Towards User-Independent Classification of Multimodal Emotional Signals

Jonghwa Kim, Elisabeth André, Thurid Vogt
Augsburg University
Eichleitnerstr. 30, D-86159 Augsburg, Germany
http://mm-werkstatt.informatik.uni-augsburg.de/index.html

Abstract

Coping with differences in the expression of emotions is a challenging task not only for a machine, but also for humans. Since individualism in the expression of emotions may occur at various stages of the emotion generation process, human beings may react quite differently to the same stimulus. Consequently, it comes as no surprise that recognition rates reported for a user-dependent system are significantly higher than recognition rates for a user-independent system. Based on empirical data we obtained in our earlier work on the recognition of emotions from biosignals, speech and their combination, we discuss which consequences arise from individual user differences for automated recognition systems and outline how these systems could be adapted to particular user groups.

1. Introduction

Emotion is a function of time, context, space, culture, and person and therefore the appraisal of emotion-eliciting events and the manner of emotional expression may widely differ from user to user and from situation to situation. There are still plenty of debates in psychology on the dominance comparison between situational variables and stable individual differences, such as traits encompassing personality, that affect person's behavior and emotional expression [11]. Since individualism in the expression of emotions may occur at various stages of the emotion generation process, human beings may react quite differently to the same stimulus [4]. For instance, extravert people tend to show their emotional state more overtly than introvert people [12]. Moreover one and the same person may express emotions differently depending on the context [6]. For example, anger about the behavior of a person with a higher status is usually not displayed in the presence of this person. Overall, it can be concluded that there exist various factors, such as age, gender, social functioning, physical functioning, or psychological well-being, having strong influence on individual difference in expression and experience of emotions as well.

Over the last decade, significant research efforts have been devoted to automatic emotion recognition towards affective computing and advanced human-computer interaction [5, 14, 19]. However, it is still quite difficult to precisely appraise how suitable the recognition systems reported in literature are for practical applications. Basically the main difficulty lies in the fact that it is almost impossible to uniquely map behavior patterns onto specific emotion types because of the complex multifaceted nature of human emotion. Another reason for inconsistent recognition accuracy is that it depends on the nature of the used dataset and the personality of the subjects. Most of recognition systems in the literature employ machine learning algorithms to classify emotional patterns in audiovisual recordings or physiological measurements. Therefore it is inevitable to train such systems using a given dataset. Based on the splitting type of a given dataset into train and test samples, the system performs user-dependent or user-independent [1] classification. In the former case, classifiers are trained and tested on data of an individual user. In the latter, classifiers are trained from the data of various users and then tested on the data of a new user.

In fact, individual differences in emotional expression observed in psychological studies based on subjects' self-report scores are well supported by the significantly lower rates of user-independent recognition compared to user-dependent recognition. So far, many researchers have investigated the possibility of automatic emotion recognition from facial expression, speech and biosignals. Probably, the next step to improve accuracy rates for user-independent recognition should be to analyze individual differences in experiencing emotion-eliciting events and expressing emotions. In particular, we have to consider that people may (1) respond to stimuli in a different way and (2) may express emotions differently.

[1] Strictly speaking, the term of "user-independent classification" should mean to test a new user by a classifier trained through a dataset from other users. We note that the term in this paper includes "mixed-user classification" in which training samples contain the data from all users and a certain user among the users is tested.

Furthermore, a number of approaches established in this research area need to be revisited. Most emotional data sets so far failed in providing detailed subject information, in particular, information about long-term mood, perceived health, social functioning, disease profile etc. For a user-independent recognition process, it seems to be promising to identify user-specific characteristics prior to the classification process and train the system using the samples that are collected from the subjects having similar characteristics as the test user. Such an approach would also improve the performance of user-dependent classification if the classifier is parametrized by individual information. Of course, it is assumed that sufficient data sets from various individuals are available to be viewed in dense clusters representing individual differences.

