COMBINING MIMIC AND PROSODIC ANALYSES FOR USER DISPOSITION CLASSIFICATION

Ronald Böck[*] *Kerstin Limbrecht*[†] *Ingo Siegert*[*] *Stefan Glüge*[*]
Steffen Walter[†] *Andreas Wendemuth*[*]

[*]*Chair of Cognitive Systems, Otto von Guericke University Magdeburg, Germany*
[†]*Medical Psychology, Ulm University, Germany*
ronald.boeck@ovgu.de

Abstract: Automatic classification of the users' internal affective and emotional states is to be considered for many applications, ranging from organisational tasks to health care. To develop automatic technical systems suitable training material is necessary and an appropriate adaptation towards users is needed. In this work, we present preliminary but promising results of our research focusing on emotion classification by visual and audio signals. This is related to a semi-automatic and cross-modality labelling of data sets which will help to establish a kind of ground truth for labels in the adaptation process of classifiers. In our experiments we showed that prosodic features, especially, higher order ones like formant's three bandwidth are related to visual/mimic expressions.

1 Introduction

Emotions, dispositions, and feelings are substantial elements of daily life as they are able to interrupt ongoing activities as well as they influence communication and interactions and can induce the willingness to act in a specific way. Humans are able to analyse specific cues (mimic, gesture, speech, physiological reactions, etc.) in order to interpret emotional states in themselves and others. The interpretation of emotional states is influenced by a mirroring process, i.e. a user projects the emotions onto another person as he is believing the counterpart is reacting. Modern digital technical systems increasingly occupy a wide range of daily activities. They are used for organisational tasks, navigation systems, calendar synchronization, entertainment, etc. Researchers intend to optimize the usability of such cognitive-technical systems [21] in such a way that they will provide people not only with helpful information, but also to support them during their decision making processes, and communicate their intentions on a social level to their users. To achieve this goal, interfaces are needed that analyse the user's emotional feedback and respond without any difficulty towards the complex demands of human communication. Further, technical systems have to identify emotional cues sufficiently/properly during human-computer interaction (HCI) and discriminate these to interpret humans' intentions and needs [21]. As emotions are expressed by verbal and non-verbal speech, speech content, facial expression, and gesture the focused interface has to consider a large amount of data [9]. Emotion assessment is often carried out through analysis of users' mimic, audio signals, and/or physiological signals [8, 22]. And most research paradigms so far focus only on visual or auditory human emotion detection. The research community is on the way to use multi-modal feature sets to investigate the user's behaviour in an appropriate way [22, 19, 6].

In this paper our main focus is on describing the effects received by integrating information of visual and audio material in emotion classification in HCI. In recent years an explosion

of user-generated multimedia data can be observed, therefore, a strong need for efficient tagging/labelling strategies is described [9]. In contrast to labour-intensive manual annotation commonly used, this approach may offer the possibility of implicit tagging without the attention of the user. In our research, we are looking for a semi-automatic annotation of data sets which provides us with a pre-classification of multi-modal data. This is inspired by forced alignment in speech processing, which is an automatic approach to process huge data sets of audio recordings to get an annotation with less human effort. In automatic detection, classification, and annotation of emotions, classifiers can be trained and good results can be achieved [15]. So far, it is the case for acted, pre-annotated material. Classification results get worse in case of naïve material [13]. In general, better results can be achieved with personalised classifiers [1]. Therefore, and also to access huge data sets, we need a ground truth to train proper classifiers. Thus, we propose a multi-modal approach to explore the material. In particular, it is to apply Facial Action Coding System (FACS) coded video sequences to label audio material and the other way around, which is the more interesting case. Especially, FACS coding is quite time consuming and therefore, a semi-automatic pre-classification is helpful. The main idea of the research is to use audio analyses to identify relevant sequences in the video material and provide a pre-classification for FACS (cf. Fig. 1). Thus, human annotators are just asked to label debatable sequences. The framework was influenced by [3] where mimicry cues, e.g. poses, were to relate with verbal utterances.

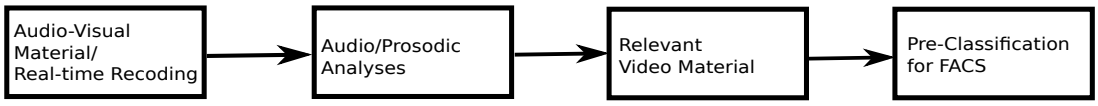


Figure 1 - Workflow to establish a pre-classification of video material.

