Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus

Jörg Frommer (2), Bernd Michaelis (3), Dietmar Rösner (1), Andreas Wendemuth (3), Rafael Friesen (1), Matthias Haase (2), Manuela Kunze (1), Rico Andrich (1), Julia Lange (2), Axel Panning (3), Ingo Siegert (3)

1: Otto-von-Guericke Universität, Institut für Wissens- und Sprachverarbeitung (IWS)
Postfach 4120, D-39016 Magdeburg
{roesner, friesen, makunze, andrich}@ovgu.de
2: Otto-von-Guericke-Universität, Universitätsklinik für Psychosomatische Medizin und Psychotherapie
Leipziger Straße 44, D-39120 Magdeburg
{joerg.frommer, matthias.haase, julia.lange}@med.ovgu.de
3: Otto-von-Guericke Universität, Institut für Elektronik, Signalverarbeitung und Kommunikationstechnik (IESK)
Postfach 4120, D-39016 Magdeburg
{bernd.michaelis, andreas.wendemuth, axel.panning, ingo.siegert}@ovgu.de

Abstract

The LAST MINUTE corpus comprises multimodal recordings (e.g. video, audio, transcripts) from WOZ interactions in a mundane planning task (Rösner et al., 2011). It is one of the largest corpora with naturalistic data currently available. In this paper we report about first results from attempts to automatically and manually analyze the different modes with respect to emotions and affects exhibited by the subjects. We describe and discuss difficulties encountered due to the strong contrast between the naturalistic recordings and traditional databases with acted emotions.

Keywords: User-Companion-Interaction, Affect, Multimodal

1. Introduction

1.1. Design and creation of the LAST MINUTE corpus

The LAST MINUTE corpus contains multimodal recordings from a WOZ experiment that allows to investigate how users interact with a companion system in a mundane situation with the need for planning, re-planning and strategy change. The design of this experiment is described in (Rösner et al., 2011). Further specifications and the hardware configuration will be presented in *LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions* submitted to LREC 2012 as well. This work describes the ongoing analyses of the corpus.

1.2. Multimodality

Unimodal recordings or systems that get their information about the user from just one modality have the disadvantage of missing possibly significant signals of users' affect when it is not or can not be transmitted in that modality. So analyzing multiple modalities provides a more continuous stream of observable information but bears the need for fusion of different information sources.

1.2.1. Audio

The recorded audio contains speech and nonlinguistic content which are transcribed. Emotions in speech content can be expressed in many ways. Thus classification can be done with different results, e.g. discrete classes for speech content and continuous arousal values for prosody.

1.2.2. Video

The video recordings with many channels can be analyzed for gestures and mimics which provide e.g. useful information about the emotions' valence.

1.2.3. Biopsychological data

The skin reductance, heartbeat and respiration were also recorded and can be analyzed for affective signals which we expect to be expressed uncontrolled. Although these modalities seem to be reliable, it is hard to classify changes in the signals as emotions correctly.

1.2.4. Questionnaires

Linguistic, prosodic and facial expression of emotion depends very much on individual factors such as personality (Cohn et al., 2002). For this reason, various psychological constructs were measured using psychometric questionnaires (Lienert, 1961).

1.2.5. Interviews

After the WOZ experiment 73 of the participants underwent a semi-structured interview to determine the subjective experience of the experiment. In detail, these interviews focus on

- occurred user emotions,
- intentional ascriptions towards the system to explain and predict the system's behaviour (like characteristics, aims, emotions etc.; (Dennett, 1987)),
- the speech based interaction,
- the intervention (if given),

- the role of technical systems in autobiography and
- the general evaluation of the system.

The interviews, in which the participants can reflect on the interaction in a free and nonrestricted way, took about 30 to 160 minutes (in sum 93 hours) and are audio recorded.

2. Emotion detection in the corpus

The experiment provides 4 major events (Baseline, Listing, Weight limit, Waiuku) (Rösner et al., 2011), where we expect certain emotions. Starting from these events we define an observation time window of 110 seconds. For three of these timeslots there should be arousal (for now, we do not distinguish between positive or negative valence, however negative is rather expected). The remaining slot is the baseline and considered to represent the subject in a neutral state.

