How Low Level Observations Can Help to Reveal the User's State in HCI

Stefan Scherer[1,2], Martin Schels[1], and Günther Palm[1]

[1] Institute of Neural Information Processing, Ulm University, Germany
[2] Speech Communication Lab, Trinity College Dublin, Ireland
stefan.scherer@gmail.com

Abstract. For next generation human computer interaction (HCI), it is crucial to assess the affective state of a user. However, this respective user state is – even for human annotators – only indirectly inferable using background information and the observation of the interaction's progression as well as the social signals produced by the interlocutors. In this paper, coincidences of directly observable patterns and different user states are examined in order to relate the former to the latter. This evaluation motivates a hierarchical label system, where labels of latent user states are supported by low level observations. The dynamic patterns of occurrences of various social signals may in an integration step infer the latent user's state. Thus, we expect to advance the understanding of the recognition of affective user states as compositions of lower level observations for automatic classifiers in HCI.

Keywords: human computer interaction, annotation schemes, affective state, multiparty dialog.

1 Introduction

Current human machine interaction only takes place on a crude explicit question-answer level, whereas human human interaction is multifaceted, consisting of manifold interactive feedback loops between interlocutors, comprising social components, moods, feelings, personal goals, nonverbal and paralinguistic conversation channels and the like [1,5]. In order to close this gap it is crucial for a machine to perceive and understand the user's current interaction and affective state. Most of the research aiming towards recognizing the user's state focuses on the recognition of emotions [2], often the so called big six introduced by [3] and [4]. However, it is not entirely clear what is meant by the word emotion nor what types of emotion or states are relevant for human machine interaction. Further, as stated in [9], traditional theory on emotion includes extremes experienced throughout human lives that never occur in human computer interaction.

In this work, we elaborate on the set of labels, describing user dispositions in HCI as introduced in [13]. These labels comprise categories of different complexity: several are directly inferable (e.g. a subject is laughing), while others are only accessible, when provided with context of the interaction – even for

S. D´Mello et al. (Eds.): ACII 2011, Part II, LNCS 6975, pp. 81–90, 2011.

human annotators. In this context, the aim of this paper is to connect simpler observations to a higher level subject state by evaluating and revealing the pairwise coincidences of labels using t-tests. These coincidences and their dynamic patterns of occurrences in turn support the idea that social signals can serve as basic building blocks that help to infer the latent subjects' state within HCI.

The remainder of this paper is organized as follows: in Section 2, the data collection and the used annotation schemes are described. In Section 3, the interconnection between the labels are evaluated and the results are discussed. Finally, Section 4 concludes.

2 The PIT Corpus of German Multi-party Dialogs

The data collection used for the evaluation is is the "PIT corpus of German multi-party dialogs" [15,14], which is recorded using a WOZ approach. The scenario is a restaurant search, which is composed of three dialog participants: two human subjects (U1 and U2), discussing their choice of a restaurant, and one computer (S) assisting them. This supporting system acts as an independent dialog partner and only turns active when addressed by the main user U1. The system S itself acts as an independent dialog partner and becomes active as soon as the users start to speak about the specified domain.

In Figure 1, the utilized setup of the system and a typical scene of the interaction is shown. Each dialog involves two human participants, who interact with the system operated by the wizard. The system reacts to questions or gives hints about possible restrictions or search queries.

The acquired corpus consists of 36 dialogs with 72 participants, between 19 and 51 years of age (on average 24.4 years); 31 of them female. The shortest recorded dialog lasts 2:43 minutes, the longest lasted 18:24 minutes. For an exact distribution of dialogs and dialog duration please refer to [14].

