Is There a Biological Basis for Success in Human Companion Interaction?

Results from a Transsituational Study

Dietmar Rösner[1](✉), Dilana Hazer-Rau[2], Christin Kohrs[3], Thomas Bauer[1], Stephan Günther[1], Holger Hoffmann[2], Lin Zhang[2], and André Brechmann[3]

[1] Institut für Wissens- und Sprachverarbeitung (IWS),
Otto-von-Guericke Universität, Postfach 4120, 39016 Magdeburg, Germany
{roesner,tbauer,stguenth}@ovgu.de

[2] Medical Psychology, Ulm University, Frauensteige 6, 89075 Ulm, Germany
{dilana.hazer,holger.hoffmann,lin.zhang}@uni-ulm.de

[3] Special Lab Non-Invasive Brain Imaging, Leibniz Institute for Neurobiology,
Brenneckestr. 6, 39118 Magdeburg, Germany
{christin.kohrs,andre.brechmann}@lin-magdeburg.de

Abstract. We report about a transsituational study where a representative subsample of twenty of the subjects from the LAST MINUTE experiment underwent two additional independent experiments: an fMRI study and a psychophysiological experiment with emotion induction in the VAD space (Valence, Arousal, Dominance). A major result is that dialog success in the naturalistic human machine dialogs in LAST MINUTE correlates with individual differences in brain activation as reaction to delayed system responses in the fMRI study and with the classification rate for arousal in the emotion induction experiment.

1 Introduction

Empirical research in HCI is conducted by different disciplines with a multitude of approaches. Such empirical investigations range from in-depth analysis of full fledged human dialogs and naturalistic dialogs between humans and Wizard of Oz simulated systems (e.g. [11,22,23]) to psychophysiological and neurophysiological studies utilizing controlled stimuli, e.g. for the induction of emotional responses (e.g. [6,21]) or the neural responses elicited by computer feedback in dialog-like situations (e.g. [8]).

Despite the relative merits of such experiments, their outcome can not easily be combined and generalized when the experiments are performed completely independent and with different cohorts. This changes remarkably in a transsituational setting with a common cohort of subjects undergoing a series of different experiments.

In the following we report about such a transsituational study. A representative subsample of $N_{exps} = 20$ subjects from the LAST MINUTE experiment (LME) [5,15,16] has undergone two additional independent experiments:

M. Kurosu (Ed.): HCI 2016, Part I, LNCS 9731, pp. 77–88, 2016.
DOI: 10.1007/978-3-319-39510-4_8

an fMRI study [9] and a psychophysiological experiment [18,24]. We will present and discuss what the transsituational analysis of data from all three experiments revealed about correlations between dialog success in LME and results from fMRI and analysis of biopsychological data.

2 The Experiments

2.1 The LAST MINUTE Experiment

The LAST MINUTE corpus (LMC) is derived from a large scale Wizard of Oz (WoZ) experiment – the LAST MINUTE experiment (LME) – that required users to solve a mundane task with the need for planning, replanning and strategy change (cf. [5,15]). The LMC comprises multimodal recordings (audio, transcripts, video, biopsychological data, ...) from a cohort of $N_{total} = 133$ subjects. The cohort was balanced in gender (68 women and 65 men) and age group (72 subjects aged between 18 and 28 vs. 61 aged above 60 years).

The LMC has been intensively investigated with respect to differences in dialog success (e.g. [14,15]). For example, significant differences between the age groups could be found, whereas global differences in gender were insignificant [17].

Dialog Acts and Interaction Success. In the transcripts of the LMC all user and system utterances are semi-automatically enriched with dialog act labels in the format of the dialog act representation (DAR) [15]. The DAR employs triples that first encode the speaker (i.e. S for subject, W for wizard), then the dialog act (e.g. R for REQUESTs, A for ACCEPTs or Rj for REJECTs, cf. [3]) and finally an optional subtype (e.g. a REQUEST for an action may be subtyped with the shorthand code for the action: P for packing, U for unpacking, C for changing category, ..., [15]).

