Audio-Based Pre-classification for Semi-automatic Facial Expression Coding

Ronald Böck[1], Kerstin Limbrecht-Ecklundt[2], Ingo Siegert[1],
Steffen Walter[2], and Andreas Wendemuth[1]

[1] Cognitive Systems Group, Otto von Guericke University Magdeburg, Universitätsplatz 2,
39106 Magdeburg, Germany
[2] Medical Psychology, Ulm University, Frauensteige 6, 89075 Ulm, Germany
ronald.boeck@ovgu.de
http://www.cognitive-systems-magdeburg.de

Abstract. The automatic classification of the users' internal affective and emotional states is nowadays to be considered for many applications, ranging from organisational tasks to health care. Developing suitable automatic technical systems, training material is necessary for an appropriate adaptation towards users. In this paper, we present a framework which reduces the manual effort in annotation of emotional states. Mainly it pre-selects video material containing facial expressions for a detailed coding according to the Facial Action Coding System based on audio features, namely prosodic and mel-frequency features. Further, we present results of first experiments which were conducted to give a proof-of-concept and to define the parameters for the classifier that is based on Hidden Markov Models. The experiments were done on the *EmoRec I* dataset.

1 Introduction

Dispositions and emotions are substantial elements of daily life as they influence the communication and way of interaction as well as they can induce the willingness to act in a specific way. Humans are able to analyse specific cues (facial expression, gesture, speech, etc.) in order to interpret emotional states in themselves and others, i.e. mainly generating hypotheses on how another person might feel or react. Modern technical systems increasingly occupy a wide range of daily activities like organisational tasks, calendar synchronisation, entertainment, etc. Therefore, researchers intend to optimise the usability of such cognitive, technical systems [22] in such a way that they will provide people not only with helpful information, but also to support them during their decision making processes. Hence, technical systems have to identify emotional cues properly during Human-Computer Interaction (HCI) [22].

Emotions are usually expressed by multiple modalities like verbal and paralinguistic speech expression, speech content, facial expressions, and gestures. Thus, a cognitive technical system has to consider a large amount of data [9]. In recent years, various emotion recognition methods and an enormous number of multimodal datasets were generated (cf. [8, 14, 24]) and therefore, a strong need for efficient labelling strategies arises [9, 16, 17]. In this paper, we are presenting first steps towards a semi-automatic annotation framework for multimodal datasets. The main idea is to use audio analyses to

M. Kurosu (Ed.): Human-Computer Interaction, Part V, HCII 2013, LNCS 8008, pp. 301–309, 2013.

identify relevant affective sequences which can be aligned with the corresponding video material and hence, provide a pre-classification for the Facial Action Coding System (FACS) (cf. Fig. 1). From the schematic workflow of the framework we can see that the recodings have to be multimodal, for the framework itself, but as well such modalities vary from user to user and from situation to situation. In fact, this is the case for almost all corpora which are currently recorded and we assume that this will be true for those which will be generated. Further, the utilised audio features should be relatively general, so that a wide range of domains and audio conditions are covered. This was investigated in parts in [1, 2]. The relevant video sequences can be afterwards determined by marking the time stamps of the video material; an idea influenced by forced aligment in speech recognition. The identification of the sequence further provides the opportunaty to pre-classify the FACS. Thus, human annotators are just asked to label debatable sequences which reduces the manual effort of annotation. A detailed visualisation of the pre-selection process is given in Fig. 2. The general idea was influenced by [5] where mimicry cues, e.g. poses, were related with verbal utterances. The audio features used here were previously identified in [1, 2]. So far, we do not intent to and are not able to specify a set of pre-selected Action Units (AUs) from audio analysis.