One thing should be in mind here, for databases with acted emotions the subjects generally do what is expected. In the performed WOZ experiments it is the plan to bring the subjects in arousal, but it is not guaranteed that they will experience the expected emotions. Therefore the validity of the ground truth is rather uncertain, which also spoils down classifier results. Summed up, in the step from databases with acted emotions to a real affected corpus we face three challenges at the same time: (1) more difficult measurement, (2) less expressional and less simultaneous facial actions and (3) uncertain ground truth.

2.1. Emotion expression in User-Companion Interaction (UCI)

In databases with acted emotion expressions very often the classical emotions of Ekman (Cohn et al., 2002) are dominant and available in high intensity. In naturalistic records – like in the LAST MINUTE corpus – a completely different set of emotions and affective states is relevant. In addition the intensity is typically lower and there may be long phases of neutral states. This provides a challenge for data analysis. A simple transfer of methods developed for acted material is not adequate and sufficient.

What emotions and affects are most relevant in user companion interaction (UCI) is part of the questions that are intensively investigated with the LAST MINUTE corpus. A still incomplete list of candidates includes e.g. surprise, boredom, irritation, frustration, helplessness, pride, shame, superiority, astonishment, feeling under time pressure, ...

2.2. Prosody

For the experiment four events where defined, that are investigated further. The observation length than was set to 110 seconds. In a first round a manual labelling process was started, to identify observations within the different speakers. We concentrated on following vocal cues, as they were stated as the most informative emotional cues (see (Hietanen et al., 1998; Owren and Bachorowski, 2007)): (a) breathing (b) pauses (c) tongue-clicks (d) interjections (e) harrumphes (f) laughter (g) off-talk (h) energy (i) loudness (j) pitch (k) MFCCs. As a first measure we calculated the relation of speaking-time for the system and the user (see Table 1) to get a feeling about expected amount of speech signals. Because the system was designed to allow no barge-in and furthermore nearly nobody tries this, the overall length of utterances is oversee-able and thus acoustic cues appearing in a cumulated manner. For further investigation we only concentrate on the Baseline and the Challenge event, as only for them a sufficient amount of speech-data is available.

Table 1: Average wizard and user speaking-time in relation to the length of specific event (110 seconds).

Event	Wizard	User
Baseline	32.6% $\pm$ 6.3	14.4% $\pm$ 3.9
Listing	55.9% $\pm$ 16.8	7.3% $\pm$ 3.1
Challenge	58.6% $\pm$ 8.6	9.7% $\pm$ 4.9
Waiuku	67.7% $\pm$ 19.9	3.9% $\pm$ 2.5

In a first screening of voice cues, it is noticeable, that big individual differences exist between speakers. We consider off-talk and pauses as one of the most informative events occurring in this sub-part of the investigated material. But it is hard to generate automatic classifiers because this cues are strongly individual.

Besides this acoustic analysis, we also developed a classifier based on previously defined classes, baseline and challenge. We picked the same speakers that were chosen for facial expression analysis (see 2.3.), but had to skip one speaker (20110117bsk) because of missing audio. We extracted all utterances from the baseline and challenge interval for all selected speakers, the length of a single utterance varies between 0.5s and 2s. A detailed summary can be found in Table 2. We used HTK (Young et al., 2006)

Table 2: Detailed summary of number of utterances for Baseline and Challenge for the selected speakers.

Speaker	# Baseline	# Challenge	Age	Gender
20101117auk	13	11	24	f
20110110bhg	10	9	26	m
20110112bkw	9	9	23	m
20110124bsa	11	6	25	m
20110126bck	15	7	27	m
20110209bbh	15	6	66	m
Total	73	48	31.5 $\pm$ 16,8	

for training and testing our models and extracted different acoustic features automatically. As features we used various combinations of 12 MFCCs (`MFCC`), their Deltas (`_D`) and Accelerations (`_A`), Energy (`_E`) and the Zero-mean coefficent (`_Z`), in total we used a vector of 36 or 39 features. We used a three-state left-right Hidden Markow Model with 81 gaussian mixture distributions. Additionally we compared two different methods for Training, leave-one-speaker-out (`LOSO`) and ten-fold-cross-validation with 90% of the merged data for training and 10 % of the merged data for test (`CV`). An overview of used feature and training combinations with overall recognition results can be found in Table 3. Comparing the results in Table 3, it is obvious, that the leave-one-speaker-out strategy fails in this task. But using a ten-fold-cross-validation the models are able to dis-