These local DAR annotations are exploited for defining measures for global dialog success by integrating over local interaction success or failure in problem solving [15]:

- **DSM1**: ratio between accepted subject requests and total number of subject requests;
- **DSM2**: ratio between accepted subject requests and total number of turns (i.e. not only subject requests).

By definition the following holds for all transcripts: $0 \leq \text{DSM2} \leq \text{DSM1} \leq 1$. In the following evaluations we work with these measures.

Subphases. As a key aspect, an inherent need for re-planning and strategy change was built into the WoZ scenario of LAST MINUTE (cf. [5]). Therefore problem solving was divided into three major subphases which where demarcated by the weight limit barrier (**WLB**, after the eighth of twelve consecutive selection categories) and the weather info barrier (**WIB**, after the tenth selection category).

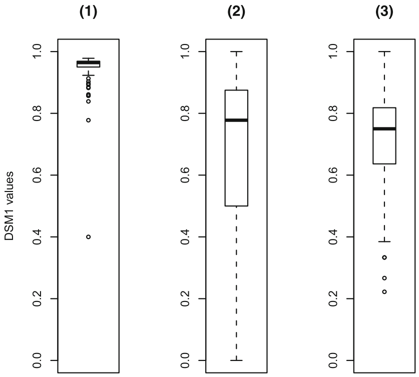

Fig. 1. DSM1 values in the three subphases of problem solving ($N_{total} = 133$): (1) before weight limit barrier (WLB), (2) from WLB to weather info barrier (WIB), (3) from WIB to end of experiment. Please note outliers in (1) and (3).

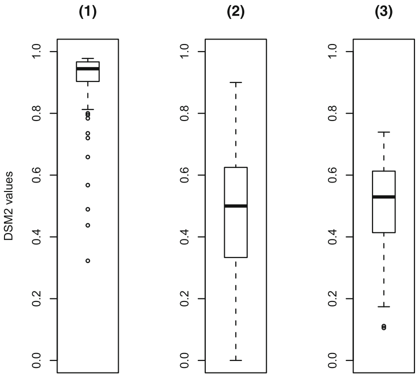

Fig. 2. DSM2 values in the three subphases of problem solving ($N_{total} = 133$). Please note outliers in (1) and (3).

The WLB enforces re-planning. Up to the WLB there is no need for unpacking, but thereafter it becomes crucial, for no progress can be achieved without at least one successful unpacking attempt.

At the WIB the subject receives a deliberately delayed weather information about the target location. This enforces a strategy change. Now items for cold and rainy weather are needed in exchange for e.g. bathing suits and other summer items.

Reasons to Fail. At first glance packing a suitcase by selecting items from menus seems to be a fairly easy task. How can the resp. dialogs become problematic? The weight limit barrier (WLB) poses a major challenge for many subjects. They have to successfully unpack items in order to make room for additional items. What can prevent subjects from successful unpacking? There are a variety of potential failures:

- Subjects may try to unpack items that are not in the suitcase.
- A variant of this failure is when subjects employ synonyms or hypernyms for items and when the wizards are not accepting these terms. Such rejections often proof to be very puzzling for subjects.
- In its verbal explanation of the weight limit the system has offered the option to enumerate the suitcase contents (verbally). Such a listing may be very helpful. However, subjects have to be aware of this option. Quite a number of subjects seem to be unaware of this option, at least they do not make use of it.

Examples of Transcripts. In the following we shortly present excerpts from transcripts illustrating these points.

The transcript example in Table 1 with DAR labels is taken from a dialog segment where a subject (20110117bsk) hits the weight limit again and successfully overcomes it by unpacking several items (SRU WAU pairs).

The excerpt in Table 2 illustrates that unpacking may fail: the unpacking requests (SRUs) of subject 20110126afw are rejected several times (WRjUs). Therefore the packing request (SRP) after the category change (WIC) is doomed to fail as well (WRjP).

Table 1. Transcript example with DAR tags (subject 20110117bsk; cf. text). Following the GAT-2 minimal standard [19] short pauses are noted as (.) and (-).