Is user-independent emotion recognition feasible? In this paper, we discuss this question by reviewing our previous works on emotion recognition from multichannel biosignals, speech and combined analysis of them. We briefly describe the settings of data collection experiments and revisit recognition results in terms of the question.

2. General Approach and Corpora

The major steps in emotion recognition are signal segmentation, which means finding appropriate segments as emotion classification units, feature extraction to find those characteristics of the signal that best describe emotions and to represent each segmented unit as a (series of) feature vector(s), and lastly the actual classification of the feature vectors into emotional states. A generic architecture of a pattern recognition system is shown in Figure 1.

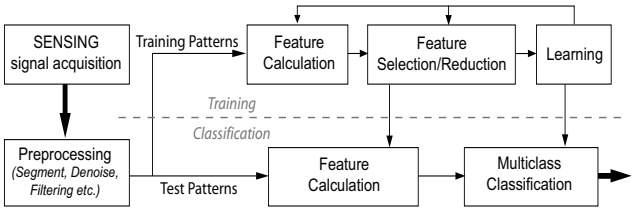

Figure 1. General framework of recognition system

Databases with emotional behavior are not only essential for psychological studies, but also for automatic emotion recognition, as standard methods are statistical and need to learn by examples. Generally, research deals with databases of acted, induced or completely spontaneous emotions. In the case of acted corpora, subjects are usually instructed to express a certain emotional state. Differences in emotional behaviors result from the subjects' acting talent. An example of such a corpus includes the Berlin database [3] of acted emotional speech. In the case of induced emotions, subjects are confronted with stimuli. The stimuli are chosen either by themselves or by an experimenter. If subjects choose the stimuli themselves, there is a higher likeliness that subjects actually feel the emotions, as a consequence, observed differences should be mainly due to different ways of expressing emotions. Examples of a corpus based on individually picked stimuli include the Augsburg biosignal corpus and to some extent the Augsburg student speech corpus. If the experimenter chooses the stimuli, it might happen that subjects respond differently to stimuli. For example, one and the same stimulus might make one subject very angry, but not affect the other. Examples of a corpus based on stimuli chosen by an experimenter include the SmartKom corpus and the Augsburg Quiz corpus.

In this paper, we concentrate on emotion recognition from speech and biosignals. Two kinds of data are investigated here:

- corpora of people that were told to express/feel emotional states AuBT music corpus (biosignals), EmoVoice student corpus (speech)
- corpora of elicited emotions in WOZ experiment SmartKom (speech) and Quiz corpus (biosignals and speech)

We first study the single modalities separately discussing user-dependent and user-independent recognition results. After that, we investigate how to fuse speech and biosignals focusing on user-dependent and user-independent recognition results. In particular, we focus on differences in expressive behaviors by examining for different users and user groups:

- How do people express emotions using a particular channel (expressivity of features)?
- How do people express emotions using multiple channels (expressivity of channels)?

3. Biosignals

Work done in psycho-physiology provides evidence that there is a strong relationship between physiological reactions and emotional/affective states of humans. Physiological reactions should be more robust against possible artefacts of human social masking since they are directly controlled by the human autonomous nervous system. As a consequence, individual differences resulting from deliberately hiding emotions should be smaller for biosignals than for other channels of expression, such as speech or facial expressions.

Recently, a number of studies on engineering approaches to automatic emotion recognition from physiological data have been published [8, 9, 13, 16], although research in that field is relatively new compared to the long history of emotion research in psychology and psychophysiology. Of particular relevance to this paper is a study by Villon and

Lisetti [18] which compares intra-individual differences of 40 subjects in psychological and physiological responses to stimuli. They showed that different subjective responses to stimuli will lead to a lower recognition performance. That is the feasibility of a user-independent recognition approach depends on whether subjects respond to stimuli in a similar way or not.

