Jonghwa Kim and Pasi Karjalainen

Bio-inspired Human–Machine Interfaces and Healthcare Applications

Proceedings of B-INTERFACE 2010
1st International Workshop on Bio-Inspired Human–Machine Interfaces and Healthcare Applications
In conjunction with BIOTEC 2010
Valencia, Spain, January 2010
Jonghwa Kim and
Pasi Karjalainen (Eds.)

Bio-inspired Human-Machine Interfaces and Healthcare Applications

Proceedings of the
1st International Workshop on
Bio-inspired Human-Machine Interfaces
and Healthcare Applications
B-Interface 2010

In conjunction with BIOSTEC 2010
Valencia, Spain, January 2010

INSTICC PRESS
Portugal
Foreword

This volume contains the proceedings of the First International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications (B-Interface 2010). The Workshop was held in Valencia, Spain, 21-22 January 2010, under the umbrella of BIOSTEC 2010. The objective of B-Interface is to assemble researchers from diverse backgrounds together to discuss their ideas and solutions and to build a new vision in biosignal analysis, physiological pattern recognition, biomedical engineering, neural human-machine interface, affective human-machine interface, home sensor networks, and healthcare applications. The peer-reviewed papers in this volume cover both fundamental and applied topics in these multidiscipline research fields.

We would like to take this opportunity to thank all the international program committee members for their valuable reviews and help on launching this new workshop successfully. We gratefully acknowledge the professional support of INSTICC and BIOSTEC team for all organizational processes. Special thanks are due to Ms. Vera Coelho for preparing the workshop proceedings and continuous help throughout the workshop organization. No workshop would be successful without outstanding papers and inspiring presentations. We sincerely thank all the authors for their submissions and participation in B-Interface 2010.

January 2010,

Jonghwa Kim
University of Augsburg, Germany

Pasi Karjalainen
University of Kuopio, Finland
Workshop Chairs

Jonghwa Kim
University of Augsburg
Germany

and

Pasi Karjalainen
University of Kuopio
Finland

Program Committee

Elisabeth André, University of Augsburg, Germany
Egon L. van den Broek, Human-Centered Computing Consultancy, The Netherlands
Anthony Burkitt, University of Melbourne, Australia
Giuseppe Fico, Universidad Politécnica de Madrid, Spain
Claudia Igney, Philips Technologie GmbH, Germany
Pasi Karjalainen, University of Kuopio, Finland
Desok Kim, KAIST, Korea, Republic of
Jonghwa Kim, University of Augsburg, Germany
Benjamin Knapp, Queen’s University Belfast, U.K.
Toshio Moritani, Kyoto University, Japan
Marimuthu Palaniswami, University of Melbourne, Australia
Alois Schloegl, Graz University of Technology, Austria
Jindong Tan, Michigan Tech University, U.S.A.
Pekka Toivanen, University of Kuopio, Finland
## Table of Contents

Foreword ......................................................... iii

Workshop Chairs ............................................... v

Program Committee ........................................... v

**Full Papers**

Identifying Psychophysiological Correlates of Boredom and Negative Mood Induced During HCI .......................... 3
  *Dimitris Giakoumis, Athanasios Vogiannou, Ilkka Kosunen, Kostantinos Moustakas, Dimitrios Tzovaras and George Hassapis*

The Smart Sensor Integration Framework and its Application in EU Projects ....................................................... 13
  *Johannes Wagner, Frank Jung, Jonghwa Kim, Thurid Vogt and Elisabeth André*

A Spectral Mapping Method for EMG-based Recognition of Silent Speech ......................................................... 22
  *Matthias Janke, Michael Wand and Tanja Schultz*

**Short Papers**

Facial Features’ Localization using a Morphological Operation .......................................................... 35
  *Kenz A. Bozed, Ali Mansour and Osei Adjei*

Start and End Point Detection of Weightlifting Motion using CHLAC and MRA ........................................... 44
  *Fumito Yoshikawa, Takumi Kobayashi, Kenji Watanabe, Katsuyoshi Shirai and Nobuyuki Otsu*

Prerequisites for Affective Signal Processing (ASP) – Part IV .......................................................... 51
  *Egon L. van den Broek, Marjolein D. van der Zwaag, Jennifer A. Healey, Joris H. Janssen and Joyce H. D. M. Westerink*
Biometrics for Emotion Detection (BED): Exploring the combination of Speech and ECG .......................... 59
Marleen H. Schut, Kees Tuinenbreijer, Egon L. van den Broek and Joyce H. D. M. Westerink

Contagion of Physiological Correlates of Emotion between Performer and Audience: An Exploratory Study ............ 67
Javier Jaimovich, Niall Coghlan and R. Benjamin Knapp

Motion and Single-trial Biosignal Analysis Platform for Monitoring of Rehabilitation .......................... 75
Perttu Ranta-aho, Stefanos Georgiadis, Timo Bragge, Eini Niskanen, Mika P. Tarvainen, Ina M. Tarkka and Pasi A. Karjalainen

AffectPhone: A Handset Device to Present User’s Emotional State with Warmth/Coolness .......................... 83
Ken Iwasaki, Takashi Miyaki and Jun Rekimoto

Author Index ................................................................. 89
FULL PAPERS
Identifying Psychophysiological Correlates of Boredom and Negative Mood Induced During HCI

Dimitris Giakoumis$^{1,3}$, Athanasios Vogiannou$^{1,3}$, Ilkka Kosunen$^2$
Kostantinos Moustakas$^1$, Dimitrios Tzovaras$^1$ and George Hassapis$^3$

$^1$Centre for Research and Technology Hellas, Informatics and Telematics Institute
6th Km Charilaou-Thermi Road Rd., 57001, Thermi, Thessaloniki, Greece
{dgiakoum, tvog, moustak, tzovaras}@iti.gr
$^2$Helsinki Institute for Information Technology, Pilotti Building
Metsän Meinonkuja 4, 02100 Espoo, Finland
ilkka.kosunen@hiit.fi
$^3$Aristotle University of Thessaloniki
Department of Electrical and Computer Engineering, Greece
ghass@auth.gr

Abstract. This paper presents work conducted towards the automatic recognition of negative emotions like boredom and frustration, induced due to the subject’s loss of interest during HCI. Focus was on the basic pre-requisite for the future development of systems utilizing an “affective loop”, namely effective recognition of the human affective state. Based on the concept of “repetition that causes loss of interest”, an experiment for the monitoring and analysis of biosignals during repetitive HCI tasks was deployed. During this experiment, subjects were asked to play a simple labyrinth-based 3D video game repeatedly, while biosignals from different modalities were monitored. Twenty one different subjects participated in the experiment, allowing for a rich biosignals database to be populated. Statistically significant correlations were identified between features extracted from two of the modalities used in the experiment (ECG and GSR) and the actual affective state of the subjects.

1 Introduction

The development of automatic affect recognition systems based on biosignals has attracted much attention recently. The Jamesian theory [1] emphasizes the importance of peripheral signals in affect recognition, as it suggests there are specific patterns of physiology that relate to different emotions. During the last years, several important attempts have been made towards this direction; e.g. [2], [3], [4], [5], [6], [7], underlining the usefulness of peripheral activity for emotion assessment in diverse conditions.

The potential development of future game-playing systems which, based on an affective loop [8], will be able to adapt on the basis of the player’s emotions seems very interesting. Such systems will be able to identify whether the player is getting bored of the game and then adapt the playing context accordingly, in order to draw her/his attention again and induce more positive emotions. The first step towards this direc-
tion is the development of appropriate classifiers, able to effectively identify the user’s affective state of boredom, induced while playing. Previous work [9] has already shown that playing simple games like Tetris at different levels of difficulty gives rise to different emotional states that can be defined as boredom, engagement and anxiety. That specific work aimed at the automatic recognition of the player’s state of boredom from biosignal features; the desired emotional states were induced by playing Tetris game versions of different difficulty.

Moving towards more typical game-playing scenarios, in this work we focus on the identification of the most appropriate biosignal features to use for the effective classification of negative emotions like boredom and frustration, during playing typical 3D video-games. For this purpose, we have examined a set of features extracted from various biosignal modalities monitored during a negative emotion-induction experiment, based on repetitive playing of a 3D Labyrinth game. The aim of this analysis (Fig. 1) was to identify correlations between the extracted biosignal features and the actual affective state of the player, as the latter changed during the experimental session. For the purpose of the experiment, data was collected from four different biosignal modalities (EEG, ECG, GSR and EMG). However, since the data analysis is a work in progress, in this paper we focus on the two modalities that have until now produced the most significant correlations to the ground truth data, namely ECG and GSR.

Fig. 1. Overview of Experiment and Data Analysis presented in this paper.

In the following of this paper, the monitoring framework’s background is provided in Section 2, followed by the description of the methods used for feature extraction regarding each modality (Section 3). The experiment conducted for data collection is presented in Section 4. Finally, Section 5 presents statistically significant results from the analysis of the data collected, followed by conclusions, presented in Section 6.

2 Monitoring Framework Background

In this work, Electrocardiogram was used in order to assess the subject’s Heart Rate Variability (HRV). HRV describes the variations between consecutive heartbeats. The regulation mechanisms of HRV originate from the sympathetic and parasympathetic nervous systems and thus HRV can be used as a quantitative marker of the autonomic nervous system’s operation [10]. Features extracted from ECG, reflecting
the subject’s HRV have already been used together with features derived from other modalities in a number of studies targeting automatic emotion recognition e.g. [3], [4], [5], [6], [11]. Most commonly used HRV analysis methods are based on the time and frequency domains [12].

Time-domain HRV parameters are the simplest ones, calculated directly from the RR interval (or Inter-Beat Intervals) time series. These are the time series produced from the time intervals between the consecutive “R-peaks” of the raw ECG signal. The simplest time domain measures are the Mean and Standard Deviation of the IBIs. Commonly used HRV features are also the RMS of the IBI Sequential Differences (RMSSD) and the percentage within a time period of sequential differences that are over 50 milliseconds (pNN50). These features provide additional information about large-amplitude beat-to-beat changes in HR. In the frequency-domain analysis, power spectral density (PSD) of the IBI series is usually calculated. The commonly used frequency bands for HRV are Very Low Frequency (VLF, 0-0.04 Hz), Low Frequency (LF, 0.04-0.15 Hz), and High Frequency (HF, 0.15-0.4 Hz). The most common frequency-domain HRV features include the powers of VLF, LF, and HF bands and the LF to HF ratio.

Galvanic Skin Response (GSR), also referred to as Electrodermal activity (EDA), is a measure of skin conductance, which can be seen as an indirect measure of sympathetic nervous system activity [13]. The outer level of skin is highly resistive while the deeper layers of skin are highly conductive. These levels are “connected” by sweat glands, that when opened, create a pathway from the surface of the skin to the deeper, conductive level of the skin [14]. There are two main types of fluctuations of EDA that occur with stimulation: the momentary phasic responses and the more stable tonic level. Both phasic and tonic GSR features are commonly used towards automatic affect recognition [2], [3], [4], [5], [6], [11], [15]. GSR features commonly extracted and used in the literature are the Mean level of the GSR signal and the skin conductivity startle responses (Skin Conductance Response - SCR). SCRs are distinctive short waveforms (for a description see [4]) found inside the GSR signal that signify responses to internal or external stimuli.

3 Feature Extraction

In an effort to identify the best features to use for the development of proper classifiers regarding our specific application scenario, various features were extracted from the recorded signals and analyzed. The features used in the present analysis were checked for robustness to potential noise that could appear in the recorded signals.

Regarding the ECG modality, we considered the extraction of features from the subject’s Inter Beat Intervals (IBI) time series. ECG data were collected at a sampling rate of 256 Hz. Inter-Beat Intervals were calculated from the subject’s recorded Electrocardiogram, directly by our monitoring device’s (Procomp5) software. Prior to feature extraction, IBI artifacts were removed by a filter excluding IBIs over 1200 and under 500 ms. This filtering was applied in order to exclude IBI values which could not be normal, given our specific application scenario. Thresholds were set at
the values 500 and 1200 ms since an IBI outside this range would mean that the subject suddenly had a Heart Rate over 120 or under 50 beats/minute respectively.

Table 1. Features extracted from the Inter Beat Intervals time series of the ECG modality.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBI Mean per trial</td>
<td>[ ibi_{mean} = \frac{1}{n} \sum_{i=1}^{n} ibi_i ]</td>
<td>The average duration of the Inter-Beat Intervals during each trial</td>
</tr>
<tr>
<td></td>
<td>( n ) = number of IBIs during the trial</td>
<td></td>
</tr>
<tr>
<td>IBI SD per trial</td>
<td>[ ibi_{SD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (ibi_i - ibi_{mean})^2} ]</td>
<td>The Inter Beat Intervals Standard Deviation during a trial</td>
</tr>
<tr>
<td>IBI LF/HF per Trial</td>
<td>[ i bi_{LF/HF} = \frac{\sum_{i=1}^{n} lfp_i}{\sum_{i=1}^{n} hfp_i} ]</td>
<td>The average ratio of the Low Frequency band power to the High Frequency band power during a trial</td>
</tr>
<tr>
<td></td>
<td>( lfp_i ) = Low Frequency band power</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( hfp_i ) = High Frequency band power</td>
<td></td>
</tr>
<tr>
<td>IBI RMSSD</td>
<td>[ ibi_{RMSSD} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (ibi_{i+1} - ibi_i)^2} ]</td>
<td>RMS of the sequential differences of the IBI calculated for the whole trial</td>
</tr>
<tr>
<td></td>
<td>( \forall (ibi_{i+1} - ibi_i) \neq 0 )</td>
<td></td>
</tr>
<tr>
<td>IBI pNN50</td>
<td>[ ibi_{pNN50} = \frac{nd_{\geq50}}{nd} ]</td>
<td>Percentage of the number of sequential IBI differences that are over 50 ms during a trial</td>
</tr>
<tr>
<td></td>
<td>( nd_{\geq50} ) = number of sequential IBI differences that are over 50 ms</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( nd ) = total number of sequential differences during the trial</td>
<td></td>
</tr>
</tbody>
</table>

The time-domain (TD) and frequency-domain (FD) features shown in Table 1 were extracted from the IBI time series:

Regarding the GSR modality, we examined both the tonic and phasic electrodermal activity. The following features were extracted from the recorded GSR signals, sampled at a rate of 256 Hz:

The average value of the GSR signal during each trial (feature **GSR Mean per Trial**) was calculated with the formula:

\[ gsr_{mean} = \frac{1}{n} \sum_{i=1}^{n} gsr_i \]  \( \tag{1} \)

Where \( n \) = Total number of GSR samples during the trial

For the extraction of features related to the phasic electrodermal activity the subject’s Skin Conductance Responses (SCRs) during each trial were identified. Due to
the fact that the majority of trials were about half minute long, only the first twenty
five seconds of each trial were taken into account for the identification of SCR occur-
cences. Initially, the 1st derivative of the GSR signal values was calculated:

$$g_{sr} - d1_{raw} = \frac{g_{sr_{i+1}} - g_{sr_i}}{t_{i+1} - t_i}.$$  \hspace{1cm} (2)

Where $g_{sr}$ = Value of the $i^{th}$ GSR sample, $t_i$ = Timestamp of the $i^{th}$ GSR sample

The resulting time-series was convoluted with a 255-point Bartlett window. As a
result, the time series of the smoothed GSR 1st derivative values, $g_{sr\_d1}$ was produced.
Similarly to the procedure applied in [4], the occurrence of an SCR was detected by
finding two consecutive zero crossings, from negative to positive and from positive to
negative within the time series of the GSR smoothed first derivative ($g_{sr\_d1}$). The
maximum amplitude of the detected SCR was obtained by finding the maximum
value of the actual GSR signal between these two zero-crossings. Detected SCRs with
maximum amplitude smaller than the 10% of the maximum SCR amplitude detected
within each trial were excluded. After all SCRs were identified, together with their
maximum amplitude and duration, the features Number of SCRs, Average Amplitu-
ude of SCRs and Average Duration of SCRs were calculated for the first 25
seconds of each trial. The average value of the smoothed GSR first derivative ($g_{sr\_d1}$)
per trial was also extracted as feature (GSR 1st Derivative).

In order to perform proper group analysis, between-subject normalization was ap-
plied to the data collected from the ECG and GSR modalities, following two different
normalization methods: The first method (N1) produced the ratio of each feature to its
value obtained from the rest period of the specific subject. The second method (N2)
was based on the transformation of each sample into a percentage of the span for that
particular signal, similarly to the procedure applied in [16]: For each signal (GSR and
IBI), a global minimum ($x_{min}$) and maximum ($x_{max}$) were obtained for each participant
using all game playing trials, and the same global values were used for normalizing
each sample of the specific signal within each trial with the formula:

$$x_{N2}(i) = \frac{x(i) - x_{min}}{x_{max} - x_{min}}.$$ \hspace{1cm} (3)

Where $x$ = Samples of the GSR or IBI signals, $x_{N2}$ = Normalized samples of the GSR
or IBI signals

4 Experimental Setup

The aim of the experiment was to monitor the subject’s biosignals while the state of
boredom due to loss of interest was induced from a repetitive HCI task, namely the
repetitive playing of the same 3D Labyrinth game. The subject’s actual affective state
during the experimental session was assessed with the use of questionnaires, filled in
after each trial.
4.1 Stimuli

A basic 3D labyrinth game (Fig. 2) was developed for the purpose of the experiment. In order to complete the game, the players had simply to find the exit. The player could walk through the mazy corridors of the labyrinth using a 3D first person camera which is controlled by the WASD/Arrow keys and the mouse, a standard method used in commercial games. The game was developed in C++ using OGRE (http://www.ogre3d.org/) for graphics and the “Bullet” physics library (http://www.bulletphysics.com) for physics simulation. The tests were performed on a Laptop PC with an Intel Core 2 Duo T7700@2.40GHz CPU, 2 GBs of RAM and a NVIDIA GeForce 8600M GT graphics card. The game ran steadily on a 60 frame/sec rate.

Fig. 2. Screenshot of the VR Labyrinth game.

In order to induce drowsiness due to loss of interest effectively, the Labyrinth was designed to be a very simple one. Furthermore, in all repetitions the player started from the same point and the Labyrinth exit was always at the same place. Usually, after the third or fourth trial, the subject had learnt the shortest path to the exit. Thus, even though in the beginning (first two trials), the game was kind of exciting, as soon as the subject had learned the shortest path to the exit, the stimuli became an absolutely repetitive HCI task, ideal to induce negative emotions (e.g. boredom) due to loss of interest.