Table 3: Detailled Summary of number of utterances for Baseline and Challenge for selected speakers

Combinations of features	LOSO	CV
MFCC_D_A	47.10%	83.48%
MFCC_D_A_Z	57,69%	83.91%
MFCC_E_D_A	52.33%	85.65%
MFCC_E_D_A_Z	51.41%	83.45%

tinguish both classes. This leads to the assumption, that the acoustics of the selected speakers varies too much. This is supported by the informations about age and gender of selected speakers, one elder person, one female. So we either need individual models that could be derived applying adaption techniques, or have to use speaker-group dependent models for age and gender to improve the recognition results. Further investigations are needed to prove this assumption.

2.3. Facial expressions

In a first step we applied a SVM trained classifier which was trained on a default acted emotion dataset, determining 4 of the 6 standard Ekman emotions (Anger, Happiness, Disgust, Surprise and Neutral). We used following input:

- Geometrical features: Distances between landmark-points as shown in (Niese et al., 2010). In this method the general approach is to detect certain landmark points on the face which are prominent (Mouth corners, Eyebrows, Eye-Corners).
- An estimation of the absolute head movement (direction independent) and
- the eye blink rate.

For the analyses we chose one image of each subject to represent the neutral state. Calculating the distance between a feature vector and the neutral state vector, we can derive facial actions. Additionally we gather hand movement using the 3D camera images and analyze it for the most simple feature "hand touches face" (cf. Fig. 2). In the context of the WOZ experiment, we expected a change from calm to aroused periods ("Baseline" vs. "Weight limit"/"Waiuku"). For impartial analysis we performed a manual simplified FACS coding, to avoid any issues in measurement of raw features. We focused here on two predefined timeslots (Weight Limit and Waiuku), where emotional reactions were expected. Our FACS coding was done by a certified FACS coder. The coding is guided by the general FACS rules but does not scale the different levels A-E because of time effort. Thus it is a simple binary coding where intensity levels A-E are summed to value 1 while value 0 indicates no action of the Action Unit. Since manual FACS coding is a very time consuming process at the current state we labeled the three timeslots for 6 persons only. It is difficult to find obvious rules or differences in the general pattern (Figure 1).

We found that very prototypical, strongly expressed patterns as displayed in acted databases can be found less in real affected data. Expected emotion classes had been