In this paper, we present results of a preliminary study which uses FACS coded sequences to analyse audio material. It is a first step to generate a framework by finding relations between known FACS labels and surrounding audio material and afterwards use audio analyses to identify relevant video sequences for FACS classification. So far, both analysing process was done manually as a case study.

2 Data Set

2.1 Experimental Overview

The simulation of the natural verbal human-computer interaction was implemented as a Wizard-of-Oz (WoZ) experiment. This kind of experiment allows the simulation of computer's or system's properties in a manner, such that subjects have the impression that they have a completely natural verbal interaction with a computer-based memory trainer. The design of the trainer followed the principle of the popular game "Concentration". The variation of the system behaviour in response to the subjects was implemented via natural spoken language, with parts of the subject's reactions taken automatically into account. However, the system does not work with automatic speech recognition and response control, but is controlled by an experimenter in an adjoining room. This method allows simulating natural interactions that permit systematic variation of different emotional states. The procedure of emotion induction included differentiated experimental sequences during which the user passed through specific pleasure/arousal/dominance (PAD) [10] octants in a controlled fashion. The workflow of this experiment is given in Fig. 2 where each Experimental Sequence (ES) represents a certain octant in the PAD space. In our experiment, we investigate only ES-2 that is assumed to be positive and ES-5 which is negative. For a detailed description of the experimental setup we refer to [19].

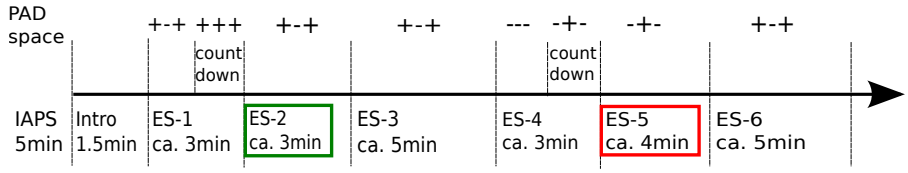


Figure 2 - Workflow of EmoRec-Woz I experiment.

Table 1 - Basic emotions combined by Action Unit where a number represents the muscle activity and a letter the strength of it (A . . . low; E . . . high).

Emotion	Action Unit Combination
Happiness	[6 and/or 7] with 12CDE
Sadness	1 + 4 + [6 and/or 7] + 15ABC + 64
Disgust	9 + [10 and/or 16] + 19 + 26
Anger	4CDE + 5CDE + 7CDE + 17 + 23 + 24
Fear	1 + 2 + 4 + 5ABCDE + 7 + 20ABCDE + 26
Surprise	1CDE + 2CDE + 5AB + 26

2.2 Subjects

In general, 125 subjects participated on the experiment and hence, on the corpus called EmoRec-WOZ I+II [19]. For this case study we analysed the video material of a subset of 20 subjects (EmoRec-WOZ I) which is so far manually FACS coded. The subset is composed by 10 women (5 of them were younger [M = 27.6] and 5 were older [M = 46.4]; split-half: 40 years) and 10 men (5 of them were younger [M = 25.2] and 5 were older [M = 56.2]; split-half: 40 years). The subjects received an expense allowance.

3 Methods

3.1 Facial Action Coding System

Observational coding systems, provide the possibility to identify facial expressions. The annotation is not automatically associated to an emotion. Since facial expressions can be defined as a sequential set of facial movements caused by the underlying muscle activities, Ekman and Friesen [5] developed the Facial Action Coding System (FACS) to code these patterns. The facial expressions created by muscles are defined by (a set of) Action Units (AU). Overall 40 AUs describe nearly all possible facial movements in an objective way. The stimuli used in this experiment were selected based on EmFACS [4]. Facial expressions are probably the most informative emotion signals as they are easy to observe and well-known for their relevance in human interactions. In Tab. 1 an overview of the AU combinations indicating muscular activity in different parts of the face are presented. The numbers indicate the part of the face in which the muscular activity occurs, whereas characters refer to the intensity of the expression.