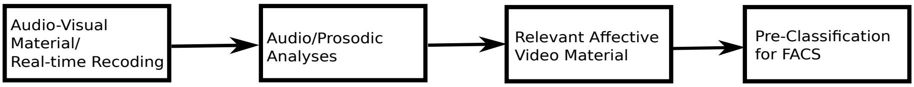
Audio-Visual Material/ Real-time Recoding
Audio/Prosodic Analyses
Relevant Affective Video Material
Pre-Classification for FACS

Fig. 1. Workflow to establish a pre-classification of video material as it is proposed

Certified human FACS coders achieve a hit rate of at least 76% when annotating facial expressions manually [10]. But unfortunately this is a very time consuming task. Therefore, researchers are very interested in developing a computer based approach for facial expression analysis. From literature, we know that there are systems which already deal with an automatic analysis of video material to support FACS coding, for instance [4, 12, 18]. In contrast to those systems our framework overcomes several disadvantages. Most systems fail if additional characteristics are in the face, for instance, glasses, fringe, or sensors. Further, video based facial analysis can just work if the face is oriented towards the camera. In natural HCI this is not the case all the time. Therefore, we apply a different modality, namely audio, to get a selection of relevant sequences, independent from the video features.

In this paper, we present the results for the automatic analyis of audio sequences to identify relevant parts related to FACS in the audio modality. As this paper presents a kind of proof-of-concept, we used material which is already manually annotated and thus, are able to evaluate our framework as well as the applied audio classifiers. Further, to show the generalisation of the framework we already applied the methods in a Leave-One-Speaker-Out (LOSO) validation (cf. Sect. 3.2 and 4). In future non-annotated material has to be processed; not considering any kind of evaluation.

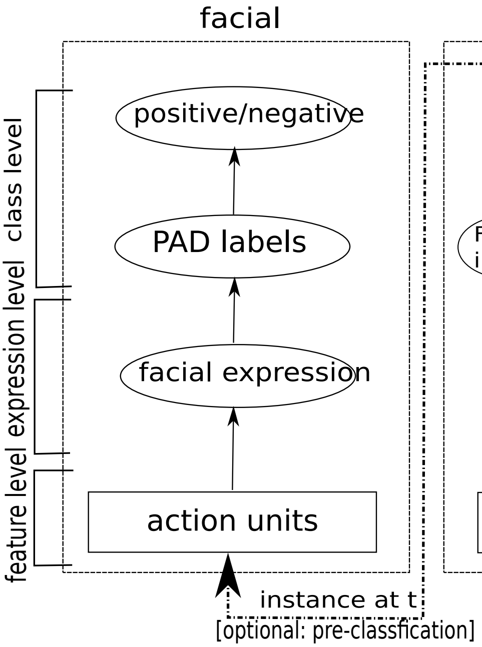

Fig. 2. Flow of the features to get a classification. The speech part presents the steps getting a classification based on audio features. Pre-selected instances are fed into the facial annotation part. Using Action Units and facial expressions a classification can be derived.

2 Dataset

To evaluate the framework, we applied it on the EmoRec I experiment, a realistic speech driven HCI introduced in [21]. The material was collected in a Wizard-of-Oz experiment which represents an interaction with a computer-based memory trainer relying on the design of the game “Concentration”. In this experimental setup the user was passed through specific pleasure-arousal-dominance (PAD) space [11] octants, representing the user’s emotions, in a controlled fashion. Although the emotional state was induced in the experiment by design, a prediction of the specific user reaction, no matter whether by way of interaction with the system or by emotional expressiveness, was not possible. Hence, we assume that the dataset represents a natural, realistic HCI. The experiment’s workflow is given in Fig. 3, where each Experimental Sequence (ES) represents certain octants in the PAD space. To see the differences in the user reactions, we investigated so far only ES-2 that is assumed to be positive and ES-5 which is negative that are interpretations of the PAD values. For a detailed description of the experiment and its parameters see [21].

In general, 125 subjects participated in the experiment. For this case study, we analysed the video material of the *EmoRec I* dataset - 20 subjects (ten female, ten male) - which is so far manually FACS coded. As not everybody showed facial expression in

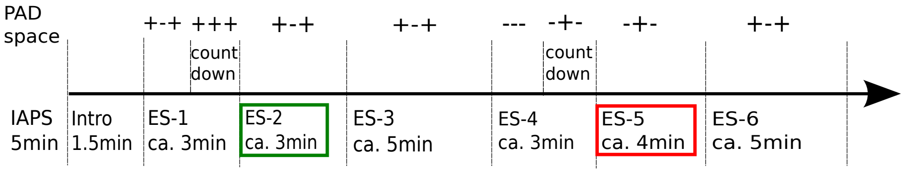

Fig. 3. Workflow of *EmoRec I* experiment

the ESs we reduced the number of subjects to 13 individuals who provided material for training and testing of classifiers. Despite performance loss, this enabled us to evaluate the framework's performance.