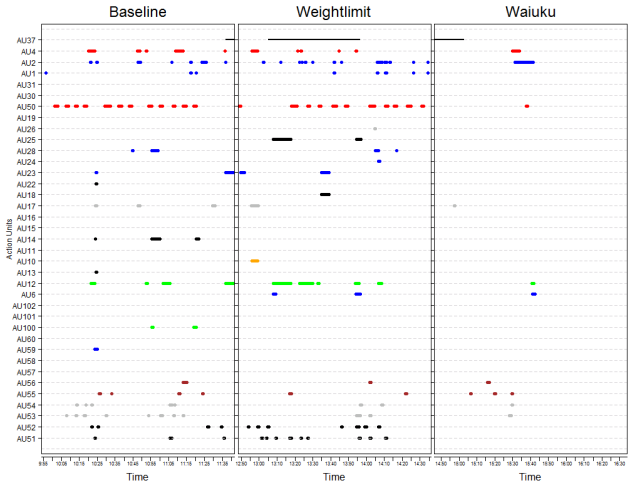

Figure 1: Patterns of facial actions at different times

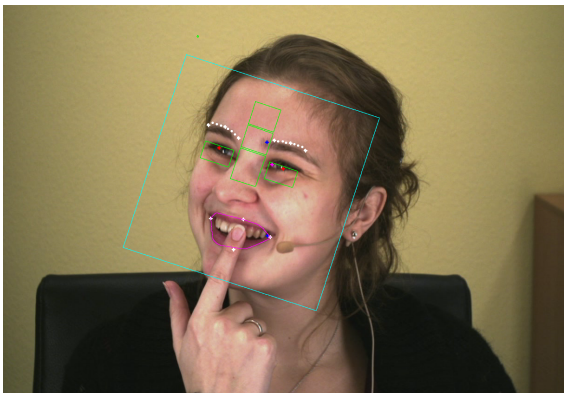

Figure 2: Example of subject touching her face

"anger","surprise" or "happiness". The only findings were some genuine happy smiles for very short periods of frames. Ekman and Friesen provide a certain pattern for each emotion class, but they did not match in full assembly for any emotion except for happiness (Figure 1). They rather seem to be presented only partially or completely replaced by temporal patterns which spread over longer time. Real data is influenced by a lot of measurement noise, occlusions (hand covering mouth, large rotations, hair covering eyebrows and eyes, unfortenately prominent glasses etc.) which almost never occur in ideal databases. Thus classifiers, which classify acted emotions with more than 90% accuracy, are not applicable on this data.

Nevertheless, we could achieve motivating results using the geometrical features when training an optimal filter approach on "Baseline" and "Weight Limit" events with leave one out strategy. Even with lots of noise in the measurement classification results around 70% for the "Baseline" event and around 60% for the "Weight Limit" event which are substantial better than chance, but far from the 90% rates we know from acted databases. However, for the next steps we will improve this measurement, refine the feature choices (absorb features which are obviously broken like mouth parameters in case of hand is partially covering the mouth) and the choice of more sophisticated classifiers instead of simple linear classifiers. We expect better results

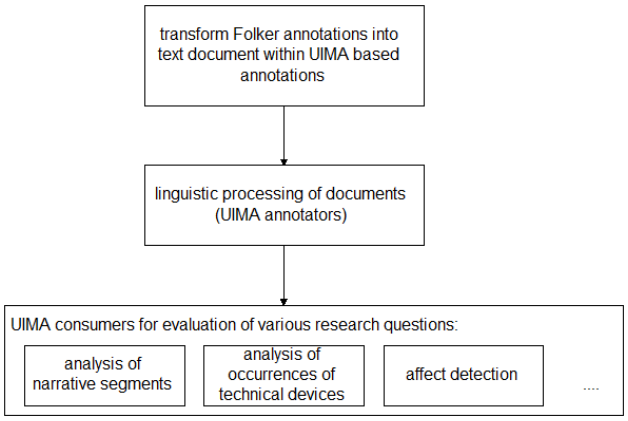

Figure 3: Processing pipeline

here which can reach acceptable classification rates.
We can derive so far, that the general amount of action during the arousal window was increased. For validation we also labeled the second arousal window ("Waiuku") of 2 other subjects. To our surprise the result was a contradiction to the prior finding. The action was dramatically decreased even compared to the baseline. The current hypothesis is here that the kind of arousal was slightly different. While the subject in the first arousal ("Weight Limit") event was actively managing the arrangement of baggage the subject was in a passive state (only listening to the WOZ-Speaker) during the "Waiuku" event.

2.4. Explicit linguistic content

To get emotion information on a higher level the audio records were transcribed following the GAT2 standard using Folker (Schmidt and Schütte, 2010).

2.4.1. Processing pipeline

For processing of the Folker based transcripts we used the UIMA framework (`uima.apache.org`). This framework contains various Java interfaces for each processing step (reading a document, analysing it, and printing results). The modular structure of the framework makes it easy to use tools (e.g. analysis engines) in different applications, while the concept of annotations defined in UIMA makes it possible to exchange results between different applications. Fig. 3 gives a small overview about the processing pipeline. The first step is to transform Folker format into UIMA based annotations. After this, we initiate a number of linguistic and dialogue based analyses (see Fig. 4). For these analyses, we used internal and external tools and resources. For example, we integrated resources of GermaNet (`http://www.sfs.uni-tuebingen.de/lsd/`), LIWC (Wolf et al., 2008) and of the project Senti Wortschatz (SentiWS) (Remus et al., 2010). The results of these analyses were exported as XMI documents within UIMA based annotations. These data were exploited for specific research questions realized by specific UIMA consumers.
The WOZ experiments have sections (e.g. the packing and unpacking phases) where vocabulary (and syntax) is rather restricted. This contrasts with other sections (e.g. those starting with open questions that stimulate narratives) where content, vocabulary and syntax are quite unrestricted. In the following example we concentrate on these latter sections in the transcripts.

