Cross-modal Elicitation of Affective Experience

Christian Mühl and Dirk Heylen
Human Media Interaction
University of Twente, NL
c.muehl@utwente.nl

Abstract

In the field of Affective Computing the affective experience (AX) of the user during the interaction with computers is of great interest. Physiological and neurophysiological sensors assess the state of the peripheral and central nervous system. Their analysis can provide information about the state of a user. We introduce an approach to elicit emotions by audiovisual stimuli for the exploration of (neuro-)physiological correlates of affective experience. Thereby we are able to control for the affect-eliciting modality, enabling the study of general and modality-specific correlates of affective responses. We present evidence from self-reports, physiological, and neurophysiological data for the successful induction of the affective experiences aimed for, and thus for the validity of the elicitation approach.

1. Introduction

Affective computing aims at an enrichment of HCI by taking the user's affective state into account [29]. Thereby, applications can unfold their functions in the context of user experience, ideally leading to the increase of the bandwidth and naturalness of interaction.

To achieve such enhanced interactions a robust automatic recognition of the user state is a necessary prerequisite. In the past years the automatic analysis of affect-related behaviours, especially those evident in facial expression or voice, yielded promising results [10, 41]. Still, the classification of affective user state is no trivial endeavor, as the subjective state, the experience of the user, is not necessarily observable by external means as cameras or microphones.

The analysis of physiological responses during affective experience offers an alternative to the analysis of behavioural responses [6–8,20,23,24,30,39]. However, whilst observable behaviour as facial expressions or voice, can be conveniently studied in the field, physiological and neurophysiological responses are less readily available. Therefore the elicitation of affective experience in the laboratory is still a necessary step to acquire physiological and neurophysiological databases. This data can then be analysed in order to extract features capable of discriminating between affective experience. These are then the basis to develop and refine suitable classification methods using those features.

To explore the generalisation of physiological and neurophysiological correlates of affective experiences we developed a cross-modal elicitation method. Specifically, we constructed a set of audiovisual stimuli to be able to elicit emotions either from the auditory or from the visual modality. This study presents evidence, based on the subjects' self-assessments, and on preliminary physiological and neurophysiological results, for the induction of different emotions, and thereby for the validity of the approach.

Before we outline our research questions in more detail, we will introduce the reader to the issue of the validation of emotion elicitation approaches, and to our specific approach.

1.1. The validation of an elicitation method

One can discriminate between endogeneous and exogeneous elicitation methods [30]. The former require the subject to induce affective experiences by remembering or imagening emotional episodes [1, 6, 30, 31]. The latter approach makes use of affective stimuli or tasks to elicitate corresponding experiences. A wide variety of affective stimuli has been used for this purpose, among them pictures [5, 7, 26], naturally occurring sounds [3], music pieces [13, 22, 33], films [15, 19, 23, 39], manipulated applications [20], and computer games [8, 40]. In our approach, outlined below, we will use affective stimuli.

A general problem accompanying the induction of affective experience is the validation of the induction method [14, 38]. Fairclough [14] discusses several methods that can be applied to ensure this *concurrent validity* of the elicitation approach in the context of psychophysiological measurements.

For the use of stimuli or tasks one has to be fairly confident that they indeed induce the target states. The use of

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

normed stimulus sets, as the IADS [4] or IAPS [25] can make this more likely. Similarly when using tasks one can use standardized tasks developed within the field of experimental psychology. Alternatively, one might use tasks that have known effects on the user, for example manipulated computer games. However, as Fairclough points out, the use of these latter approaches is close to a natural context, but also prone to confounds due to the complexity of real-world situations.

Another method for the labeling and validation of the data, especially in the domain of facial expression or voice analysis, are observer ratings of the participants behaviour. However, the occurrence frequency of behaviour might be low in the cases that we are considering where the participant is restrained by recording equipment.

Alternatively, one can apply self-assessment methods as the Self-Assessment Manikin [2], to ensure the success of the elicitation method. This, of course, comes for the price of a possible interference with the target behaviour, and an added risk of artifact production. Additionally, self-assessments are not free of bias and dependent on a truthful report.

Furthermore, one might record physiological or neurophysiological data and contrast the different conditions. Should one find a difference between conditions, this can be taken as evidence that indeed different states were induced. To ensure that the desired states were induced one would have to measure and contrast physiological variables that were shown before to vary with the target state or dimension. Those variables, for example, could be chosen by an extensive literature review or the consultation of an expert.

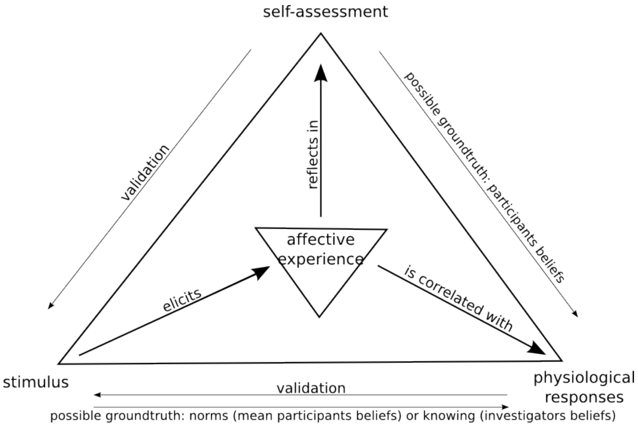
self-assessment
validation
reflects in
possible groundtruth: participants beliefs
affective experience
elicits
is correlated with
stimulus
validation
physiological responses
possible groundtruth: norms (mean participants beliefs) or knowing (investigators beliefs)

Figure 1. The relationship between stimuli, self-assessment, and physiological data, and the elicitated affective experience.

Figure 1 illustrates the above described relationships between administered stimuli, logged self-assessments, and physiological data. The elicited affective experience can be validated by each of these. However, it has to be mentioned that none of the validation methods is perfect and thus a combination of different validation procedures might yield the most insight into the concurrent validity of the elicitation approach.

In this paper we will analyse the results of the first series of elicitation experiments that we carried out with our approach. We look at how self-assessments relate to the affective states that we intend to elicit and to what extent (neuro-)physiological measures can discriminate between the various stimuli groupings.