Tag	German text	English gloss
	...	...
SRP	((schnalzt)) (-) zwei bücher	((flicks)) (-) two books
WRjP	der artikel buch kann nicht hinzugefügt werden (.) anderenfalls würde die maximale gewichtsgrenze des koffers überschritten werden	the item book cannot be added (.) otherwise the weight limit of your suitcase will be exceeded
SRU	schnorchel entfernen	remove snorkel
WAU	ein schnorchel wurde entfernt	a snorkel has been removed
SRU	taucherbrille entfernen	remove diving googles
WAU	eine taucherbrille wurde entfernt	diving googles have been removed
SRU	kleid entfernen	remove dress
WAU	ein kleid wurde entfernt	a dress has been removed
SRU	((schnalzt)) badelatschen entfernen	((flicks)) remove beach slippers
WAU	ein paar badelatschen wurden entfernt	a pair of beach slippers have been removed
SRP	buch hinzufügen	add book
WAP	ein buch wurde hinzugefügt	a book has been added
SRC	nächste kategorie	next category
WIC	sie können jetzt aus der rubrik technische geräte auswählen ...	you may now choose from the category technical devices ...

Dialog Success in Subphases. The DSM values differ remarkably within the subphases of problem solving (cf. Figs. 1, 2).[1] As shown in Fig. 1, the DSM1 values lie above 0.95 for more than 75 % of the subjects before the WLB. In subsequent phases the medians drop to 0.78 and 0.75, while the means go down to 0.69 and 0.71 respectively.

These differences can be explained as follows: In the first subphase the vast majority of subjects has hardly any problems. The weight limit barrier (WLB)

[1] The distributions are visualized — here and in other figures — as trellis box plots: the rectangles represent the interquartile range (i.e. the range of 25 % of the values above and below the median resp.), the black bar gives the median, the whiskers extend the rectangle to the range of values, but maximally to 1.5 of the interquartile range, outlier values beyond the maximal whisker range are given as unfilled dots (cf. [1]).

Table 2. Transcript example with DAR tags (subject 20110126afw; cf. text). Following the GAT-2 minimal standard [19] short pauses are noted as (.) and (-), longer pauses with their duration in brackets, e.g. (3.0).

Tag	German text	English gloss
	...	...
SRP	ein reiseführer	a guidebook
WRjP	der artikel reiseführer kann nicht hinzugefügt werden (.) anderenfalls würde die maximale gewichtsgrenze des koffers überschritten werden	the item guidebook cannot be added (.) otherwise the weight limit of your suitcase will be exceeded
SRU	hosen herausnehmen	remove trousers
WRjU	der gewünschte artikel ist nicht im koffer enthalten	requested item is not contained in the suitcase
SRU	fünf socken herausnehmen	remove five socks
WRjU	der gewünschte artikel ist nicht im koffer enthalten	the requested item is not contained in the suitcase
SRU	einen hut herausnehmen	remove a hat
WRjU	der gewünschte artikel ist nicht im koffer enthalten	the requested item is not contained in the suitcase
SRU	sonnenhut (3.0) herausnehmen	sunhat (3.0) out
WRjU	der gewünschte artikel ist nicht im koffer enthalten	the requested item is not contained in the suitcase
	(---)	(---)
WIT	die auswahl von artikeln aus der rubrik reiselektüre muss jetzt beendet werden (.) um die aufgabe in der zur verfügung stehenden zeit beenden zu können	selecting items from category travel reading needs to be finished now in order to complete the task in time
	(--)	(--)
WIC	sie können jetzt aus der rubrik technische geräte auswählen	you may now choose from the category technical devices
SRP	fotoapparat	camera
WRjP	der artikel fotoapparat kann nicht hinzugefügt werden (.) anderenfalls würde die maximale gewichtsgrenze des koffers überschritten werden	the item camera cannot be added (.) otherwise the weight limit of your suitcase will be exceeded
	...	...

changes this drastically because without success in unpacking, no further packing is possible. Thus if unpacking is not attempted or is not successful a downward spiral with a series of subsequent rejections – and thus low DSM values – may result. As already discussed above, the task of unpacking demands to remember already packed items (or - as an alternative - to remember that a listing of the suitcase contents can be asked for), to decide which packed items to sacrifice and to request the respective unpacking action.