One of the challenges, dealing with unscripted and naturalistic interactions, as available in the PIT corpus, is the lack of knowledge about the actual ground truth of the participant's affective states. In contrast to acted emotional data it is not possible to fully control the behavior. On the other hand, this lack of control provides naturalistic behavior of users while interacting with machines. The available labels, developed in this work, for the naturalistic interaction data are shown in Table 1. The annotations in this work are provided in different levels: subject state, talk style, events, focus of user, and dominant dialog role. These levels group similar categories used for the annotation: several of these labels are directly inferable from the data, e.g. laughter or the different talk styles. Others are more complex and require context of the conversation for the annotator, such as the subject state. Additionally, in all layers the annotations are temporal attributes and can be assigned with varying lengths and offsets. Using this annotation approach, 15 out of the previously mentioned 36 dialogs were annotated using the well known labeling tool ANVIL [6]. In order not to introduce any bias, the annotators had to annotate the subject state layers of the dialogs in first screenings before knowing about the objective layers. Each

Table 1. Extended list of label groups and organization in layers as introduced in [13]. The top and most abstract level is the subject state layer. Lower levels are more objective observations and comprise social signals (i.e. paralinguistic cues and nonverbal behavior). The rightmost column indicates a short version of the bullet points provided to the annotators.

Level	Label	Meaning
Subject state	Interested	listening (not active), showing interest, reading (silent/loud)
	Uninterested	distracted, uninterested, not paying attention
	Surprised reacting	surprised, facial expression, utterance of surprise
	Embarrassment	embarrassment, insecure, blushing, confused
	Impatient	commenting waiting, impatient movement
	Stressed	seeming stressed (work, appointments), hasty behavior
	negative accepting	may be compromising, disappointment
	positive accepting	pleased with outcome, neutral acceptance
	Disagreement	disagreeing with the outcome but not accepting yet
Talk style	Commanding	non-natural command style talk, imperative speech
	Off-Talk	non-related to topic or HCI
	Ironic	speaking ironically about something
	Explaining	pedagogical, arguing, giving facts
	Active Listening	nodding, back-channeling, non-verbal communication
	Question	posing a question go get information
	Thinking	loud thinking, pausing, hm... what shall we do?
Events	Laughs	loud laughters, silent ones, prominent smiles
	Silence	agreeing or disagreeing silences
	Exciting Moments	for the participants
	Topic Shifts	change of topic
	Waiting	waiting for a reply (mostly due to WOZ lags)
Focus of user U1	User U2	the focus lies on person B
	System	the focus lies on the system
	Others	the focus is something else
	Changing focus	there is a shift of focus (i.e. phase of head or eye movement)
Dominant dialog role	User U1/U2	one person is dominant (longer periods)
	Eq. Active	lively conversation, back and forth between participants
	Eq. Passive	slow and boring conversation

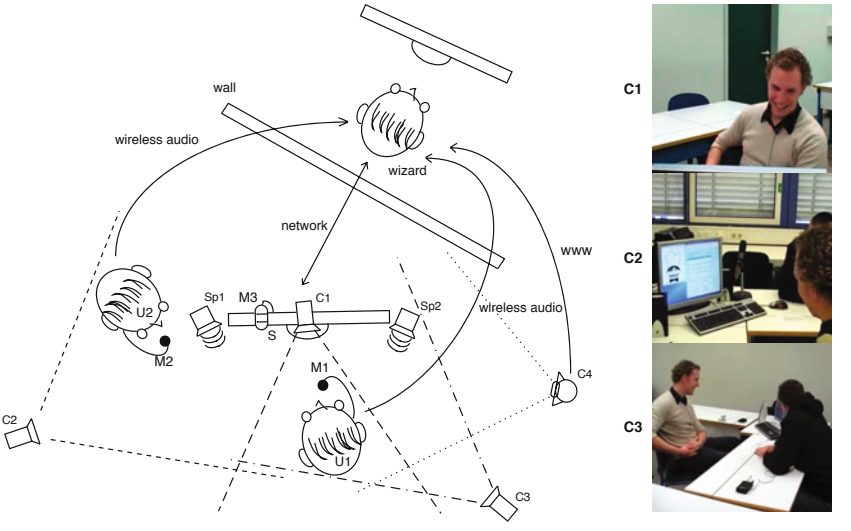

Fig. 1. Schematic view of the Wizard-of-Oz recording setup. The primary user marked as U1 interacts with the secondary user U2 and the System S. The wizard is located in another room and receives real time input from camera C4, and microphones M1 and M2. The subjects receive audiovisual output from the speakers Sp1-2 and from the screen of S. Camera C1 records the face of U1 directly and cameras C2-3 record the scenery. The figure is adapted from [15]. On the right side, a typical scene taken from a recording from all three different camera angles is shown.

lower level was then annotated separately in consecutive annotation runs. The distribution of labels over the dialogs and subject roles is listed in Table 2. Additionally, the average durations and the standard deviations are listed there. Further, alongside the listed labels, the focus of attention and the actual gaze direction of the primary subject was annotated and analyzed [16].