3.1. Corpus

In our earlier work [8], we conducted an experiment in which we used a musical induction method to put subjects in four different emotional states: high arousal/positive (Hi/Po), high arousal/negative (Hi/Ne), low arousal/negative (Lo/Ne), and low arousal/positive (Lo/Po). The subjects were three males (two students and an academic employee) between 25-38 years old who enjoyed listening to music. The subjects individually handpicked four music songs by themselves that should spontaneously evoke their emotional memories and certain moods corresponding to the four target emotions. Generally, emotional responses to music vary greatly from individual to individual depending on their unique past experiences. This is why we advised the subjects to choose themselves the songs that recall their individual special memories with respect to the target emotions. Four-channel biosensors were used to measure electromyogram, electrocardiogram, skin conductivity, and respiration changes. During three months, a total of 360 samples (90 samples for each emotion) from the three subjects were collected. The signal length of each sample was between 3-5 minutes depending on the duration of the songs.

3.2. Approach

From the four channel signals we calculated a total of 110 features from various analysis domains including conventional statistics in time series, frequency domain, geometric analysis, multiscale sample entropy, subband spectra, etc. For the signals with nonperiodic characteristics, such as EMG and SC, we focused on capturing the amplitude variance and localizing the occurrences (number of transient changes) in the signals. For classification we used the pseudoinverse linear discriminant analysis (pLDA) [23], a natural extension of classical LDA, combined with the sequential backward selection (SBS) [10] to select significant feature subset.

3.3. Results

The confusion matrix in Table 1 presents the correct classification ratio (CCR) of subject-dependent (Subjects A, B, and C) and subject-independent (All) classification where the features of all of the subjects are simply merged and normalized. The recognition results show that overall accuracy differs from one subject to the next and the CCR of the single emotions varies as well with significant disparity. For example, Hi/Ne was perfectly recognized for Subject C, while it caused the highest error rate for Subject B. Particularly in this work, we proposed best emotion-relevant features (see Table 3 in [8]) that are extracted from variety range of feature domains. It turned out that the relevant features varied according to individual differences in the four emotions. The CCR of subject-independent classification was not comparable to that obtained for subject-dependent classification. When considering CCR difference of 32% between both cases, even with relative small number (three) of subjects, it is obvious that there is much to be improved by resolving individual differences.

Subject A ($\mathcal{CCR}$ % = ***81%***)

	Hi/Po	Hi/Ne	Lo/Ne	Lo/Po	*total**	*error*
Hi/Po	***22***	4	1	3	*30*	*0.27*
Hi/Ne	3	***26***	1	0	*30*	*0.13*
Lo/Ne	1	2	***23***	4	*30*	*0.23*
Lo/Po	3	0	1	***26***	*30*	*0.13*

Subject B ($\mathcal{CCR}$ % = ***91%***)

	Hi/Po	Hi/Ne	Lo/Ne	Lo/Po	*total**	*error*
Hi/Po	***27***	3	0	0	*30*	*0.10*
Hi/Ne	3	***25***	1	1	*30*	*0.17*
Lo/Ne	0	2	***28***	0	*30*	*0.07*
Lo/Po	0	1	0	***29***	*30*	*0.03*

Subject C ($\mathcal{CCR}$ % = ***89%***)

	Hi/Po	Hi/Ne	Lo/Ne	Lo/Po	*total**	*error*
Hi/Po	***28***	0	2	0	*30*	*0.07*
Hi/Ne	0	***30***	0	0	*30*	*0.00*
Lo/Ne	0	0	***24***	6	*30*	*0.20*
Lo/Po	0	0	5	***25***	*30*	*0.17*

All: Subject-independent ($\mathcal{CCR}$ % = ***65%***)

	Hi/Po	Hi/Ne	Lo/Ne	Lo/Po	*total**	*error*
Hi/Po	***62***	9	8	11	*90*	*0.31*
Hi/Ne	15	***57***	13	5	*90*	*0.37*
Lo/Ne	9	6	***58***	17	*90*	*0.36*
Lo/Po	8	5	21	***56***	*90*	*0.38*

*: Actual total # of samples

Table 1. Recognition results in correct classification ratio ($\mathcal{CCR}$ *100%= error 0.00%*) achieved by using pLDA with SBS and leave-one-out cross validation. # of samples: 120 for each subject and 360 for All

4. Speech

Generally, research on vocal emotion recognition deals with databases of acted, induced or completely spontaneous emotions.