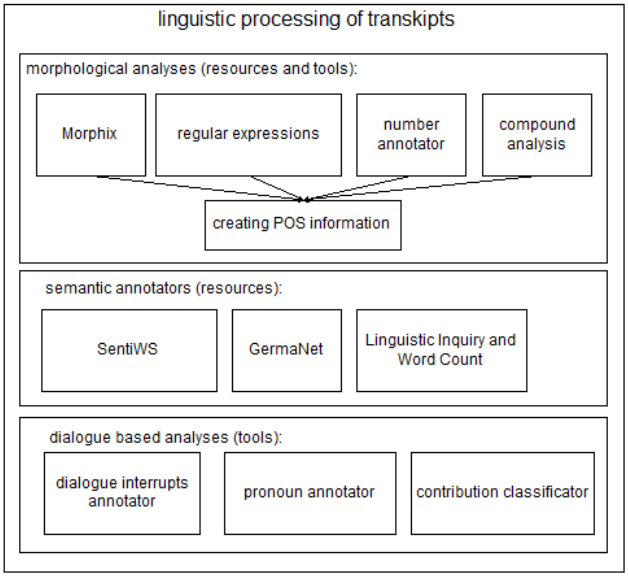

Figure 4: Tools and resources

Extraction of Holiday Plans In the final phase of the experiments users are stimulated to describe their plans for the holidays. Here the answers are analysed under the question which holiday activities can be found or inferred based on GermaNet entries. We selected specific GermaNet synsets which describe holiday activities and summarize these to a holiday activity category. For example, we defined the following categories *sport activities*, *chillout*, *excursions* and so on. Each of these categories was assigned to a specific GermaNet synset. Token in the corpus which are annotated with a GermaNet synset, which is a hyponym of the specific category synset, will be annotated with the holiday category. The results were used for a shallow classification of the answers.
For example, the user answer *ähm ne sightseeing tour und ansonsten entspannen (um 'n excursion and otherwise relax)* results in assigning the categories *excursions* and *relax*.
Only in 6 transcripts, we found no categories about holiday plans and in 5 transcripts we ignored negations in the answers. In the rest of the corpus (51 transcript), we got the correct categories. In some of these transcripts, we haven't annotate all kind of categories, because not all tokens were covered by the resources of GermaNet (e.g. GermaNet contains *spazierengehen* (to go for a walk) but it doesn't contains the noun *Winterspaziergang* (walk in Winter).

Affect Detection In the WOZ experiment half of the subjects (N=63) get an intervention with a stimulus to report on possible feelings that users might have experienced when the information about the target location of the trip was given so late. We analyse the users' answers with LIWC to derive the general tendency of the reply (i.e. positive, negative or neutral). The analysis of answers for this question (for the subgroup of 26 with transcripts already available at the time of writing) results in 11 users which describe neg-

ative feelings while 10 other users describe that they got no negative feelings. 5 answers could not be classified. Only one answer was classified incorrectly (negative instead of positive).
We used the categories of LIWC about affective and emotional processes. The token in the answers were annotated with their specific LIWC categories. We analysed for each answer the occurrences of positive and negative LIWC categories.
The following excerpt shows an example for a negative reply (for readability without GAT encoding):
unangenehme gefühle nich nein ich hätte eben nur anders eingepackt aber das wetter wär mir egal gewesen (unpleasant feelings not no I would have packed different things but i wouldn't have cared about the weather)
LIWC results for this answer: unangenehme[Affect; Negemo] nein[Negate] ich[Pronoun; I; Self] eben[Time] eingepackt[Past; Motion] aber[Cogmech; Discrep; Excl] das[Article] wetter[Tentat; Present; Money] mir[Pronoun; I; Self] gewesen[Past]