As it was intended to classify emotions of positive and negative valence, facial expressions occurring during the interaction with the technical system were checked for AUs indicating expressions of happiness and anger. The analysis revealed that only for a selection of AUs enough occurrences were available. AU12 (smiling) was selected to indicate positive valence as well as in combination with AU17/23/24 (AU17: chin raiser, AU23: lip tightener, AU24: lip presser). All facial expressions were coded by a certified FACS coder not involved in the experiment. It can be assumed that facial activity is less expressive in HCI as no additional

value for the interaction partner (in this case a technical system) is expected [20].

3.2 Prosodic Features

According to the FACS analysis we investigated prosodic features extracted from the audio recordings. The procedure to get the speech samples was as follows: According to the manually FACS labelled video parts speech snippets from the audio stream were cut. As we are interested in a correlation, which is also biologically plausible, we used only these samples that ended two seconds before the mimic starts at the latest (cf. Fig. 3). Thus, we assume that the subject is in the same emotional state for the mimic and acoustic expression, since both characteristics are close together in time.

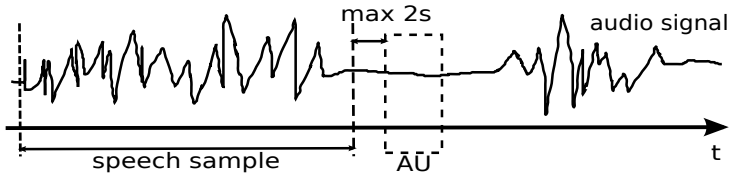


Figure 3 - Selection of audio material according to FACS coded AUs.

As prosodic features we applied the first to third formant and their corresponding bandwidth, pitch, intensity, and jitter to the audio material. For this, we applied PRAAT [2] to all speech samples. Especially, formants, pitch, and intensity are quite expressive in emotion analyses [11, 14, 15, 18]. So far, such analytical descriptions were mainly done on vowels. In contrast to these investigations, we applied the methods to a word sequence with two parts, which is a letter and a number, e.g. “C 2”, “C 4”, “A 1”. Such a combination results directly from the rules of the “Concentration” game. Moreover, Scherer [11] discussed the characteristics of prosodic features in correlation with appraisals. Up to now, it is unclear whether appraisals in HCI really occur or in which strength. We know from [11] that as AUs in FACS appraisals just appear in groups to get a meaningful interpretation. In addition to formants and their bandwidth, we further investigated pitch, intensity, and jitter. From the literature, [16, 15, 3], we know that such features represent or are related to negative and high aroused emotions.

4 Results

In this paper we present the results of a case study which is based on a set of 19 subjects. Originally, in total we had 20 subjects in the set, but for one only positive facial expressions were annotated. The material was labelled and the audio samples were cut manually; 125 samples in total (ES-2: 55, ES-5: 70). The results, which are given in Fig. 4 and Fig. 5, are the mean of all samples of a subject combined as an average of all subjects. Hence, in general, the length of an utterance is not suitable to extract information whether a sample belongs to ES-2 or ES-5 because of a large variation in the sample’s duration (ES-2: $\mu = 1.24s$ ($\sigma = 0.86s$); ES-5: $\mu = 1.32s$ ($\sigma = 0.738s$)).

In Fig. 4 the formants and the corresponding bandwidths are given. In fact and in accordance with [11, 18] we got a shift in the formants, especially, in the parts related to phoneme representations of vowels. This is the same with the bandwidths. However, this shift is complementary to the expected one, i.e., the more negative a word is uttered, the more the discussed features should be shifted towards higher frequencies. Except the second formant Fig. 4(c), ES-5 has lower frequencies. From our point of view, this is related to the characteristics of HCI, i.e., emotions were not expressed in such an extent as in human-human communication. Nevertheless,

and this is important for our purpose, formant 2 and its bandwidth as well as the bandwidth of formant 3, in particular, gave indications to separate ES-2 and ES-5. So far, usually formant's 3 bandwidth was not observed in analyses.

The expected shift of frequencies could be seen in pitch (cf. Fig. 5(b)). Moreover, the graph shows also the two parts of the word group, e.g. "C 4". In general, pitch is a good indicator to distinguish the two Experimental Sessions, especially in the beginning of an utterance.

In contrast to literature [15], with intensity a separation is not possible (cf. Fig. 5(a)). This is quite interesting as, in general, the intensity raises in negative utterances. In the data set at hand there is almost no difference in the emotions. This is the same with jitter, since it is usually related to emotions which cause tremble in voice like fear. We assume that such an emotion or reaction is unlikely in HCI.