3 Methods

3.1 Facial Action Coding System

Ekman & Friesen [6] developed the FACS since facial expressions can be defined as a sequential set of facial movements caused by the underlying muscle activities. Facial expressions are defined by Action Units. The stimuli used in this *EmoRec I* experiment were selected based on EmFACS [7]. The manual labelling process for the 20 subjects took 70-80 hours in total. In particular, for the two ESs it took 13-19 hours for a video time of roughly $20 \cdot 7\text{min} = 140\text{min}$ (cf. Figure 3) which indicates the necessity of a semi-automatic annotation.

As it was intended to classify and identify positive and negative emotions, facial expressions occurring during the interaction with the technical system were checked for AUs indicating expressions of happiness and anger (cf. Table 1). The analysis revealed that especially negative emotions were shown most frequently in HCI. Therefore, we decided to identify negative reactions first. Moreover, this is feasible as for positive expressions almost no material was available, neither as facial expression nor as speech samples. Furthermore, only for a subset of AUs enough material occurred in the dataset. Hence, we opt for a combination of AUs and a selection of those which are mainly exhibited in the sequences. 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 [21].

3.2 Audio Features

Based on previous work [1], we selected prosodic features which are quite expressive in emotion analyses: the first to third formant and their corresponding bandwidth, pitch, jitter, and intensity [13, 14, 19]. Further, those features represent or are related to negative and high aroused emotions [5, 14, 15]. Moreover, we combined this feature set of prosodic features with Mel Frequency Cepstral Coefficients (MFCCs) to enrich the

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) (cf. [1])

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

set as a positive effect was figured out in preliminary tests. To handle time dependencies derivatives were added, too; namely Δ and $\Delta\Delta$. Those features are meaningful and widely used in the speech emotion recognition community (cf. e.g. [2, 15, 24]).

To extract the features we applied PRAAT [3] and the Hidden Markov Toolkit (HTK) [23] on frame-level to speech samples which are a combination of letters and numbers, for instance, “C 2”, “C 4” representing the commands of the “Concentration” game. For classification purposes we used Gaussian Mixture Models (GMMs) as they are commonly used in the community.In total, we had three classes (cf. Fig. 2): *FACS in ES-2*, *FACS in ES-5*, and *no-FACS* (cf. Fig. 2). For each class a GMM is trained on audio features on utterances which are 2seconds before the facial expression (cf. Fig. 4). Again, so far we concentrated on the negative user’s reactions in each ES, only. In testing we classified audio samples according to defined classes and compared these to the FACS coded facial expressions (cf. [1]). For training and testing the HTK [23] was used. The evaluation strategy was LOSO that means one speaker’s material was left out from training and used only in testing. With this method, the generalisation purpose of the framework can be shown.

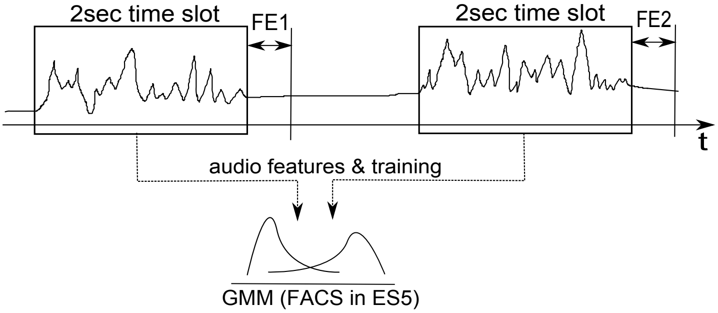

Fig. 4. A 2seconds slot before the facial expression (FE) starts is used to define an utterance. Its extracted features are used to train GMMs. In the figure it is visualised exemplarily with *FACS in ES5*.