2.2 The fMRI Experiment

The subcohort of twenty subjects was recruited to participate in a functional magnetic resonance imaging (fMRI) experiment carried out in a 3 Tesla scanner (Siemens Trio, Erlangen, Germany). Three participants had to be excluded due to contraindications for MRI, excessive head motion, or abortion of the experiment before completion. Methodological details of the scanning parameters for anatomical and functional imaging, presentation of acoustic and visual stimuli, and task procedure are published in [8]. In short, the participants had to perform an auditory categorization task on frequency modulated sounds. In 300 trials they had to indicate by left or right button press if a sound was rising or falling in pitch. In 85 % of all trials they received immediate feedback, i.e. visual presentation of a green checkmark for correct responses and a red cross for false responses. In 15 % this feedback was delayed by 200, 400, or 600 ms. The fMRI data were analyzed with a region of interest (ROI) analysis of variance (ANOVA) as implemented in BrainVoyagerQX using the four different feedback times as predictors. As regions of interests, we used five brain areas identified to be significantly activated by feedback that was delayed by 500 ms as compared to immediate feedback [9], i.e. posterior medial prefrontal cortex, bilateral anterior insula, left inferior parietal lobe, and right inferior frontal cortex. Within each of these areas, we identified the beta values resulting from the ANOVA regarding feedback delays of 400 ms and 600 ms and subtracted the beta values for immediate feedback in order to determine the increase of neural activity elicited by delayed feedback as compared to immediate feedback. We did not analyze activity elicited by 200 ms delays because such short delays do not have a significant impact on brain activity and are usually not perceived as delayed [8]. We then calculated a two sided Pearson correlation between the participants' dialog success rate from the LAST MINUTE experiment and the activation data of the participants in each of the five brain areas.

2.3 The Psychophysiological Experiment

In the psychophysiological experiment, the same representative subcohort of twenty from the LME recruited at the university of Magdeburg participated in a controlled emotion induction setting. Numerous studies on emotion recognition based on facial expression, speech, body language, contexts and physiological signals have been performed in the past few decades [2]. Among them, physiological signals have considerable advantages, for example, as honest signals [13], they cannot be easily triggered by any conscious or intentional control. Various classifications, feature selection and evaluation algorithms are currently used for the emotion recognition from physiological data [21,24].

In this experiment, emotions were induced by using standardized stimuli from the International Affective Picture System (IAPS) to represent the VAD (Valence, Arousal, Dominance) space. The advantage of using standardized stimuli relies in the reliability of the induction of a specific VAD value. Prolonged presentations consisting of 10 pictures à 2 s (total of 20 s) are used to intensify the elicitation. A total of 10 picture-presentations à 20 s each were presented to induce a total of 10 VAD-states. 20 s neutral fixation crosses were introduced as baseline between 2 different presentations. The induced VAD-space for the 10 picture-presentations included positive/negative/neutral (+/-/0) Valence, positive/negative (+/-) Arousal, and positive/negative (+/-) Dominance values. For the classification of the emotional states, picture-presentation with similar ratings in terms of Valence (+/-/0) and/or Arousal (+/-) and/or Dominance were combined in one category. In total we evaluated the emotion recognition rates of 5 different category-classes: V(+/-/0), A(+/-), D(+/-), VA(0-/++/-+/+-/–), VAD(10 different picture presentations) [24].

We processed the emotion recognition rates by fusing four physiological signals including skin conductivity, respiration and 2x electromyography signals (corrugator & zygomaticus). The evaluation was conducted using the Augsburg Biosignal Toolbox (AuBT, [20]). The AuBT provides Matlab-based tools to analyze physiological signals for the emotion recognition. The emotion recognition rates were computed for each subject individually. Therefore, for the evaluation of the individual classification rates for each subject, 10 samples from the subject itself were used as test-set and 190 samples from the 19 subjects left were used as training-set. In total 20 different classifications were conducted.