5 Conclusion

In our study, we explored mimic and audio data derived from a natural interaction with a technical system. None of the expressions was posed. Results show differences concerning positive ES-2 and negative ES-5, especially in higher order prosodic features. From the results we see and expect that pitch, formant 2, and the bandwidth of formant 3 are usefull to distinguish ES-2 and ES-5 in this context. A shift in the formants could be found as well as in bandwidth, but unexpectedly, results tend contrarily to the supposed direction. Two reasons may be true: humans express emotions different in a communication with a human than with a technical system. Additionally, the interaction with the technical system in this study offered a high amount of standardisation, but therefore there was no speech apart of the commands given by the subjects. Hence, this contrary behaviour of formants will be in focus of our further research. Additionally, from our previous research [1] we found that in HCI the expression of basic emotions is not very common. Especially strongly expressed emotions are seldom shown in such interactions. For this, the intention of research has to be led towards an adapted and personalised scheme of classification. The authors' will focus on frustration as this emotion is supposed to be shown quite often in an interaction with a technical system [17, 12, 7]. It is intended to consider these aspects in further research with the complete data set EmoRec-WOZ I+II.

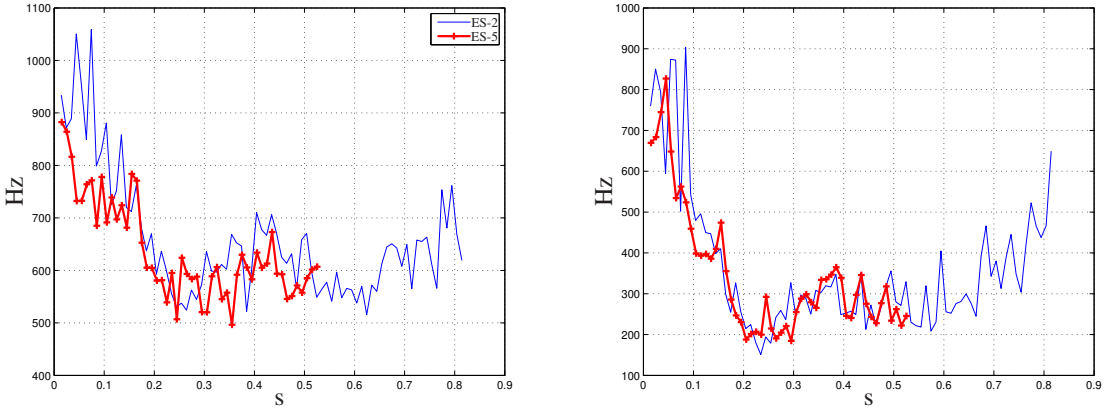
6 Acknowledgement

We acknowledge continued support by the Transregional Collaborative Research Centre SFB/-TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). We also acknowledge the DFG for financing our computing cluster used for parts of this work.

References

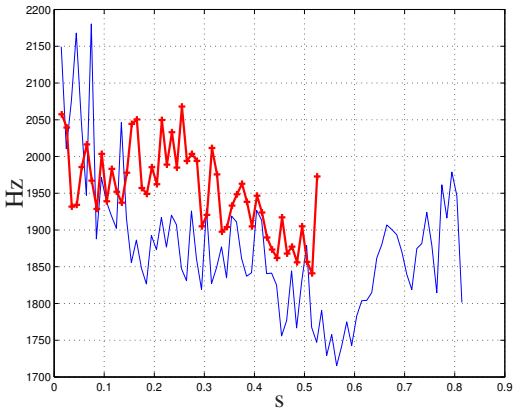
[1] BÖCK, R. ; GLÜGE, S. ; WENDEMUTH, A. ; LIMBRECHT, K. ; WALTER, S. ; HRABAL, D. ; TRAUE, H. : Intraindividual and interindividual multimodal emotion analyses in human-machine-interaction. In: *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, CogSIMA 2012*, 2012, S. 59–64

[2] BOERSMA, P. ; WEENINK, D. : Praat: Doing Phonetics by Computer. (2011)

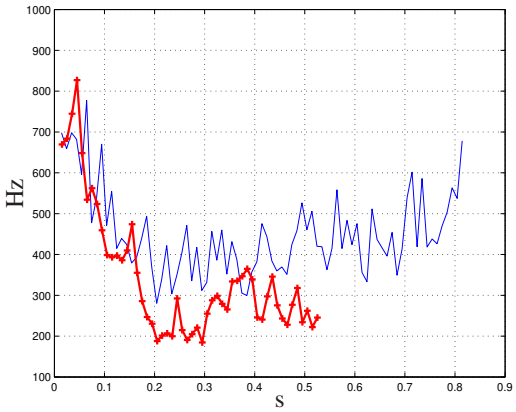


(a) Formant 1 (mean)

