Biosignal Based Emotion Analysis of Human-Agent Interactions

Evgenia Hristova[1], Maurice Grinberg[1], and Emilian Lalev[1]

[1]Central and East European Center for Cognitive Science, New Bulgarian University,
1618 Sofia, Bulgaria
{ehristova@cogs.nbu.bg, mgrinberg@nbu.bg, elalev@cogs.nbu.bg}

Abstract. A two-phase procedure, based on biosignal recordings, is applied in an attempt to classify the emotion valence content in human-agent interactions. In the first phase, participants are exposed to a sample of pictures with known valence values (taken from IAPS) and classifiers are trained on the physiological data recorded. During the second phase, biosignals are recorded for each participant while interacting with an embodied conversational agent (ECA) and the classifiers trained in the first phase are applied. The results from the procedure are promising and are discussed in the paper together with the problems encountered and the suggestions for possible future improvement.

1 Introduction

Emotional reactions and users' satisfaction are an important factor in usability evaluation of human-agent interactions. A commonly used method for gathering such information is by means of self-administered questionnaires. An additional method that is used to assess the emotional reactions is by means of biosignals. Traditionally, these measures are used to study the attitudes and emotional experiences in various situations [1], [2]. Recently, they are more and more often used to study the emotional experiences of users during game playing, watching videos, visiting web sites or working with software applications [3], [4], [5], [6]. Among the most commonly used biosignals are the galvanic skin response, cardiovascular measures, respiration, and facial electro-myogram. Using such measures, one can gain information not only about emotions experienced consciously, which are generally stronger, but also about weaker emotions.

At the same time, the problem of identifying emotions using biosignals is notoriously difficult, although recently there have been many attempts to even build automatic emotion recognition systems [7], [8]. In refs. [7], [8], [9], [10], various data processing and classification methods have been compared trying to achieve high sensitivity and precision of emotion classification. While the results for one subject seem quite good with over 95% of recognition rate, subject independent classification seldom reach 70 %. Another unsolved problem is generalization over more than one task, i.e. applyin classifiers trained in one situation (e.g. music listening) to another situation (e.g. video watching).

In the present paper, we have two goals. First of all, we propose and assess a methodology for emotion valence assessment based on classifier training (calibration) in one task (phase 1) and application in the actual task (phase 2). The second goal is to apply this methodology in a real usability study aimed at analyzing the emotional content of the interaction with an ECA and being able to choose the interface with the more positive valence. In order to achieve these goals biosignals have been recorded and subjective ratings collected. Taking into account the problems mentioned above, our ambition was to try to classify emotional response only with respect to the valence of emotions and not with respect to arousal or distinction of various emotions.

The ECA interfaces used have been developed for the RASCALLI multi-agent architecture [11]. The agents can, via a multimodal system, interact with the user in various ways – text, speech, and gesture. The believability of the agents and the quality of the interaction are crucial for the usability of the platform as users are expected to interact with the agent as with a personal assistant for a long period of time. So, the emotions provoked by the ECA, if measured, can be a good basis for choosing or improving the interface and are thus a very important task during the usability evaluation.

More specifically, we apply the proposed methodology to compare users' emotional reactions towards two versions of the ECAs developed for the RASCALLI platform [11].

2 Physiological Data in the Study of Emotions

User satisfaction questionnaires are self-report measures that are administered at the end of the interaction session. As such, they have two possible disadvantages [3], [4]. First, sometimes it is not possible to consciously recollect the exact emotional experience during the interaction. Second, the questionnaires provide measures only at the end of the interaction – so we have only one measure for the whole interaction process.

To overcome these drawbacks, physiological recordings were used. Their first advantage is that they provide continuous measure of the experience emotions – so we can study the emotions over time. Second, by using such measures, we can have access not only to consciously experienced emotions.

There is a huge body of research that uses biosignals to study emotions. Some of them try to find several physiological features that can differentiate between different emotions. This line of research encountered serious problems, so now the efforts have been redirected to finding patterns of many features, extracted from various biosignals, in order to study emotions. Attempts to classify emotions on the basis of biosignals are performed applying different statistical methods for feature extraction, feature combination, dimensional projection, etc. [3], [7], [8], [10], [12], [13].