3 Results

3.1 The Subcohort in the LAST MINUTE Experiment

The LAST MINUTE experiments with the total cohort were performed over a time period of nearly a year. The transsituational experiment with the subcohort of twenty subjects was performed in a compact subinterval of two months. The participants in the subcohort were randomly chosen during this interval with the only restriction that finally the subcohort was roughly balanced with respect to gender (9 women vs. 11 men) and age groups (10 young subjects aged between 18 and 28 years vs. 10 elderly subjects aged above 60 years).

When one tests dialog success values (as measured with DSM1 and DSM2) for representativity, the distribution in the subcohort of twenty does not differ significantly from that of the total cohort (Wilcoxon tests; DSM1: W = 1105, p = 0.2243; DSM2: W = 1051, p = 0.1317; see also Figs. 3, 4). Having a representative subsample with respect to dialog success motivates the following investigations about correlations between dialog success in LME and relevant outcomes in the two other experiments.

The discourse success measures DSM1 and DSM2 are strongly correlated. This holds for the whole cohort of $N_{total} = 133$ (Pearson's product-moment correlation 0.85, $t = 18.1972, df = 131, p < 2.2e - 16$) as well as for the sample of twenty (Pearson's product-moment correlation 0.88, $t = 8.022, df = 18, p = 2.357e - 07$). But there is considerable variance with respect to individual differences between the two values. When we take the quotient $DSM2/DSM1$ as a percentage we get distributions as summarized in Table 3 for the whole cohort and for the subcohort of 20.

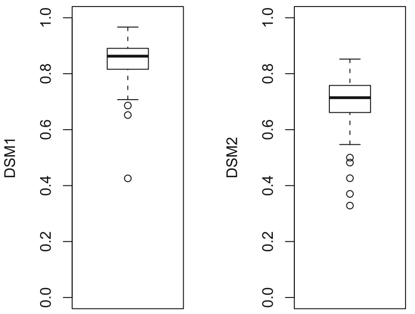

Fig. 3. Distribution of DSM1 and DSM2 values for complete cohort ($N_{total} = 133$)

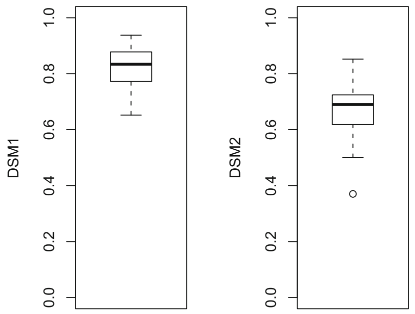

Fig. 4. Distribution of DSM1 and DSM2 values for subcohort of $N_{exps} =$ 20

Table 3. DSM2 values as percentage of DSM1 values per transcript

cohort	Min.	1st Qu	Median	Mean	3rd Qu.	Max.
133	55.29	80.60	83.10	82.34	85.92	92.65
20	56.79	79.50	82.68	80.29	85.45	90.91
Elder	55.29	78.95	82.14	80.79	84.62	89.58
Young	62.12	81.23	84.53	83.66	86.12	92.65

For the whole cohort the age groups differ significantly and with a medium effect size with respect to this quotient, with elderly subjects having greater differences between DSM2 and DSM1 than the younger (Wilcoxon rank sum test W = 1531, p = 0.002695, $d_{Cohen} = 0.503$).