4.1. Corpora

As a basis for studies described in this paper, we made use of two different speech corpora:

- *EmoVoice Student Corpus* Using the EmoVoice Framework for corpus creation and classifier building [22], 29 students of computer science (8 females, 21 males, aged 20 to 28) recorded sentences for four emotional classes: positive-high, positive-low, negative-high and negative-low. The emotion elicitation was inspired by the Velten mood induction technique [17] where subjects had to read out loud a set of 20 emotional sentences for each emotional class that should set them into the desired emotional state. We have prededefined a set of such sentences for the four emotional states. However, the users were encouraged to change sentences according to their own emotional experiences. Students did the recordings at home, so the audio quality and equipment were not controlled, but all students were told to use a head-set microphone.

- *SmartKom Database of Spontaneous Emotional Speech* This database was recorded within the SmartKom project at the University of Munich [15], from persons interacting with a multi-modal dialogue system. We evaluated only those utterances that were recorded in the mobile setting with a head-set microphone. This subset was splitted into a training set with 56 speakers (24 male and 32 female) and a test set with 14 speakers (7 male and 7 female). The original annotation comprises 12 emotional states. Since 12 emotional classes in spontaneous speech are for the current state-of-the-art in speech emotion recognition a too complex task, we merged them into a 4-class problem applying a scheme suggested in [2]: neutral and unidentifiable utterances to neutral, strong and weak joy and strong and weak surprise into joy, strong and weak pondering/reflecting and strong and weak helplessness into helplessness and strong and weak anger into anger.

4.2. Approach

Our basic emotion recognition system, which is in detail presented in [20], works as follows: From the series of energy, MFCCs, the center of gravity of the spectrum, duration and pause values of a time unit of emotional analysis, further series such as the series of maxima or minima, distances, differences or slope between adjacent extrame are derived. For every series, mean, minimum, maximum, range, variance, median, first quartile, third quartile and interquartile range are computed. From the resulting 1289 features, the most relevant ones for the given task are chosen. The optimized feature subsets are chosen by a best-first search through the feature space according to the classification performance achieved by a Naïve Bayes classifier, which is also used for the final classification of utterances.

4.3. Results

In the following, we report on individual differences in vocal emotion expression. Furthermore, we illustrate how recognition rates for emotional speech can be improved by training classifiers for individual user groups.

Individual differences in the Augsburg Student Corpus Offline speaker-dependent accuracies in 10-fold cross-validation for all 4 classes varied a lot among speakers and ranged from 24 % to 74 %, with an average of 55 %. Thus, accuracy is in some cases lower or not significantly higher than the chance recognition rate of 25 %, though on average, this is the case. This great variation is not only due to the uncontrolled audio recordings, but also due to differences in the expressivity of speech. From all test persons, we selected 10 speakers (5 female, 5 male) that were German native speakers, whose speaker-dependent accuracy was not below 40 % and where audio quality was satisfactory, to train a speaker-independent classifier. This user-independent classifier resulted in a recognition accuracy of 41 % (compared to 25 % chance recognition rate). The result could be improved by more homogenous recording conditions. Furthermore, for good results in a realistic setting and online recognition, only 2 or 3 of these classes should be used. For example, we obtained recognition rates between 60 % and 70 % for the speaker independent system when leaving two classes out where the chance recognition rate is 50 %.