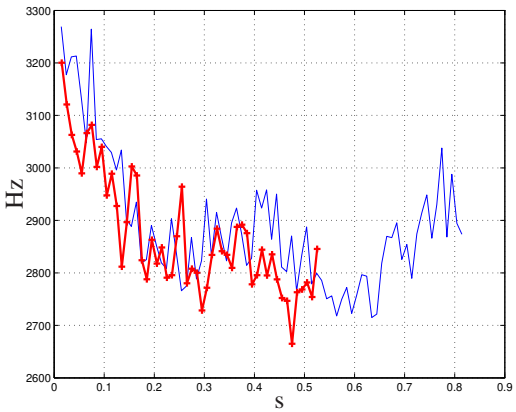
(b) Formant 1 Bandwidth (mean)



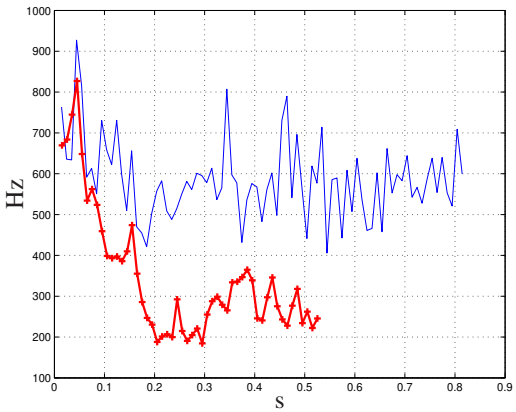
(c) Formant 2 (mean)



(d) Formant 2 Bandwidth (mean)

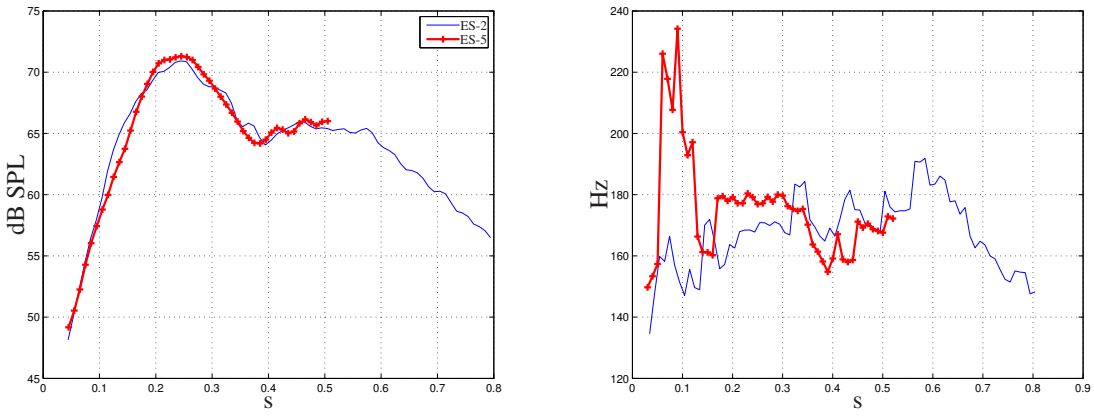


(e) Formant 3 (mean)



(f) Formant 3 Bandwidth (mean)

Figure 4 - Global means of formants 1 to 3 ((a), (c), (e)) and corresponding bandwith ((b), (d), (f)); blue curve presents ES-2 and red with + corresponds to ES-5.



(a) Intensity (b) Pitch

Figure 5 - Global means of intensity (a) and pitch (b); blue curve presents ES-2 and red with + corresponds to ES-5.