3 Evaluation and Discussion

In the following three labels, namely interested, positive accepting[1], and negative labels[2] are statistically analyzed in order to find significant correlates between the lower and higher levels.

In order to measure the coincidence of the simpler annotations with the subject state, the relative overlap of these lower level labels with the subject state was measured for all the annotated dialogs. The relative overlap r is calculated as the overlapping length o of the lower objective label with respect to the length of the subject state annotation l: $r = \frac{o}{l} \in [0, 1]$. The result is evaluated using

[1] An offer or suggestion of the system is perceived positively by the subject.

[2] All negative subject states combined, i.e. uninterested, embarrassed, impatient, stressed, negative accepting, disagreement.

Table 2. Number of occurrences and durations of labels for user U1 and U2. All four annotation layers are listed with their respective labels. The durations are listed in seconds.

Subject state	**Avg. Length**		**Std. Deviation**		**Duration**		**Occurrences**	
	U1	**U2**	**U1**	**U2**	**U1**	**U2**	**U1**	**U2**
Interested	13.1	13.5	11.0	11.5	4012.6	3599.2	306	266
Uninterested	11.4	11.9	4.5	8.1	91.2	261.7	8	22
Surprised reacting	6.6	4.7	7.6	2.4	159.1	32.9	24	7
Embarrassment	9.4	9.4	7.5	7.0	366.4	103.2	40	11
Impatient	7.6	5.6	6.5	4.5	175.2	90.2	23	16
Stressed	5.8	3.5	2.7	2.3	69.5	7.0	12	2
neg. accepting	4.9	5.1	4.0	3.3	173.0	111.5	35	22
pos. accepting	6.1	6.1	5.0	4.7	904.5	725.9	149	119
Disagreement	5.5	9.3	3.4	5.1	82.9	102.0	15	11
Talk style								
Commanding	4.1	4.8	2.5	4.2	252.8	124.6	61	26
Off-Talk	10.3	9.9	6.7	6.0	227.7	138.9	22	14
Ironic	5.4	4.0	4.9	2.8	75.2	56.0	14	14
Explaining	8.5	5.0	7.5	3.6	1296.7	474.9	153	95
Active Listening	9.1	14.2	5.7	15.6	2731.5	4071.9	299	287
Question	4.3	4.4	2.5	4.3	595.6	352.0	137	80
Thinking	4.5	4.6	3.2	2.4	90.7	68.4	20	15
Reading	9.6	n/a	5.3	n/a	105.2	n/a	11	n/a
Event								
Laughs	3.2	2.9	1.5	1.7	352.9	306.8	112	107
Silence	9.0	9.3	7.2	9.4	135.1	74.6	15	8
Exciting Moments	7.1	3.7	7.5	1.4	56.5	14.8	8	4
Topic Shifts	2.1	2.1	1.3	1.3	21.2	21.2	10	10
Waiting	4.8	8.0	3.5	5.1	135.7	88.4	28	11

Dom. dialog role	**Average**	**Standard dev.**	**Duration**	**Occurrences**
User U1	10.9	11.2	1460.4	134
User U2	7.7	5.0	271.2	35
System S	10.5	4.8	1988.3	189
Eq. Active	18.8	15.6	1765.2	94
Eq. Passive	10.7	6.6	363.0	34

box plots where brackets with * or ** indicate significant ($p < .05$) and highly significant ($p < .01$) differences in the overlaps calculated using paired t-tests. The boxes denote 50% of the data and the median value is shown as the middle line of the plot. Whiskers include 1.5 times the standard deviation of the data and outliers marked as crosses are further away from the median.