Gender-specific differences in the SmartKom Corpus In the introduction, we suggested to improve emotion recognition rates by computing different classifiers for different types of user. This idea was explored in [21] by training gender-specific classifiers for the SmartKom Corpus. Differences in features for male and female speakers are a well-known problem and it is established that gender-dependent emotion recognizers perform better than gender-independent ones. However, even if not known before, gender can be detected very reliably and thus in a two-stage recognition process, a gender-dependent emotion recogniser can be applied after a gender recogniser. Besides the SmartKom corpus, we tested this approach also on the Berlin database of emotional speech produced by actors [3]. Gender detection achieved an accuracy of about 90 % and the combined gender and emotion recognition system improved the overall recognition rate of a gender-independent emotion recognition system by 2-4 %. The better performance of the gender-specific classifier may be explained by differences in relevant features for the classification of emotions in male and female emotions. For example, in the SmartKom corpus three times as many features were selected for classifying male emotions than for classifying

female emotions. Furthermore, it could also be that the gender detection actually performs a favorable division of the speakers along voice characteristics and that thus even speakers whose gender was incorrectly recognised are now assigned to a more suitable classifier.

5. Fusion of Biosignals and Speech

In order to improve the recognition accuracy obtained from the unimodal recognition system, many studies attempted to exploit the advantage of using multimodal information [1]. In the following, we report on results we obtained by fusing biosignals and speech. In particular, we discuss user-specific differences in the expressivity of channels.

5.1. Corpus

Since there was no multimodal corpus available, we conducted a Wizard-of-Oz study to acquire a bimodal corpus containing emotional speech and biosignal data [7]. The Wizard-of-Oz study was inspired by the quiz "'Who wants to be a millionaire?"' The three test subjects of our experiment were all students - three males in their twenties. All subjects were native speakers of German, which was also the language for the experiment. Each of the sessions took about 45 minutes to complete. The subjects were equipped with a directed microphone to interact with a virtual quiz master via spoken natural language utterances. While the subjects interacted with the system, their physiological feedback was monitored by 4-channel biosensors to record electromyogram (EMG) at the nape of the neck, electrocardiogram (ECG), skin conductivity (SC) and respiration change (RSP). In addition, we recorded the interaction between the user and the quiz master and captured a visual impression of the user on video. In the experimental setting, the agent was controlled through a Wizard-of-Oz interface by a human quiz master who guided the quiz, following a working script to evoke situations that lead to a certain emotional response. The wizard's working script was roughly divided into four situations which served to induce certain emotional states in the user: phase 1: low arousal, positive valence, phase 2: high arousal, positive valence, phase 3: low arousal, negative valence, phase 4: high arousal, negative valence. Thus, the experiment was based on the same emotional states as the music induction experiment described in Section 3. As noted earlier, we could not take it for granted that the subjects responded to quiz events in the intended manner. To cope with this problem, we annotated the recordings based on the audiovisual interpretation of two labelling experts.

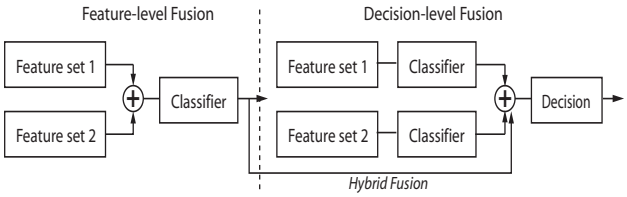

Figure 2. Considered fusion schemes for integrating bimodal information

5.2. Approach

Three different fusion approaches were implemented to exploit the advantage of using two modalities for emotion recognition. In feature-level fusion, the features of both modalities were simply merged and provided as input to a single classifier. Thereby, we also attempted to extract the most significant features from the fused features by SFS to compare the results. In decision-level fusion, the outputs of two unimodal classifiers for speech and biosignals were integrated employing a posteriori probabilities. As a further variation, we applied majority voting to the decision process, according to the recognition rates of each emotion from unimodal classifiers. Finally, we employed a new hybrid scheme of the two fusion methods in which the ouput of feature-level fusion is also fed as an auxiliary input to the decision-level fusion stage (see Figure 2).