[3] DE LOOZE, C. ; OERTEL, C. ; RAUZY, S. ; CAMPBELL, N. : Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In: *17th International Congress of Phonetic Sciences*, 2011

[4] EKMAN, P. ; FRIESEN, W. V.: EmFACS. In: *Unpublished Document. San Franciso, USA* (1982)

[5] EKMAN, P. ; FRIESEN, W. : *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, Palo Alto, CA, 1978

[6] GLODEK, M. ; TSCHECHNE, S. ; LAYHER, G. ; SCHELS, M. ; BROSCH, T. ; SCHERER, S. ; KÄCHELE, M. ; SCHMIDT, M. ; NEUMANN, H. ; PALM, G. ; SCHWENKER, F. : Multiple classifier systems for the classification of audio-visual emotional states, 2011 (Lecture Notes in Computer Science PART 2), S. 359–368

[7] HOQUE, M. ; PICARD, R. : Acted vs . natural frustration and delight : Many people smile in natural frustration. In: *Computer* 17 (2011), S. 354–359

[8] KIM, J. ; ANDRÉ, E. : Emotion Recognition Based on Physiological Changes in Listening Music. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (2008), Nr. 12, S. 2067–2083

[9] In: KOELSTRA, S. ; MUHL, C. ; PATRAS, I. : *EEG analysis for implicit tagging of video data*. IEEE, 2009, S. 1–6

[10] MEHRABIAN, A. : Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. In: *Current Psychology* 14 (1996), Nr. 4, S. 261–292

[11] SCHERER, K. R.: Appraisal considered as a process of multilevel sequential checking. In: *Appraisal processes in emotion: Theory, methods, research* (2001), S. 92–120

[12] SCHERER, K. R. ; ELLGRING, H. : Multimodal expression of emotion: Affect programs or componential appraisal patterns? In: *Emotion* 7 (2007), Nr. 1, S. 158–171

[13] SCHULLER, B. ; VALSTAR, M. ; EYBEN, F. ; MCKEOWN, G. ; COWIE, R. ; PANTIC, M. : AVEC 2011 - The first international audio/visual emotion challenge, 2011 (Lecture Notes in Computer Science PART 2), S. 415–424

[14] SCHULLER, B. ; VILLAR, R. ; RIGOLL, G. ; LANG, M. : Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* Bd. I, 2005, S. I325–I328

[15] SCHULLER, B. ; VLASENKO, B. ; EYBEN, F. ; RIGOLL, G. ; WENDEMUTH, A. : Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In: *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2009*. Merano, Italy, 2009, S. 552–557

[16] SCHULLER, B. ; VLASENKO, B. ; EYBEN, F. ; WOLLMER, M. ; STUHLSATZ, A. ; WENDEMUTH, A. ; RIGOLL, G. : Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. In: *IEEE Transactions on Affective Computing* I (2010), S. 119–131. – ISSN 1949–3045

[17] UNZ, D. C. ; SCHWAB, F. : Viewers viewed: Facial expression patterns while watching TV news. In: ANTOLLI, L. (Hrsg.) ; DUNCAN JR., S. (Hrsg.) ; MAGNUSSON, M. (Hrsg.) ; RIVA, G. (Hrsg.): *The hidden structure of interaction: From neurons to culture patterns*. IOS Press, 2005, S. 254–262

[18] VLASENKO, B. ; PHILIPPOU-HÜBNER, D. ; PRYLIPKO, D. ; BÖCK, R. ; SIEGERT, I. ; WENDEMUTH, A. : Vowels formants analysis allows straightforward detection of high arousal emotions. In: *2011 IEEE International Conference on Multimedia and Expo (ICME)*, 2011

[19] WALTER, S. ; SCHERER, S. ; SCHELS, M. ; GLODEK, M. ; HRABAL, D. ; SCHMIDT, M. ; BÖCK, R. ; LIMBRECHT, K. ; TRAUE, H. ; SCHWENKER, F. : Multimodal Emotion Classification in Naturalistic User Behavior. In: *Proc. of the 14th International Conference on Human-Computer Interaction*. Orlando, USA, 2011

[20] WALTER, S. ; WENDT, C. ; LIMBRECHT, K. ; GRUSS, S. ; TRAUE, H. : Comparison of subjectively experienced emotions and dispositions in man-machine and man-man scenarios. In: *In Proceedings of the 41th Conference of the Gesellschaft für Informatik*, 2011

[21] WENDEMUTH, A. ; BIUNDO, S. : A Companion Technology for Cognitive Technical Systems. In: *COST 2012 Conference "Cross Modal Analysis of Verbal and Nonverbal Communication"*. Dresden, Germany, 2011

[22] ZENG, Z. ; PANTIC, M. ; ROISMAN, G. I. ; HUANG, T. S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (2009), Nr. 1, S. 39–58