Emotion studies using biosignals have demonstrated that there are huge inter-individual differences, evinced by the fact that one and the same emotion can lead to different patterns of physiological reactions. And it is very difficult to find patterns of features (or even classifiers) that work at the group level [8].

3 Design of the Study

The testing comprises two major phases. During the first phase, called the **calibration phase**, subjects are presented with stimuli with known normative ratings of emotional impact. The stimuli belong to several different emotional categories. While users were watching the pictures with a task to get immersed in the emotion related to each picture their biosignals were recorded. We tried to find individual physiological patterns that differentiate well between different emotional categories for each subject. Using these data we tried to assess the emotional reactions of the user in the second phase.

In the second phase, called the **main phase**, we studied the emotional reaction of the users while interacting with the embodied conversational agent. The interaction was simulated by showing video clips of the agent engaging in conversation and answering questions. The users' task was to watch the videos and try to imagine that they were actually interacting with the agent. During this second phase the same biosignals as the ones in the previous phase were recorded.

3.1. Stimuli and Procedure: Calibration Phase

During this session users saw several pictures with emotional content. These pictures have been chosen from the International Affective Pictures System (IAPS) [14] in order to have standardized measures of the valence of the elicited emotions. We used 32 pictures from the IAPS – 16 pictures with average valence ratings defining the 'neutral' condition, 8 pictures with high positive valence ratings defining the 'positive' condition, and 8 pictures with high negative valence ratings defining the 'negative' condition. The picture numbers in IAPS and their valence and arousal ratings from the database are presented in Table 1.

Pictures were arranged in blocks of 4 pictures with the same valence (negative, neutral, or positive). Positive and negative blocks were separated by a neutral block in order to have a more clear distinction in the physiological data.

The study began after users gave their informed consent and were acquainted with the instructions. Each picture was presented for 20 seconds. The participants looked at the picture with a task to imagine the depicted emotion and to try to experience it. After the 20 s interval, two 5-point rating scales appeared on the screen one after the other: the first for rating the valence of the emotion (from 1 = 'completely negative' to 5 = 'completely positive') and the second for rating the arousal (from 1 = 'very low' to 5 = 'very high'). As a graphical representation of the scales we used the SAM scales [14].

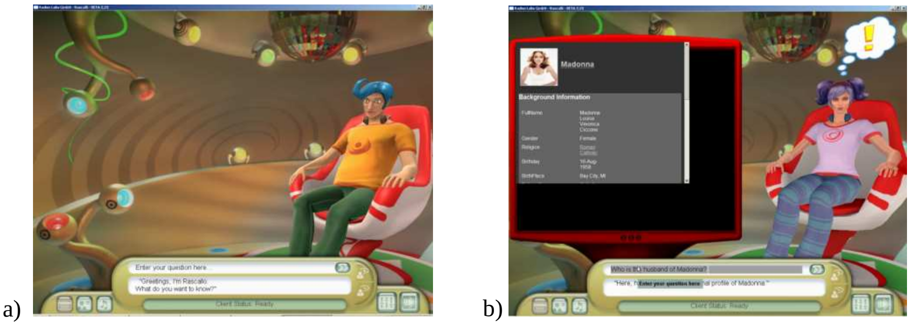
Ratings were used for two purposes. First, in the case of very big discrepancies between the normative and the subjective ratings, we could exclude the data from further analysis. Second, subjective ratings were made available for comparison and consistency checks (e.g. to what extent participants gave ratings consistent with the IAPS ones). Moreover, in such a way the participants are more motivated to follow the instructions to experience the emotions.

Table 1. Pictures used in the calibration phase. Standardized ratings from the IAPS database [14] are presented. Valence is rated on a 9 point scale (1 = ‘extremely negative’ to 9=’extremely positive’) and arousal is rated on a 9 point scale (1 = ’extremely calm’ to 9 = ‘extremely aroused’).