5.3. Results

As shown in Table 2, the disparity of recognition accuracy between user-dependent and user-independent classification is not improved by exploiting bimodal fusion but became worse. The discrepant rates among the subjects and modalities indicate the fact that some subjects are more expressive in their physiological reactions and others more in their speech. However, no suggestively dominant modality has been revealed neither in user-dependent nor in user-independent classification.

Contrary to the experiment described in Section 3, the subjects were not told to put themselves into a particular emotional state, but they had to respond to stimuli provided by an experimenter. As a consequence, the quiz experiment introduces more uncertainty regarding the ground truth. For example, some subjects might have been more ambitious in making profit in the game than others. Furthermore, some subjects might have tried to hide their emotional state while others might not have cared revealing it. The large differences in the recognition results for individual users participating in the quiz experiment reveal these uncertainties. Nevertheless, the quiz corpus is of higher interest when coping with everyday emotional states due to its higher ecological validity.

Different accuracy rates were also obtained by using fusion methods. Overall, we obtained the best results

System	Hi/Po	Hi/Ne	Lo/Ne	Lo/Po	*Average*
Subject A					
Biosignal	0.95	0.92	0.86	0.85	0.90
Speech signal	0.64	0.75	0.67	0.78	0.71
Feature Fusion	0.91	0.92	1.00	0.85	**0.92**
Decision Fusion	0.64	0.54	0.76	0.67	0.65
Hybrid Fusion	0.86	0.54	0.57	0.59	0.64
Subject B					
Biosignal	0.50	0.79	0.71	0.45	0.61
Speech Single	0.76	0.56	0.74	0.72	0.70
Feature Fusion	0.71	0.56	0.94	0.79	**0.75**
Decision Fusion	0.59	0.68	0.82	0.69	0.70
Hybrid Fusion	0.65	0.64	0.82	0.83	0.73
Subject C					
Bio Single	0.52	0.79	0.70	0.52	0.63
Speech Single	0.55	0.77	0.66	0.71	0.67
Feature Fusion	0.50	0.67	0.84	0.74	**0.69**
Decision Fusion	0.32	0.77	0.74	0.64	0.62
Hybrid Fusion	0.40	0.73	0.86	0.71	0.68
All: Subject-independent					
Bio Single	0.43	0.53	0.54	0.52	0.51
Speech Single	0.40	0.53	0.70	0.53	0.54
Feature Fusion	0.46	0.57	0.63	0.56	**0.55**
Decision Fusion	0.34	0.50	0.70	0.54	0.52
Hybrid Fusion	0.41	0.51	0.70	0.55	0.54

Table 2. Recognition results in rates (1.0=100% accuracy) achieved by using SBS, LDA, and leave-one-out cross validation.

from feature-level fusion, although it is generally known that feature-level fusion is more appropriate for combining modalities with analogous characteristics. As a result, it turned out that modality fusion does not necessarily imply a potential solution for the individual difference.

6. Conclusion

Coping with individual differences in the expression of emotions is a difficult task not only for a machine, but also for humans. In order to identify potential problems caused by individual differences of users, we reviewed our earlier works on automatic emotion recognition from biosignals, speech and their combination. To assess the performance of a user-independent recognition approach, we validated our recognition approaches using a one-person-leave-out validation approach. Due to the small number of subjects employed to train general classifiers, it is not surprising that the user-independent approaches could not compete with the user-dependent approaches. As a first step towards a user-independent approach, we suggest the collection of detailed subject information when acquiring corpora of emotional data. Such information would help classify users based on specific characteristics and then train user-specific classifiers. At runtime, a test user would first be assigned to a particular user class and then the corresponding classifier be applied. By developing a gender-specific vocal emotion classifier, we demonstrated the potential of such an approach. A more fine-grained classification of users might lead to further improvements provided that a sufficient amount of data has been collected. Furthermore, a deeper analysis of the correlation between emotion-eliciting events and observed or self-reported emotions depending on user characteristics should be conducted in order to shed light on individual appraisal processes. When combining a recognition approach with a predictive emotional model, information on individual appraisal processes could be used to improve accuracy rates.