4.2 Hardware Setup

Both ECG and GSR biosignals were recorded using a Procomp5 Infiniti device (Fig. 3b). One three-electrode ECG sensor was placed at the subject’s forearms, or in cases that the subject had very low cardiac pressure, on its chest (Fig. 3a). Autoadhesive Ag/AgCl bipolar surface electrodes (bandwidth 10-500 Hz, pickup surface 0.8 cm2, inter-electrode distance 2 cm) were used for the ECG signal acquisition. Furthermore, one two-electrode GSR sensor placed at the subject’s left hand ring and small fingers (Fig 3c). The synchronization of the measurements and the VR Labyrinth game was based on a custom-made application, using the Network Time Protocol (NTP).
4.3 Participants and Procedure

The experiment was performed with twenty one subjects chosen among the participants of the eNTERFACE’09 summer workshop held in Genova, Italy. All participants frequently used computers in their work and were between 23 and 44 years old with 48 percent of them being 25 and 26 years old, 14 were males and 7 females. Forty two percent of the subjects were already familiar with 3D maze games but only nineteen percent of them played this type of game frequently (more than one hour per week). Trials from four different subjects were excluded from data analysis due to artifacts.

Initially, the subjects were asked to sign a consent form. After that, the sensors were installed, while the subject answered questions regarding personal details and previous gaming experience in the pre-questionnaire. At this point, the proper sensor placement was ensured, by checking carefully the robustness of the signal delivered from each monitoring modality. The recorded signals were checked on line for artifacts due to external noise or mechanical causes (e.g. subject’s motion). The preparation was renewed when severe artifacts were observed.

Once the sensors were properly placed, the subject was asked to relax with eyes-closed for one minute in order for the signal to stabilize and calibration data to be recorded (rest period). After the end of the rest period, the VR Labyrinth game was presented to the subject. From this point, the subject would play the Labyrinth game repeatedly. Each experimental session constituted of at least ten trials. Each trial started when the subject started playing the Labyrinth game and stopped as soon as the subject had found the exit, or a 10 minute time-limit had expired. Trials usually lasted from half to eight minutes. A mid-trial relaxation period of one minute was assigned between each trial. During this period, subjects had to fill in the mid-trials questionnaire. Using a Likert scale ranging from 1 to 5, they had to answer whether they would like to play the Labyrinth game again, whether they felt frustrated or bored of it and whether they were concentrated during the trial or focused on external events and/or personal thoughts. The experiment continued until the subjects had
played a minimum of ten trials and had signaled drowsiness/boredom in the questionnaire at least two times in a row. At the end of the experiment, subjects were asked to fill in a post-questionnaire, which was used for the assessment of their overall level of immersion during the entire experiment. Additional stages were also included in the experimental protocol; however these did not interfere with the induction of boredom and are outside the scope of the data analysis presented in this paper.

5 Results

In order to identify correlations between biosignal features and the subject’s actual affective state, we followed an analysis based on the Kendall’s tau correlation coefficient. In particular, the correlation between the subject’s Likert-scaled answers to specific questions of the mid-trial questionnaires, and each of the features extracted from the biosignals was calculated. Significance level was set at \( p=0.05 \) (*) and \( p=0.01 \)**. Questions considered in this analysis assessed the player’s tendency to stop playing the game (Q1), frustration (Q2) and boredom (Q3). Furthermore, we considered two more questions, assessing factors of the player’s affective state which can be thought opposite to boredom, like the player’s level of immersion/flow (Q4) and concentration (Q5). Several statistically significant correlations were identified and are summarized in Table 2.

![Table 2: Statistically significant correlations of Features extracted from the ECG and GSR modalities (N=Number of cases).](image)

<table>
<thead>
<tr>
<th>Question</th>
<th>Feature</th>
<th>Correlation (Kendall’s tau)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 (Play again)</td>
<td>IBI Mean per trial (N2)</td>
<td>( \tau=-0.284** ) ( p&lt;0.001 ) ( N=221 )</td>
</tr>
<tr>
<td>Q2 (Frustration)</td>
<td>Number of SCRs per trial</td>
<td>( \tau=0.133^* ) ( p=0.014 ) ( N=221 )</td>
</tr>
<tr>
<td>Q3 (Boredom)</td>
<td>IBI Mean per trial (N2)</td>
<td>( \tau=0.120^* ) ( p=0.019 ) ( N=221 )</td>
</tr>
<tr>
<td>Q4 (Flow / Immersion)</td>
<td>IBI LF/HF per Trial (N1)</td>
<td>( \tau=0.104^* ) ( p=0.042 ) ( N=221 )</td>
</tr>
<tr>
<td>Q5 (Concentration)</td>
<td>Number of SCRs per trial</td>
<td>( \tau=-0.199** ) ( p&lt;0.001 ) ( N=221 )</td>
</tr>
</tbody>
</table>

Frustration (Q2), boredom (Q3) and a tendency to stop playing the game (Q1 reversed) were found to correlate positively to the IBI Mean per trial feature. This indi-
cates a tendency of the subject’s Heart Rate to decrease, when a negative mood is induced from the interaction. Furthermore, frustration was also found to correlate to the LF to HF Average Ratio per trial and the Average value of the subject’s GSR signal per trial ($\tau = 0.193$, $p<0.001$). These features are indicative of the subject’s sympathetic nervous system activation and thus, her/his overall level of arousal, which is expected to increase with frustration. Boredom was found to correlate negatively to the subject’s number of SCRs per trial ($\tau = -0.199$, $p<0.001$), in accordance to the fact that increased numbers of SCRs are connected to higher levels of arousal. Furthermore, boredom was also found to correlate negatively to the average value of the GSR first derivative per trial ($\tau = -0.166$, $p<0.001$).

Regarding the questions assessing factors opposed to boredom, flow and immersion was found to positively correlate to the IBI Standard Deviation, LF to HF Average Ratio, RMSSD and pNN50, features connected to higher levels of immersion and arousal. Finally, concentration correlated positively to the subject’s Heart Rate (Inverse of IBI Mean), and negatively to the IBI Standard Deviation and RMS of Sequential Differences. The GSR Mean per trial feature correlated negatively ($\tau = -0.203$, $p<0.001$) to the subject’s concentration level.

Summarizing, several features extracted from ECG and GSR biosignals were found to correlate significantly to the subject’s actual affective state during the experimental session. These identified correlations could be used in the future as a guide for effective feature selection towards automatic emotion recognition, although the Kendall correlation coefficient did not reach very high values (up to ~0.25) in general. The EEG and EMG modalities used in the experiment have not produced equally significant results until now; however we strongly believe that more sophisticated preprocessing, analysis and fusion of all monitored modalities can lead to better results in the future.

6 Conclusions

This work presents a biosignals-based experiment, which focused on the identification of psychophysiological correlates of the changes in the user’s affective state during repetitive tasks in HCI. Data was collected from 21 subjects who played the same video game repeatedly, while their EEG, EMG, ECG and GSR signals were recorded. Various features were extracted from the biosignals and examined with the aim to identify statistically significant correlations between them and various Likert-scaled questions assessing the player’s affective state. The analysis was based on the Kendall’s tau correlation coefficient.

Various features extracted from ECG and GSR biosignal modalities were analyzed, so as to identify significant ones that could be used in the future for the automatic classification of negative emotions and mood, induced during 3D video-game playing. This work is planned to continue working towards the development of classifiers for the effective recognition of boredom, induced due to the player’s loss of interest. The major future goal is the development of a real-time monitoring framework for affective state classification, towards the realization of Human-Machine Interfaces based on affective loops.
References

The Smart Sensor Integration Framework and its Application in EU Projects

Johannes Wagner, Frank Jung, Jonghwa Kim, Thurid Vogt and Elisabeth André

Multimedia Concepts and Applications, University of Augsburg
Universitätsstr. 6a, D-86159 Augsburg, Germany
{wagner,jung,kim,vogt,andre}@informatik.uni-augsburg.de

Abstract. Affect sensing by machines is an essential part of next-generation human-computer interaction (HCI). However, despite the large effort carried out in this field during the last decades, only few applications exist, which are able to react to a user’s emotion in real-time. This is certainly due to the fact that emotion recognition is a challenging part in itself. Another reason is that so far most effort has been put towards offline analysis and only few applications exist, which can react to a user’s emotion in real-time. In response to this deficit we have developed a framework called Smart Sensor Integration (SSI), which considerably jump-starts the development of multimodal online emotion recognition (OER) systems. In this paper, we introduce the SSI framework and describe how it is successfully applied in different projects under grant of the European Union, namely the CALLAS and METABO project, and the IRIS network.

1 Introduction

Next generation human-computer interaction (HCI) claims to analyze and understand the way users interact with a system in a more sophisticated and smarter way than traditional systems do. No longer should it be the user who adapts to the system, but the system that adjusts itself to the user. This requires the system to be not only aware of the users’ goals and intensions, but also their feelings and emotions. For this reason, during the last decade, plenty of methods have been developed to detect a user’s emotions from various input modalities, including facial expressions [12], gestures [2], speech [9], and physiological measurements [7]. Also, multimodal approaches to improve recognition accuracy are reported, mostly by exploiting audiovisual combinations [1].

To date, however, most of the systems have been developed for offline processing and are not yet ready to be used under real-time conditions. This, of course, hampers their usefulness in practical applications. We believe that this is due to the varied difficulties real-time capability implies. On the one hand, an online system needs to deal with additional requirements, such as automatic segmentation, normalization issues, or the constraint to build on low-cost algorithms. On the other hand, there are certain implementation hurdles arising from the parallel processing of the input modalities, e.g. sensor data must be permanently captured and processed, while at the same time classification has to be invoked on detected segments. While for the offline analysis of
emotional corpora a number of powerful tools are available, such as Anvil\(^1\) for data annotation, Matlab for signal processing and Weka\(^2\) for classification, only little support is given for Online Emotion Recognition (OER) systems.

This paper reports on a framework called Smart Sensor Integration (SSI), which we have developed at our lab in order to support the building of OER systems. Originally designed to tune our own efforts on offline processing to the more challenging task of online processing, we have now made it available to the public, hoping to contribute to the implementation of OER systems in the future. In the following, we will shortly introduce the concept and architecture of the SSI framework, followed by a description how SSI is successfully applied in a number of projects funded by the European Union.

2 Smart Sensor Integration (SSI)

Online Emotion Recognition (OER) deals with two tasks: a learning phase, which involves data acquisition and training of a model from the data, and setting up an online recognizer, which is able to track a user’s affective state. A framework meant to support the creation of OER systems must consider both tasks. To this end, SSI, which in the first place is a framework made of tools for setting up an online recognition pipeline, also includes a graphical user interface for data acquisition and training.

2.1 Building Pipelines for Online Emotion Recognition

As depicted in the left part of Figure 1 the SSI framework has a three-layered architecture. The two lower levels, called data and communication layer, are responsible for data buffering and access. They temporarily store the processed signals. A synchronization mechanism handles the simultaneous access to the data and enables us to request snapshots of different signals for the same time slot, e.g., fetch video frames and the

---

\(^1\) Anvil is a video annotation tool offering hierarchical multi-layered annotation driven by user-defined annotation schemes. http://www.anvil-software.de/

\(^2\) Weka is an open source software, which offers a large collection of machine learning algorithms for data mining tasks. http://www.cs.waikato.ac.nz/ml/weka/
according portion of audio. An internal clocking mechanism takes care of data synchronization and restores it if necessary.

A developer, however, only has to deal with the top layer of the framework, called service layer. It allows the easy integration of sensors, processing algorithms, triggers, and output components to a signal processing pipeline capable of live input. The advantage for the developer is that he is not struggled with the usual problems online signal processing involves. To build the pipeline he can either integrate own code, or choose from a large number of available components.

Available components fall into three categories: Components related to data segmentation responsible to automatically detect chunks of activity. Components, which apply necessary pre-processing and extract from the detected chunks meaningful features that express the changes in the affective state of the user. And finally, the models, which are used to map from a continuous feature space to discrete emotion categories.

The right part of Figure 1 shows the basic configuration of an emotion recognition system implemented with SSI. A further advantage of using a generic framework like SSI is the great amount of flexibility and reusability that is gained. Since the processing units work independently of each other, they can be easily re-assembled in order to experiment with different kind of settings or to fit new requirements. In an earlier paper we show examples, how a basic OER system can be stepwise extended to a more complex one, e.g. by fusing information from multiple modalities [11]. The SSI framework has been published under LGPL license and can be freely obtained from https://mm-werkstatt.informatik.uni-augsburg.de/ssi.html.

2.2 Data Acquisition and Classifier Training

The model that a classifier uses to map continuous input to discrete categories is initially unknown and has to be learned from training data. This task falls into two main parts. The first part, referred to as data acquisition, involves the collection of representative training samples. Here, representative means that the picked samples should render the situation of the final system as accurate as possible. The second part concerns the actual training of the classifier. Basically, this is related to the problem of training a model that gives a good separation of the training samples, but at the same time is generic enough to achieve good results on unseen data.

Offline classification is usually evaluated on a fixed training and test set collected under similar experimental setting and by tuning the model parameters until they yield optimal recognition results. In contrast, the success of an online system depends on the ability to generalize on future data, which is not available to evaluate the classifier. Hence, special attention has to be paid that the training data is obtained in a situation, which is similar to the one it will be used in. To this end, we have developed a graphical user interface (GUI) on top of SSI, which gives non-experts the possibility to record emotional corpora and train personalized classifiers, which can be expected to give considerably higher accuracy than a general recognition system.

The GUI will be introduced by means of a concrete application within the CALLAS project in Section 3.1.
3 SSI in Practical Applications

A main motivation for building the SSI framework has been our participation in different EU projects, which are concerned with the analysis and development of novel user interaction methods. They all require to some extent online detection of the user’s affective state. In the following we explain how SSI is applied to this task.

3.1 The CALLAS Project

The CALLAS\(^3\) (Conveying Affectiveness in Leading-edge Living Adaptive Systems) project aims to develop interactive art installations that respond to the multimodal emotional input of performers and spectators in real-time. Since the beginning of the project a various number of showcases have already been created, such as the E-Tree\(^4\), which is an Augmented Reality art installation of a virtual tree that grows, shrinks, changes colours, etc. by interpreting affective multimodal input from video, keywords and emotional voice tone, or Galassie\(^6\) by Studio Azzurro, which creates stylized shapes similar to galaxies depending on the users’ emotional state detected from the voice.

One of the research questions raised by CALLAS is the interpretation, understanding, and fusion of the multimodal sensor inputs. However, since multimodal data corpora with emotional content are rather rare, effort was made to create a setup, which simplifies the task of data acquisition. In one experiment, which targets the mapping between gestures/body movements and emotion, and their relation to other modalities, such as affective speech and mimics, user interaction is captured with two cameras, one focused on his head and one on his whole body, and a microphone near the head. Additionally the users interact with different devices, such as Nintendo’s Wii Remote or a data glove by HumanWare\(^4\). To elicit the desired target emotion a procedure inspired by the Velten emotion induction method is used: first, a sentence with a clear emotional message is displayed and the user is given sufficient time to read it silently. Then the projection turns blank and the user is asked to express the according emotion through gesture and speech. It is up to the user to use own words or to say something, which is similar to the displayed sentence. Figure 2 illustrates the setting.

A main challenge regarding the experimental design is the proper synchronization between the different modalities. This, however, is an important requirement for the further analysis of the data. To obtain synchronized recordings we use the SSI framework. SSI already supports common sensor devices, such as webcam/camcorder, microphone and the Wii Remote, and can also record from multiple devices of the same kind. To connect more exotic devices, such as the data glove, SSI offers a socket interface, which can be used to grab any signal stream and feed it into the processing pipeline. SSI takes care of the synchronization by constantly comparing the incoming signals with an internal clock. If a sensor breaks down or does not deliver the appropriate amount of data, SSI compensates the lack in order to keep the stream aligned with the other channels. If it is not possible to capture all sensors with the same machine - which was actually the case with the two high-quality cameras, due to the vast amount of data they produce

\(^3\) http://www.callas-newmedia.eu
\(^4\) http://www.hmw.it/
- SSI offers the possibility to synchronize recordings among several computers using a broadcast signal, which is sent through the network.

For the accomplishment of the experiment we use the graphical interface of SSI, called SSI/ModelUI (see Figure 3). It works on the top of SSI and allows the experimenter to display a sequence of HTML documents, which contain the according instructions or stimuli. When the recording is finished it is added to the database. However, the functionality of the tool goes beyond this. During the recording SSI already tries to detect the interesting parts in the signals, e.g. when a user is talking or performing a gesture. These events are stored and can be reviewed together with the videos and the other raw signals. An annotator can now crawl through the events and adjust these pre-annotations. Finally, the tool automatically extracts feature vectors for the labelled segments and uses them to train a model for online classification. This way, the tool combines the tasks of recording, annotation and training in one application.

At the moment, we have conducted the presented experiment in two countries: Greece and Germany. We have recorded 30 subjects (10 Greek, 20 German) of which half were male and the other half female. The total length of recorded data sums up to almost 9h. The study will be repeated in Italy and possibly in countries of other partners. The analysis of the corpus has recently started and first results can be soon expected. The corpus will also allow us to analyze similarities/differences between individuals and between cultural groups in terms of selected features, recognition results and use of modalities.

3.2 The METABO Project

METABO\(^5\) is a European collaborative project with the aim to set up a platform for monitoring the metabolic status in patients with, or at risk of, diabetes and associated

\(^5\) http://www.metabo-eu.org/
The hike was fantastic! You won’t believe it! But we made it to the top!

Fig. 3. The Figure shows two screenshots of the graphical interface we have developed on top of SSI. It helps the experimenter to record the user interaction (left image) and let him review and annotate the recordings (right image). It also leads through the other steps of the training procedure. The GUI uses a DLL to communicate with SSI (see sketches below the screenshots).

metabolic disorders. The platform serves as bridge between physicians and patients to exchange information, but at the same time also provides recommendations for their clinical treatment. Recommendations are generated individually based on the patient’s metabolic behaviour model, which is learned from the patient’s history, and the current context derived from the patient’s physiological state. An essential requirement for such a system is the online acquisition and the prompt analysis of user data measured from different sensor devices. For this purpose a special case study is carried out, called the in-vehicle hypoglycemia alerting system (IHAS).

IHAS is an emotion monitoring system, which analyses a driver’s behaviour and emotional state during driving. Based on the measurements the system predicts hypoglycemia events and alerts the driver. The setting is motivated by the critical role fluctuant emotions play for diabetic drivers [8]. To derive the emotional state, the driver is equipped with several biosensors, including electromyogram (EMG), galvanic skin response (GSR), electrocardiogram (ECG), and respiration (RSP). The captured data is analyzed using an online recognition component implemented with the SSI framework.