4 Results

In these experiments we applied a LOSO strategy on the three class issue introduced in Sect. 3.2 and in Fig. 2 as so far this is the only possibility for us to assess our framework. To evaluate the classification results, we rely on the Weighted Average accuracy (WA) which reflects the class-wise accuracy calculated for each class separately and afterwards averaged over all classes. This can be compared to the Unweighted Average accuracy (UA) calculated based on all samples and therefore, a representation of the sample's distribution is not considered. Thus, the WA is the more meaningful measure as we have just a low amount of samples for relevant sequences but a lot for no-FACS.

At first, we examined the parameter setting for the classifier. Based on previous experiments reported in [20] we opt for GMMs. Further, the training is executed according to specifications in [2]; that is, having five iterations. The number of mixtures is varied in the range of $[6, 15]$ which was driven by [20] and the search for the number of mixtures was stopped when the performance decreased whereas the performance was evaluated in a LOSO manner. The corresponding accuracy values are given in Tab. 2. From this, we will operate the recognisier in future applications with the following parameters: 9 Gaussian Mixtures, 5 training iterations. Nevertheless, a kind of tuning towards specific conditions like noise, echo, etc. might be necessary.

Table 2. Classification results according to the number of mixtures used in a GMM in percent

Number of mixtures	Weighted Average accuracy	Unweighted Average accuracy
6	72.6	69.9
9	**73.9**	**75.6**
12	73.4	84.2
15	70.4	84.2

Table 3. Classification results in percent as Weighted Average accuracy over all LOSO experiments combining ES-2 and ES-5. The bold value represents the false acceptance rate whereas the italic on is the false rejection rate.

Classifier \ true	FACS in ES	no-FACS
FACS in ES	61.9	**18.8**
no-FACS	*38.1*	81.2

To evalute the system, we are interested in the performance as a notifier; this means, we are observing the false acceptance and false rejection rates (cf. Tab. 3). False acceptance indicate how many samples are identified as a *FACS in ES* though it was *no-FACS*; false rejection is defined the otherway around. Despite we use three GMMs we combined the two classes of ES as we do not give any indication to the coder to keep him unaffected in the annotation process. Therefore, we discuss the acceptance rates based on two classes. From Tab. 3 it can be seen that the distinct classes can be distinguished.

Further, the false acceptance rate is 18.8%. So far, the false rejection rate is relatively high with 38.1%. This has tow reasons: 1) the number of samples in ES is low in comparison to *no-FACS* which has influence of the performance and 2) in some LOSO experiments the distinguishing of the classes is confused.

For this, we inspected the confusion matrices of each run of LOSO manually as we were interested in the distribution of recognition results. From this we found that the class *FACS in ES-5* was recognised reliably. On the other hand, no-FACS and FACS in ES-2 are often confused. In particular, for a few number of subjects FACS in ES-2 was not recognised at all which also results in a low performance of the system. To evaluate this issue we looked into the recordings and found that the particular subjects showed only slight facial expressions and, from a human's point of view, almost no reactions in speech. As even for humans the recognition of *FACS in ES-2* is quite hard, we will concentrate our research on finding suitable features and methods which can deal with such slight, but natural emotional reactions, and further handle positive reactions properly as well.

5 Conclusion

In this paper, we presented and discussed a framework towards audio-based semi-automatic selection of video sequences for labelling of emotional facial expressions. Based on classified emotional utterances a pre-selection of sequences is given that afterwards have to be annotated according to the regulations of FACS. Having this selection process the manual effort of annotation is reduced. For instance, watching 30-40 minutes per subject for each subexperiment in *EmoRec I* is reduced to a few minutes (varies according to the certain subject) in total which have to be viewed and annotated. Furthermore, automatic recogniser can also handle slight and natural reactions of subjects which are even hard to realise by humans. Results from an evaluation were presented and discussed.