1.2. The cross-modal elicitation of affective experiences

As already outlined above, physiological and neurophysiological signals carry information about the affective state of the user. While relatively many studies explored the potential of physiological features to differentiate affective experiences [8,20,23,24,30,39], only few studies looked at the suitability of neurophysiological sensors [6, 7]. Most studies were conducted under controlled circumstances. This is especially true for the EEG studies. The tight control of experimental protocols is a necessary prerequisite to disentangle the manifold physiological processes occurring in real-world environments, and thus to avoid confounding variables. However, it is also impeding the ecological validity of the psychophysiological inferences made on the basis of such simple elicitation paradigms [14]. To ensure the generalisation of the feature-experience relationships found in the controlled laboratory experiments to real-world applications, a slow increase in the complexity of the experiments and finally the step into the field seems advisable.

Our motivation for the development of the elicitation method introduced here is to make a modest step in this direction, exploring the generalisation of psychophysiological inferences over different affect elicitation modalities, but still staying inside the laboratory. We are aiming for the controlled induction of affective experiences via the visual or auditory stimulus modality. Specifically, we want to manipulate experience along the valence dimension of the dimensional emotion model according to Russel [32], that is to elicitate negative, neutral, and positive affective experiences. For this purpose we combine affective neutral and valence-carrying stimuli from different unimodal stimulus sets to a new multimodal stimulus set.

We chose sound (IADS) and picture (IAPS) stimulus sets to construct new audiovisual stimuli. One advantage of those specific sets is that they are normed by hundreds of participants according to their effect on the participants' affective experience. The knowledge about mean valence and arousal responses for a given stimulus guides our combinations of neutral and valence-carrying unimodal stimuli to one multimodal. However, one should take in mind the standard deviations of the norm ratings are quite big and show a large spread of responses from different subjects to a given stimulus over the arousal-valence plane. This indicates a subject- and context-specific response to the stimuli. Furthermore, to construct our stimulus set we make combinations of different stimuli from the original databases. It cannot be assumed that the combination of different affective stimuli has a linear effect on the affective response. The combination of picture and sound might produce a different context, changing or even inverting the original affective response caused by the valence-carrying stimulus. Despite our intentions to control for that by a careful choice of combinations of auditory and visual stimulus parts, the elicitation of the target emotions has to be shown to ensure the validity of our elicitation method. In the following section we will describe the construction of the new multimodal stimulus set in detail.

1.3. Stimuli construction

To study the effects that the different modalities have on neurophysiological affective responses, 180 multimodal stimuli were constructed from the auditory and visual affective stimuli sets IADS and IAPS.

From each stimulus set, IADS and IAPS, we chose 30 stimuli from the positive and 30 stimuli from the negative side of the valence dimension. Additionally, we chose 60 neutral auditory and 60 neutral visual stimuli from each modality. Note that we employed each neutral unimodal stimulus twice (due to the low number of IAPS stimuli). Each neutral stimulus from one data set would appear one time in combination with another neutral stimulus from the other data set and one time in combination with a valence-carrying stimulus of the other data set. The three different valence intervals, positive, neutral, and negative, were defined according to the mean ratings on the valence scales. The 9-point valence Likert scale the norm-ratings are based on are ranging from 1(feeling unhappy) to 9 (feeling happy). Therefore, we required positive stimuli to have a mean rating above 6.5, negative stimuli to have a mean rating below 3.5, and neutral stimuli to lie in between these two groups.

We constructed five groups of auditory-visual stimuli: (1) *auditory negative*, (2) *auditory positive*, (3) *visual negative*, (4) *visual positive*, and (5) stimuli that were neutral both auditory and visually, referred to as *multimodal neutral*. An auditory negative stimulus consisted of a negative auditory stimulus and a neutral visual stimulus. An auditory positive stimulus contained a positive auditory and neutral visual stimulus. This way the affect elicitation was supposed to result from the auditory stimulus. Correspondingly, the visual negative and positive stimuli were created from a neutral auditory and a valence-holding visual stimulus. The multimodal neutral stimuli consisted of a neutral auditory and a neutral visual stimulus. This group was important as a control condition, which enables the analysis of the specific effects of positive and negative stimulation, respectively. While the grouping was based on the distribution of the stimuli on the valence axis, we tried to keep the group differences on the arousal axis comparable to avoid confounding effects. Specifically, we tried only to use stimuli that had a relatively high arousal value, i.e. higher than 3.5. Because of a bias in the original sets, we were not able to do this.

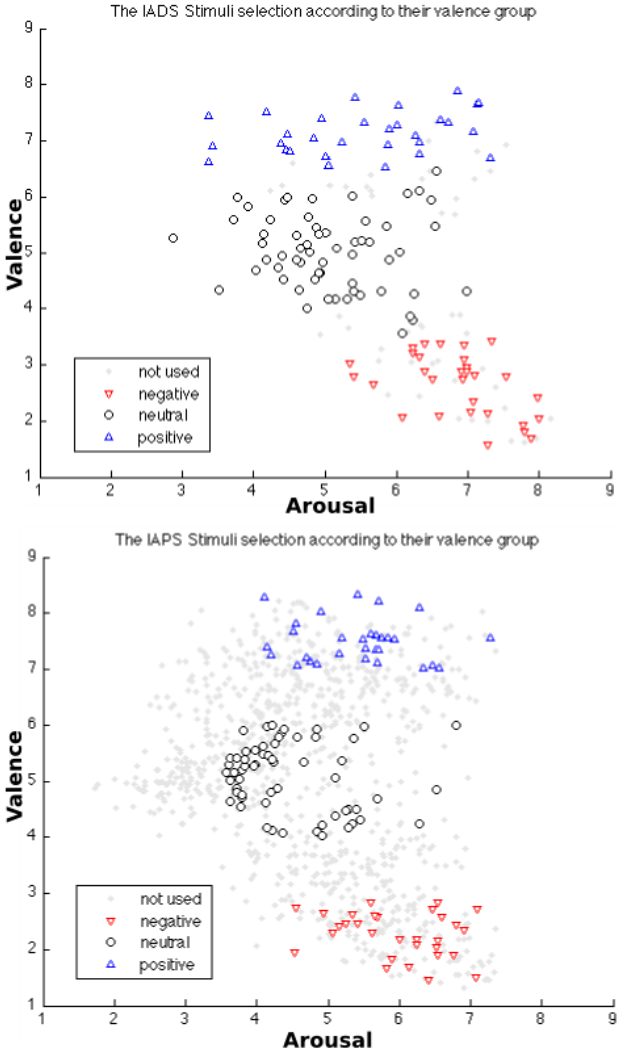

Figure 2. The position of the selected auditory (above) and visual (below) stimuli in the valence-arousal space.