In our previous studies we have mainly focused on the offline analysis of physiological data, where we have tested a wide range of physiological features from various analysis domains including time, frequency, entropy, geometric analysis, sub-band spectra, multi-scale entropy, and HRV/BRV. When we applied these features to different data sets recorded at our lab, we were able to discriminate basic emotions, such as joy, anger or fear, with an accuracy of more than 90% [10, 7]. For the sake of real-time analysis, large parts of the code that has been originally developed in Matlab, were ported to C++ and incorporated into the SSI framework and offline algorithms were replaced by corresponding real-time versions. That recognition results do not necessarily drop when
a generic set of recursively calculated real-time features is used instead of specialized offline features has been shown by Hönig et al. [5]. Many of the new real-time features are oriented to their proposed feature set. Figure 4 shows the state flow of the online recognition component we have developed with the SSI framework and which has been integrated into the METABO platform.

### 3.3 The IRIS Project

The IRIS project is concerned with the development of novel technologies for interactive storytelling. Recognition of affect is one of the novel techniques to be integrated into virtual storytelling environments. One of the showcases developed by Teesside University, EmoEmma[3], is based on Gustave Flaubert’s novel “Madame Bovary”. Here, the user can influence the outcome of the story by acting as one of the characters and their interaction mode is restricted to the emotional tone of their voice.

In a first step, we have integrated the EmoVoice system [9] into the SSI framework. EmoVoice has been developed at our lab as a tool for the recognition of affective speech. It provides tools for acoustic feature extraction (no semantic information is used) and building an emotion classifier for recognizing emotions in real-time. In total, a set of 1451 features can be derived from pitch, energy, voice quality, pauses, spectral and cepstral information as conveyed in the speech signal. The performance of the system has been evaluated on an acted database that is commonly used in offline research [7](http://iris.scm.tees.ac.uk)
emotion classes, 10 professional actors) and on a speech database we recorded for online classification (4 emotion classes, 10 German students) and achieved an average recognition accuracy of 80% and 41% respectively.

With the integration we have set the ground for a more complex analysis of the user interaction as now additional sensors can be added with minimal effort. To analyze the user’s affective and attentive behaviors, we use SSI in two ways: firstly, it offers the possibility to collect synchronized sensor data from users interacting with the system. In order to get realistic data the system response will be simulated at this point. Afterwards we analyze the recordings and based on the observations we use SSI to implement a pipeline that extracts the observed user behaviour in real-time. This information is then used by the planner to automate the system response. The architecture of the system is shown in Figure 5.

4 Conclusions

In the past pages, we have introduced our approach to contribute the building of OER systems, a framework called Smart Sensor Integration (SSI). First, we have discussed the multiple problems concerned with the implementation of such systems and how SSI faces them. After a short introduction of the framework architecture, we have moved on to explain how SSI is applied in a number of projects funded by the EU. SSI is freely available under LGPL license from the following address: https://mm-werkstatt.informatik.uni-augsburg.de/ssi.html.
Acknowledgements. The work described in this paper is funded by the EU under research grants CALLAS (IST-34800), IRIS (Reference: 231824) and Metabo (Reference: 216270).

References

A Spectral Mapping Method for EMG-based Recognition of Silent Speech

Matthias Janke, Michael Wand and Tanja Schultz
Cognitive Systems Lab, Karlsruhe Institute of Technology
Karlsruhe, Germany
matthias.janke@student.kit.edu, mwand@ira.uka.de
tanja@ira.uka.de

Abstract. This paper reports on our latest study on speech recognition based on surface electromyography (EMG). This technology allows for Silent Speech Interfaces since EMG captures the electrical potentials of the human articulatory muscles rather than the acoustic speech signal. Therefore, our technology enables speech recognition to be applied to silently mouthed speech. Earlier experiments indicate that the EMG signal is greatly impacted by the mode of speaking. In this study we analyze and compare EMG signals from audible, whispered, and silent speech. We quantify the differences and develop a spectral mapping method to compensate for these differences. Finally, we apply the spectral mapping to the front-end of our speech recognition system and show that recognition rates on silent speech improve by up to 12.3% relative.

1 Introduction

Automatic Speech Recognition (ASR) has matured to a point where it is successfully applied to ubiquitous applications and devices, such as telephone-based services and mobile personal digital assistants. Despite their success, speech-driven technologies still face two major challenges: recognition performance degrades significantly in the presence of noise, and confidential or private communication in public places are jeopardized by audible speech. Both of these challenges are addressed by Silent Speech Interfaces (SSI). A Silent Speech Interface is an electronic system enabling to communicate by speech without the necessity of emitting an audible acoustic signal.

In this paper, we present our most recent investigations in electromyographic (EMG) speech recognition, where the activation potentials of the articulatory muscles are directly recorded from the subject’s face via surface electrodes\(^1\). This approach has two major advantages: firstly, it is able to recognize silent speech, where not even a whispering sound is uttered. Secondly, the required technology is mobile, lightweight, and comes at very reasonable costs.

The use of EMG for speech recognition dates back to the mid 1980s, when Sugie and Tsunoda published a first study on Japanese vowel discrimination [1]. Competitive performance was first reported by [2], who achieved an average word error rate of 7%\(^1\) Strictly spoken, the technology is called surface electromyography, however we use the abbreviation EMG for simplicity.
on a 10-word vocabulary of English digits. A good performance could be achieved even when words were spoken non-audibly, i.e. when no acoustic signal was produced [3], suggesting this technology could be used to communicate silently. While the former approaches used words as model units, [4] successfully demonstrated that phonemes can be used as modeling units for EMG-based speech recognition, paving the way for large vocabulary continuous speech recognition. Recent results include advances in acoustic modeling using a clustering scheme on phonetic features, which represent properties of a given phoneme, such as the place or the manner of articulation. In [5], we report that a recognizer based on such bundled phonetic features performs more than 30% better than a recognizer based on phoneme models only.

While reliable automatic recognition of silent speech is currently heavily investigated and recent performance results come within useful reach, little is known about the EMG signal variations resulting from differences in human articulation between audible and silent speech production. Therefore, this paper studies the variations in the EMG signal caused by speaking modes. We distinguish audible EMG, i.e. EMG signals recorded on normally pronounced speech, whispered EMG, i.e. EMG signals recorded on whispered speech, and silent EMG, i.e. signals from silently mouthed speech.

Maier-Hein [6] was the first to investigate cross-modal speech recognition performance, i.e. models were trained on EMG signals from audible speech and tested on EMG signals from silent speech, and vice versa. The results suggested that the EMG signals are impacted by the speaking mode. Also, it was found that performance differences were lower for those speakers who had more practice in speaking silently while using the system.

Since the capability to recognize silent speech is the focus of Silent Speech Interfaces in general, and EMG-based speech recognition in particular, we consider it very crucial to investigating how the difference between speaking audibly or silently affects the articulation and the measured EMG signal. Furthermore, it is of very high interest to the silent speech research community how to compensate for these differences for the purpose of speech recognition.

In [7] we performed first experiments on cross-modal recognition of continuous speech based on units which were smaller than words. We showed that the difference between audible and silent speaking modes has a significant negative impact on recognition performance. We also conducted preliminary experiments on comparing the differences between recordings of audible and silent EMG and postulated a correlation between signal energy levels and cross-modal recognition performance. The current study is a continuation of these initial experiments. Here, we investigate the spectral content of the EMG signals of audible, whispered and silent speech, showing that there is a correlation between similar spectral contents and good recognition performance across different speaking modes. We then present a spectral mapping method which serves to reduce the discrepancies between spectral contents in different speaking modes. We perform additional experiments on whispered speech, since this speaking mode can be seen as an in-between of audible and silent speech: On the one hand, it is generally softer than audible speech and does not involve any vocal chord vibration, on the other hand, whispered speech still provides acoustic feedback to the speaker.
The remainder of the paper is organized as follows: In section 2 we describe the data corpus which we used for this study. Section 3 documents our EMG-based speech recognizer. Section 4 specifies our analytic experiments, and in section 5 we apply the results to our EMG-based speech recognizer. Section 6 concludes the paper.

2 Data Corpus

For our experiments, we recorded a corpus of EMG signals of audible, whispered, and silent speech of seven male speakers and one female speaker, aged between 24 and 28 years. Each speaker recorded between one and six sessions. The recording protocol was as follows: In a quiet room, the speaker read 50 English sentences for three times, first audibly, then in whispered speech, and at last silently mouthed. In each part we recorded one BASE set of 10 sentences which were identical for all speakers and all sessions, and one SPEC set of 40 sentences, which varied across sessions. In each session, these sentence sets were the same for all three parts, so that the database covers all three speaking modes with parallel utterances. The total of 50 BASE and SPEC utterances in each part were recorded in random order. In all recognition experiments, the 40 SPEC utterances are used for training, and the 10 BASE utterances are used as test set.

For EMG recording we used a computer-controlled 6-channel EMG data acquisition system (Varioport, Becker-Meditec, Germany). All EMG signals were sampled at 600 Hz and filtered with an analog high-pass filter with a cut-off frequency at 60Hz. We adopted the electrode positioning from [6] which yielded optimal results. Our electrode setting uses five channels and captures signals from the levator angulis oris, the zygomaticus major, the platysma, the anterior belly of the digastric and the tongue. In the audible and whispered parts, we parallelly recorded the audio signal with a standard close-talking microphone connected to an USB soundcard.

The collected data is intended to be comparable to the EMG-PIT corpus of EMG recordings of audible and silent speech [5]. However, the EMG-PIT corpus lacks whispered recordings, which we perceived as essential for our investigation. Figure 1 shows the electrode positioning and the final corpus of utterances which we used for this study.

<table>
<thead>
<tr>
<th>Speaker</th>
<th># Sessions</th>
<th>Average Session Length in [sec] (Training/Testing)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>audible</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>190/54</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>162/44</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>189/52</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>169/44</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>182/49</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>151/43</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>188/51</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>158/41</td>
</tr>
<tr>
<td>TOTAL</td>
<td>16</td>
<td>2753/751</td>
</tr>
</tbody>
</table>

Fig. 1. Electrode positioning (left), properties of the data corpus (right).
3 The EMG-based Silent Speech Recognizer

For all recognition experiments we apply our EMG speech recognition system based on three-state left-to-right fully continuous Hidden-Markov-Models [5], which are used to model phonetic features (PFs) representing phoneme-based properties. The modeling details are not required for the understanding of the remainder of this paper and are therefore omitted. The interested reader is referred to [5].

3.1 Feature Extraction

Figure 2 gives an example for a raw EMG signal (one channel) of the utterance “We can do it”. At the bottom of the signal the phoneme alignment is displayed.

The feature extraction is based on time-domain features [4]. Here, for any given feature $f$, $\bar{f}$ is its frame-based time-domain mean, $P_f$ is its frame-based power, and $z_r$ is its frame-based zero-crossing rate. $S(f, n)$ is the stacking of adjacent frames of feature $f$ in the size of $2n + 1$ (−$n$ to $n$) frames.

For an EMG signal with normalized mean $x[n]$, the nine-point double-averaged signal $w[k]$ is defined as

$$w[n] = \frac{1}{9} \sum_{n=-4}^{4} v[n], \quad \text{where} \quad v[n] = \frac{1}{9} \sum_{n=-4}^{4} x[n].$$

The rectified high-frequency signal is $r[n] = |x[n] - w[n]|$. In baseline experiments with audible EMG, the best word error rate is obtained with the following feature, which we use in this study as well:

$$TD15 = S(f_2, 15), \text{ where } f_2 = [\bar{w}, P_{w}, P_{r}, z_r, f].$$

As in [7], frame size and frame shift were set to 27 ms resp. 10 ms. In all cases, we apply LDA on the TD15 feature to generate a final feature with 32 coefficients.
3.2 Cross-modal Initialization

Initializing an EMG-based Continuous Silent Speech recognizer is a challenging task since in order to initialize acoustic models representing sub-word units (phonemes or phonetic features), one needs a time-alignment of the training material, i.e. information about the phoneme boundaries in the training utterances. Our previous works on audible EMG data used a conventional speech recognizer on the parallel-recorded audio stream in order to create such a time-alignment and then forced-aligned the training sentences. However, this method is infeasible for silent EMG, and information on the phoneme boundaries is not readily available.

We employ two kinds of initialization methods for the silent EMG recognizer [7]. Both methods rely on the existence of a “base recognizer”, which must be trained in advance on audible or whispered EMG by using the parallel-recorded audio stream. These methods are as follows:

**Cross-modal Testing.** We directly use the base recognizer to decode the silent EMG test set.

**Cross-modal Labeling.** We use trained models from the base recognizer to create a time-alignment for the silent EMG data. Then we forced-align the silent EMG data and do a full training run. This means that we create specific acoustic models for silent EMG.

For decoding, we use the trained acoustic model together with a trigram Broadcast News language model giving a perplexity on the test set of 24.24. The decoding vocabulary is restricted to the words appearing in the test set, which results in a test vocabulary of 108 words.

4 Spectral Analysis of Audible, Whispered and Silent EMG

4.1 Spectral Density Comparison

As a first experiment, we computed the power spectral density (PSD) of the EMG recordings of each session on a per-utterance and per-channel basis and then averaged over the utterances of each session and each speaking mode. For the PSD computation, we used Welch’s method [8], which works as follows:

- The input signal is divided into an integer number of segments with a 30 samples window length with 67% overlap.
- Each segment is windowed with one Hamming window to reduce spectral distortion.

Thus for each session we obtained the average spectral contents of the audible, whispered, and silent EMG recordings. Since the EMG signals were sampled with 600 Hz, the frequency range is between 0 and 300 Hz.

The left-hand side of figure 3 shows the average power spectral density of EMG channel 1 for the three sessions of Speaker 1, who has moderate experience in speaking silently. The curve shapes look similar, but the amplitudes differ for the speaking modes.
This speaker has an average Word Error Rate (WER) of 32.3% on audible EMG, while the Cross-Modal Labeling experiment gives a WER of 87.8% on silent EMG. Whispered speech is recognized with 51.2% WER. The right-hand side of figure 3 charts the PSD of a speaker well practiced in speaking silently, with good recognition rates for all speaking modes. The shape of the PSD curves is very similar to those of speaker 1, and the ratio between the energy contents per frequency is much closer to one than for speaker 1. Also note that the average signal energy is an order of magnitude higher than for speaker 2. For speaker 2 the WERs are 34.3% for audible EMG (which is about the same as for speaker 1), but only 6.1% for whispered EMG, and 26.3% for silent EMG. The good performance for whispered speaking mode might be related to the higher energy.

Except for some recordings which contained 50 Hz power line noise, the PSD shapes look quite similar across channels. Based on physiological differences and individual articulation characteristics the maximal values of the PSD differ. The maximum PSD on channel 1 varies between 0.017 and 0.22 for the eight speakers.

Fig. 3. PSD (Channel 1) of three Sessions of Speaker 1 (left) and Session 1 of Speaker 2 (right) with audible, whispered and silent speech.

These examples suggest that the recognition rates on silent speech and the similarity of the PSD curves are related to each other. In order to quantify this statement, we performed the following computations, individually for each session:

1. We computed the ratio of audible EMG and silent EMG PSD of each channel for each frequency bin and took the mean of this ratio over the frequency bins. This gives a value representing the spectral energy discrepancy between audible and silent EMG.
2. We calculated the WER difference between audible EMG and silent EMG (Cross-Modal Labeling Experiment) as a measure for the disparity of EMG recognition performance on audible and silent speech. Note that except for one session, all speakers perform better results with audible than silent speech.

A scatter plot of these values for each of the 16 sessions and for EMG channel 1 is shown in figure 4. The correlation coefficient is 0.67 and indicates that speakers with
high spectral energy difference between speaking modes tend to have a high difference in audible and silent speech recognition performance. The scatter plots for EMG channels 2 to 6 show quite similar characteristics, the correlation coefficients for the other channels are 0.44 for channel 2, 0.66 for channel 3, 0.74 for channel 4, and 0.75 for channel 6. In some recordings of sessions 1-2, channel 2 shows noticeable power line noise, which appears in the PSD graph as a high peak centered around 50 Hz. This explains why the correlation coefficient for channel 2 is much lower than for the other channels: When session 1-2 is removed from the computation, the correlation coefficient for channel 2 increases to 0.72.

![Fig. 4](image)

**Fig. 4.** Scatter plot comparing the ratio between power spectral density (PSD) of audible EMG and PSD of silent EMG and the difference of word error rates (WER) on silent and audible EMG for each session, with regression line. The PSD is for EMG channel 1 only and was maximized over frequency bins. The WER for silent EMG is from the “Cross-Modal Labeling” system.

### 4.2 Comparison based on Phoneme Classes

As a final experiment, we investigated the relationship between the spectral contents of audible and silent speech in more detail. It may be assumed that computing the spectral contents of the EMG signal on a per-utterance basis is a rather coarse way to analyze the highly time-variant EMG signal. Therefore we split up the signals into frames with a length of 27ms and a frame shift of 10ms, which is the same windowing used by the EMG recognizer. We then grouped the frames according to the phonetic features they represented. As our first experiment, we compared the spectral components of vowels and consonants taken from the audible EMG signal and the silent EMG signal. Figure 5 shows the PSDs for consonants and vowels of session 3 of speaker 2.

While the PSD shapes of silent and audible consonants differ only little, there is a noticeable difference in the vowel PSD chart. A reason for the higher vowel PSD in audible EMG could be the fact that (English) vowels are syllable peaks and thus major articulatory targets. When the speaker lacks acoustic feedback while articulating, this might have the consequence that acoustic targets are not fully reached any more, thus causing a less intense vowel articulation. However, further research on this question is necessary.
Spectral Mapping and Recognition Experiments

With the information we obtained about the relationship between the spectral contents of EMG signals of audible, whispered and silent speech and the recognition performance of the respective cross-modal recognizers, we developed a spectral mapping method which is applied to the silent EMG signals. This mapping is applied independently to each channel and each session and works as follows:

1. We define a “base speaking mode” (audible or whispered) for this experiment.
2. We compute the PSD ratio between the target silent EMG and the EMG signal of the base speaking mode (as a function of the frequency). The result is averaged over all utterances of one session. We call this ratio mapping factor.
3. Each utterance is transformed into the frequency domain by the Fast Fourier Transform (FFT), then each resulting frequency component is multiplied by the corresponding mapping factor, and the resulting frequency representation of the signal is transformed back into the time domain by application of the inverse FFT.
4. After this procedure, the transformed signal is used for the training and testing process as usual. We tested both Cross-Modal Testing and Cross-Modal Labeling on the transformed silent EMG, where the transformation was computed with the same base speaking mode which was also used to train the base EMG recognizer.