3.2 Results from the fMRI Experiment

Consistent to the previous study, we found that delays of 400 and 600 ms result in significant activation of the selected brain regions. However, we found considerable interindividual variance of the activation increase elicited by delayed feedback in each of the selected brain regions. Pearson correlations showed a significant positive relation between dialog success in the LME and activation increase only in the anterior insula (left insula: DSM1: $p = 0.005$, $r = 0.65$, DSM2: $p = 0.005$, $r = 0.65$; right insula: DSM1: $p = 0.014$, $r = 0.65$, DSM2: $p = 0.018$, $r = 0.57$). The lower significance for the right insula may be due to the fact that the ROIs selected from a previous study with different participants and thus different anatomy matched the location of activation resulting from the group level analysis of the actual participants less well (see overlap of transparent ROI with significant voxels coded in yellow to red in the right panel of Fig. 5). After the fMRI experiment, the subjects filled in a questionnaire whether or not they have noticed delays in feedback and if so how many different delays. Five subjects

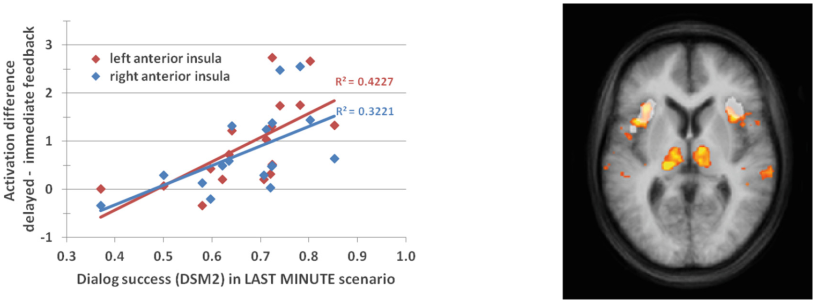

Fig. 5. Correlation between the dialog success (DSM2) in the LAST MINUTE experiment and the increase in fMRI activation in the anterior insula elicited by delayed vs. immediate feedback (left panel). The right panel shows the location of the region of interest (transparent grey cluster) extracted from a previous study [9] and the significant fMRI activation from the group level analysis of the current study (Color figure online).

reported not to have noticed any delay at all, and these subjects are among the lower performers (rank 10 to 14) in the LAST MINUTE experiment.

3.3 Results from the Psychophysiological Experiment

For the analysis of the correlation, the two-category-classes Arousal (+/-) and Dominance (+/-) as well as the three-category-class Valence (+/-/0) are considered in this study. For each subject, the individual recognition rates are correlated with the individual dialog success ratios. The classification data are normally distributed and the Pearson correlation coefficients are used to assess the correlation between the emotion recognition rates and the dialog success DSM1 and DSM2 ratios. Two strongly positive correlations between the emotion recognition rates of the classifiers and the dialog success from the LME were found: (1) A strong positive correlation between the emotion recognition rate of the

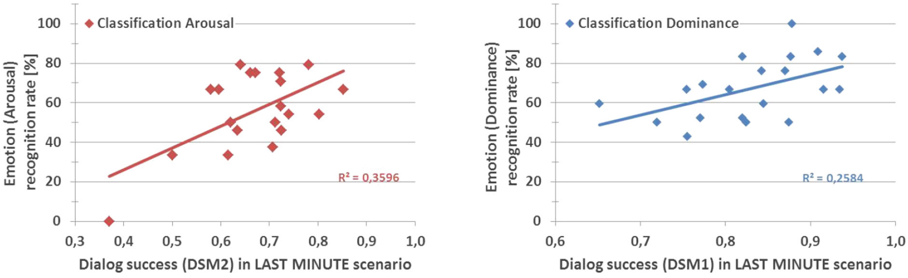

Fig. 6. Correlation between the dialog success DSM2 and the psychophysiological Arousal recognition rate (left-panel) and between the dialog success DSM1 and the psychophysiological Dominance recognition rate (right-panel).

two-category-class Arousal (+/-) and the dialog success rate DSM2 ($r = 0.600$, $p = 0.005$), and (2) A strong positive correlation between the emotion recognition rate of the two-category-class Dominance (+/-) and the dialog success rate DSM1 ($r = 0.503$, $p = 0.024$). The correlations results are illustrated in Fig. 6.

The biophysiological recognition rate results were obtained using the Sequential Forward Selection (SFS) feature selection and the Linear Discriminant Analysis (LDA) model for the two-category-class Arousal classification and using the Sequential Forward Selection (SFS) feature selection and the k-Nearest Neighbor (kNN) model for the two-category-class Dominance classification. No correlation was found between the dialog success and the emotion recognition of the three-category-class Valence (+/-/0) classification.