Condition	IAPS mean rating on 'valence' dimension	IAPS mean rating on 'arousal' dimension	Pictures used
Negative	2,35	6,42	2095; 2352,2; 2683; 2730; 6313; 6550; 8485; 9050
Positive	7,66	5,63	1710; 2216; 2345; 2352,1; 5623; 5833; 7502; 8210
Neutral	5,00	2,65	2102; 2235; 2440; 2575; 7000; 7004; 7010; 7020; 7035; 7059; 7080; 7175; 7217; 7224; 7235; 7491

3.2. Stimuli and Procedure: Main Phase

As mentioned above, the testing was performed on the agent interfaces, developed in the RASCALLI multi-agent architecture [11]. In the RASCALLI platform, the agents can interact with the user in various ways – using text, speech, and gestures. The agent is a personal assistant, acting in a scenario where it assists the human user in gathering and organizing information related to music.



"Greetings, I'm Rascalli. What do you want to know?"

Who is th] husband of Madonna?

"Here, h Enter your question here al profile of Madonna."

Fig. 1. Two versions of the agent are tested during the study: a) male and b) female agent.

As stimuli in this phase, we used videos representing interactions between a user and the Rascalli ECA. We used two versions of the agent: male and female (see Fig. 1). For each agent, six videos representing interactions with duration 20-30 s each were used. The first video shows the agent just sitting when the user enters the system. The next 5 videos had the following structure: the user asks a question, the agent answers, the user attends to the answer given. In three of these videos the agent answers correctly. In two of them the agent answers incorrectly. The agent answers by text (in the dialogue bar) and by voice. The agent also uses various gestures to emphasize something or to point at information presented on the screen. The questions and answers were the same for the male and the female agent. The use of video clips instead of real interactions allowed for a more controlled environment with the same behavior of the agent for all users.

Each user watched 6 videos with either the male or the female agent. The users had the instruction while watching the videos to imagine they were interacting with the system. After the end of each video, the participants had to rate the experienced emotions on the same two rating scales described above – the first for rating the valence of the emotion and the second for rating the arousal. Again, as a graphical representation of the scales we used the SAM scales [14].

While users were watching the videos, the biosignals used in the calibration phase and described in Section 3.3, were recorded.

3.3. Biosignals Used in the Study

The following biosignals were recorded during the calibration and the main phases: electrocardiogram (ECG), photoplethysmogram (PPG), galvanic skin reaction (GSR), and electromyogram (EMG).

ECG was recorded using the Biopac's ECG100 amplifier. Two LEAD110S electrode leads with EL503 electrodes were attached to the participant’s left and right hand. ECG is recorded with sampling rate of 200 sample/s.

PPG was recorded using the Biopac's PPG100C amplifier with the TSD200 transducer. The TSD200 consists of an infrared source and photo diode, which records the infrared reflectance thus reflecting changed resulting from varying blood flow. The PPG signal is recorded with sampling rate of 200 sample/s.

GSR was recorded using Biopac GSR100C amplifier. The GSR100C uses a constant voltage (0.5 V) technique to measure skin conductance. The GSR100C amplifier was connected one set of the TSD203 Ag-AgCl, unpolarizable, finger electrodes. GSR is recorded with sampling rate of 200 sample/s.

Facial EMG was measured over Corrugator Supercilii and Zygomaticus Major using surface electrodes using the scheme suggested by [15]. The EMG was recorded using surface electrodes with sampling rate of 250 samples/sec.

3.4. Participants

19 participants, 7 men and 12 women (age varied between 20 to 45) were invited as test users. Two of the users were not able to follow the instructions in the calibration phase, so their data were excluded from further analysis. From the remaining 17 participants, 5 were male and 12 – female.

4 Analysis of Physiological Data: Calibration Phase

The data collected during the preliminary session are used in order to extract patterns of physiological responses for emotions with different valence. As stated above, we were interested in assessing the overall interaction with the agent and we didn't try to identify specific emotions but just to identify the emotion valence.

