Lindemann for finding the examples.

5. 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.

Douglas-Cowie, E., Cowie, R., & Schröder, M., 2000. A New Emotion Database: Considerations, Sources And Scope. In R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.): *Proc. of the ISCA Workshop on Speech And Emotion*. Belfast: Textflow.

Cowie, R., 1999. What a neural net needs to know about emotion words. In N. Mastorakis (ed.): *Computational Intelligence and Applications. Word Scientific Engineering Society*, pp. 109-114.

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

Jönsson, A., & Dahlbäck, N., 1988. Talking to a computer is not like talking to your best friend. *Proc. of the First Scandinavian Conference on Artificial Intelligence*, Tromso, Norway, p. 297- 307.

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

Scherer, K. R. (1999). Appraisal theory. In T. Dalgleish & M. Power (Eds.): *Handbook of Cognition and Emotion*. New York: John Wiley.

Schiel, F., Steiniger, S., & Türk, U., 2002a. The SmartKom Multi-modal Corpus at BAS. To appear in the *Proc. of the 3rd Int. conf. on Language Resources and Evaluation*, Las Palmas, Spain.

Schiel, F., Steininger, S., Beringer, N., Türk, U., & Rabold, S., 2002b. 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 Multimodal 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., Rabold, S., Dioubina, O., & Schiel, F., 2002b. Development of the User-State Conventions for the Multimodal Corpus in SmartKom. To appear in the *Proc. of the 3rd Int. conf. on Language Resources and Evaluation, Workshop On Multimodal Resources and Multimodal Systems Evaluation*. Las Palmas, Spain.

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