The resulting WERs for silent EMG and an audible EMG base system are charted in Figure 6. One can see that particularly for Cross-Modal Testing, spectral mapping indeed yields a large improvement: The average WER without spectral mapping is 62.13%, which by our mapping approach is improved to 54.58%. So we achieved a relative improvement of 12.3%. Note that the improvements for Speaker 2 tend to be slightly smaller, which can be explained by the fact that this speaker is most experienced in speaking silently and tends to have good recognition rates for both silent and audible speech. The Cross-Modal Labeling System yields an average WER of 54.36%; with spectral mapping this is improved by 6.74% relative to 50.7%.

The overall improvement for cross-modal labeling is smaller than for cross-modal testing. This may be explained by the fact that cross-modal testing directly applies the
recognizer trained on audible or whispered EMG to silent EMG, whereas cross-modal labeling means that a full training is performed on the silent EMG data, which should lessen the influence of the difference between the speaking modes.

![Fig. 6. Word Error Rates for the silent/audible mapped data.](image1)

As a next step, we performed the same set of experiments with whispered EMG as the base system. The resulting word error rates can be seen in figure 7. As with the audible EMG base system, spectral mapping improves the resulting WER in the cross-modal testing experiment from 55.88% to 51.07%, which is a relative improvement of 8.6%. However, in the case of Cross-Modal Labeling, the Word Error Rate increases from 50.31% to 52.77%. The resulting Word Error Rates for the spectral mapping experiments are summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Cross-Modal Testing</th>
<th>Cross-Modal Labeling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline With Mapping</td>
<td>Baseline With Mapping</td>
</tr>
<tr>
<td>audible/silent</td>
<td>62.13% 54.48%</td>
<td>54.36% 50.70%</td>
</tr>
<tr>
<td>whispered/silent</td>
<td>55.88% 51.07%</td>
<td>50.31% 52.77%</td>
</tr>
</tbody>
</table>

![Fig. 7. Word Error Rates for the silent/whispered mapped data.](image2)
In particular, it can be seen that for three out of four experiments, particularly when no spectral mapping is performed, using whispered EMG is superior to using audible EMG as the base signal. This supports our earlier claim that whispered speech is an in-between of audible (normal) and silent speech.

6 Conclusions

Using EMG signals captured from the articulatory muscles, we compared the discrepancies between different speaking modes, analyzing corresponding audible, whispered and silent utterances. We showed that EMG signals from audible speech generally have a higher spectral power than EMG signals from silent speech, and that this power ratio and the Word Error Rate difference between speech recognition on audible and silent EMG correlate with a correlation coefficient of up to 0.75. With this information a mapping from silent to audible/whispered EMG could be achieved, which resulted in 12.3% relative improvement in our audible-to-silent Cross-Modal Testing System.

References

SHORT PAPERS
Facial Features’ Localization using a Morphological Operation

Kenz A. Bozed, Ali Mansour and Osei Adjei

Department of Computer Science and Technology, University of Bedfordshire
Park Square, Luton, LU1 3JU, U.K.
(kenz.bozed,ali.mansour,osei.adjei)@beds.ac.uk

Abstract. Facial features’ localization is an important part of various applications such as face recognition, facial expression detection and human computer interaction. It plays an essential role in human face analysis especially in searching for facial features (mouth, nose and eyes) when the face region is included within the image. Most of these applications require the face and facial feature detection algorithms. In this paper, a new method is proposed to locate facial features. A morphological operation is used to locate the pupils of the eyes and estimate the mouth position according to them. The boundaries of the allocated features are computed as a result when the features are allocated. The results obtained from this work indicate that the algorithm has been very successful in recognising different types of facial expressions.

1 Introduction

Developments in digital image processing have grown with different algorithms for various applications of Computer vision techniques. Such applications have been reported by Lekshmi et al. [7] for face detection, Hannuksela et al. [3] for facial feature extraction and Mohamed et al. [10] for face recognition.

Face detection is one the active research applications in these areas. Face detection is defined by Yang et al. [13] to find the location and the size of the face in the input image. Some of face detection approaches do not have any assumptions regarding the number of faces in the image but they assume that a face exists in the image in order to classify as face and non-face regions [1], [4]. In the facial localization, it is normally assumed that the input image has at least one face. Generally, facial recognition problems are based on the features in the face. Salient features can be recognized easily by human eyes but it is challenging to locate and extract these features using a machine. The challenges of these applications are associated with pose, structural components, facial expressions, illumination, occlusion, and image quality of the subjects [11], [13]. Previous research has been concerned with the applications of face detection and recognition [9]. Many methods have been developed to locate and extract facial features. These methods classify into two categories: Feature based and Holistic. In the feature-based method, face recognition relies on the detection and localization of facial features and their geometrical relationships [1]. In a holistic method, a full face image is transformed to a point on a high dimensional space such as Active Appearance Model (AAM) [8], neural nets [5].
The morphological operation is a well-known technique used in image processing and computer vision for manipulating image features based on their shapes [2]. However, some methods need a considerable amount of computational or intensive memory to implement, and improve the speed and accuracy [7]. Our research aims to develop a simple and an accurate method that can be used in facial systems such as emotional detection.

In facial detection systems, eyes detection is a significant feature in the human face, where the detected eyes are easier to locate than other features. Also, the localization of the eyes is a necessary stage to help in the detection of other facial features which can be used for facial expression analysis as they convey the human expressions.

Although research have been done in this area, the process of solving the problem of facial features’ detection is still incomplete due to its complexity [6], [7]. For example, face posture, occlusions and illumination have effects on the performance of the features’ detection.

In this paper, a facial localization algorithm for salient feature extraction is presented. The algorithm consists of three steps: (1) a morphological process is applied to search the darkest parallel features in the upper face as a result of eyes localizations; (2) the distance between the estimated pupils is used to locate the mouth. (3) Localization of the salient features is used to compute their boundaries.

2 Facial Feature Localization

Features that are commonly used to characterise the human face are the eyes and mouth. It is normally assumed that the facial region is present in the input image and the features are searched within this region. The algorithm is based on the observation that some features such as the pupils of the eyes are darker than other facial features. Therefore, morphological operations can be used to detect the location of the eyes. The morphological operations are compatible with rough feature extraction for their fast and robust nature [3], [12].

The method proposed in this paper involves the morphological technique to detect the pupil of the eyes, and then the distance between them is used to detect the position of the mouth. The method is also simple and less computationally intensive. It has the advantage of using three facial features points instead of using the holistic face such as Active Shape Model with 58 facial feature points to locate the features [14].

The morphological erosion operation is applied on a grayscale face, using this operation to remove any pixel that is not completely surrounded by other pixels. The operation is applied when assuming 8-pixels are connected in order to reduce the unnecessary pixels in the boundaries of the face. Fig. 1 shows the some faces after applying the erosion operation. The eyes localization is determined based on the darkest pixels that are close to each other. The positions of the eyes allow the distance between them to be computed and also to locate the mouth.
2.1 Eyes Detection

The upper face is scanned individually to search for the pupils of eyes. However, when the darkest pixel is obtained for every eye, the algorithm is searched again for all the pixels that have the same value as the darkest one. Fig. 2 shows correct eyes detection where the search algorithm of the darkest pixel is satisfied. Also, Fig. 3 shows the final eyes detection where the pupil of the eyes estimated the darkest pixels of each eye.
The location of the pupils is calculated based on the average of the darkest pixel of each eye. Some experiments gave unsuccessful eyes location detection. These were corrected by adjusting the distance between them in the order of 15 to 20 pixels. Fig. 4 shows some unsuccessful eyes detection, and correcting this fault detection based on the distance between the averaged dark pixels.

![Fig. 4. Re-correct eyes detection.](image)

2.2 Mouth Detection

The mouth detection algorithm is presented when the localization of the eyes is known. Otherwise, the algorithm ignores this face. The mouth position is calculated according to the distance between the estimated pupils of both eyes.

The $L_{left}$ represents the computed centroid point $(V_{left_x}, V_{left_y})$ of the left eye (i.e. pupil of the left eye), and the $R_{right}$ represents the computed centroid point $(V_{right_x}, V_{right_y})$ of the right eye (i.e. pupil of the right eye).

The distance between the eyes $D_{eyes}$ is computed as follows:

$$ D_{eyes} = (V_{left_x}, V_{left_y}) - (V_{right_x}, V_{right_y}) $$

(1)
The average of the pupils $\overline{D_{\text{Eyes}}}$ is used to estimate the mouth position that represents the middle of distance between the eyes illustrated in equation (4). Therefore, it is computed by averaging the $L_{\text{Eyes}}$ and $R_{\text{Eyes}}$ as shown in equations (2) and (3).

$$X_{\text{Eyes}} = \mu(X_{\text{Left}} - X_{\text{Right}})$$  \hspace{1cm} (2)

$$Y_{\text{Eyes}} = \mu(Y_{\text{Left}} - Y_{\text{Right}})$$  \hspace{1cm} (3)

$$D_{\text{Eyes}} = (X_{\text{Left}} - X_{\text{Right}})$$  \hspace{1cm} (4)

The centroid point of the mouth $(X_{\text{M}} , Y_{\text{M}})$ is computed based on equations (1), (2) and (3) as follows:

$$X_{\text{M}} = X_{\text{Eyes}} + D_{\text{Eyes}}$$  \hspace{1cm} (5)

$$Y_{\text{M}} = Y_{\text{Eyes}}$$  \hspace{1cm} (6)

In this work, the facial features are segmented from the face image based on the $D_{\text{Eyes}}$.

Fig. 5 illustrates the centroid point of the left eye $(X_{\text{Left}}, Y_{\text{Left}})$, the centroid point of the right eye $(X_{\text{Right}}, Y_{\text{Right}})$, the distance between the eyes $D_{\text{Eyes}}$, and the centroid point of the mouth $(X_{\text{M}}, Y_{\text{M}})$.

**Fig. 5.** Final eyes detection with eyes distance and mouth position estimated.

### 3 Facial Feature Boundaries
After the possible facial features are detected, the distances $D_{\text{eyes}}$, $(X_{\text{left eye}}, Y_{\text{left eye}})$, $(X_{\text{right eye}}, Y_{\text{right eye}})$ are applied to evaluate the features’ boundaries.

The boundaries are determined according to the $D_{\text{eyes}}$ based on the experimental evaluation. The width and height of the facial features are calculated, where $W_\text{e}$ and $H_\text{e}$ are the width and height of the rectangles of each eye illustrated in equations (7) and (8) respectively. Also, equations (9) and (10) show $W_\text{m}$ and $H_\text{m}$ represent the width and height of the rectangles of the mouth.

$$W_\text{e} = \frac{2}{3} D_{\text{eyes}}$$

(7)

$$H_\text{e} = \frac{2}{3} D_{\text{eyes}}$$

(8)

$$W_\text{m} = D_{\text{eyes}}$$

(9)

$$H_\text{m} = \frac{2}{3} D_{\text{eyes}}$$

(10)

The left eye coordinate can be calculated as:

$$X_L = X_{\text{left eye}} - \frac{D_{\text{eyes}}}{3}$$

(11)

$$Y_L = Y_{\text{left eye}} - \frac{D_{\text{eyes}}}{3}$$

(12)

where $(X_L, Y_L)$ is the upper left corner coordinate of the left eye. In the same way, the right eye coordinate can be calculated as:

$$X_R = X_{\text{right eye}} - \frac{D_{\text{eyes}}}{3}$$

(13)

$$Y_R = Y_{\text{right eye}} - \frac{D_{\text{eyes}}}{3}$$

(14)

where $(X_R, Y_R)$ is the upper left corner coordinate of the left eye.

Furthermore, the upper corner of the mouth coordinate $(X_M, Y_M)$ can be calculated as:

$$X_M = X_{\text{left eye}} - X_{\text{right eye}} - X_{\text{left}}$$

(15)
The boundaries and the centroid points are illustrated in Fig. 6.

Briefly, once the eyes are identified correctly, the mouth is detected from the distance between the eyes. Then, the boundaries are computed as the following Fig. shows.

4 Experimental Results

The efficiency of this algorithm was tested on individuals’ images captured as frontal faces using a digital camera from the same distance and with normal room lighting conditions. It is well known that the difficulties to locate the features exactly on each face due to the face structure and the difference of face features. The proposed algorithm located the face features based on the erosion operation on the greyscale facial
cropped image and distance between the pupils of the eyes were computed and ignored any image that did not satisfy the location of the eyes correctly. The eyes detection based on darkest pixels of each eye. As the localization of the eyes is identified, the mouth location is computed based on the distance between the estimated pupils of both eyes. The boundaries of facial salient features are computed according to position based on calculated distance.

This algorithm needs adjusting due to the presence of some incorrect detection of the location of the eyes as a result of lighting and some occlusions such as glasses. A success rate of over 91% has been achieved based on a sample rate of 318 images. The following table shows the ratio detection for every facial feature.

<table>
<thead>
<tr>
<th>Features</th>
<th>Ratio of Feature detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left Eye</td>
<td>96.3%</td>
</tr>
<tr>
<td>Right Eye</td>
<td>95.6%</td>
</tr>
<tr>
<td>Mouth</td>
<td>91.9%</td>
</tr>
</tbody>
</table>

The experiments’ results show that locating the eyes is more accurate compared to the mouth. Therefore, further work is needed to increase the accuracy of features location.

5 Summary and Future Work

This work presents a new algorithm based on morphological process to detect the eyes localization and use the distance between them to locate the mouth position. The method defines a morphological operation to extract the important contrast regions of the face. These features are robust to lighting changes.

Future work will concentrate on improving the mouth detection to reduce the false rate detection. The false eyes detection can be enhanced as well, which will increase the ratio of features detected. The outcome of this algorithm can be used in other facial detection systems such as the analysis of facial expressions.

Acknowledgements. We would like to thank Mr. Xiao Guo, a PhD researcher at University of Bedfordshire University, Computer Science and Technology, who kindly allowed the use of his photos in this paper.

References


Start and End Point Detection of Weightlifting Motion using CHLAC and MRA

Fumito Yoshikawa¹, Takumi Kobayashi¹, Kenji Watanabe¹
Katsuyoshi Shirai² and Nobuyuki Otsu¹

¹ AIST Tsukuba, C2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
{yoshikawa.fumito, takumi.kobayashi
kenji.watanabe, otsu.n}@aist.go.jp
² Japan Institute of Sports Sciences, 3-15-1 Nishigaoka, Kita-ku, Tokyo 115-0056, Japan
shirai.katsuyoshi@jiss.naash.go.jp

Abstract. Extracting human motion segments of interest in image sequences is essential for quantitative analysis and effective video browsing, although it requires laborious human efforts. In analysis of sport motion such as weightlifting, it is required to detect the start and end of each weightlifting motion in an automated manner and hopefully even for different camera angle-views. This paper describes a weightlifting motion detection method employing cubic higher-order local auto-correlation (CHLAC) and multiple regression analysis (MRA). This method extracts spatio-temporal motion features and learns the relationship between the features and specific motion, without prior knowledge about objects. To demonstrate the effectiveness of our method, the experiment was conducted on data captured from eight different viewpoints in practical situations. The detection rates for the start and end motions were more than 94% for 140 data in total even for different angle views, 100% for some angles.

1 Introduction

Detecting and segmenting human action and behavior of interest in video sequences is necessary in various applications such as quantitative motion analysis and video browsing. It is a fundamental procedure for understanding the motions in question.

In sport motion analysis, especially for weightlifting, athlete’s sport-motions are often analyzed for improving their sport performance by using video sequences captured during competition and training. In biomechanical studies of weightlifting, much research efforts have been made to explore the relations between several kinematic parameters, such as time from barbell lift-off to its maximum height, and the winning lifts by conventional approaches involving manual indexing operations [1, 2]. In practical situations, however, quick feedback of the resultant quantitative data and/or videos is expected to be conducted for the relevant coaches and athletes without using any intrusive manner in data acquisition, for example by placing markers on the human body. In order to reduce the burden imposed on human operators for more detailed analysis, automation of motion analysis is required. Thus, first of all, automatic detection of the start and the end of predefined single motions
the snatch, or the clean and jerk) in weightlifting is essential for both the quantitative analysis and the effective video handling.

For the task to detect and recognize the full-body human motions, many researchers have investigated the performance of various video-based motion analysis methods [3 - 6]. These methods require segmentation of the target objects such as persons, in which the segmentation error tends to affect final recognition. In addition, the conventional approaches include sequential and procedural processes require too special and tedious steps. These make it difficult to design adaptive and real-time systems.

In recent years, on the other hand, a scheme of adaptive vision system has been presented, which comprises two stages of feature extraction, namely, higher-order local auto-correlation (HLAC) or its extension CHLAC (Cubic HLAC) [7] and multivariate analysis [8, 9]. Concerning human motion and behaviour analysis, CHLAC approach has been successfully applied to motion recognition [7], unusual motion detection [10, 11] and motion segmentation [12]. The CHLAC approach, however, has not been applied to detection of predefined motions. In [12], the task of segmenting single weightlifting motions into detailed primitive motions is addressed in the experiment; however, the segmentation methods proposed there need image sequences to be clipped so as to include the entire weightlifting action of interest in advance.

In this paper, we applied CHLAC and MRA to the start and end time point detection in weightlifting, obeying the simple statistical scheme framework [9]. In the experiment we used the dataset which had been acquired by capturing national elite athletes’ lifts from eight different viewpoints in practical situations. The experimental results demonstrated the effectiveness of the present method.

2 Method of Weightlifting Motion Detection

The proposed method consists primarily of three steps; preprocessing, motion feature extraction, and linear regression. In this section, we begin by making brief explanation of the input images and the tasks to be addressed.

2.1 Input Image and Task

Weightlifting videos captured during competition and training usually contain both of transient barbell-lifting segments (“work”) in question and the others (“rest”) including setup of barbell weight. The “work” and “rest” segments are alternatively concatenated. In addition, the videos are often captured from different viewpoints in practical environments. Note that they also contain background noise derived from other moving persons. Fig. 1 shows examples of actual still images in weightlifting videos captured from different viewpoints during training.

In order to clarify the task addressed in this paper, examples of multi-viewed image sequences around the start and the end motion in weightlifting are illustrated in Fig.1. These motion segments are defined in this study as follows; the start motion is from the time when the barbell plate lifts off the floor until the bar reaches maximum
height above the platform surface while the lifter is moving into squat position to catch it, on the other hand, the end motion is from the time when the barbell starts to descend until it once reaches the floor. In practical applications for in-depth motion analysis, detecting and indexing the time just when the barbell plates are lifted off is the primary procedure. The rest of this section describes the proposed approach to automatic detection of these start and end motions in weightlifting.