As large inter-individual differences were expected, the analysis was performed for each user separately. As a first step in the analysis we used the AuBT software [16**Error! Reference source not found.**] to extract features from the 5 bio-signals recorded. From the ECG signal 80 features were extracted, from the PPG signal – 67 features, from the SGR signal – 19 features, and 21 features are extracted from each of the EMG signals. The complete list of the features extracted could be found in [16**Error! Reference source not found.**]. This is a large number of features to deal with especially taking into account the fact that for each user we have only 32 stimuli for which physiological data were recorded. To cope with this problem we used Fisher transformation as a dimension reduction technique. In such a way we obtained a reduced representation of the high-dimensional data set of physiological features. As pointed in [17], using the reduced data representation, classification procedures can very often give better results. Fisher transformation looks for a data representation that distinguishes between different categories. The goal was to find data projection on a low-dimensional space where the classes are well separated [18].

After applying the Fisher transformation, we obtained projection weights for each of the dimensions (respectively for each feature extracted). Using the projection weights, we were able to project the data from the pictures in a 2-dimensional space and this should be a dimension reduction that distinguishes between three emotion categories that we used. An example result (for an individual user) is presented in Figure 2.

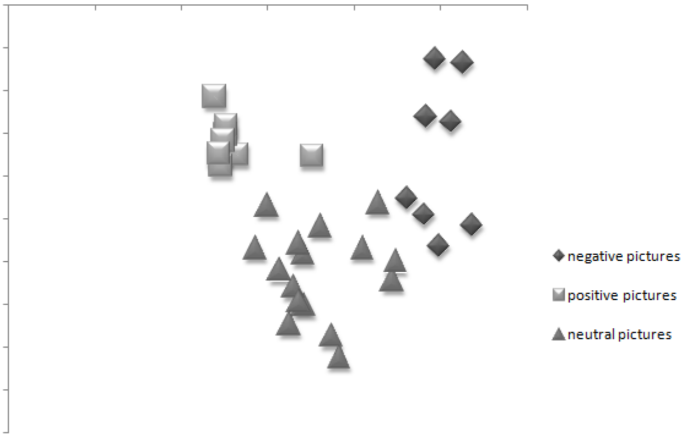


Fig. 2. Fisher projection of physiological data for three classes of emotional pictures used – negative, positive, and neutral (the presented data are for a single user).

The projection weights are different for each user and because of this the projection dimensions are different for each user. We did not try to find if these dimensions correspond to some psychological constructs. We used them as a basis for representation of the physiological data and which basis can be used next to classify new data during the main phase (physiological signals recorded during the video clips with interactions with the agents).

To verify the quality of classification achieved, we use LDA (linear discriminant analysis). We assess how well the pictures are classified (as a result of the Fisher transformation applied) in three categories corresponding to the categories they belong to. The percentage of correctly classified pictures using LDA for all participants is 99.3% correct for the negative pictures, 97.4% for the positive pictures, and 93.4 % for the neutral pictures (96.9% on average). The discriminant functions for each user are saved for further use in the next step (see section 5).

5 Classification of Biosignals During the Main Phase

Biosignals recorded during the main phase are analyzed again individually for each user. First, for each video we extracted the same features from the physiological signals recorded (the same features extracted in the calibration phase). Then we used the projections weights from the Fisher transformation used for the pictures' data to be able to represent the new stimuli in the same two-dimensional space. As was stressed in Section 4, the analysis is carried on separately for each individual; e.g. different projection weights are extracted and used for each user. As a result, for each user we are able to project both pictures' and videos' data in a 2-dimensional space. The result of this procedure for one participant is presented in Fig. 3.

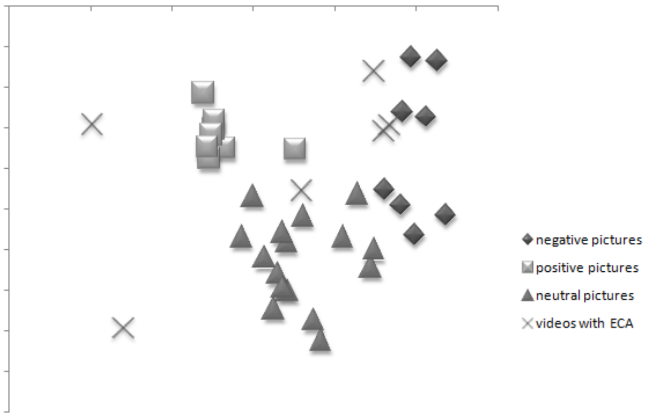


Fig. 3. Fisher projection of physiological data for three classes of emotional pictures used (negative, positive, and neutral) and for the ECA tested (the presented data are for a single user).