As described above, the stimuli were constructed from carefully selected subsets of the IADS and IAPS. Figure 2 shows the positive, neutral, and negative stimuli groups chosen from the IADS and IAPS sets, and their locations in the valence-arousal plane. The stimulus sets used for our stimulus construction are not evenly distributed in the valence-arousal plane, but describe a C-form. There are almost no stimuli that have low arousal and high valence values, or high arousal and medium valence. On the other hand are the negative stimuli in general more arousing than positive or neutral stimuli. In consequence a selection of stimuli that differ in valence, but not in arousal is only possible to a limited degree. The higher the number of stimuli is, the stronger is the selection influenced by the mentioned IAPS and IADS characteristics. The most limiting factor is the small size of the IADS, which makes it difficult to select more than 30 negative and 30 positive stimuli.

Group	Valence mean (std)	Arousal mean (std)
A positive	7.14 (0.38)	5.51 (1.16)
A neutral	5.01 (0.67)	5.06 (0.85)
A negative	2.65 (0.55)	6.84 (0.71)
V positive	7.49 (0.38)	5.40 (0.77)
V neutral	5.08 (0.60)	4.46 (0.77)
V negative	2.27 (0.40)	5.97 (0.73)

Table 1. The mean valence and arousal ratings per modality and stimulus group.The value in brackets is the standard deviation.

Table 1 gives an overview over the group's valence and arousal means according to the norm ratings from the IADS and IAPS manuals. The valence means of the groups are all significant different. Despite our efforts to keep the arousal equal over the groups, also the arousal means are significantly different. However, as the norm values of the IAPS and IADS are already characterised by a big standard deviation, it was not assumed to be able to predict the precise effect of the stimuli on a particular group of participants. In that respect the valence and arousal values were only used as an initial strategy to select the optimal stimuli for our purpose.

1.4. Research questions

To validate our elicitation approach, we are interested in the effect that our stimulation has on the participant's experience. As described above there are different strategies that one can use to ensure that the affective experience of interest was induced. Therefore, we analysed the participants' self-assessments according to the different stimuli categories employed, irrespective of the elicitation modality. Furthermore, we analysed the (neuro-)physiological responses to the different stimulus categories employed. Finally, we explored the effect of the choice of an alternative ground truth, that is the (neuro-)physiological responses to different groupings of the trials according to the self-assessments of the subjects.

Our main question was whether the target emotion is indeed induced by our elicitation paradigm. We expected that in the comparison of the self-assessments given after each stimulus presentation, the valence judgements over stimuli and subjects would be significantly different between conditions. On the other hand, arousal should ideally be comparable, as we aimed for similar arousal values during the construction of the stimulus groups.

A further expectation was that the comparison of the physiological responses during the presentations of the different stimulus groups yields significant differences. Especially, we expected differences for those physiological and neurophysiological sensors implied before in valence-related nervous system responses. Physiological correlates of valence manipulations have been found for electromyographical responses recorded from the facial muscles (EMG) [5, 39], in electrocardiographical recordings (ECG), specifically heart rate [31, 33], and for blood pressure [36]. Neurophysiological correlates, specifically those derived from electroencephalography (EEG), include the asymmetry of alpha power between the left and right hemisphere of the brain [11], and frontomedial theta power [33].

Significant differences in other (neuro-)physiological signals not directly implied in valence-, but other affect-related experiments might also offer evidence about different states induced, though they could not be used as evidence for the elicitation of the target states. Physiological signals implied in arousal-related manipulation of affective experiences are the galvanic skin response (GSR) [3,9,22,26], the respiratory sinus arrhythmia (RSA) derived from the heart rate [15], and respiration [13, 17]. Neurophysiological arousal-specific responses include a decrease of the overall level of power in the alpha band [27] and an increase of power in the gamma band [21, 28].

Finally, the question most relevant for the determination of a ground truth for later classification approaches is, if there is a more favourable grouping (of trials into conditions) possible according to the self-assessments. That there is a significant difference between classification results achieved via a norm based and a self-assessment based ground truth was shown by Chanel et al. [7]. Therefore, we resorted the stimuli, and thus the trials, into positive, neutral, and negative affect conditions according to self-assessment. The trivial assumption was that this regrouping would create more homogeneous conditions, with condition means further apart, and smaller standard deviations. Furthermore, we expected more significant (neuro-)physiological differences, especially for sensors implied in valence manipulation before.

2. Methods

2.1. Participants

14 participants (7 men and 7 women) took part in the experiment. Due to incomplete recordings the data of two participants was not analysed. The participants were aged between 19 and 53 (mean age 28) and all except one indi-

cated their right hand as the dominant hand.

2.2. Stimuli

For the experiments the newly constructed audiovisual stimulus set as described in section 1.3 was used. To avoid eye-movements during the stimulus presentations the pictures were decreased in size to 400 x 300 pixels. Primary stimulus characteristics as overall loudness of sounds or brightness of visual stimuli may have significant effects on neurophysiological data. To minimize the risk of a confound by stimulus-related non-affective characteristics we tested the group differences of mean subjective loudness and mean luminance. No significant differences between the visual parts of the positive, negative, and neutral group was found in terms of mean luminance. Similarly, no significant differences between the auditory parts of the positive, negative, and neutral group in terms of mean subjective loudness could be detected.

2.3. Equipment and signal acquisition

2.3.1 Presentation and recording hardware

The stimuli were presented on a dedicated stimulus PC (P4 3.2GHz), which sent markers according to stimulus on- and offset to the EEG system (Biosemi ActiveTwo system, www.biosemi.com). For the stimulus presentation we used ”Presentation” (Neurobehavioral systems, www.neurobs.com). The visual parts of the stimuli were presented on a 20 inch monitor (Samsung SyncMaster 203B). The auditory parts of the stimuli were presented via a pair of custom computer speakers (Phillips Multimedia Speaker System). The distance between participants and monitor/speakers was about 70 cm.

The physiological and neurophysiological signals were recorded with 512 Hz on a dedicated recording PC (P4 3.2GHz) running Actiview software (BioSemi).

We recorded from 64 active silver-chloride electrodes placed according to the the 10-20 system. Additionally, 4 electrodes were applied to the outer canti of the eyes and above and below the right eye to derive horizontal EOG and vertical EOG, respectively.

Besides recording neurophysiological signals by electroencephalography we assessed also the state of the peripheral nervous system via several physiological sensors.