4 Discussion and Future Work

As a summary we have the following: dialog success in the naturalistic human machine dialogs in LAST MINUTE correlates with individual differences in brain activation as reaction to delayed system responses in the fMRI study and with the classification rate for arousal in the emotion induction experiment.

What do different – and especially low – values of the dialog success measures mean? Not surprisingly, there are quite different dialog courses in the LMC but generally speaking, subjects with low DSMs are locally unsuccessful repeatedly in interaction and they generally take longer (or completely fail) to overcome challenging situations in the dialogs where – by design [5] – re-planning or even strategy change is needed. In contrast high values of the DSMs go with avoiding to repeat errors once encountered and with high flexibility in adapting to unforeseen situations in the dialog course.

The original objective of the psychophysiological experiment was to find feature combinations for classifying emotional states by psychophysiological responding [24]. On an individual level, high classification rates for emotional states mean a consistency between the psychophysiological response and a certain emotion stimulation, e.g. a high arousing IAPS picture series would than lead to high amplitudes of the skin conductance level (SCL). Given that such an interpretation is correct, the obtained psychophysiological correlation implicates that persons with high DSM in a HCI setting show a high correspondence between their dialog activity and their psychophysiological responding and vice-versa. This would render the classification of the arousal state based on psychophysiology more easy in high expressive persons [24]. Since the correlation between classification rates and brain activation was not significant, further analyses e.g. of personality traits are needed to better understand the biological basis of individual success in HCI.

The successful participants in the LAST MINUTE experiment show strong fMRI activation of the anterior insula after delayed SRT. The original aim of the fMRI experiment was to study the effects of breaching a general rule of communication, namely the subjective sense of completion of an action [12]. We have shown that unexpected delays in feedback elicits an emotional response that can be classified as "suspense" according to the accompanying psychophysiological effects of a decelerating heart rate together with an increased skin conductance response [10]. Thus, communication has subjectively been perceived as unsuccessful. The anterior insula has recently been suggested to play a more general role in awareness (beyond interoception), and as neural correlate of consciousness [4]. Since an increase in activity has been suggested as sign of a conscious perception of an error [7], participants with a strong activation in the fMRI experiment may have perceived the delay as irritating. Participants with low activity, however, are less irritated and indeed four subjects with low DSM2 values and low activation of the anterior insula could not remember to have encountered delayed SRT after the MRI scan. Taken together, participants with a higher degree of conscious awareness of maladaptive dialog acts of technical systems seem to be more successful in challenging situations of the LAST MINUTE experiment

Acknowledgements. The presented study is performed in the framework of the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). It is also supported by a doctoral scholarship funded by the China Scholarship Council (CSC) for Lin Zhang and a Margarete von Wrangell (MvW) habilitation scholarship for Dilana Hazer-Rau. The responsibility for the content of this paper remains with the authors.

References

1. Baayen, R.: Analyzing Linguistic Data - A Practical Introduction to Statistics using R. Cambridge University Press, Cambridge (2008)
2. Calvo, R.A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans. Affective Comput. **1**(1), 18–37 (2010)
3. Core, M., Allen, J.: Coding dialogs with the DAMSL annotation scheme. In: AAAI fall symposium on communicative action in humans and machines, pp. 28–35 (1997)
4. Craig, A.: How do you feel – now? the anterior insula and human awareness. Nat. Rev. Neurosci. **10**, 59–70 (2009)
5. Frommer, J., Rösner, D., Haase, M., Lange, J., Friesen, R., Otto, M.: Früherkennung und Verhinderung negativer Dialogverläufe - Operatormanual für das Wizard of Oz-Experiment. Pabst Science Publishers (2012)
6. Hazer, D., Ma, X., Rukavina, S., Gruss, S., Walter, S., Traue, H.C.: Transsituational individual-specific biopsychological classification of emotions. In: Stephanidis, C. (ed.) Proceedings of the HCI International 2015, pp. 110–117 (2015)
7. Hester, R., Foxe, J., Molholm, S., Shpaner, M., Garavan, H.: Neural mechanisms involved in error processing: a comparison of errors made with and without awareness. NeuroImage **27**(3), 602–608 (2005)
8. Kohrs, C., Angenstein, N., Brechmann, A.: Delays in human-computer interaction and their effects on brain activity. PLoS ONE **11**(1) (2016). doi:10.1371/journal.pone.0146250
9. Kohrs, C., Angenstein, N., Scheich, H., Brechmann, A.: Human striatum is differentially activated by delayed, omitted, and immediate registering feedback. Frontiers Human Neurosci. **6**, 00243 (2012)