Fig. 1. Examples of video captured during weightlifting training: eight angle-views (upper), image sequences around the start (middle) and the end (bottom) of lifting motions.
2.2 Preprocessing

In the preprocessing, we apply frame differencing and then automatic thresholding [13] in order to detect and binarize motion pixels as in [7]. These processes filter out both inherent noise and brightness information, which are irrelevant to the motion itself. Consequently, pixel values in each frame become either 1 (moved) or 0 (static). The examples of the binary images are shown in Fig.2 at the same scenes as the middle and bottom of Fig.1.

![Fig. 2. Examples of the preprocessed image sequences around the start and the end of lifting motions.](image)

2.3 Motion Feature Extraction

In the stage of feature extraction, we employ Cubic Higher-order Local Auto-Correlation (CHLAC) [7]. CHLAC enables simultaneous extraction of spatio-temporal features from the motion image. Let \( f(r) \) be three way data defined on the region (cubic data) \( D : X \times Y \times 3 \) with \( r = (x, y, t)^T \), where \( X \) and \( Y \) are the width and height of image frame and \( T \) is the length of a time-window. Then, the \( N \)-th order auto-correlation can be defined as,

\[
\int_{-\Delta r}^{\Delta r} \int_{-\Delta r}^{\Delta r} \int_{-\Delta r}^{\Delta r} f(r)f(r+a_1)f(r+a_2) \cdots f(r+a_N) dr
\]

where the \( a_i (i = 1, \cdots, N) \) are displacement vectors from a reference point \( r \). Since Eq. (1) can take many different forms by varying \( N \) and \( a_i \), we limit \( N \leq 2 \) and \( a_i \) to a local region: the configurations of \( r \) and \( a_i \) are represented as mask patterns shown in Fig.3. The motion features are extracted by scanning the entire data set \( D \) with local cubic mask patterns. Thus, CHLAC feature corresponds to a histogram of local configuration patterns (auto-correlation) of moving points (pixels) found by frame difference. The dimension of CHLAC of up to the second order within the local 3x3x3 region is 251 for the binary data. CHLAC has a parameter denoted by \( \Delta r \) which is the spatial interval of the mask patterns along the \( x \)- and \( y \)- axes in the image frame.

CHLAC features possess important properties of shift invariance (rendering the method segmentation-free) and robustness to noise in data. Moreover, this method requires no prior knowledge or heuristics about objects. These favourable properties can benefit all aspects of approach to adaptively detect weightlifting motions.
including possible variability in terms of their appearances due to difference in lifter’s physical attribute, kinematic profile and camera angle.

Fig. 3. Examples of mask patterns: (left) $N=0$; (middle), $N=1$, $a_1 = (Δr, Δr, 1)^T$; (right), $N=2$, $a_1 = (-Δr, -Δr, -1)^T$, $a_2 = (Δr, Δr, 1)^T$.

2.4 Linear Regression

In the training phase, effective features for the start and end motion detection are extracted from the given training example. The pairs of the motion feature vector $x_i$ and the teacher signal $c_i$ at time $i$ are given. We apply multiple regression analysis (MRA), which determines the optimal linear coefficients $a$, to estimate $c$ from $x$:

$$\hat{c} = a' x + b$$

where $b$ is constant, $a = [a', b']$, $x = [x', 1]$. In this study, the teacher signals are binary, assigning 1 at times during the start and the end motions, and otherwise assigning 0.

Given the motion feature $x$, the existence of the target motion segment can be estimated by $c = a'x + b$. In the method, the target motions are finally identified by detecting the local peak along the time axis and thresholding it after applying moving average to the estimated values over a time-window $T$.

3 Experiments

The proposed method was applied to automatic detection of weightlifting motions from the image sequences. The dataset utilized in this experiment comprises 140 video sequences of the successful snatch and clean and jerk performed by national top-level athletes in different categories according to their bodyweight. The dataset had been acquired by filming lifts from eight different viewpoints in practical situations, as shown in Fig. 1. These data were captured at 30 frames per second (fps) and 320 x 240 pixels (QVGA).

For evaluating the performance of the proposed method, a leave-one-out scheme was applied to video sequences captured from the same camera angle, respectively, and then precision rates were calculated for both each target motion, i.e. start and end motion and each camera angle. In this evaluation, the detected point was regarded as correct if it was within each time duration of the target motion which was strictly determined by hand as ground truth.

CHLAC features are obtained by using all mask patterns of $Δr = 1, 3, 5, 7, 9$. The time-window $T$ for smoothing the estimation results is 27 and 16 for the start and end detection, respectively, based on the averaged time interval of the target motions in
the dataset. The results are shown in Table 1. The proposed method produces favorable results on every angle views and different kinds of weightlifting, the snatch and the clean and jerk. These results can indicate that our method can address the corresponding needs in the practical situations of weightlifting. The degradation of the precision rate for the end motion detection in V1 was largely due to that the vertically higher-positioned barbell was out of frame-view in several input images. In addition, some sample in the dataset includes an incomplete single lifting movement because the start or end of a single weightlifting motion is just near the corresponding start or end of a clipped video sequence.

Table 1. Detection rates (%) over different angle-views for each start and end motion in weightlifting. Angle views in this experiment nearly correspond to those illustrated in Fig. 1.

<table>
<thead>
<tr>
<th>Angle-View</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>100</td>
<td>100</td>
<td>94.7</td>
<td>94.1</td>
<td>100</td>
<td>94.4</td>
<td>100</td>
<td>94.7</td>
</tr>
<tr>
<td>End</td>
<td>71.4</td>
<td>100</td>
<td>89.5</td>
<td>100</td>
<td>88.9</td>
<td>90.0</td>
<td>93.8</td>
<td></td>
</tr>
</tbody>
</table>

From the viewpoint of kinematics, the entire single lift can be subdivided into several phases and the profiles in each phase are different among athletes, as indicated in [1]. In order to cope with the diversity in the kinematic profiles, our method employs various sizes of mask patterns for motion feature extraction, which can contribute to the performance. On the other hand, the motion orientation in the pulling phase after the start of each lift is similar to that after squat position to catch the barbell and that during jerk thrust, and consequently the spatio-temporal features of these movements extracted locally along the time axis can be not largely varied in some cases.

We applied this method to other sport motion, such as service detection of badminton, and obtained the similar results, which shows the validity and generality of our method [14].

4 Concluding Remarks

We have presented CHLAC approach to automatic detection of weightlifting motions, which can be mentioned as typical examples of predefined transient motions. The present method consists of motion feature extraction by CHLAC and prediction by MRA, and yields favorable detection performances for the start and end motions in weightlifting. By detecting these two motions, the whole weightlifting motion can be roughly segmented. Then, the weightlifting motion can be finely segmented, such as by using the methods proposed in [12]. It is expected that the integration between these approaches can contribute to more precise analysis of single transient motions of interest which are not limited to weightlifting motions.

Acknowledgements. We thank Mr. Miyoji Kikuta and Mrs. Kumiko Haseba at Japan Weightlifting Association for his cooperation on data gathering.
References

Prerequisites for Affective Signal Processing (ASP) –
Part IV

Egon L. van den Broek¹, Marjolein D. van der Zwaag², Jennifer A. Healey³
Joris H. Janssen²,⁴ and Joyce H. D. M. Westerink²

¹ Human-Centered Computing Consultancy, Austria
http://www.human-centeredcomputing.com
vandenbroek@acm.org

² User Experience Group, Philips Research Europe
High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
{marjolein.van.der.zwaag,joris.h.janssen,joyce.westerink}@philips.com

³ Future Technology Research, Intel Labs Santa Clara
Juliette Lane SC12-319, Santa Clara CA 95054, U.S.A.
jennifer.healey@intel.com

⁴ Department of Human Technology Interaction, Eindhoven University of Technology
P.O. Box 513, 6500 MB Eindhoven, The Netherlands
j.h.janssen@tue.nl

Abstract. In [1–3], a series of prerequisites for affective signal processing (ASP) was defined: validation (e.g., mapping of constructs on signals), triangulation, a physiology-driven approach, contributions of the signal processing community, identification of users, theoretical specification, integration of biosignals, and physical characteristics. This paper defines three additional prerequisites: historical perspective, temporal construction, and real-world baselines.

1 Introduction

In his book *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*, [4] stated: “...emotion is one of those suitcase-like words that we use to conceal the complexity of very large ranges of different things whose relations we don’t yet comprehend. Five pages later, he suggests to replace “old questions like, “What sorts of things are emotions and thoughts?” by more constructive ones like, “What processes does each emotion involve?” and “How could machines perform such processes?” Affective computing (AC) aims to answer these questions through processing signals that correlate with emotions: affective signals processing (ASP).

ASP can be employed from (a combination of) biosignals, movement analysis, computer vision, and speech processing. However, the techniques other than biosignals have major disadvantages [1–3]. In contrast, such issues have been resolved for biosignals in recent years: currently, it is easy to obtain, high fidelity, cheap, and unobtrusive biosignal recordings; e.g., see [5]. Moreover, the recording devices can be easily integrated in various products [6]. Therefore, this paper focuses on biosignals. For an overview of the most commonly used biosignals and their features, we refer to [1].
This prerequisites paper is designed to discuss unsolved issues related to ASP and to introduce a framework for future research. It is not designed as a paper on novel methods in signal processing, but rather on the specific issues on applying those methods to the problem of ASP. A particular focus is on the problem of ASP in the real world, with long latency signals (e.g., electrodermal activity; EDA), and affective responses that are ambiguously defined in time and that often depend on previous events and are, therefore, neither linear nor time invariant in their responses. Much of (traditional) signal processing relies on the linear time invariant assumption. Real affective responses do not fit this description. Consequently, ASP requires its own set of prerequisites as they are denoted in this paper and the other prerequisites papers of [1–3].

For AC, a broad plethora of classifiers is used as part of the ASP. The classification performances are hard to compare since the emotion classes used are typically defined in different ways. Additionally, the number of emotion classes to be discriminated is small, it ranges from 2 to 6. Nevertheless, the results are behind that of other classification problems. With AC recognition rates $< 90\%$ are common, where in most other pattern recognition problems, recognition rates of $> 90\%$ and often $> 95\%$ are often reported. This illustrates AC’s complex nature and the need for a comprehensive review of the prerequisites involved.

To force a breakthrough in results on AC we propose a set of prerequisites for ASP, before starting with AC in practice. The first three parts of these prerequisites were introduced in [1–3]. In the next section, the fourth part is introduced. Together, these prerequisites should form the foundation for more successful ASP and AC. We end this paper with a brief conclusion.

2 Prerequisites – Part IV

In [1–3], the following prerequisites for ASP were introduced: validity, triangulation, a physiology-driven approach, contributions from signal processing, user identification, theoretical specification, integration of biosignals, and physical characteristics. While each of these is still of the utmost importance for ASP, we will now denote three additional ones: historical perspective, temporal construction, and real-world baselines.

2.1 History: Lessons to be Learned and Experiences to Remember

Centuries ago, the relation between physiological reactions, as expressed through biosignals, and emotions was already mentioned by poets and ancient philosophers. This resulted in a plethora of definitions, almost impossible to list and illustrates the complexity of the concept emotion; cf. [4].

Although much knowledge on emotions is gained over the last centuries, researchers tend to ignore this up to a high extent and stick to some relatively recent theories; e.g., the valence and arousal model or the approach avoidance model. This holds in particular for ASP and AC, where an engineering approach is dominant and a theoretical framework is considered of lesser importance [2]. Consequently, for most engineering approaches, the valence-arousal model is applied as a default option, without considering other possibilities.
It is far beyond the scope of this paper to provide a complete overview of all literature relevant for ASP and AC. For such an overview, we refer to the various handbooks and review papers on emotions, affective sciences, and affective neuroscience; e.g., [7–9]. In this section, we will touch some of the major works on emotion research, which origin from medicine, biology, physiology, and psychology.

Let us start with one of the earliest works on biosignals: De l’Électricité du corps humain of M. l’Abbé Bertholon (1780), who already described human biosignals. One century later Darwin (1872) published his book The Expression of Emotions in Man and Animals [8]. Subsequently, independently of each other, William James and C. G. Lange revealed their theories on emotions, which were remarkably similar [8]. Consequently, their theories has been merged and were baptized the James-Lange theory.

In a nutshell, the James-Lange theory argues that the perception of our own biosignals is the emotion. Consequently, no emotions can be experienced without these biosignals. Two decades after the publication of James’ theory, it was already seriously challenged by [11, 12] and [13, 14]. They emphasized the role of subcortical structures (e.g., the thalamus, the hypothalamus, and the amygdala) in experiencing emotions. Their rebuttal on the James-Lange theory was founded on five notions:

1. Compared to a normal situation, experienced emotions are similar when biosignals are omitted; e.g., as with the transection of the spinal cord and vagus nerve.
2. Similar biosignals emerge with all emotions. So, these signals cannot cause distinct emotions.
3. The bodies internal organs have fewer sensory nerves than other structures. Hence, people are unaware of their possible biosignals.
4. Generally, biosignals have a long latency period, compared to the time emotional responses are expressed.
5. Drugs that trigger biosignals to emerge do not necessarily trigger emotions in parallel.

We will now address each of Cannon’s notions from the perspective of ASP. As will become apparent, considering these notions with current ASP is of importance. To the authors knowledge, the first case that illustrated both theories weaknesses was that of a patient with a lesion, as denoted in Cannon’s first notion. This patient reported: Sometimes I act angry when I see some injustice. I yell and cuss and raise heel, because if you don’t do it sometimes, I learned people will take advantage of you, but it just doesn’t have the heat to it that it used to. It’s a mental kind of anger (p. 151) [15]. Moreover, this case clearly illustrated the use of such special cases, as is denoted in [2].

The second notion of the Cannon-Bard theory strikes the essence of ASP. It would imply that the quest of affective computing is deemed to fail. According to Cannon-Bard, ASP is of no use since no unique sets of biosignals exist that map to distinct emotions. Luckily, nowadays, this statement is judged as coarse [8]. However, it is generally acknowledged that it is very hard to apply ASP successfully [7]. So, (at least) to a large extent Cannon was right.

It was confirmed that the number of sensory nerves differs in distinct structures in human bodies (Cannon’s notion 3). So, indeed people’s physiological structures determine their internal variations to the emotional sensitivity. To make ASP even more
challenging, cross-cultural and ethnic differences exist in people’s patterns of biosignals, as was already shown by [16].

The fourth notion concerns the latency period of biosignals, which Cannon denoted as being ‘long’. In the next section we address this problem.

The fifth and last notion of Cannon is one that is not addressed so far. It goes beyond biosignals since it concerns the neurochemical aspects of emotions. Although this component of human physiology can indeed have a significant influence on experienced emotions, this falls far beyond the scope of this paper.

It should be noted that the current general opinion among neuroscientists is that the truth lies somewhere between the theories of James-Lange and Cannon-Bard [8]. However, the various relations between the latter notions and the set of prerequisites, illustrates that these notions, although a century old, are still of interest for current AC and ASP.

2.2 Temporal Construction

There are many temporal aspects in biosignals that should be taken into account in ASP. These aspects can be categorized in three classes: psychological, physiological, and signal processing aspects.

The psychological aspect has to do with habituation; in general, every time a stimulus is perceived one’s reaction to it will get smaller. With large delays between the stimuli, one recovers from the habituation effect. There are several ways of dealing with this in ASP. One way is to keep track of the moments in which stimuli were present. This information can then be used to predict how strong the effect of a similar stimulus will be. Alternatively, in applications where stimuli presentation can be controlled, the variety of the stimuli can be directed such that habituation effects are canceled.

The first physiological aspect deals with the fact that the affective signals can be processed in different time windows. For instance, we can look at parts of 30 minutes but also at 10, 30, or 60 seconds. There are many challenges in modeling the temporal aspects of emotion. One is the annotation challenge of determining when the emotion begins and when it ends. Another is the sensor fusion problem of determining how to window individual signals within the emotional event since different signals have different latencies. In response to a high arousal event, an instant gasp may occur in respiration and a tensing of muscles, heart rate will then increase in the next few seconds and EDA should start to rise and may continue to rise for several minutes. Using the same window and offset for all signals would not capture the most salient discriminating features of the experience. So, in general, biosignal features calculated over time windows with different length cannot be compared with each other.

That being said, time window selection is often done empirically; i.e., many different time windows are tried and those leading to the best results are used in the final models [17, 6]. Other automatic options include finding the nearest significant local minima or making assumptions about the start time and extend of the emotion; e.g., an average over previous emotions. In addition, another empirical solution is to ask the

Author’s note. Nowadays, this paper would run up to resistance, as it denotes both ethnical issues and as its subjects. Perhaps that is why so little work is done on this topic.
user to define the window of interest. This can be done through sliders, as for real world research can be presented on a PDA. However, also this approach has its downside: in general, people’s introspection is not good and you do not want to bother users with these tasks. The physiological response to an emotion may have started well before the person realized that they were in this state, so if a single annotation is used, it will definitely come after the start of the experience. Moreover, in the real world the temporal nature of the reaction to the stimulus is undetermined. A uniform window may not be appropriate.

There are also valuable theoretical considerations. Different psychological processes develop over different time scales. On the one hand, emotions lead to very short and fast phasic changes and, thus, require short time windows. On the other hand, changes in mood are more gradual tonic and, so, require broader time windows. In general, the time window used should depend on the psychological construct studied. Furthermore, there is always a lag between the psychological change and the physiological change. These lags differ per signal: heart rate changes almost immediately while skin temperature can take more than a minute to change. Skin conductance is somewhere in between. This shows the need for different time windows for different signals.

A second physiological aspect stems from the idea that physiological activity tends to move to a stable neutral state; i.e., when the physiological level is high, it tends to decrease; whereas, when the physiological level is low, it tends to increase. Hence, the effect of a stimulus on physiology depends on the physiological level before stimulus onset; i.e., the principle of initial values [9]. When you perceive a scary stimulus and your heart rate is at 80 it might increase by 15 beats, however, when you heart rate is at 160 it is unlikely to increase at all. As this is found to be a linear relationship, it can be modeled by linear regression. The first step is to assess the regression line, which is different per feature and person. Next, this regression line can be used to correct each feature by computing its residualized value.