To obtain the electrocardiogram we placed an electrode at the inner side of the left arm of the participant. A plethysmograph was clipped to the left index finger to assess blood volume pulse. A temperature sensor was placed on the distal phalange of the small finger of the left hand to measure peripheral temperature. Respiration was assessed via a respiration belt placed around the chest just over the stomach. To assess the activity of the somatic nervous system we applied electrodes to two facial muscles, the right corrugator supercilii (implied in frowning) and the left zychomatic major (implied in smiling). The EMG sensor placement over the zygmaticus major and the corrugator supercilii muscle was done via two electrodes for each muscle and according to the guidelines from [16] on the left cheek and over the right brow, respectively.

2.4. Procedure

The Participants were seated in a comfortable chair in front of monitor and speakers. They read an informed consent form and user instructions before the experiment. After filling in a questionnaire and signing the informed consent the EEG cap and the physiological sensors were placed according to the descriptions above. Before the start of the experiment the participant was introduced to the Actiview online view of her EEG signals to make her conscious of the influence of movement artifacts. She was instructed to restrict movements to the periods between trials. Finally, the SAM scales were explained, so that a good understanding of the concepts of arousal and valence could be assured. Participants were advised to give a ”gut response” to emphasise the importance of their subjective feeling and to avoid a more cognitive judgement of the stimuli themselves.

2.5. Experiment Design

The stimulus presentation was done in 4 blocks with 45 stimuli each. The order of the stimuli presentation was randomised for each participant. To avoid tensions or fatigue, in the breaks the participant could correct seating position, drink, and relax until she felt ready to continue. Figure 3 depicts the trial structure employed. Below we will outline each of the trial periods and its functions.

Pre-stimulus phase Two seconds before a stimulus is presented a fixation cross is blended into the middle of the screen. This cross is supposed to limit eye movement during stimulus presentation and will be kept on the screen until the self-assessment phase.

Stimulus phase The stimulus is presented for six seconds, which is the length of the auditory stimulus. The visual counterpart is shown during the time the sound is played.

Post-stimulus phase Between the stimulus offset and the begin of the self-assessment the fixation cross is further visible on a black background for two seconds. This phase is intended to serve as a stimulus free period in which the independence of a potential affect-related neurophysiological response from the stimulus characteristics can be shown.

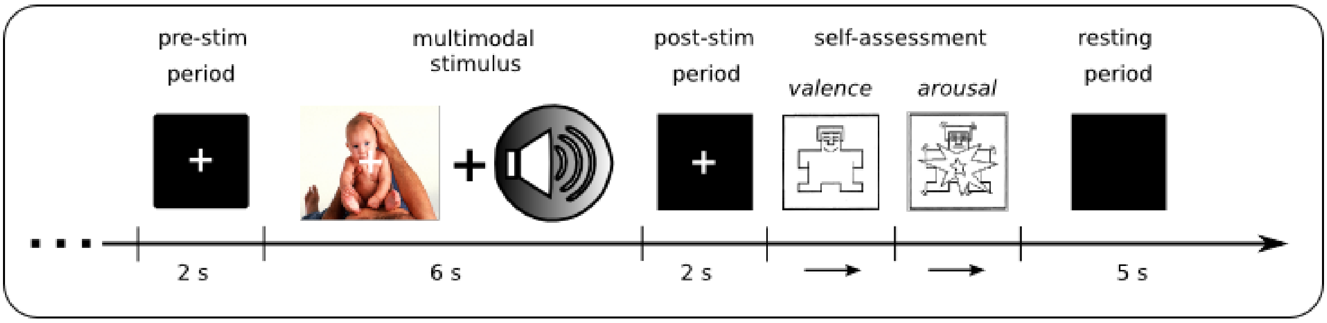

Figure 3. Example trial with the six trial periods and their duration (arrows indicate self-paced rating phases).

Self-assessment phase The norm ratings of the IAPS and IADS are characterised by a considerable variance per stimulus. Thus a given stimulus might induce different affective states in different subjects. To study the effectiveness of our affective stimulation and to explore alternative groupings for the signal samples in positive, neutral and negative trials, a self-report in form of the self assessment manikin (SAM; see [2]) is employed after each stimulus presentation. The duration of the rating phases for arousal and evaluation is not limited. It ends as soon as the user finishes the self-assessment. However, the subject is instructed to answer by a fast intuitive judgement.

Resting phase The physiological response is known to be relatively slow, peaking around five seconds after stimulus presentation [37]. To reduce the contamination of the samples by prior samples, the rating is followed by an inter-stimulus interval of averagely five seconds. The participants are also instructed to blink and move preferably in this period, to decrease the contamination of the trials by movement artifacts.

2.6. Preprocessing of EEG data

We used EEGlab [12] to preprocess the EEG data. Specificly, we computed the common average reference (CAR), downsampled the data to 256 Hz, and high passed it with an infinite impulse response filter at 1 Hz. Then we extracted epochs of six seconds, from stimulus onset to stimulus offset. We computed the absolute frequencies for the theta (θ, 4 - 7 Hz), alpha (α, 9 - 12 Hz) via a FFT with a sliding window length of 128 samples and 50% overlap.

Furthermore, we computed the asymmetry for each pair of the left and right frontal channels, that is AF3 and AF4, and F3 and F4, and F5 and F6, in the alpha frequency band by formula 1.

$$X_{asym} = \frac{(X_{left} - X_{right})}{(X_{left} + X_{right})} \qquad (1)$$

As we did not remove potential artifacts from the data, we restricted our analysis to the alpha and theta frequency bands. Furthermore, we focused on the analysis of anterior regions of interest, as we expected modality-related variations in EEG power in the posterior modality-specific regions. Figure 4 shows the electrode layout for the frontal cortex. We extracted the power of the alpha band for the left and right frontal regions, and the power of the theta band for the fronto-medial region.

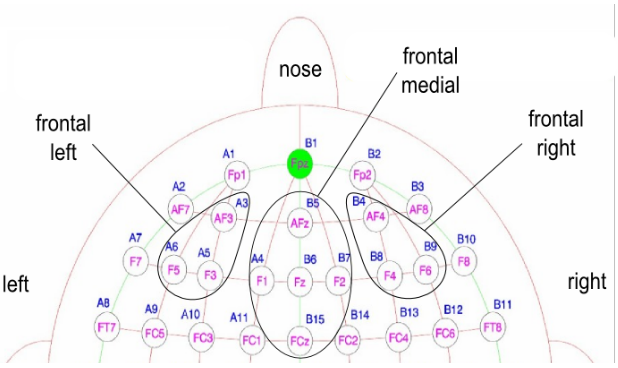

Figure 4. The regions of interest for the preliminary analysis of EEG signals.