10. Kohrs, C., Hrabal, D., Angenstein, N., Brechmann, A.: Delayed system response times affect immediate physiology and the dynamics of subsequent button press behavior. Psychophysiology **51**(11), 1178–1184 (2014)
11. Legát, M., Grůber, M., Ircing, P.: Wizard of Oz data collection for the Czech senior companion dialogue system. In: Fourth International Workshop on Human-Computer Conversation, pp. 1–4. University of Sheffield (2008)
12. Miller, R.B.: Response time in man-computer conversational transactions. In: AFIPS Conference Prodeedings, pp. 267–277. Thompson Book Company, Washington (1968)
13. Pentland, A., Pentland, S.: Honest Signals: How They Shape Our World. MIT Press, London (2008)
14. Prylipko, D., Rösner, D., Siegert, I., Günther, S., Friesen, R., Haase, M., Vlasenko, B., Wendemuth, A.: Analysis of significant dialog events in realistic human-computer interaction. J. Multimodal User Interfaces **8**(1), 75–86 (2014)
15. Rösner, D., Friesen, R., Günther, S., Andrich, R.: Modeling and evaluating dialog success in the LAST MINUTE Corpus. In: Proceedings of LREC 2014. ELRA, Reykjavik, May 2014
16. Rösner, D., Haase, M., Bauer, T., Günther, S., Krüger, J., Frommer, J.: Desiderata for the Design of Companion Systems - Insights from a Large Scale Wizard of Oz Experiment. Künstliche Intelligenz (2015), 28 October 2015. doi:10.1007/s13218-015-0410-z
17. Rösner, D., Andrich, R., Bauer, T., Friesen, R., Günther, S.: Annotation and analysis of the LAST MINUTE corpus. In: Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology. pp. 112–121. Gesellschaft für Sprachtechnologie and Computerlinguistik e.V. (2015)
18. Rukavina, S., Gruss, S., Walter, S., Hoffmann, H., Traue, H.C.: Open_emorec_ii-a multimodal corpus of human-computer interaction. World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng. **9**(5), 1135–1141 (2015)
19. Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J.R., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., et al.: Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). Gesprächsforschung-Online-Zeitschrift zur verbalen Interaktion 10 (2009)
20. Wagner, J., Kim, J., André, E.: From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In: IEEE International Conference on Multimedia and Expo (ICME) (2005)
21. Walter, S., Kim, J., Hrabal, D., Crawcour, S.C., Kessler, H., Traue, H.C.: Transsituational individual-specific biopsychological classification of emotions. Systems, Man, and Cybernetics: Systems, IEEE Transactions **43**(4), 988–995 (2013)
22. Webb, N., Benyon, D., Bradley, J., Hansen, P., Mival, O.: Wizard of Oz Experiments for a Companion Dialogue System: Eliciting Companionable Conversation. In: Proceedings of LREC 2010. ELRA (2010)
23. Wolters, M., Georgila, K., Moore, J., MacPherson, S.: Being old doesn't mean acting old: how older users interact with spoken dialog systems. ACM Trans. Access. Comput. **2**(1), 2:1–2:39 (2009)
24. Zhang, L., Rukavina, S., Gruss, S., Traue, H.C., Hazer, D.: Classification analysis for the emotion recognition from psychobiological data. In: International Symposium on Companion-Technology (ISCT) (2015)