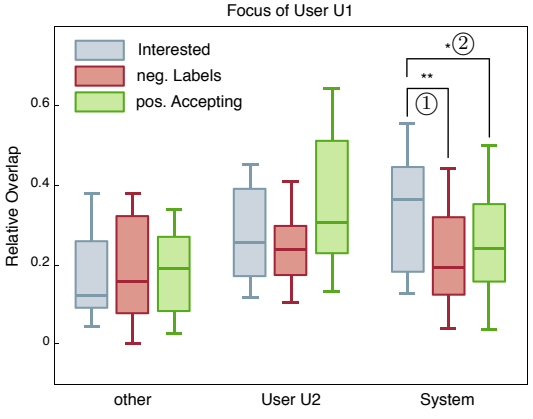

Fig. 2. Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 with his focus towards the system, user U2, or elsewhere (other)

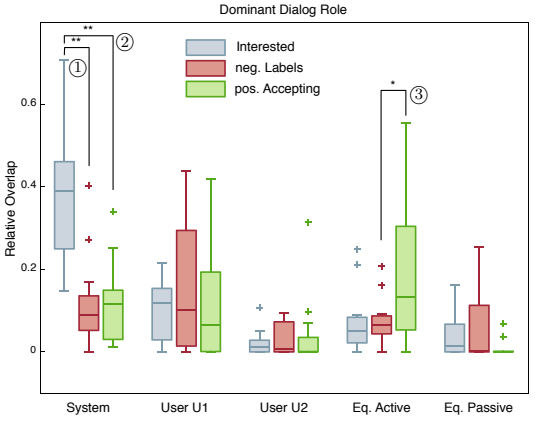

Fig. 3. Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 with the dominant dialog role annotations

In Figure 2, it is seen that the focus of attention towards the system differs significantly over the three targeted subject states. In detail, U1 is labeled significantly ($p = .003$, ①) more as focusing on the system while he is labeled

as interested contrary to negative labels. Further, he significantly ($p = .031$, ②) focuses the system more while interested in contrast to the label positive accepting.

In Figure 3, the dominant dialog role annotations are compared to U1's subject states: if the system takes over the dominant role in the conversation, highly significant support for the state interested is found (vs. positive accepting $p = .001$, ①; vs. negative labels $p < .001$, ②). It is also seen that if all participants are equally active in the dialog the state positive accepting is significantly ($p = .043$, ③) overlapped to a higher extent.

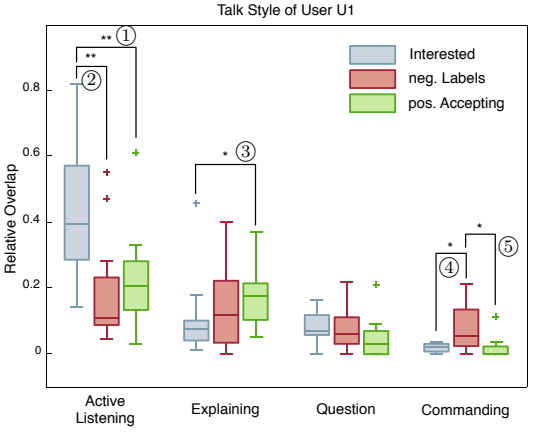

Fig. 4. Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 with the talk style/utterances of user U1

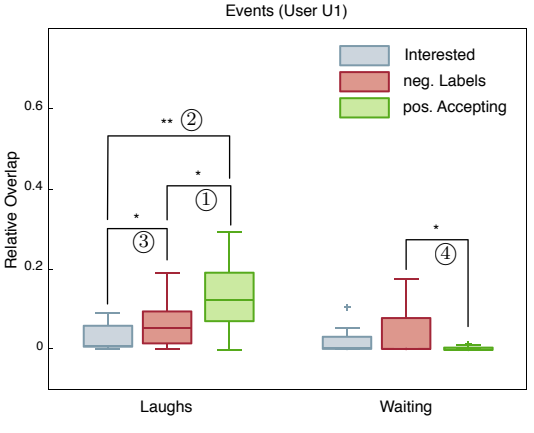

Fig. 5. Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 with the two most frequent U1 related events (laughs and waiting)