2.7. Preprocessing of physiological data

As most physiological sensors are known to have a slow response to stimulation and thus a long response latency, we extracted long epochs of ten seconds for each trial. An epoch contained the pre-stimulus period, the stimulation period and the post-stimulus period. From the signal part that contained the stimulus and post-stimulus period we obtained several features for each of the measured biosignals, while the pre-stimulus part of the signal was used for baseline removal for sensors that are susceptible to stimulus-independent long-term variations. We sampled the physiological signals down to 256 Hz. Below we describe the extracted features for the cardiovascular signals, the galvanic skin response, and the facial EMG sensors in detail.

Cardiovascular features In Table 2 the extracted cardiovascular features are described. For the extraction of the heart beats and the computation of the highest frequency of the heart rate variability, the respiratory sinus arythmia, the BIOSIG toolbox for Matlab was used ([35]). To eliminate the effect of stimulus-independent, low frequency fluctuations in the blood volume pulse data, we subtracted the baseline mean from each trial.

Feature	Description
$E\{h\}$	mean heart rate
HF	highest frequency of the heart rate variability
$E\{b\}$	mean of the blood volume pulse
$\sigma\{b\}$	standard deviation of the blood volume pulse
$min\{b\}$	minimum of the blood volume pulse
$max\{b\}$	maximum of the blood volume pulse
$\delta^b_{\lvert 1 \rvert}$	mean of the abs. of the 1. difference of BVP
$\delta^b_{\lvert 2 \rvert}$	mean of the abs. of the 2. difference of BVP
$E\{t\}$	mean T
$\sigma\{t\}$	standard deviation of T

Table 2. The cardiovascular features derived from the electrocardiogram (ECG), blood volume pulse (BVP) and skin temperature (T) sensors and their description.

Galvanic skin response To analyse the galvanic skin response we first low-pass filtered the signal at 0.05 Hz via an infinite impulse response filter of length 4. To further reduce the stimulus independent variance of the data, we de-trended each trial and subtracted the baseline mean. Table 3 shows the features extracted from the filtered signal.

Feature	Description
$E\{s\}$	mean skin conductance
$\sigma\{s\}$	standard deviation of the SC
$\delta^s_{\lvert 1 \rvert}$	mean of the abs. of the 1. difference of SC
$\delta^s_{\lvert 2 \rvert}$	mean of the abs. of the 2. difference of SC

Table 3. The features derived from skin conductance sensors (SC).

Facial electromyography According to [39] the two electrode pairs placed over the right corrugator supercilii and the left zychomatic major were subtracted, yielding the EMG signals for each muscle, from which we extracted the first four statistical moments, as enlisted in Table 4.

Feature	Description
$E\{c\}$	mean CS
$\sigma\{c\}$	standard deviation of the CS
kurt$\{c\}$	kurtosis of the CS
skew$\{c\}$	skewness of the CS
$E\{z\}$	mean ZM
$\sigma\{z\}$	standard deviation of the ZM
kurt$\{z\}$	kurtosis of the ZM
skew$\{z\}$	skewness of the ZM

Table 4. The EMG features derived from the right corrugator supercilii (CS) and the left zychomatic major (ZM).

3. Results

In a preliminary analysis we studied the recorded data to gain insights into the validity of our approach. Furthermore we hoped to learn which grouping method, according to stimulus norms or according to self-assessment, would be better suited as ground truth for future in-depth study of the physiological and neurophysiological correlates. We first will present the self-assessment data for the different grouping methods. Then, we will examine the physiological and neurophysiological differences between the 3 conditions, positive, negative, and neutral emotions, for the different grouping methods.

3.1. Analysis of the self-assessment data

The evaluation of the self-assessment is not only a mean to validate our emotion induction method, but also gives us the possibility for an alternative grouping of the stimuli according to the individual affective response toward each multimodal stimulus. The grouping of the stimuli establishes the ground truth in the search for physiological and neurophysiological differences and for a later classifier training.

The analysis of the mean stimulus valences suggested that different stimulus groups (positive, neutral, negative) resulted in different affective experiences. The mean values behaved according to the group membership. However, for many stimuli the induced emotions differed from the emotions the stimuli were supposed to induce. This was also reflected by participants informal reports after the experiments. For example, a starving child on a blue blanket was perceived by one participant as cared for and elicited a calm and rather positive response, while it was intended to elicit a negative reaction. Figure 5 shows the distribution of valence and arousal ratings over all stimuli and subjects for the five conditions, also taking the modality of the affect eliciting stimulus into account. Despite the clear differences that are visible between the groups, the distributions are overlapping to a large degree. That is, some of the stimuli had not the intended effect on some subjects, but instead elicited another affective state.

These deviations of the individual affective experience from the target affective experience of the stimuli are natural taking the individual differences between participants

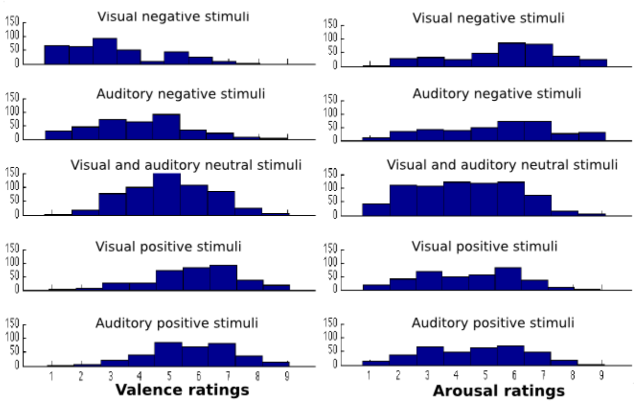

Figure 5. The distributions of the SAM ratings for the original groupings for valence (left) and arousal (right).

	N+	Nn	N-	sum
S1+	270	104	38	412
S1n	386	519	313	1218
S1-	60	97	369	526
sum	716	720	720	2156
S2+	426	216	92	734
S2n	162	308	140	610
S2-	128	196	488	812
sum	716	720	720	2156

Table 5. The contingency table shows the relationship of the stimulus grouping into the affect conditions (+ = positive, n = neutral, - = negative stimuli) according to the IADS and IAPS stimuli norms (N group) and to the self-assessment with normal-sized (S1 group in upper table) and small-sized (S2 group in lower table) neutral condition.

into account. They are already reflected in the variances that characterise the norm ratings of the individual stimuli in the IAPS and IADS. Therefore, the question arises, if another sorting of the trials into the positive, neutral, and negative experience conditions could result in more homogeneous conditions of affective experiences. This is especially interesting for a later use of the data for the identification of physiological and neurophysiological features able to differentiate between the affective states.