Next, using the LDA functions from the calibration phase (see section 4), we classify the emotions experienced during each video. Using these functions we can classify each user's experience in each video as positive, negative, or neutral. Summarized data for all users are presented in Table 2.

Table 2. Classification of videos using physiological data and LDA functions. Data are in percentages over all users and all videos for the corresponding agent (male or female).

Agent	**Classification of videos based on biosignals**		
	Negative (%)	Neutral (%)	Positive (%)
Female	37,0	31,5	31,5
Male	31,9	55,3	12,8
Total	34,7	42,6	22,8

The emotional reactions from videos with the female agent are classified as negative, neutral, and positive in roughly equal proportions (37,0 %, 31,5 %, and 31,5 %, respectively). The pattern in the classification of the videos involving the male agent is different: negative – 31,9%, neutral – 55,3 %, and negative – 12,8%.

To assess the quality of classification, we compared the classifications based on physiological data with those based on valence subjective ratings. Subjective ratings are made using the scale: 1 = 'completely negative', 2 = 'somewhat negative', 3 = 'neutral', 4 = 'somewhat positive', and 5 = 'completely positive'. For the purpose of the present analysis ratings '1' and '2' are coded as 'negative', rating '3' is coded as 'neutral', and ratings '4' and '5' are coded as 'positive'. On the basis of these subjective ratings, 32 % of the videos are classified as 'negative', 28 % – as 'neutral', and 40 % – as 'positive'.

Comparison between these two classification methods is shown in Table 3. In total, 36,6 % of the videos are classified with one and the same label using the two classification methods.

Table 3. Comparison between the classification of videos on the basis of the subjective ratings and classification based on the physiological data.

Classification based on subjective ratings	**Classification of videos based on biosignals**		
	Negative (%)	Neutral (%)	Positive (%)
Negative	**34,4**	50,1	15,6
Neutral	35,7	**46,4**	17,8
Positive	34,1	34,2	**31,7**

As seen from the data in Table 3, the match between rating-based evaluation and physiological data based is the largest, about one half, for classification 'neutral' (46,4 %) while the match is only about one third of the videos classified either as 'negative' or 'positive' (34,4 % and 31,7%, respectively). As discussed above, the stimuli were not supposed to elicit strong emotions and tend to be quite neutral and this can explain 'low' physiological response and higher error in the assignment. This match is quite low and suggests that the results obtained from biosignal classification can be considered only as a tendency. In order to assess the reliability of mixing physiological data with subjective ratings, more elaborate experiments are needed with careful choice of the basis set of picture from IAPS set and testing stimuli with much more clearer valence of the emotional response, than the one chosen in this study.

6. Summary and Future Work

In this paper, we proposed a two-phase procedure, based on biosignals, aimed at classifying users' emotional experience while interacting with embodied conversational agents.

In the first phase of the procedure images from IAPS, standardized with respect to emotional response, were shown to participants. In the second phase, participants were shown a sequence of interaction episodes with the ECA. Biosignals were recorded during both phases. Additionally, subjective ratings have been gathered using a 5 point scale (form negative to positive).

Our more general goal was to test the applicability of this procedure, viewed as a possible tool in usability studies, and check to the possibility to do automatic emotion valence assessment during human-agent interaction given up-to-date signal processing and classifier techniques. Our specific goal was to assess and differentiate several versions of ECAs digital characters and choosing those eliciting more positive response.

The results confirm the general finding in the field of automatic emotion recognition that distinguishing emotional responses even with respect only to valence is a difficult task. Especially in our settings in which emotion recognition from one task (looking at pictures) had to be used in another task (interacting with an agent). Despite these difficulties, at least for part of the subjects, the method generated results which allow to discern the general tendencies as far as valence is concerned. The comparison with the data from explicit rating of emotion valence while quite consistent for the IAPS images, was not satisfactory for the interaction episodes. This situation is of course related partly to the complexity of the latter situations and to the implication of a lot of additional factors in the rating process (wishful thinking, reasoning, etc.). Moreover, many questions can be asked about the relation between biosignal based data and subjective ratings. One such question, in the first place, is about the limitations of using subjective ratings for training classifiers for physiological signals.