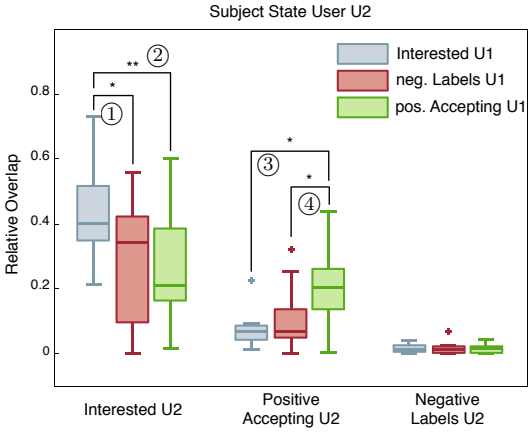

Fig. 6. Comparison of the relative overlaps for the subject states interested, positive accepting and negative labels of the users U1 and U2

Single turns, utterances or paralinguistic cues of subject U1 can be investigated in Figure 4. It is seen that the overlap for active listening, including many feedback and back-channeling utterances and paralinguistic cues, such as "um", or "hm", is highly significantly larger for the state of interested as opposed to the other two categories (vs. positive accepting $p = .001$, ①; vs. negative labels $p < .001$, ②). Further, the amount of overlap with respect to the talk style explaining for positive accepting is significantly higher compared to interested ($p = .028$, ③). The overlap of questions posed is not significantly higher if the user is interested, whereas the annotations of commanding ares significantly overlapping more with the negative labels as with the two other categories (vs. interested $p = .012$, ④; vs. positive accepting $p = .019$, ⑤).

Additionally, Figure 5 allows the examination of the relevance of the labels subsumed in "events" for the identification of the user's state in the interaction. Laughter overlaps significantly more in the state of positive accepting as opposed to the negative annotations ($p = .012$, ①) and highly significantly more in contrary to interested ($p < .001$, ②). This finding supports the idea that U1 is commenting positively perceived suggestions with a surprised or pleased smile or laughter. Further, the overlap of laughs with the negative labels in comparison to interested is significant too ($p = .018$, ③). Figure 5 also shows that the relative amount of overlap of the annotation of waiting with negative subject state labels is significantly higher than the one for positive accepting ($p = .046$, ④).

In Figure 6, a comparison of the subject state of U1 to the one of U2 is shown. The labels interested and positive accepting correlate between both users forming some sort of common interactional state. The relative overlap of the interested state of U2 with the state of interested of U1 is significantly higher ($p = .022$, ①) than the negative label overlap and highly significant ($p = .001$, ②) for positive accepting. Further, if U1 is in the state of positive accepting we can observe significantly higher overlaps for the subject state positive accepting of U2 (vs.

interested: p = .017, ③; vs. negative labels p = .010, ④), which indicates some sort of mimicry behavior of U2.

4 Conclusions and Future Directions

This paper links the different categories of labels, first introduced in [13], in order to reveal the dynamics of patterns and correlations among them. These annotations are grouped in various sets combining labels of similar character, i.e. subjective user state labels, as well as coarse and fine grained observations, such as dialog roles or laughter (compare Table 1), arranged in a hierarchical structure. The analysis revealed several significant dependencies between the subjectively annotated user's state and the low level observations (i.e. social signals, etc.). These dependencies support the hierarchical approach that social signals and the patterns in which they can be observed can serve as basic building blocks that help infer the current users' interactional state.

These dynamic multimodal patterns could now in turn be used to automatically detect the user's state: Work, such as [10,12], have for example shown that it is possible to recognize laughter in naturalistic conversations, or in [8] head poses and the focus of a subject is recognized. These automatic detection for the lower level observations could now be integrated in hierarchical classification approaches. The inherent temporal structure of this application demands classifiers that are able to incorporate these dynamics, such as hidden Markov models or echo state networks. Further, such a hierarchical architecture needs to be robust with respect to uncertainty and possible sensory outage or failure.