A consideration specific to ASP is that emotional responses are likely comprise a layered response involving components that have different time periods including: disposition (long term - years), circumstance (days), mood (hours) and emotion (seconds). An accurate model of an individuals affective response to these varying time influences is difficult to determine, even the totality of influences are difficult to catalog in the real world. Therefore, also for this reason, a major consideration in ASP is choosing a window length appropriate to the type of affective response you are considering.

### 2.3 Real-world Baselines

Baselining is the process of correcting the biosignal to a standard level that is comparable over users and/or sessions (also called normalization/standardization). Finding an appropriate baseline is both important and difficult for sensors whose readings depend on factors that can easily change on a daily basis; e.g., sensor placement, humidity, temperature, and the use of contact gel [3]. Still, baselining over multiple people or multiple days is required for ASP in order to compare and combine data from these different sources in a meaningful way. Many different approaches to baselining are known in the literature. However, as will be shown, we require affect specific approaches.
Table 1. Seven methods of using baseline information to normalize the signal. \( x \) denotes the original signal and \( \tilde{x} \) the corrected signal. \( \mu_B, \min_B, \max_B, \) and \( \sigma_B \) are respectively the mean, minimum, maximum, and standard deviation of the baseline. Sources of information: 1:[18], 3,5,6:[19] and 7:[20].

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \tilde{x}_i = x_i - \mu_B )</td>
<td>Standard correction, often used in psychological experiments.</td>
</tr>
<tr>
<td>2</td>
<td>( \tilde{x}_i = x_i - \min_B )</td>
<td>Useful alternative to the first method when there is no relaxation period and a lot of variance in the signal.</td>
</tr>
<tr>
<td>3</td>
<td>( \tilde{x}_i = (x_i - \mu_B) / \sigma_B )</td>
<td>Strong baselining method; works best for continuous signals.</td>
</tr>
<tr>
<td>4</td>
<td>( \tilde{x}_i = (x_i - \mu_B) / \mu_B )</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>( \tilde{x}_i = (x_i - \min_B) / (\max_B - \min_B) )</td>
<td>Sensitive to outliers.</td>
</tr>
<tr>
<td>6</td>
<td>( \tilde{x}_i = x_i / \max_B )</td>
<td>Used for Skin conductance responses features.</td>
</tr>
<tr>
<td>7</td>
<td>( \tilde{x}_i = (x_i \times 100) / \mu_B - 100 )</td>
<td>Used for facial EMG measurements.</td>
</tr>
</tbody>
</table>

With ASP we have to handle long term continuous (re-)baselining of biosignals. There exists no guideline on how to apply these methods to continuous physiological data in the real world. An exception to this is [21], which discusses ECG recording in ambulatory settings; however, it does not focus on ASP. In this section, we discuss how to apply known methods from the laboratory to a new situation, continuous ambulatory monitoring in the real world. We also discuss how to apply these methods to affective reactions of varying length and intensity in the presence of noise. Some of the methods commonly used in other types of signal processing, such as “zeroing the mean” and “dividing by the variance” do not work for long term physiological records, which is the problem we are trying to bring to light. In the following paragraphs, we will try to give an overview of the different baselining approaches and explain when they are appropriate. We also call for empirical comparisons of different baseline methodologies specific to ASP in the real world, as this is still lacking.

The two main issues with baselining are (1) the selection of a suitable correction method and (2) the selection of a period over which to calculate the parameters of the correction method (the baseline period). The correction methods are summarized in Table 1. Once the baseline is removed, it becomes the new base (or zero) and the original value is lost. Each baseline has different merits. Taking the minimum baseline is more equivalent to taking the resting EDA that would normally be used in a laboratory experiment. This is the best method if a consistent minimum seems apparent in all data being combined. The problem is that for each data segment, a minimum must be apparent. It is straightforward to eliminate point outliers such as those at 3.7 hours and 3.9 hours and find a more robust minimum for the baseline. An other often used method for continuous signals like EDA and skin temperature is called standardization (method 3 in Table 1) [19]. This is probably the most powerful correction method and is applied very often. It corrects not only for the baseline level but also for the variation in the signal, making it more robust. Other correction methods are tailored to specific features; e.g., the amplitude of skin conductance responses is often corrected by dividing by the maximum amplitude. Taken together, different signals and situations require different correction methods, which should be chosen carefully.

The second issue in baselining is the selection of an appropriate time-window over which the parameters for the correction are calculated. For short term experiments, a single baseline period is usually sufficient. However, when monitoring continuously,
the baseline may have to be re-evaluated with greater frequency. The challenge here is to find a good strategy for dividing the signal into segments over which the baseline should be re-calculated. A simple solution is to use a sliding window; e.g., where the last 30 minutes are taken into account. In this case, it seems obvious that the segments should be considered independently, since the time period between the two is long and it may be that the electrodes fell off and may have been re-applied. To be able to conduct proper normalization when signal loss is shorter than the baselining window, the period right before the signal loss can for example be used to complement the baseline window of the current signal. However, in general, data segmentation has not had much attention in ASP and long term continuous (re-)baselining of biosignals is still an open problem.

As with all pattern recognition pipelines, baselining (or normalization) is of utmost importance. Most efforts towards AC and ASP have not paid much attention to this. We hope that this prerequisite is a start for the development of more sophisticated algorithms that can deal with the difficult problems of data segmentation, as they have been dealt with in other research fields like computer vision.

3 Conclusions

This paper explains the importance of prerequisites specifically for ASP and introduces the fourth part of a series of such prerequisites for ASP. The prerequisites foundation in historical perspective, adequate temporal construction, and well-chosen real-world baselines are introduced. These prerequisites are complementary to those presented in [1–3]: validity, triangulation, the physiology-driven approach, and contributions of signal processing, identification of users and theoretical specification, and physical characteristics and integration of biosignals.

The review and the prerequisites, both illustrate and explain the complexity of ASP and its limited progress. Therefore, we advise to incorporate these prerequisites for successful ASP, instead of running forward and ignoring the problems encountered in previous studies. We hope that the prerequisites can contribute to or even guide future research in ASP.

References


Biometrics for Emotion Detection (BED\textsuperscript{1}): Exploring the combination of Speech and ECG

Marleen H. Schut\textsuperscript{1}, Kees Tuinenbreijer\textsuperscript{1}, Egon L. van den Broek\textsuperscript{2} and Joyce H. D. M. Westerink\textsuperscript{3}

\textsuperscript{1} Philips Consumer Lifestyle Advanced Technology
High Tech Campus 37, 5656 AE Eindhoven, The Netherlands
{marleen.schut,kees.tuinenbreijer}@philips.com
\textsuperscript{2} Human-Centered Computing Consultancy, Austria
http://www.human-centeredcomputing.com
vandenbroek@acm.org
\textsuperscript{3} User Experience Group, Philips Research Europe
High Tech Campus 34, 5656 AE Eindhoven, The Netherlands
joyce.westerink@philips.com

Abstract. The paradigm Biometrics for Emotion Detection (BED) is introduced, which enables unobtrusive emotion recognition, taking into account varying environments. It uses the electrocardiogram (ECG) and speech, as a powerful but rarely used combination to unravel people’s emotions. BED was applied in two environments (i.e., office and home-like) in which 40 people watched 6 film scenes. It is shown that both heart rate variability (derived from the ECG) and, when people’s gender is taken into account, the standard deviation of the fundamental frequency of speech indicate people’s experienced emotions. As such, these measures validate each other. Moreover, it is found that people’s environment can indeed influence experienced emotions. These results indicate that BED might become an important paradigm for unobtrusive emotion detection.

1 Introduction

A goal of human-centered computing is computer systems that can unobtrusively perceive and understand human behavior in unstructured environments and respond appropriately. \cite{1}

Three of the issues raised in this quote formed the starting point of the research described in this paper: 1) unobtrusively perceive, 2) understand, and 3) unstructured environments. Regrettably, in practice, the combination of these issues is rarely taken into account in research towards human-centered computing. This paper describes research that took into account all three aspects that were just denoted. To enable unobtrusive recordings of humans, we exploit the combination of the electrocardiogram (ECG) and

\textsuperscript{1} In parallel with the abbreviation for Biometrics for Emotion Detection, BED denotes a second aspect of the paradigm: a supporting surface or structure or a foundation (source: http://www.merriam-webster.com) for unobtrusive emotion detection.
speech. The combination of speech and biosignals (in general) to determine emotions is rare. However, some research has been conducted; e.g., [2, 3]. This research concluded that they . . . did not achieve the same high gains that were achieved for audio-visual data which seems to indicate that speech and physiological data contain less complementary information. [2] (p. 63). This paper present a study that challenges this conclusion. In combination with people’s subjective experiences, these two signals should be able to unravel generic principles underlying humans’ emotions, at least partly.

It has been widely acknowledged that emotion elicitation is as crucial as is it complex for affective computing purposes. Moreover, context has been frequently mentioned as a factor of influence. This study explores whether or not context has an influence on emotion elicitation, as it comprises two identical studies in different settings. This is expected to provide a little grip on the influence of environmental factors / context. Since we want to explore generic principles behind human emotions, we ignore inter-personal differences. Please note that with this position, we do not challenge the notion that significant differences among humans exist; cf. [4]. Ultimately, a personalized approach will need to be adopted; however, before that, generic principles have to be identified to enable more efficient processing of cues.

First, we will briefly introduce the construct emotion (Section 2) and the model of emotion used. Section 4 describes the two studies conducted, with which we explore the use of the paradigm BED. This includes a description of the speech and ECG signals. In Section 5, the analysis of the two studies is presented. These comprise only a limited set of features, as an exhaustive search in feature space was not the aim of this study. Last, in Section 6, the implications of this research will be discussed and future directives will be provided.

2 Emotions

As John F. Cohn states [1]: Efforts at emotion recognition [...] are inherently flawed unless one recognizes that emotion - intentions, action tendencies, appraisals and other cognitions, physiological and neuromuscular changes, and feelings is not readily observable. In other words, emotions are a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems. Consequently, emotions can [5]:

1. cause affective experiences such as feelings of arousal and (dis)pleasure;
2. generate cognitive processes; e.g., emotionally relevant perceptual effects, appraisals, labeling processes;
3. activate widespread physiological adjustments to arousing conditions; and
4. lead to behavior that is often expressive, goal directed, and adaptive.

Much can be said both in favor and against this definition. However, with this work we do not aim to provide an exhaustive elaboration on the definition of emotions. Therefore, we adopt the previous definition of emotions as working definition.

In his seminal study, Russell (1980) introduced the circumplex model of emotion, which claims that all emotions can be characterized by two independent bipolar dimensions: judged valence (i.e., pleasure/positive or displeasure/negative) and arousal [6].
The circumplex model allows a representation of emotions in a 2D space, which provides a visualization of emotions and their relation in a comprehensible way. Parallel to Russell’s model various other models of emotion have been introduced, using 1) categories, 2) unipolar or bipolar dimensions, 3) some other simple structure, and 4) a hierarchy. For an overview on models of emotion, we refer to [6].

For the current research, we adapted Russell’s circumplex model [6]. The bipolar valence dimension was replaced by two unipolar valence dimensions, untying positive and negative valence. The arousal dimension was not altered. A similar approach have been used earlier; e.g., see [7, 8]. The 3D model we propose tackles a major disadvantage of the circumplex model: the disability to handle mixed emotions; i.e., parallel experience of positive and negative valence.

3 Signals of Emotion

People’s emotional state can be accessed through processing various of their signals. When reviewing literature, it becomes apparent that these signals can be assigned to two groups: 1) A broad range of physiological measures signals [9] and 2) Specialized areas of signal processing: speech processing, movement analysis, and computer vision techniques [10–12]. These distinct measurement methods are seldom combined; where, on the one hand, several physiological measures are frequently combined (e.g., [4, 8]) and, on the other hand, speech processing, movement analysis, and computer vision are frequently combined (e.g., [11]).

Physiological measures are often obtrusive and, hence, disregarded for user-centered applications, as AmI is. However, wearable computing and wireless sensing technologies relief this problem [13]. In contrast, speech and computer vision are unobtrusive but very noise sensitive. The audio recordings used for speech processing suffer various types of noise. However, with no need for speech recognition, the remaining problem is binary: a speech signal or no speech signal, which makes it feasible. Computer vision techniques, although appealing, are only usable for emotion recognition in very stable environments; e.g., without occlusion and with fixed light sources.

Speech and physiological measures, in particular the ECG, are rarely combined to access the emotional state of users, although especially their combination is promising. A possible explanation is the lack of knowledge that exists on the application of this combination of measures for emotion measurement; cf. [10, 11] and [8, 9].

From features of both the speech and the ECG signal, we expect to extract cues on people’s experienced valence and arousal. Since this study is (one of) the first to employ the combination of speech and ECG, we chose for a controlled study to assess their feasibility for human-centered computing purposes. However, before the study is described, each of the signals used are introduced.

4 Principles and Implementation

Let us quote John F. Cohn [1], once more: Emotion can only be inferred from context, self-report, physiological indicators, and expressive behavior. The four factors ad-
dressed in this quote are incorporated in this research on BED. For each of them will be explained how they are embedded in BED and are implemented in the research.

4.1 Context

As was stated in the introduction, context is of the utmost importance. However, field research introduces a broad range of factors that cannot be controlled. These factors can be considered as severe forms of noise, when analyzing the data obtained. Most often, this results in qualitative data gathering and its analysis. In contrast, with the signals ECG and speech, we aim to gather quantitative data. Therefore, we chose to conduct a semi-controlled study.

40 people (average age: 27) voluntarily participated in this research. Half of them was assigned to a living room environment and half of them was assigned to an office environment. Except for this difference, both groups participated in the same study, under the same conditions. All participants had (corrected to) normal vision and hearing and none of them had cardiovascular problems.

To trigger emotions, the participants watched six scenes (length: 3.18 min), adopted from [7, 8]. For a more exhaustive description of the film scenes, we refer to [7]. The length of the film scenes enables a reliable extraction of HR variability from the ECG [14]. The film scenes were presented in a random order. While watching the film scenes, their ECG signal was recorded using a modified Polar ECG measurement belt that was connected to a NI USB-6008 data acquisition device. Its output was recorded in a LabVIEW program (sample rate: 200 Hz). After the film scene was watched, the people were asked to talk about it. For this a standard microphone was used (sample rate: 44.100 Hz; sample size: 16 bit).

After instructions and a control of the equipment, the people read aloud a non emotional story. In this way it was checked if the people had understood the instructions, whether or not the equipment worked, and people’s personal baseline for both the speech and the ECG signal was determined.

4.2 Self-report

After all film scenes were shown, the people rated the films, using 11 point Likert scales. Hence, their subjective attitudes were determined. Separate Likert scales were presented for positive and negative affect and for arousal; see also Section 2.

4.3 Physiological Indicator: Electrocardiogram (ECG)

The electrocardiogram (ECG) is an autonomic signal that cannot be controlled easily, just like the electrodermal activity. ECG can be measured directly from the chest. Alternatively, the periodic component of the blood flow in the finger or in an ear which can be translated into the ECG. From the ECG, the heart rate (HR) can be easily obtained; e.g., [9, 13]. Research identified features of HR as indicators for both experienced valence and arousal [15].

In addition to the HR, a range of other features can be derived from the ECG. The most frequently used one is HR variability (HRV). HRV decreases with an increase in
mental effort, stress, and frustration [9]. Moreover, some indications have been found that HRV is also influenced by the valence of an event, object, or action [15]. Taken together, HRV is one of the most powerful features to discriminate among emotions.

4.4 Expressive Behavior: The Speech Signal

Speech processing, speech dialog, and speech synthesis can exhibit some form of intelligent, user-perceived behavior and, hence, are useful in designing human-centered computing environments [1]. However, speech comprises another feature: emotion elicitation [10, 11].

The human speech signal can be characterized through various features and their accompanying parameters. However, no consensus exists on the features and parameters of speech that reflect the emotional state of the speaker. Most evidence is present for the variability (e.g., standard deviation; SD) of the fundamental frequency (F0), energy of speech, and intensity of air pressure [10, 11]. Therefore, this feature is derived from speech as its emotion indicator.

5 Analysis

5.1 Data Reduction

Before the actual data reduction could take place, noise reduction was applied for both the ECG and the speech signal. First, recording errors were removed. Of 11 participants either the ECG signal, the speech signal, or both was distorted (e.g., the microphone was positioned too close to the mouth, coughing, yawning) or not recorded at all. The data of these participants was omitted from further analyses.

The measurement belt for the ECG signal appeared to be sensitive to movements of the participant. For this type of noise was controlled and, if present, the ECG signal was corrected for it. Using the corrected ECG signal, people’s HR was determined. Subsequently, HRV (i.e., SD of HR) was calculated.

The speech signal was segmented in such a way that for each film scene a separate speech signal was determined. Moreover, silences and utterances such as ‘euh’ were removed. This resulted in noise-free speech signals. In order to cope with interpersonal differences in speech, the signals were normalized by subtracting a baseline from the original signal. Subsequently, the speech processing environment Praat [16] was used to extract the SD F0 and the intensity of air pressure.

5.2 Subjective Measurements

The people rated the film scenes on their experienced positive and negative valence and on arousal. For each film scene, peoples’ average ratings on each of these scales were calculated. Together, these two categorized dimensions of the valence–arousal model depict six emotion classes.

Each of the six emotion classes is represented in this research through one film fragment. The emotion classes with the values on the three dimensions, their categorization in valence and arousal, and the accompanying film fragment can be found in [7].
5.3 Feature Extraction

From both the speech signal and the ECG signal a large number of features can be derived [14, 15]. This research did, however, not aim to provide an extensive comparison of speech and ECG features. Instead, the combination of these two signals was explored. From the ECG signal, the intervals between the R-waves (R-R intervals) as well as their mean were determined. Subsequently, the HRV was defined as the SD of the R-R intervals.

Although no general consensus exists concerning the parameters of speech to be used for emotion detection, much evidence is present for the SD F0 and the Intensity of air pressure. They are useful for measuring experienced emotions. For a domain \([0, T]\), the intensity of air pressure in the speech signal is defined as:

\[
10 \log_{10} \frac{1}{T P_0^2} \int_0^T x^2(t) \, dt,
\]

where \(x(t)\) is the amplitude or sound pressure of the signal in Pa (Pascal) and \(P_0 = 2 \cdot 10^{-5}\) Pa is the auditory threshold [16]. It is expressed in dB (decibels) relative to \(P_0\).

The F0 of pitch was extracted from the corrected speech signal through a fast Fourier transform over the signal. Subsequently, its SD is calculated. We refer to the documentation that accompanies [16], for a more detailed description of the extraction of F0 of pitch from the speech signal.