2.5. Fusion of multimodal data

Extracting the speakers arousal out of the audio-signal is easy, but getting the measure for valence is hard. On the other hand it is hard to get information about arousal out of mimics, but mapping simple FACS coding units to positive or negative emotions it is easy to get a valence value out of mimics. This leads to a model that combines the arousal from speech and the valence from mimics to a first two-dimensional emotion. The emotions' authenticity can next be confirmed by classification of the biosignals. It could also be classified finer if the speech content describes the emotion in more detail.

3. Discussion

3.1. Problems encountered

Due to the fact that the wizard has the largest part of the speech during the interaction and also does not consider and rarely allows comments, we find a quite small amount of speech data for the users in this corpus. But we could show, that even here acoustic features could suffice to solve the baseline/challenge classification task, when training models are speaker dependent. An analysis of paralinguistic signals could improve this classification.
Since the ground truth is far from being precisely defined the video analysis focuses to seperate only two classes for now. Here encouraging results of >60% accuracy are achieved. For more sophisticated information about ground truth this recognition rates should increase. In ongoing work the differences between subjects are taken into account. Different groups of users will be identified and the results will hopefully allow better recognition rates.

3.2. Future work

We will concentrate on identifying the best features from prosody, paralinguistic signals, facial expressions and gestures for emotion recognition. We will also focus combinations of features from different modalities that improve the prediction of negative dialog courses.

Acknowledgment

The presented study is performed in the framework of the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems" (SFB website) funded by the German Research Foundation (DFG). The responsibility for the content of this paper lies with the authors.

Availability

The LAST MINUTE corpus is available for research purposes upon written request from project A3 of SFB TRR 62 (heads: Prof. Frommer and Prof. Rösner). For the reviewers a sample from the corpus with anonymised data is available from the following URL `http://iws.cs.uni-magdeburg.de/a3/lrec2012/index.htm` with loginname `reviewer` and password `lrec2012`.

4. References

J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman. 2002. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings of the International Conference on Multimodal User Interfaces (ICMI 2002)*, pages 491–496.

D. C. Dennett. 1987. *The Intentional Stance*. The MIT Press, Cambridge.

J. K. Hietanen, V. Surakka, and I. Linnankoski. 1998. Facial electromyographic responses to vocal affect expressions. *Psychophysiology*, 35(05):530 – 536.

G. A. Lienert. 1961. *Testaufbau und Testanalyse*. Beltz, Weinheim.

R. Niese, A. Al-Hamadi, and B. Michaelis. 2010. Emotion recognition based on 2d-3d facial feature extraction from color image sequences. *Journal of Multimedia, in print*.

M. J. Owren and J. Bachorowski. 2007. Measuring emotion-related vocal acoustics. *Handbook of emotion elicitation and assessment*, pages 239 – 266.

R. Remus, U. Quasthoff, and G. Heyer. 2010. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pages 1168–1171.

D. Rösner, R. Friesen, M. Otto, J. Lange, M. Haase, and J. Frommer. 2011. Intentionality in interacting with companion systems – an empirical approach. In J. Jacko, editor, *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, volume 6763 of *Lecture Notes in Computer Science*, pages 593–602. Springer Berlin / Heidelberg. 10.1007/978-3-642-21616-9_67.

T. Schmidt and W. Schütte. 2010. Folker: An annotation tool for efficient transcription of natural, multi-party interaction. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Website of the Transregional Collaborative Research Centre SFB/TRR 62. http://www.sfb-trr-62.de/.

M. Wolf, A. B. Horn, M. R. Meh, S. Haug, J. W. Pennebraker, and H. Kordy. 2008. Computergestützte quantitative textanalyse. In *Diagnostica*, volume Vol. 54, Number 2/2008. Hogrefe, Göttingen.

S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. 2006. *The HTK book (for HTK Version 3.4)*. Number July 2000. Cambridge University Press, Cambridge, UK.