The proposed annotation approach, however, is not exhaustive and extensions are straightforward. In [16] for instance it has been shown that voice quality can be used to infer the affective state of the speaker and [11] shows the capability to automatically recognize it. Additionally, the subject state layer comprising the so called conversational dispositions is not compulsory, but could in principal be exchanged by other schemes, such as dimensional affect annotations.

Acknowlegdements. The presented work was developed within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). The work of Martin Schels is supported by a scholarship of the Carl-Zeiss Foundation.

References

1. Campbell, W.N.: On the use of nonVerbal speech sounds in human communication. In: Esposito, A., Faundez-Zanuy, M., Keller, E., Marinaro, M. (eds.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 117–128. Springer, Heidelberg (2007)
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18(1), 32–80 (2001)

3. Darwin, C.: The expression of emotion in man and animals, 3rd edn. HarperCollins, London (1978)
4. Ekman, P.: Facial expression and emotion. American Psychologist 48, 384–392 (1993)
5. Kendon, A. (ed.): Nonverbal Communication, Interaction, and Gesture. Selections from Semiotica Series, vol. 41. Walter de Gruyter, Berlin (1981)
6. Kipp, M.: Anvil - a generic annotation tool for multimodal dialogue. In: Proceedings of the European Conference on Speech Communication and Technology (Eurospeech), Aalborg, pp. 1367–1370. ISCA (2001)
7. Layher, G., Liebau, H., Niese, R., Al-Hamadi, A., Michaelis, B., Neumann, H.: Robust stereoscopic head pose estimation in human-computer interaction and a unified evaluation framework. To Appear in 16th International Conference on Image Analysis and Processing (ICIAP 2011). Springer, Heidelberg (2011)
8. Russell, J.A., Barrett, L.F.: Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. Journal of Personality and Social Psychology 76(5), 805–819 (1999)
9. Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G.: Spotting laughter in naturalistic multiparty conversations: a comparison of automatic online and offline approaches using audiovisul data. ACM Transactions on Interactive Intelligent Systems: Special Issue on Affective Interaction in Natural Environments (accepted for publication)
10. Scherer, S., Kane, J., Gobl, C., Schwenker, F.: Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. IEEE Transactions on Audio, Speech and Language Processing (under review)
11. Scherer, S., Schwenker, F., Campbell, W.N., Palm, G.: Multimodal laughter detection in natural discourses. In: Ritter, H., Sagerer, G., Dillmann, R., Buss, M. (eds.) Proceedings of 3rd International Workshop on Human-Centered Robotic Systems (HCRS 2009). Cognitive Systems Monographs, pp. 111–121. Springer, Heidelberg (2009)
12. Scherer, S., Trentin, E., Schwenker, F., Palm, G.: Approaching emotion in human computer interaction. In: International Workshop on Spoken Dialogue Systems (IWSDS 2009), pp. 156–168 (2009)
13. Strauss, P.-M., Hoffmann, H., Minker, W., Neumann, H., Palm, G., Scherer, S., Traue, H.C., Weidenbacher, U.: The PIT corpus of german multi-party dialogues. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, pp. 2442–2445. ELRA (2008)
14. Strauss, P.-M., Hoffmann, H., Neumann, H., Minker, W., Palm, G., Scherer, S., Schwenker, F., Traue, H.C., Weidenbacher, U.: Wizard-of-oz data collection for perception and interaction in multi-user environments. In: Proceedings of the Fifth International Language Resources and Evaluation (LREC 2006), pp. 2014–2017. ELRA (2006)
15. Strauss, P.-M., Scherer, S., Layher, G., Hoffmann, H.: Evaluation of the PIT corpus or what a difference a face makes? In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, pp. 3470–3474. ELRA (May 2010)
16. Yanushevskaya, I., Gobl, C., Chasaide, A.N.: Voice quality and loudness in affect perception. In: Proceedings of Speech Prosody 2008, Campinas, Brazil, pp. 29–32. ISCA (2008)