Part of these questions will be explored in future work together with further elaboration of the methodology with respect to training material, correlation between subjective ratings and biosignal analyses, etc. Moreover, on the usability evaluation side, the comparison between ECA with emotional expressions and gestures and 'simpler' versions, as the ones presented here, will be investigated. The latter will be done in combination with eye-tracking studies in order to attempt to assess more precisely the reasons for changes in emotional response.

Acknowledgments. The work described here has been supported by the EC FP6 project RASCALLI. We would like to thank the AuBT team (University of Augsburg) and especially Johannes Wagner, for making available the AuBT program for data analysis and the help and advices concerning our study. We want to thank Alexander Gerganov and Ivo Popivanov for their great help in the lab setup and physiological recordings, Gergana Kuzmova for her help in the preparation of the calibration phase of the study, and Dimiter Atanasov for his advices about the data analysis.

References

1. Bradley, M.: Emotion and motivation. In: John T. Cacioppo, J, Tassinary, L. G., Berntson, G., Handbook of Psychophysiology, Cambridge University Press (2000)
2. Levenson, R.: Autonomic nervous system differences among emotions. Psychological science, vol. 3, no. 1 (1992)
3. Mandryk, R., Atkins, M.: A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human Computer Studies, vol.65 (4), 329-347 (2006)
4. Benedek, J., Hazlett, R.: Incorporating facial EMG emotion measures as feedback in the software design process. In Proc. Human Computer Interaction Consortium (2005)
5. Ward, R., Mardseen, P.: Psychological responses to different WEB page designs. International Journal of Human-Computer Studies, 59, 199-212 (2003).
6. Wilson, G., Sasse, M.: Do users always know what's good for them? Utilising physiological responses to assess media quality. In McDonald, S., Waern, Y., and Cockton, G. [Eds.]: People and Computers XIV - Usability or Else! Proceedings of HCI 2000 (September 5th - 8th, Sunderland, UK). Springer (2000)
7. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: First steps towards an automatic system. LNCS 3068: 36–48 (2004)
8. Kim, J., André, E.: Emotion Recognition Based on Physiological Changes in Listening Music, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30 (12), pp. 2067-2083 (2008)
9. Wagner, J., Kim, J. and André, A.: From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification. In IEEE International Conference on Multimedia & Expo (2005)
10. Nasoz, F., Alvarez, K., Lisetti, C., Finkelstein, N.: Emotion recognition from physiological signals for presence technologies. International Journal of Cognition, Technology, and Work – Special Issue on Presence, vol. 6 (1) (2003)
11. Krenn, B.: RASCALLI. Responsive Artificial Situated Cognitive Agents Living and Learning on the Internet, in Proc. of the International Conference on Cognitive Systems/University of Karlsruhe, Karlsruhe, Germany, April 2 – 4 (2008)
12. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: analysis of affectivephysiological state. Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 23 (10), 1175 – 1191 (2001)
13. Christie, I.C., & Friedman, B.H.: Autonomic specificity of discrete emotion and dimensions of affective space: A multivariate approach. International Journal of Psychophysiology, 51, 143-153 (2004)
14. Lang, P.J., Bradley, M.M., & Cuthbert, B.N.: International affective picture system (IAPS): Digitized photographs, instruction manual and affective ratings. Technical Report A-6. University of Florida, Gainesville, FL (2005)

15. Tassinary, L., Cacioppo, J., Geen, T.: A psychometric study of surface electrode placements for facial electromyographic recording: I. The brow and cheek muscle regions. Psychophysiology, 26, 1, 1-16 (1989)
16. Wagner, J.: Augsburg Biosignal Toolbox (AuBT): User Guide (2005)
17. Cunningham, P.: Dimension reduction . Technical report UCD-CSI-2007-7, August 8th, 2007, University College Dublin (2007)
18. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press, Inc, (1990)