To explore the potential of alternative ways to assign the trials to the conditions we made use of the gathered self-assessment data. We will refer to the alternative ways of sorted trials into conditions as grouping approaches, resulting in different groupings of the trials.

The most obvious grouping approach, and the one used so far, is to sort the recorded trials according the norm ratings of the affect-eliciting stimuli, coinciding with the stimuli conditions we constructed. In the following we will refer to this grouping approach as NORM, or as N in the tables.

When the self-assessment values are used as the basis for the stimuli- and thus trial-grouping, we obtain the SAM1 grouping (S1 in the tables). The deviations from the intended grouping become obvious, when directly comparing the overall number of trials per condition intended (positive: N+, neutral: Nn, negative: N-) with those derived from the new grouping approach (S1+, S1n, S1-) in a contingency table 5. Due to a rating trend towards the middle, thus toward the neutral condition, the positive and negative conditions are underrepresented in the number of trials.

However, by assuming each rating that deviates from the middle of the Likert scale by one scale unit towards one end of the scale to result from a negative or positive affective response, we obtain the SAM2 grouping (S2 in the tables). Here the responses are equally distributed over all three conditions (S2+, S2n, S2-), as the neutral condition is narrowed down to one Likert point. The relationship between the intended grouping, by use of IAPS and IADS norm values, and this this less strict grouping due to self-assessment can again be seen in table 5.

Group	Valence mean (std)	Arousal mean (std)
N+	5.29 (1.58)	4.01 (1.84)
Nn	4.49 (1.35)	3.75 (1.86)
N-	3.04 (1.70)	5.02 (2.03)
S1+	6.79 (0.64)	3.71 (1.94)
S1n	4.4 (0.74)	3.76 (1.78)
S1-	1.79 (0.75)	5.88 (1.59)
S2+	6.22 (0.81)	3.79 (1.89)
S2n	4.47 (0.13)	3.41 (1.70)
S2-	2.37 (0.98)	5.34 (1.79)

Table 6. The mean valence and arousal ratings per group and grouping method. The value in brackets is the standard deviation.

Table 6 enlists the means and standard deviations for the positive, neutral, and negative conditions according to the different grouping methods. The grouping of the stimuli based on the self-assessment leads to a clearer distinction of the conditions in terms of valence means and to a smaller standard deviation. As a consequence of the SAM2 grouping variation, however, the differences between condition means are decreasing again and the standard deviations of positive and negative condition are increasing. A Wilcoxon signed-rank test on the valence ratings revealed statistical significant differences ($p \leq 0.001$) for all emotion contrasts within all three grouping approaches. The same was observed for the arousal ratings, except for the contrasts of positive and neutral conditions. These differences were due to a higher arousal induced by the negative stimuli.

Summarising, the analysis of self-assessment rating means of the conditions of the NORM suggests that indeed different affective experiences were elicited. Furthermore, it was shown that the alternative grouping according to the self-assessments, to make the conditions more homogeneous in terms of elicited emotion, results in an imbalance of trials per conditions. This can be remedied by the limitation of the neutral condition to those trials that were not accompanied in a deviation from the central, and thus most neutral bin, of the self-assessment valence scale. Furthermore, we found differences in the arousal dimension, which have to be taken into account in the further study of the data.

To explore the effect of the different grouping methods in terms of physiological and neurophysiological differences, we analysed a subset of the available sensor information.

3.2. Analysis of the physiological data

We conducted a preliminary analysis of the physiological signals according to the NORM, SAM1 and SAM2 grouping. We used the non-parametric Wilcoxon signed-rank test to test for differences between the extracted features, as some of the groups were not normally distributed. The features shown in table 7 are significant with a p-value $\leq$ 0.05. (For this preliminary analysis we did not correct for the multiple tests conducted.)

Contrast	Significant Features
N+ vs N-	HF , $\sigma\{b\}$, $E\{t\}$
N+ vs Nn	$E\{h\}$, $E\{s\}$,$\sigma\{s\}$
N- vs Nn	$E\{h\}$, $E\{s\}$, $\sigma\{s\}$, $E\{t\}$
S1+ vs S1-	
S1+ vs S1n	$\sigma\{s\}$, $\delta^s_{\|1\|}$, $\sigma\{z\}$
S1- vs S1n	$\sigma\{c\}$
S2+ vs S2-	$\sigma\{z\}$
S2+ vs S2n	HF , $\delta^s_{\|1\|}$, $\sigma\{z\}$
S2- vs S2n	

Table 7. The significant (p $\leq$ 0.05) physiological features for the contrasts of negative (-), neutral (n), and positive (+) stimulus groups according to the NORM (N), SAM1 (S1), and SAM2 (S2) grouping methods.

Surprisingly, the NORM grouping results in the most significant differences between the conditions. Heart rate, Heart rate variability, blood volume pulse, temperature, and skin conductance are differentiating between the conditions. Specifically heart rate and skin conductance seem to differ between the emotional and the neutral conditions, while heart rate variability, blood volume pulse and temperature are differentiating the two valenced conditions.

For the SAM1 and SAM2 grouping we observed only differences in skin conductance, heart rate variability, and muscle activity. Intriguingly, the corrugator supercilii muscle (implied in frowning) is differentiating the negative condition from the neutral condition in the SAM1 grouping, while the zychomatic major muscle (implied in smiling) is differentiating between positive and neutral condition for both SAM groupings. Unexpectedly, two of the SAM contrasts could not be differentiated in terms of physiological responses.

3.3. Analysis of the neurophysiological data

For the preliminary analysis of the neurophysiological sensors according to the NORM, SAM1 and SAM2 grouping we concentrated on the alpha and theta frequency over the lateral and medial frontal cortex. Again we used the non-parametric Wilcoxon signed-rank test, as some of the groups were not normally distributed. Table 8 shows the significant (p $\leq$ 0.05) features. (As in the previous analysis of physiological features we did not correct for the multiple tests conducted.)

Contrast	Left α	Right α	Medial Θ
N+ vs N-		AF4	
N+ vs Nn			
N- vs Nn			
S1+ vs S1-	AF3		
S1+ vs S1n	AF3, F5	AF4, F4, F6	FCz
S1- vs S1n			FCz
S2+ vs S2-	F3, F5		
S2+ vs S2n	F3, F5		
S2- vs S2n			FCz

Table 8. The significant (p $\leq$ 0.05) EEG features for the contrasts of negative (-), neutral (n), and positive (+) stimulus groups according to the NORM (N), SAM1 (S1), and SAM2 (S2) grouping methods.