5.4 Results

All data was analyzed through a RM ANOVA, with three measures: HRV determined from the ECG signal and the SD F0 and intensity of the speech signal. Two between subject factors were included in the analyses: the environment (office/living room) and gender (male/female). Age was omitted from the analysis since a preliminary analysis revealed that age was of no influence on any of the measures. First, the multivariate test will be reported, including all four measures. Next, for each measure the univariate tests will be reported. With all analyses, the interaction effects will be reported.

The multivariate analyses showed a strong effect for the emotion classes on the set of physiological parameters/measures, \(F(15,11) = 29.688, p < .001\). In addition, in interaction with both gender \((F(15,11) = 7.266, p = .001)\) and environment \((F(15,11) = 17.235, p = .000)\), an effect of the emotion classes on the measures was found. In line with these interaction effects, a three-way interaction effect between the emotion classes, gender, and the environment was found on the measures, \(F(15,11) = 8.737, p < .001\).

A strong main effect was found for the emotion classes on HRV, \(F(5,125) = 38.677, p < .001\). An interaction effect of the emotion classes and both gender \((F(5,125) = 7.678, p < .001)\) and environment \((F(5,125) = 18.856, p < .001)\) on HRV was found. In line with the two-way interaction effects on HRV, a three-way interaction effect on HRV between the emotion classes, gender, and environment was found, \(F(5,125) = 10.967, p < .001\).

An interaction effect between the emotion classes and gender on SD F0 was found, \(F(5,125) = 2.553, p = .031\). In addition, a three-way interaction effect on the intensity
of speech between the emotion classes, gender, and environment was found, $F(5,125) = 3.052, p = .013$.

6 Discussion

The emotions elicited by the film fragments are clearly discriminated by HRV. In interaction with gender, both HRV and SD F0 show to be a good discriminator among the six emotion classes. When both gender and environment are taken into account, both HRV and the intensity of speech reflect the emotions experienced.

Of the F0 of speech, it is known that male and females have different characteristics. Hence, the influence of gender was expected and will always have been taken into account. Moreover, environmental factors have to be incorporated in the processing scheme. Note that the difference between the environments assessed in this research was limited; hence, in practice this effect can be far more substantial. Further, research on the intensity of speech should reveal how robust the current findings for this parameter are.

Both HRV and SD F0 of speech showed to be good generic unobtrusive discriminators between emotions. This makes them par excellence suitable as biometrics for unobtrusive emotion discrimination. This study is rare in that it reports the use of biosignals in combination with speech to unravel user’s emotional state. However, it should be noted that the variety among emotions is rich and only six are assessed in the current research. Moreover, it is unknown how sensitive both measures are for emotion discrimination. Hence, further research is needed on this.

The results of this research should be seen in the perspective of the vast amount of research already done on emotions. This research has revealed a range of issues that can, and probably should, also be taken into account. First, the notion of time should be taken into consideration that helps to distinguish between emotions, moods, and personality [6, 17]. Second, BED distinguishes four factors: context, self-report, physiological indicators, and expressive behavior. The last three of them can be influenced by personality characteristics. For example, an introvert person will express his emotions in another way than an extrovert person. Moreover, probably also the self-reports and physiological indicators or biosignals will be influenced by this personality trait. Moreover, whether or not a person is neurotic will influence his behavior, in particular in relation to his environment (or context); e.g., see also [17].

Other more practical considerations should also be noted. For example, the advances made in wearable computing and sensors facilitates the communication between humans and systems; cf. [13]. This enables the use of more recordings of biosignals in parallel to speech recordings and ECG. In this way, an even higher probability of correct interpretation can be achieved [8, 9].

It is surprising that the combination of speech and biosignals has not been used more often to unravel user’s emotions; cf [2, 3]. Par excellence, these signals could be exploited in parallel for generic human-centered computing purposes, as is illustrated through BED. Both speech and ECG parameters can unravel users’ emotion space. Moreover, various manners of implementation of the required sensors secure an unobtrusive recording of both signals. This having said, with this article we hope to motivate
a further exploration of the combination of speech and biosignals. Possibly, they enable the significant step forward in making reliable unobtrusive emotion detection a success.

References

Abstract. Musical and performance experiences are often described as evoking powerful emotions, both in the listener/observer and player/performer. There is a significant body of literature describing these experiences along with related work examining physiological changes in the body during music listening and the physiological correlates of emotional state. However there are still open questions as to how and why, emotional responses may be triggered by a performance, how audiences may be influenced by a performers mental or emotional state and what effect the presence of an audience has on performers. We present a pilot study and some initial findings of our investigations into these questions, utilising a custom software and hardware system we have developed. Although this research is still at a pilot stage, our initial experiments point towards significant correlation between the physiological states of performers and audiences and we here present the system, the experiments and our preliminary data.

1 Introduction

As computers and mobile devices become simultaneously smaller and more powerful they are also being incorporated into everyday objects and our day to day lives. This move towards ‘ubiquitous’ computing means that these embedded devices are used in very diverse situations and environments, that may require some degree of context awareness from the device. One branch of research into context aware interactions is what is known as ‘affective’ computing, using emotion or ‘affect’ as an information and interaction channel for an electronic device or system [1]. This may allow appropriate responses from a computer system based on factors such as happiness, sadness, frustration or stress. Machine recognition of emotional state is not a trivial matter and research continues in a number of directions, such as facial emotion recognition, vocal analysis and posture analysis [2]. Our research has chiefly focused on physiological indicators of emotion (biosignals) and changes in emotional state such as patterns in heart rate variability and galvanic skin response [3]. While it is difficult to attempt to assign a given emotion to a particular physiological state, as opposed to say facial indicators of emotion, biosignals do have the advantage of being largely outside conscious control and thus may be viewed as a more ‘direct’ connection with
the subject. It is also relatively easy to detect subtle continuous changes in physiological (and by extension emotional) state, allowing for more nuanced interactions.

2 Music and Emotion

Emotions are a powerful force in driving human decision making and action, frequently overpowering intellect or logical reasoning [4] and with the capability to affect our interpretation and perception of events or content [5]. Music has been shown to have the capability to induce emotions in the listener [6] with corresponding physical [7] and physiological effects [8].

While most previous research into the emotional power of music has focused on structural and cognitive aspects, the neuropsychological underpinnings are only now being properly explored, with alternative mechanisms, such as those posited by Juslin and Västfjäll in [9], such as brain stem reflexes, visual imagery, episodic memory and musical expectancy now thought to also play a role in evocation of emotion. Also among these alternatives is the possibility of emotional contagion, in which emotion is engendered in the listener corresponding to the perceived emotional content or intent of the music, such as a dissonant piece with harsh timbres and fast tempo suggesting anger. There is some evidence to support this induction of mood through perceived affect of musical stimuli [10] and listeners often report a sensation of ‘chills’ from particular pieces of music with a strong personal or emotional cachet [11] that may also be an indicator of emotional peak experiences.

However most of these experiments have taken place in laboratory settings using pre-recorded musical examples and on a one-to-one basis, with little consistency in subjects responses to given pieces of music. So far little work has been done (on a physiological level) in examining group experiences of musical performance in a concert setting and it is our belief that more data gathered simultaneously from multiple participants in this ecological setting may shed some light on what causes and modulates our emotional responses to music.

We also hypothesise that there is a degree of emotional contagion between the performer and the audience, with a performer/players affective state influencing the affective state of the audience. This may be observed through channels such as the previously mentioned facial or posture indicators, through affective modulation of performance style and technique or through as yet unknown channels of affective communication.

It is important to stress that using low level biosignals like GSR or HR we are unable to definitively infer a given affective state in the subject monitored, such as happiness or boredom. We are however able to detect gross changes in state and to suggest a probability of a given state (with accuracy dependent on variables such as number of signals monitored and context). There may also be variables external to the monitoring environment (for example events that occurred during the subjects day prior to monitoring or feelings of illness) that will affect the biosignal readings.
3 Methodology

We carried out two experiments in different live performance environments, with separate subject groups. 9 random audience members were selected in each session to participate in the experiments, who were invited to sit in chairs augmented with sensors to detect physiological signals [12]. The musical program included 3 contemporary pieces during which the performers were measured with bio-sensors: a piano improvisation (12min), an interactive electronic piece ‘Stem Cells’ (12min) and an electroacoustic piece diffused by the composer ‘Imago’ (25min). The three performers’ biosignals were recorded simultaneously with the audience.

During the experiments we monitored two physiological signals known to be correlates of emotional state: Galvanic Skin Response (GSR) and Heart Rate Variability (HRV). GSR, also known as electrodemal response, is a method for measuring the conductance of the skin using an ohmmeter. Electrodes are situated in the palms or fingertips of the hands, where the eccrine glands, regulated by the sympathetic nervous system (SNS), produce sweat that varies the conductivity measured by the ohmmeter. Although one of the main evolutionary functions of the SNS is to regulate body temperature, since early studies researchers have correlated changes in GSR to different stimuli associated with emotional responses, such as film [13] and music [7].

HRV is a feature extracted from an electrocardiogram (ECG) signal, which measures the electrical impulses produced by the heart during each beat. Heart Rate Variability refers specifically to the changes in the beat-to-beat interval of the heart. In other words, a heart rate of 70 beats per minute (bpm) is an average over time of fluctuations between successive heartbeats, and these may vary significantly from the 70 bpm. Several studies have observed patterns in HRV that are associated with emotional states, yet there is much controversy in the scientific publications regarding correlation with specific emotions. Nevertheless, there is agreement that HRV patterns change when compared to neutral state [10]. In a previous study we found interesting differences in HRV patterns between different emotional states for musicians.
performing the same musical piece [14]. BioControl\(^1\) signal acquisition devices were used to capture physiological signals, which were streamed wirelessly at 250 [Hz] to signal recording computers, via the Wi-microDig\(^2\) and Arduino\(^3\) microcontrollers. In order to assure synchronicity between physiological, visual and audio signals, every sample of data was time stamped with a time code index which operated independently as part of a recording system protocol developed by the authors. The recorded physiological signals were processed offline using MATLAB and the GSR signals were low pass filtered (29th order FIR filter with 3 [Hz] cut-off frequency) and HRV was extracted, using an algorithm created by the authors, that measured the RR interval between beats (from the QRS waveform). In order to carry out a real-time evaluation of the performance, signals were analyzed using a Max/MSP\(^4\) patch that allowed the visualization of all physiological signals, audio and video material simultaneously (see Fig. 2 and [14]).

![Fig. 2. Visualization tool created in Max/MSP to provide continuous analysis of the physiological and audiovisual data. The figure shows 10 channels of ECG and GSR data (9 audience members and 1 performer) plus the audiovisual recordings.](image)

Due to the inherent problems of recording data in a live performance, the amount of viable audience biosignals captured varied between performances (see Table 1). This is principally caused by loss of signals due to movement artefacts and the non homogeneity of subject’s physiology (in a controlled lab set-up, the equipment could be calibrated and adjusted to read biosignals of subjects with different ranges). For the

---

\(^1\) [www.biocontrol.com](http://www.biocontrol.com) [accessed 05 November, 2009]

\(^2\) [www.infusionsystems.com](http://www.infusionsystems.com) [accessed 05 November, 2009]

\(^3\) [www.arduino.cc](http://www.arduino.cc) [accessed 05 November, 2009]

\(^4\) [http://www.cycling74.com/products/max5](http://www.cycling74.com/products/max5) [accessed 05 November, 2009]
preliminary results presented in the next section, a selection of the most significant physiological reactions and correlation were chosen according to the observations done with the Max/MSP patch (see Fig. 2).

Table 1. Detail of viable biosignals recorded during the experiments.

<table>
<thead>
<tr>
<th>Performance</th>
<th>Venue</th>
<th>GSR</th>
<th>HRV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piano Improvisation</td>
<td>Sonic Lab, Belfast</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Stem Cells</td>
<td>Sonic Lab, Belfast</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Stem Cells</td>
<td>School of Music, Durham</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Imago</td>
<td>School of Music, Durham</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

4 Preliminary Results

We will present a preliminary qualitative analysis of the data recorded, which at this early stage cannot be considered as conclusive due to the limited sample size of the study.

Fig. 3. Relative GSR levels of two selected audience members (bottom) and the performer (top). The plot shows 5 minutes of Imago, with strong correlation between both audience members and similarities with the changes in the performer during specific sections.

The continuous analysis of the GSR signals indicates a strong correlation with the musical characteristics of the performance. In the piano improvisation piece, the performer made several gestures in anticipation of the note that was to be played, which triggered increases in the GSR level of the audience. During the electroacoustic piece, the composition was played with strong dynamic changes, with long crescendos and
sudden silences. This also resulted in significant changes at the GSR level (See Fig. 3). The most interesting results were observed when we overlapped the performer’s GSR and HRV signals individually with each audience member. During certain passages of the musical pieces, there is strong correlation between the physiological signals. Fig. 3 to Fig. 5 show examples where this phenomenon occurred.

**Fig. 4.** HRV for performer (top) and three selected audience members (bottom) during 2.5 minutes of *Stem Cells* in Durham. The plot shows a simultaneous increase and posterior decrease in HR for both performer and audience at 450 seconds approximately.

**Fig. 5.** GSR signals for performer and one selected audience member. The plot shows a strong and continuous correlation during 5 minutes of *Stem Cells* in Durham.
5 Discussion and Conclusions

We have presented a novel approach to the study of physiological correlates of emotion between performer and audience. Preliminary results indicate significant levels of correlation, both for GSR and ECG signals. Yet, further studies are needed in order to obtain conclusive results. The use of additional physiological features, such as respiration rate and depth, has given interesting results previous studies [8] and is suggested to be incorporated in future experiments.

The actual mechanisms by which emotional contagion occurs are still largely undefined (some indicators may be found in [15] and [16]) but a theory which is currently showing promise is that of ‘mirror’ neurons in the brain, which mimic externally perceived actions or conditions with a corresponding impulse in a related part of the observers brain e.g. seeing someone running causes neurons responsible for movement to fire in the brain of the observer [17].

Auditory or visual cues are also likely to have an effect on a participant’s affective state and there are indicators in our findings suggesting correlations between visually led anticipation and changes in GSR. We have also found links between sudden or extreme auditory events and physiological changes (some of which may be explained by the ‘startle response’ [18 page 647]). Analysis of video recordings in conjunction with the time-stamped biophysical data allows us to link specific auditory or visual events with corresponding physiological changes and isolate periods in which there are physiological changes in the absence of such cues.

One of the biggest problems in working in an ecological scenario such as a live concert is the constraints imposed by time and the nature of an invited audience, which reduces the option for calibration and changes of materials in case of any technical problems. Nevertheless, we believe that methodologies as the one presented in this study are an important step towards creating a more natural environment where questions addressing the complex relationship between music, emotion and physiology are not affected by a laboratory set-up.

References

Motion and Single-trial Biosignal Analysis Platform for Monitoring of Rehabilitation

Perttu Ranta-aho¹, Stefanos Georgiadis¹, Timo Bragge¹, Eini Niskanen¹,²
Mika P. Tarvainen¹, Ina M. Tarkka³ and Pasi A. Karjalainen¹

Department of Physics, University of Kuopio, Kuopio, Finland
Department of Clinical Neurophysiology, Kuopio University Hospital, Kuopio, Finland
Department of Health Sciences, University of Jyväskylä, Jyväskylä, Finland

Abstract. Three-dimensional motion analysis is a powerful tool for the assessment of human movements during different rehabilitation applications. An adaptive virtual reality rehabilitation environment which is based on modern motion and biosignal analysis techniques is described.

1 Introduction

Noninvasive brain computer interface (BCI) has in the recent years become a highly active research topic in neuroscience, engineering and signal processing. BCIs utilize neurophysiological signals to interact with external devices and computers. Despite diverse applications that BCI technologies promise, the general methodology may open new opportunities for clinical rehabilitation, for example, by training patients with movement disabilities to control abnormal activity in selected brain regions.

Stroke can affect physical, mental and social functions. Disability or paralysis is often affected only to one hemisphere, e.g. movements of one hand can be impaired while the other hand remains intact. Some stroke survivors exhibit poor control of movement smoothness [1], and movements seem to grow more smooth with recovery [2]. In monitoring of rehabilitation of stroke patients, objective evaluation methods are required. Furthermore, evaluation of the effectiveness of rehabilitation is also crucial. However, at present monitoring can only be based on qualitative measures, such as visual interpreting of movements during specific tasks.

Three-dimensional motion analysis techniques can be a powerful and objective tool for the assessment of human movements and it can be used to monitor rehabilitation progress. With an adaptive task setting customized to individual patient’s needs and performances, motion analysis can give valuable quantitative information. Additionally, combining 3D motion analysis techniques with neurophysiological signals could provide feedback for adaptive rehabilitation tasks, thus further improving the effectiveness of the whole process. In this paper, a virtual, adaptive and controllable rehabilitation environment which uses modern motion and biosignal analysis techniques in parallel is described.
2 Multimodal Platform

Human motion, and thereby, performance in a specific rehabilitation task can be tracked by using motion analysis methods. The physiological or neurophysiological status of the patient, on the other hand, can be estimated from different biosignals acquired during the rehabilitation task. By combining these modalities, a multimodal platform for monitoring of rehabilitation can be constructed.

2.1 Motion and Performance Tracking

Motion analysis methods have been widely used to measure and model human movements. Biomechanics can be considered as the base of modern motion analysis, which aims for modeling of human body as a mechanical composition of joints and rigid segments [3]. Motion analysis can be considered to consist of three components: kinematics, kinesiological electromyography (EMG) and kinetics. Kinematics examines the motion of body segments from geometric point of view without paying attention to forces producing the movements, whereas kinetics interlinks forces and movements produced by the forces [4, 5].

In human body modeling, the body is modeled as joints and bones, and in more sophisticated models, also muscles and ligaments are included in the model. The kinematic 3D human body model describes the translational motion and orientation of different body parts. By using the model various parameters such as velocities and accelerations of body segments or joint angles can be derived for further analysis.

The most advanced methods in motion analysis, which can be used for modeling of movements of the whole human body, are based on photogrammetric methods [6]. The camera technology has advanced during last years. Cameras utilizing FireWire or Ethernet interface are nowadays available at a reasonable price. Photogrammetry can be defined as measurement of three-dimensional objects geometry through two-dimensional images. In motion analysis the photogrammetric methods are utilized for determining the temporal positions and orientations of body segments with help of markers attached on the body. When the three-dimensional point of interest, e.g. a marker, is observed simultaneously with at least two calibrated cameras, the 3D-coordinates of the point can be solved.