In future, we apply the framework to the *EmoRec II* material which is not labelled, yet, but has quite similar characteristics as *EmoRec I*. We expect to reduce the manual labelling effort drastically due to the low amount of emotional facial expressions in natural HCI.

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).

References

[1] Böck, R., Limbrecht, K., Siegert, I., Glüge, S., Walter, S., Wendemuth, A.: Combining mimic and prosodic analyses for user disposition classification. In: Wolff, M. (ed.) Proceedings of the 23rd Konferenz Elektronische Sprachsignalverarbeitung, Cottbus, Germany, pp. 220–228 (2012)

[2] Böck, R., Hübner, D., Wendemuth, A.: Determining optimal signal features and parameters for hmm-based emotion classification. In: Proceedings of the 15th IEEE Mediterranean Electrotechnical Conference, pp. 1586–1590. IEEE, Valletta (2010)

[3] Boersma, P.: Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345 (2001)

[4] Cohn, J.F., Zlochower, A.J., Lien, J., Kanade, T., Analysis, A.F.: Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. Psychophysiology 36(1), 35–43 (1999)

[5] 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, Hong Kong, China (2011)

[6] Ekman, P., Friesen, W.: Facial Action Coding System: Investigators Guide, vol. 381. Consulting Psychologists Press, Palo Alto (1978)

[7] Ekman, P., Friesen, W.: Emfacs facial coding manual. Human Interaction Laboratory, San Francisco (1983)

[8] Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions 1(1), 68–99 (2010)

[9] Koelstra, S., Muhl, C., Patras, I.: Eeg analysis for implicit tagging of video data. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–6. IEEE, Amsterdam (2009)

[10] Limbrecht-Ecklundt, K., Rukavina, S., Walter, S., Scheck, A., Hrabal, D., Tan, J.W., Traue, H.: The importance of subtle facial expressions for emotion classification in human-computer interaction. Emotional Expression: The Brain and The Face 5(1) (in press, 2013)

[11] Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. Current Psychology 14(4), 261–292 (1996)

[12] Pantic, M.: Automatic facial expression analysis and synthesis. In: Symposium on Automatic Facial Expression Analysis and Synthesis, Proceedings Int'l Conf. Measuring Behaviour (MB 2005), pp. 1–2. Wageningen, The Netherlands (2005)

[13] Scherer, K.R.: Appraisal considered as a process of multilevel sequential checking. In: Appraisal Processes in Emotion: Theory, Methods, Research, pp. 92–120 (2001)

[14] 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, pp. 552–557 (2009)

[15] Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Transactions on Affective Computing I, 119–131 (2010)

[16] Siegert, I., Böck, R., Philippou-Hübner, D., Vlasenko, B., Wendemuth, A.: Appropriate Emotional Labeling of Non-acted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins. In: Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2011, Barcelona, Spain (2011)

[17] Siegert, I., Böck, R., Wendemuth, A.: The influence of context knowledge for multimodal annotation on natural material. In: Böck, R., Bonin, F., Campbell, N., Edlund, J., de Kok, I., Poppe, R., Traum, D. (eds.) Joint Proc. of the IVA 2012 Workshops, Otto von Guericke University Magdeburg, Santa Cruz, USA, pp. 25–32 (2012)

[18] Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing 3(1), 42–55 (2012)

[19] 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), Barcelona, Spain (2011)

[20] Vlasenko, B., Prylipko, D., Böck, R., Wendemuth, A.: Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications. Computer Speech & Language (2012) (in press)

[21] Walter, S., Scherer, S., Schels, M., Glodek, M., Hrabal, D., Schmidt, M., Böck, R., Limbrecht, K., Traue, H.C., Schwenker, F.: Multimodal emotion classification in naturalistic user behavior. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part III, HCII 2011. LNCS, vol. 6763, pp. 603–611. Springer, Heidelberg (2011)

[22] Wendemuth, A., Biundo, S.: A companion technology for cognitive technical systems. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 89–103. Springer, Heidelberg (2012)

[23] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book, version 3.4. Cambridge University Engineering Department (2009)

[24] Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1), 39–58 (2009)