The most salient finding is the lower alpha power for the positive conditions. As there is a reciprocal relationship between alpha power and neural activity, this might indicate a stronger processing of positive stimuli.

The tests for alpha asymmetry between the electrode pairs AF3 and AF4, and F3 and F4, and F5 and F6 showed no significant differences.

Higher fronto-medial power in the theta band was found for neutral compared to negative conditions in the self-assessment contrasts. This relates to the study of Sammler et al. [33]. They found a fronto-medial increase in theta power for normal (positive) compared to distorted (negative) musical pieces and interpreted it as an emotional reaction associated with attentional processes. However, for the SAM1 grouping we found a decrease of theta power for positive compared to neutral trials, which seems to be a contradiction. A reconciliation is possible if one assumes that emotional stimuli in general might trigger these attentional processes observed over fronto-medial cortices.

Similar to the analysis of the physiological features we see the biggest difference between the normed grouping method on the one side and the two self-assessment based groupings on the other side. However, in contrast to the previous analysis, we now see the strongest difference for the self-assessment groupings, especially for SAM1.

4. Discussion

The analysis of the self-assessment data provided evidence for the validity of our stimulus sets. However, we observed a great variability in valence ratings for a given stimulus over subjects. This was to a certain degree expected, as individual differences already caused large variations in the ratings of the original stimulus sets of IADS and IAPS. We used multimodal stimuli, which were constructed of a valenced and a neutral unimodal stimulus. This might have weakened the effectiveness of the used stimuli further, leading to the observed trend of ratings towards the middle, i.e. the neutral condition.

Furthermore, we found significantly higher mean arousal values for the negative stimuli in the self-assessment data. This reflected the arousal bias observed in the norm ratings of the negative stimuli subsets. This effect has to be taken into account, when the (neuro-)physiological correlates of the affective experience elicited by the negative stimuli are interpreted, as a difference solely due to the experience difference in the valence dimension cannot be ensured.

As the choice of the right ground truth, the sorting of trials to the affective experience conditions, can have significant consequences for later classification attempts, we explored three ways to sort the recorded trials into positive, neutral, and negative conditions. The grouping of the trials according to the constructed stimuli groups (NORM grouping), exhibited large standard deviations, resulting from those stimuli that did not have the expected effect on the participants. To build more homogeneous conditions in terms of self-assessment ratings we grouped the trials according to those ratings (SAM1 grouping). Due to individual differences in rating styles this led to imbalanced group sizes. Specifically the negative and positive conditions contained only a small number of trials relative to the neutral condition. By a limitation of the neutral condition to the most central bin on the rating scale, we achieved a more balanced distribution (SAM2 grouping). However, also the self-assessment method is not free from biases or distortions [34]. Therefore, it is not necessarily the optimal choice for a solid ground truth construction.

A fourth sorting alternative would be a combination of stimulus reliability across participants and individual self-assessment. That is, to choose only those trials for an self-assessment based analysis, in which stimuli with relative unequivocal ratings were presented. That way we would reduce the overall number of trials, but could avoid the analysis of responses towards stimuli which might induce mixed emotions. These mixed emotions might have led to the variations of ratings for a given stimulus over subjects. By the removal of those stimuli from the data sets more homogeneous conditions could be created.

Another possibility to find suited sets of positive, neutral, and negative stimuli to build a valid ground truth is the use of physiological responses for verification. Marosi et al. [27] analysed only those trials in terms of EEG frequency activity, which were accompanied by a galvanic skin response. However, the analysis of EEG data requires a great amount of trials due to an inherent low signal-to-noise ratio. To explore the feasibility of such an approach the number of those physiological responses in the physiological data has to be determined. Furthermore, such an analysis might only find differences between trials that would theoretically be differentiable by physiological sensors, neglecting EEG features that might also differ between affective experience in the absence of physiological responses.

The preliminary analysis of physiological and neurophysiological sensors gave further evidence for a successful elicitation of different affective experiences by our approach. However, these findings were not free of contradictions. We found large differences between the NORM and the SAM grouping methods in number and type of physiological signals differing between affect conditions. We expected to find stronger differences between the conditions when grouping according to the self-assessments, as reported by Chanel and colleagues [7]. Similar to the current study, they elicited affective states (low, medium, and high arousal) via the presentation of IAPS stimuli. As we did not attempt a classification in the current analysis phase, we cannot directly compare our observed differences with their classification accuracy. However, while Chanel and his colleagues findings indicate a less robust pattern for the norm based grouping in general, we find more physiological features differentiating between conditions in the norm based grouping than in the self-assessment based grouping. For two of the SAM contrasts (S1+ vs. S1- and S2- vs. S2n) we couldn't show any significant effects for the physiological sensors at all. On the other hand, our finding that neurophysiological features do mostly differ between the self-assessment based conditions corroborates the results of Chanel et al. Here the most differences were found for the SAM1 grouping. Although we did not find the expected pattern in terms of alpha asymmetry, we observed consistent decreases of left-hemispheric alpha for the positive compared to the neutral and negative conditions and of fronto-medial theta power for the negative compared to the neutral condition.

The higher number of differences found in the EEG data for the SAM1 grouping could indicate that the emotional responses are more homogeneous for the groups established in this way. However, it might also be the result of some relatively small positive or negative groups, i.e. a small number of samples for some subjects in which possible outliers have a big effect in the statistical analysis. On the other hand, the SAM2 grouping might lead to the inclusion of neutral trials into the positive and negative conditions, and thus obscure the differences between the conditions.

Furthermore, it seems that for the analysis of temporally limited processes an analysis in shorter time windows is important. Didier et al. [18] showed that different subprocesses associated with affective responses are unfolding over different intervals of only few hundred milliseconds in the EEG. However, as auditory stimuli might have big interstimuli variations in the onset of affective response such a division of the trial into subtrials could lead to the comparison of different, unrelated parts of the emotional responses. These inadequate comparisons could lead to further variance in the signals and thus conceal the neural correlates of the emotional processes.

A further exploration of the data is needed to confirm the here presented preliminary results, resolve the contradictions, and find a reliable grouping method for the ground truth construction. The removal of artifacts will give a better insight into the true sources of the physiological and neurophysiological differences between the conditions.

5. Conclusion

We presented an analysis of an emotion elicitation experiment using multimodal stimuli and showed the validity of the experimental approach used along several dimensions. The approach will be used to study physiological and neurophysiological responses associated with affective experience while controlling the emotion-eliciting modality.