We have developed and built a flexible mobile motion analysis laboratory which consists of multiple high speed cameras, image processing system, biosignal and inertial sensor measurement system and pressure insoles. The setup is suited for various research projects as well as development of methods applied in motion analysis. As an example, marker placements and marker trajectories for tracking of hand in rehabilitation task is shown in Fig. 1[7].

In many applications, motion tracking is performed in real time. This opens new possibilities for adaptive and interactive task settings, especially in virtual or augmented reality (VR or AR) applications.

2.2 Biosignal Analysis Platform for BCI Applications

Modeling brain’s activity following environmental stimuli or in the context of dynamically changing tasks is crucial for better understanding the central nervous system
Ideally, methods for assessing brain’s ability to interact with the environment should be computationally feasible, adaptive, and sensitive to cognitive changes. The ultimate goal is to make joint inference about the CNS dynamics based on complementary information from multimodal data sets [8], by conducting experiments focusing on adaptively changing cognitive tasks, such as time-varying workload and task difficulty. Furthermore, various autonomic nervous system signals such as heart rate (HR), blood pressure (BP) and galvanic skin response (GSR) are also important for psychophysiological modeling and monitoring.

Electroencephalogram (EEG) provides information about neural dynamics on a millisecond scale. EEG’s ability to characterize certain cognitive states and to reveal pathological conditions is well documented. A significant advantage of single-trial EEG analysis is that cortical reactivity and function can be assessed with high-temporal resolution. However, the limited signal-to-noise ratio (SNR) of noninvasive brain signals, makes the detection of single-trial events a difficult estimation task. Traditional way of analyzing event related potentials (ERPs), or any other event-related biosignals, has been to use heavy averaging, and thereby loosing significant inter-trial variability. Recently, several methods for single-trial estimation of even related EEG have been proposed [9-13].

Functional magnetic resonance imaging (fMRI) is another noninvasive method for studying cognitive function by measuring the hemodynamic response related to neural activity in the brain. The blood oxygenation level dependent (BOLD) effect is used for determining where activity occurs in the brain. The relationship between stimulation, neural activation, and BOLD response has been studied since fMRI was introduced. However, it is still not yet thoroughly understood. It has been found that the shape of the BOLD response varies across subjects and also within subject depending on the type of the stimulus and active brain area. Recently, BCIs based on single-trial metabolic activity of the brain have been introduced, defining new opportunities in neu-
roscience research, for instance, for studying brain plasticity and functional reorganization following sustained training [14]. Furthermore, simultaneous acquisition of EEG and fMRI combined with single-trial analysis provides an additional monitoring tool for the investigation of brain state fluctuations [15].

**Fig. 2.** A closed-loop biosignal analysis system for BCI based on adaptive stimulation.

An illustration of a biosignal acquisition and analysis system for BCI applications is given in Fig. 2. The system is operating in two phases, namely the signal acquisition and parameter estimation phase, and the feedback and adaptive control phase. During the first phase, all relevant signals are simultaneously recorded and synchronized in relation to various tasks. Individual signals are preprocessed simultaneously or separately, depending on the type of the signal and task, for accurate noise reduction. Then, features of interest are extracted for visualization, extra analysis, or classification. This procedure is performed by combining all information extracted from multimodal measurements with all available prior information in a Bayesian mathematical framework. In the second phase, event-related information is used to define and differentiate psychophysiological states of the subject and subject’s performance. Finally, the extracted parameters are used as a feedback to the subject, for instance as a visual feedback providing a reward mechanism or within a virtual reality environment. Furthermore, the parameters can be directly used to adaptively change physical characteristics of the sensory stimulation, for instance, type, intensity and duration of the next stimulus, or even to control task difficulty for optimal subject’s performance, thus providing an adaptive control mechanism.

**Example 1: Dynamic Estimation of Event Related Potentials.** An example of single-trial estimation of evoked potentials is given in Fig. 3. In this example, measurements
were obtained from an experiment with visual stimulation. A number of fixed intensity, fixed duration flash stimuli were predefined and sequentially delivered to the subject through a monitor. A decrease in amplitude of the dominant positive peak is clearly observed, suggesting possible habituation to the stereotyped stimuli. For this particular example, amplitude information can provide an indicator for the degree of habituation, and thereafter used to adaptively change the stimuli characteristics in real time with goal of forcing stable responses.

**Fig. 3.** Tracking single-trial characteristics (amplitude and latency) of evoked potentials during visual stimulation with a Kalman filter based approach.

**Example 2: Single-trial Estimation of Multimodal Brain Responses.** In simultaneous fMRI/EEG studies, the necessity of single-trial approaches is recognized. Single-trial EEG estimates are usually used as predictors for the voxel-wise activity. However, most of the approaches do not take into account variation in the latency or shape of the BOLD response. In Fig. 4, an example of single trial fMRI/EEG analysis is illustrated. A set of simultaneous fMRI and ERP measurements was acquired, and in the approach a joint model is defined and parameter estimates are obtained through subspace regularization [16].
Fig. 4. Typical BOLD response estimates when reaction time locked ERP responses are used in the regularization. (a) Concatenated data of the 62 BOLD responses and ERPs from channel Cz (bottom) and mean of the data (top). The amplitudes of the data are arbitrary and x-scale is in points. (b) Correlation matrix of the concatenated data.

Example 3: Dynamic Estimation of Heart Rate Variability (HRV). HRV is a reliable quantitative marker of ANS activity. HRV is typically assessed with a group of time and frequency-domain methods. By using these methods, the activities of the sympathetic and parasympathetic branches of ANS can be evaluated, and thus, useful information of the (neuro)physiological state of the subject can be extracted. In Fig. 5, dynamic HRV analysis corresponding to a sudden change in physiology caused by an orthostatic test is shown. This example demonstrates how evident the changes in heart rate and also in HRV characteristics can be in case of a change in physiology.

3 Virtual Rehabilitation Environment

Three-dimensional motion analysis techniques can be a powerful and objective tool for the assessment of human movements and it can be used to monitor rehabilitation progress. With an adaptive task setting customized to individual patient’s needs and performances, motion analysis can give valuable quantitative information. Combining 3D motion analysis techniques with neurophysiological signals could provide feedback for adaptive rehabilitation tasks, thus further improving the effectiveness of the whole process. In order to be practically applicable, such a system has to be highly automatized
and robust. Furthermore, a virtual reality environment (VRE) which can be applied to various rehabilitation tasks will extend the applicability and performance of the system.

The main components of VRE are real-time motion tracker, three-dimensional VR goggles and visualization engine, EEG and other biosignal measurement system and adaptive signal feedback driven task control system. An example of such a VRE is illustrated in Fig. 6. Such approaches, when utilized for rehabilitation or clinical appli-
cations, will enable more realistic and motivating tasks for patients. Finally, VR environments are easily controlled and patient safe, e.g. crosswalk simulation.

References

AffectPhone: A Handset Device to Present User’s Emotional State with Warmth/Coolness

Ken Iwasaki\textsuperscript{1}, Takashi Miyaki\textsuperscript{2} and Jun Rekimoto\textsuperscript{2,3}

\textsuperscript{1} Graduate School of Interdisciplinary Information Studies, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku Tokyo, Japan
iwkn@acm.org

\textsuperscript{2} Interfaculty Initiative in Information Studies, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku Tokyo, Japan
miyaki@acm.org

\textsuperscript{3} Sony Computer Science Laboratory, 3-14-3 Gotanda, Shinagawa-ku Tokyo, Japan
rekimoto@acm.org

Abstract. We developed AffectPhone, a system that detects a user’s emotional state using the GSR, and conveys this state via changes in the temperature (warmth or coolness) of the back panel of the other handset. Since GSR is a good measure of a user’s level of arousal, we detect the GSR using electrodes attached to the sides of the handset. When the user’s level of arousal increases or decreases, a Peltier module in the back panel of the other device generates warmth or coolness. This system does not require special sensors to be attached to the user’s body, and therefore, it does not interrupt the user’s daily use of the mobile phone. Moreover, this system is designed to convey non-verbal information in an ambient manner, and therefore, it would be more efficient than displays or speakers. This system is expected to help enhance existing telecommunication.

1 Introduction

During face-to-face communication, non-verbal cues convey as much information as do the verbal ones.\cite{1} However, when using a computer for communication, it is difficult to convey non-verbal information. We believe that conveying such non-verbal information would improve existing telecommunication, and therefore, we have developed a system to facilitate the same.

1.1 Non-verbal Communication Channel

We aimed to develop a system that provides non-verbal communication channel in addition to the existing telecommunication. A communication channel generally includes the information to be output and a sense as the input. For example, when we communicate over a telephone, our voice is the output, and auditory sense is the input (Figure 1(a)). In this system, we selected the galvanic skin response (GSR) as the output, and temperature sensation as the input (Figure 1(b)).
1.2 Selection of Input and Output

Humans use many non-verbal cues in their day-to-day communication. These cues include apparent ones such as body language, tone of voice, and facial expressions. However, we focus on non-verbal information that is not apparent, such as physiological information in this study.

We selected GSR as the output information, because Picard et al. suggested that GSR is a good measure of a user’s arousal[2]. Moreover, the GSR can be acquired by using only two electrodes on the user’s fingertip, and no special sensors need to be attached to the user’s body.

We selected temperature sensation as the output, because warmth or coolness is often indicative of ones emotions, as borne out by expressions such as, ”he is a cold man” or ”a heart-warming story.” Moreover, such haptic feedback does not interfere with user’s hearing or sight. In other words, the user can convey his/her emotional state in an ambient manner.

2 Related Work

Wang et al. used physiological sensors and animated text to develop a system that helps in communicating emotions in an online chat.[3] This system detects a user’s level of arousal from the GSR, and presents the user’s emotional state as animated text. This system uses GSR as non-verbal information, but it also presents it as an attribute of the text. The objective of this system is to make text chat more efficient; however, verbal information is mainly used in text chats. Therefore, we focus on telephonic communication.

Brave et al. proposed inTouch, a medium for haptic interpersonal communication.[4] inTouch introduces haptic sensations in interpersonal communication and serves to enrich the communication. However, haptic sensations cannot be generated unless a user voluntarily moves the device, and therefore, this method does not suitably mimic typical day-to-day interactions. In contrast, our system can determine a user’s emotional state from physiological information that the user cannot control voluntarily. Moreover, our system is integrated in a mobile phone we usually use, and therefore, it does not interrupt a user’s daily use of the phone.

Vaucelle et al. designed haptic interfaces for therapeutic purposes.[5] A part of their research, called the Cool Me Down project uses temperature sensations for therapy.
The device developed by them can be discretely worn by the user and only activated when necessary; this would help patients self-administer soothing sensory grounding treatment. Their paper discusses the design concepts of haptic interfaces for therapeutic purposes, but not for communication. In contrast, we focus on enhancing existing telecommunication.

![Fig. 2. AffectPhone. (a) Front and rear view of the system. (b) GSR window.](image)

3 AffectPhone

3.1 Concept

GSR has been stated to be a good measure of a user’s level of arousal, which indicates the user’s level of excitement. We selected temperature sensation as the input because temperature changes are generally reflective of one's emotions. In this system, a user can feel changes in the emotional state of the person he/she is conversing with in the following manner:

- When a user’s level of arousal increases, the system makes the back panel of the other device warm.
- When a user’s level of arousal gradually decreases for approximately 30 seconds, the system makes the back panel of the other device cool.

3.2 System Configuration

AffectPhone consists of two electrodes for detecting the GSR and a Peltier module for providing information on temperature change (Figure 2(a)). This system provides a non-verbal communication channel in addition to the existing telecommunication channel in a normal mobile phone. In other words, a user can convey his/her level of arousal to the person he/she is conversing with in an ambient manner. The user’s GSR can be detected from the two fingers in contact with the phone, and temperature changes can be sensed by the palm. This system does not require special sensors to be attached to the users body, and therefore, it does not interrupt the user’s daily use of the mobile phone.
AffectPhone is designed to convey a user’s arousal in an ambient manner. Moreover, the user can be aware of the arousal level of the person he/she is talking to, by viewing the GSRs in the GSR window (Figure 2(b)).

### 3.3 Feasibility Test

In order to study the feasibility of AffectPhone, we conducted preliminary experiments. We made a subject listen to music for approximately 4 minutes, and acquired the GSR signal using AffectPhone.

Figure 3 shows the GSR signal acquired using AffectPhone. The sampling rate was 5 Hz. Because the skin resistance decreases with an increase in the level of arousal, we calculated a user’s level of arousal \( a(t) \) as follows:

\[
a(t) = 1000 - R
\]

where \( R \) is resistance of the skin. (k Ohm)

A GSR signal consists of three elements—rise time, amplitude, and half recovery time (Figure 4). As shown in Figure 3 and Figure 4, the rise time, amplitude, and half recovery time are 5-10 s, 60-90, and 20-25 s, respectively. In comparison to Gasperi’s web cite[6], we concluded that this system successfully acquires GSR signal.

![Fig. 3. GSR signal acquired from AffectPhone. Peaks indicated by arrows indicate sudden increase in the level of arousal.](image)

### 3.4 GSR Signal Analysis

The above-mentioned test confirms that the proposed system can be used to acquire GSR signals. Here, we discuss how to analyze the acquired GSR signals.

The amplitude of the GSR signal indicates a change in the user’s emotional state (get angry, happy, etc.). In such a situation, the system makes the back panel of the other device warm when the following condition is satisfied.

\[
da/dt > d_T
\]
where $a$ denotes the user’s level of arousal, $t$, the time in seconds, and $d_T$, the threshold of difference, respectively.

After the half recovery time, the system turns off the Peltier module of the other device.

$$a(t) = (a_A - a_B)/2$$  \hspace{1cm} (3)

where $a_A$ denotes the level of arousal in the amplitude and $a_B$, the basal level of arousal.

When a user’s GSR gradually decreases after recovery, we consider that the user’s level of arousal has decreased (Figure 4). In such a situation, the system makes the back panel of the other device cool when the following condition is satisfied:

$$a(t) - a_B < 0 \quad (t_R < t < t_R + 30)$$  \hspace{1cm} (4)

where $a_B$ denotes the basal level of arousal, $t_R$, the time of recovery.

4 Potential Applications of AffectPhone

AffectPhone can convey a user’s level of arousal in terms of changes in the temperature (warmth or coolness) of the back panel of a phone. This system is useful not only when a user is talking on the phone, but also when the phone is ringing. When the phone rings, the system determines the level of arousal of the caller. If this system finds widespread use, the following scenario might become possible. If the caller is upset, the receiver can be made aware of the same even before answering the phone. A user can detect whether or not a call is important, and thus deal with a difficult situation more effectively.
5 Conclusions and Discussion

In this study, we designed AffectPhone, a handset device that can be used to identify a user’s emotional state via temperature changes (warmth or coolness). This device has been designed to provide a non-verbal communication channel in addition to the existing telecommunication channel in a normal mobile phone. However, further improvements must be made to this system. In the future, we intend to focus on user evaluations of factors such as the following:

- The extent to which a user’s level of arousal changes when talking on the phone.
- The accuracy of determination of temperature changes by the user.

Acknowledgements. We are grateful for being adopted for part of this project as ”MI-TOH youth” project in Information-Technology Promotion Agency in Japan.

References

# Author Index

<table>
<thead>
<tr>
<th>Author</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjei, O.</td>
<td>35</td>
</tr>
<tr>
<td>André, E.</td>
<td>13</td>
</tr>
<tr>
<td>Bozed, K.</td>
<td>35</td>
</tr>
<tr>
<td>Bragge, T.</td>
<td>75</td>
</tr>
<tr>
<td>Broek, E.</td>
<td>51, 59</td>
</tr>
<tr>
<td>Coghlan, N.</td>
<td>67</td>
</tr>
<tr>
<td>Georgiadis, S.</td>
<td>75</td>
</tr>
<tr>
<td>Giakoumis, D.</td>
<td>3</td>
</tr>
<tr>
<td>Hassapis, G.</td>
<td>3</td>
</tr>
<tr>
<td>Healey, J.</td>
<td>51</td>
</tr>
<tr>
<td>Iwasaki, K.</td>
<td>83</td>
</tr>
<tr>
<td>Jaimovich, J.</td>
<td>67</td>
</tr>
<tr>
<td>Janke, M.</td>
<td>22</td>
</tr>
<tr>
<td>Janssen, J.</td>
<td>51</td>
</tr>
<tr>
<td>Jung, F.</td>
<td>13</td>
</tr>
<tr>
<td>Karjalainen, P.</td>
<td>75</td>
</tr>
<tr>
<td>Kim, J.</td>
<td>13</td>
</tr>
<tr>
<td>Knapp, R.</td>
<td>67</td>
</tr>
<tr>
<td>Kobayashi, T.</td>
<td>44</td>
</tr>
<tr>
<td>Kosunen, I.</td>
<td>3</td>
</tr>
<tr>
<td>Mansour, A.</td>
<td>35</td>
</tr>
<tr>
<td>Miyaki, T.</td>
<td>83</td>
</tr>
<tr>
<td>Moustakas, K.</td>
<td>3</td>
</tr>
<tr>
<td>Niskanen, E.</td>
<td>75</td>
</tr>
<tr>
<td>Otsu, N.</td>
<td>44</td>
</tr>
<tr>
<td>Ranta-aho, P.</td>
<td>75</td>
</tr>
<tr>
<td>Rekimoto, J.</td>
<td>83</td>
</tr>
<tr>
<td>Schultz, T.</td>
<td>22</td>
</tr>
<tr>
<td>Schut, M.</td>
<td>59</td>
</tr>
<tr>
<td>Shirai, K.</td>
<td>44</td>
</tr>
<tr>
<td>Tarkka, I.</td>
<td>75</td>
</tr>
<tr>
<td>Tarvainen, M.</td>
<td>75</td>
</tr>
<tr>
<td>Tuinenbreijer, K.</td>
<td>59</td>
</tr>
<tr>
<td>Tzovaras, D.</td>
<td>3</td>
</tr>
<tr>
<td>Vogianmou, A.</td>
<td>3</td>
</tr>
<tr>
<td>Vogt, T.</td>
<td>13</td>
</tr>
<tr>
<td>Wagner, J.</td>
<td>13</td>
</tr>
<tr>
<td>Wand, M.</td>
<td>22</td>
</tr>
<tr>
<td>Watanabe, K.</td>
<td>44</td>
</tr>
<tr>
<td>Westerink, J.</td>
<td>51, 59</td>
</tr>
<tr>
<td>Yoshikawa, F.</td>
<td>44</td>
</tr>
<tr>
<td>Zwaag, M.</td>
<td>51</td>
</tr>
</tbody>
</table>