Acknowledgements

This work has been funded in part by the European Commission via the CALLAS Integrated Project (ref. 034800, http://www.callas-newmedia.eu/) and the Metabo Integrated Project (ref. 216270, http://www.metabo-eu.org/).

References

[1] J. N. Bailenson, E. D. Pontikakis, I. B. Mauss, J. J. Gross, M. E. Jabon, C. A. Hutcherson, C. Nass, and O. John. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies*, 66(5):303–317, 2008.

[2] A. Batliner, V. Zeißler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth. We are not amused - but how do you know? User states in a multi-modal dialogue system. In *Proceedings of Interspeech 2003*, pages 733–736, Geneva, Switzerland, September 2003.

[3] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of German emotional speech. In *Proceedings of Interspeech 2005*, Lisbon, Portugal, September 2005.

[4] T. Canli, Z. Zhao, J. Desmond, E. Gang, J. Gross, and J. Gabrieli. An fMRI study of personality influences on brain reactivity to emotional stimuli. *Behavioral Neuroscience*, 115(1):33–42, 2001.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18:32–80, 2001.

[6] J. J. Gross and O. P. John. Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85:348–362, 2003.

[7] J. Kim. Bimodal emotion recognition using speech and physiological changes. In M. Grimm and K. Kroschel, editors, *Robust Speech Recognition and Understanding*, pages 265–280. I-Tech Education and Publishing, Vienna, Austria, 2007.

[8] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.

[9] K. H. Kim, S. W. Bang, and S. R. Kim. Emotion recognition system using short-term monitoring of physiological signals.

Medical & Biological Engineering & Computing, 42:419–427, 2004.

[10] J. Kittler. *Feature Selection and Extraction*, pages 59–83. Academic Press, Inc, 1986.

[11] H. W. Krohne. Individual differences in emotional reactions and coping. In R. J. Davison, K. R. Scherer, and H. H. Goldsmith, editors, *Handbook of Affective Sciences*, pages 698–725. Oxford University Press, 2003.

[12] R. E. Lucas and B. M. Baird. Extraversion and emotional reactivity. *Journal of Personality and Social Psychology*, 86(3):473–485, 2004.

[13] R. Picard, E. Vyzas, and J. Healy. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.

[14] R. W. Picard. *Affective computing*. MIT Press, 1997.

[15] S. Steininger, S. Rabold, O. Dioubina, and F. Schiel. Development of user-state conventions for the multimodal corpus in SmartKom. In *Proceedings Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 733–37, Las Palmas, 2002.

[16] E. L. van den Broek, V. Lisý, J. H. Janssen, J. H. D. M. Westerink, M. H. Schut, and K. Tuinenbreijer. Affective man-machine interface: Unveiling human emotions through biosignals. In A. Fred, J. Filipe, and H. Gamboa, editors, *Biomedical Engineering Systems and Technologies*, Communications in Computer and Information Science. Berlin / Heidelberg, Germany: Springer, [in press].

[17] E. Velten. A laboratory task for induction of mood states. *Behavior Research & Therapy*, 6(4):473–482, 1968.

[18] O. Villon and C. Lisetti. Toward recognizing individual's subjective emotion from physiological signals in practical application. In *IEEE Symposium on CBMS '07*, June 2007.

[19] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 1061–1070, New York, NY, USA, 2008. ACM.

[20] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *IEEE International Conference on Multimedia & Expo (ICME 2005)*, 2005.

[21] T. Vogt and E. André. Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC 2006)*, Genoa, 2006.

[22] T. Vogt, E. André, and N. Bee. Emovoice — a framework for online recognition of emotions from voice. In *Proc. Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Kloster Irsee, Germany, June 2008.

[23] J. Ye and Q. Li. A two-stage linear discriminant analysis via QR-decomposition. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(6), June 2005.