The analysis of the self-assessments of the participants emotional states in terms of valence and arousal suggested that the approach used is suitable for the induction of different affective states. However, it was also shown that the variance of the individual responses to the affective stimuli poses a great challenge in the search for (neuro-)physiological correlates of affective processes and their subsequent classification.

We studied different grouping methods to sort the acquired physiological and neurophysiological signals, that is according to the original grouping of stimuli, to the self-assessment data from the valence dimension, and to a self-assessment data with a more relaxed criterion for positive and negative conditions.

The comparison of the physiological features between the three emotion conditions for all three grouping methods revealed a variation of differentiating features over these approaches. Especially the grouping according to the normed values of the used stimuli differed from the two self-assessment groupings in number and types of distinguishing features. The analysis of EEG alpha and theta power revealed a contradictory pattern, with the self-assessment based grouping leading to the best differentiation between conditions.

A further analysis of the neurophysiological and physiological features, incorporating artifact removal and the rejection of particularly unreliable stimuli will yield a better understanding of apparently contradicting phenomena observed in this study. Finally, it will be the first step to an informed choice of features for the exploration of a multimodal affect classification.

Acknowledgements The authors gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science.

References

[1] M. Benovoy, J. Deitcher, and J. Cooperstock. Biosignals analysis and its application in a performance setting: Towards the development of an emotional-imaging generator. In *IEEE International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS)*, 2007.

[2] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[3] M. M. Bradley and P. J. Lang. Affective reactions to acoustic stimuli. *Psychophysiology*, 37(2):204–215, 2000.

[4] M. M. Bradley and P. J. Lang. The international affective digitized sounds (2nd edition; IADS-2): Affective ratings of sounds and instruction manual. Technical report, Gainesville: University of Florida, Center for Research in Psychophysiology, 2007.

[5] J. T. Cacioppo, R. E. Petty, M. E. Losch, and H. S. Kim. Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of Personality and Social Psychology*, 50(2):260–268, 1986.

[6] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607–627, 2009.

[7] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun. Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals. *Multimedia Content Representation, Classification and Security*, pages 530–537, 2006.

[8] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *MindTrek '08: Proceedings of the 12th International Conference on Entertainment and Media in the Ubiquitous Era*, pages 13–17, New York, NY, USA, 2008. ACM.

[9] M. Codispoti, M. M. Bradley, and P. J. Lang. Affective reactions to briefly presented pictures. *Psychophysiology*, pages 474–478, 2001.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.

[11] R. J. Davidson. Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20(1):125–151, 1992.

[12] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21, 2004.

[13] J. Etzel, E. Johnsen, J. Dickerson, D. Tranel, and R. Adolphs. Cardiovascular and respiratory responses during musical mood induction. *International Journal of Psychophysiology*, 61(1):57–69, 2006.

[14] S. H. Fairclough. Fundamentals of physiological computing. *Interacting with Computers*, 21(1-2):133–145, 2009.

[15] T. W. Frazier, M. E. Strauss, and S. R. Steinhauer. Respiratory sinus arrhythmia as an index of emotional response in young adults. *Psychophysiology*, 41(1):75–83, 2004.

[16] A. J. Fridlund and J. T. Cacioppo. Guidelines for human electromyographic research. *Psychophysiology*, 23(5):567–589, 1986.

[17] P. Gomez, S. Shafy, and B. Danuser. Respiration, metabolic balance, and attention in affective picture processing? *Biological Psychology*, 78(2):138–149, 2008.

[18] D. Grandjean and K. R. Scherer. Unpacking the cognitive architecture of emotion processes. *Emotion*, 8(3):341–351, 2008.

[19] J. J. Gross and R. W. Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.

[20] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007.

[21] A. Keil, M. M. Müller, T. Gruber, C. Wienbruch, M. Stolarova, and T. Elbert. Effects of emotional arousal in the cerebral hemispheres: a study of oscillatory brain activity and event-related potentials. *Clinical Neurophysiology*, 112(11):2057–2068, 2001.

[22] S. Khalfa, P. Isabelle, B. Jean-Pierre, and R. Manon. Event-related skin conductance responses to musical emotions in humans. *Neuroscience letters*, 328(2):145–149, 2002.

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

[24] K. Kim, S. Bang, and S. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, 2004.

[25] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Technical manual and affective ratings. Technical report, University of Florida, Center for Research in Psychophysiology, Gainesville, USA., 1999.

[26] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, 1993.

[27] E. Marosi, O. Bazán, G. Yañez, J. Bernal, T. Fernández, M. Rodríguez, J. Silva, and A. Reyes. Narrow-band spectral measurements of eeg during emotional tasks. *The International Journal of Neuroscience*, 112(7):871–891, 2002.

[28] M. M. Müller, A. Keil, T. Gruber, and T. Elbert. Processing of affective pictures modulates right-hemispheric gamma band eeg activity. *Clinical Neurophysiology*, 110(11):1913–1920, 1999.

[29] R. W. Picard. *Affective Computing*. The MIT Press, Cambridge, MA, USA, 1997.

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

[31] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, 61(1):5–18, 2006.

[32] J. A. Russel. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[33] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch. Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, 44(2):293–304, 2007.

[34] D. Sander, D. Grandjean, and K. R. Scherer. A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18(4):317–352, 2005.

[35] A. Schlögl and C. Brunner. Biosig: A free and open source software library for bci research. *Computer*, 41(10):44–50, 2008.

[36] R. Sinha, W. R. Lovallo, and O. A. Parsons. Cardiovascular differentiation of emotions. *Psychosomatic Medicine*, 54(4):422–435, 1992.

[37] R. M. Stern, W. J. Ray, and K. S. Quigley. *Psychophysiological Recording*. Oxford University Press, Inc., 2 edition, 2001.

[38] E. van den Broek, J. H. Janssen, J. Westerink, and J. A. Healey. Prerequisites for affective signal processing (asp). In P. Encarnao and A. Veloso, editors, *BIOSIGNALS*, pages 426–433. INSTICC Press, 2009.

[39] E. van den Broek, M. Schut, J. Westerink, J. van Herk, and K. Tuinenbreijer. Computing emotion awareness through facial electromyography. *Computer Vision in Human-Computer Interaction*, pages 52–63, 2006.

[40] C. M. van Reekum, T. Johnstone, R. Banse, A. Etter, T. Wehrle, and K. R. Scherer. Psychophysiological responses to appraisal dimensions in a computer game. *Cognition & Emotion*, 18(5):663–688, 2004.

[41] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. In *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pages 126–133, New York, NY, USA, 2007. ACM.