Thesis Master Information Studies: Human Centred Multimedia
University of Amsterdam Faculty of Science

The implementation of a real-time algorithm to measure the engagement of the audience during a performance

Alice Panza
11137800

Final version: 29/06/2016 Supervisor: Thomas Röggla

The implementation of a real-time algorithm to measure the engagement of the audience during a performance

Alice Panza
University of Amsterdam
MSc. Information Studies
alice.panza@student.uva.nl

Abstract *– Electrodermal activity (EDA) is a physiological signal which has been considered as one of the best indicators of sympathetic arousal, since it provides an indication of sweat gland activity. For this reason, it can be used to measure the engagement of the audience during a performance. This work exploits the use of wearable EDA sensors to measure the audience engagement during a live performance, testing a novel method that works in real-time. In particular, we collected data from 40 subjects participating in a jazz concert and used machine-learning based approaches to automatically identify the level of engagement during the show, minute by minute. Finally, we evaluate our algorithm using annotations reported during the live event. Our results suggest that improvements are needed in order to predict the audience's engagement accurately. However, this paper provides new insights originated from the particular experimental setting: a live music event.*

Keywords – machine learning, electrodermal activity, wearable sensor

Categories and Subject Descriptors
G.4 [**Mathematics of Computing**]: Mathematical Software – *Algorithm design and analysis*; H.1.2 [**Models and Principles**]: User/Machine Systems – *Human Information processing*

General Terms
Algorithms, Experimentation

1. INTRODUCTION

Our body reacts to external events, producing effects such as changes in heart rate, blood pressure or sweating. In particular the activity of the sweat glands, which are present throughout the human body, causes electrical variations in the skin. Since sweat is a weak electrolyte and good conductor, it impacts the conductance of an applied current [14]. It means that the level of skin conductance changes accordingly with sweat gland activity. Exploiting this phenomenon, it is possible to measure skin conductance at the surface, referred to as electrodermal activity (EDA), that provides significant information about alterations coming from the central nervous system that are associated with different types of stimuli (such as emotion, cognition and attention) [14].

Therefore electrodermal activity describes the ability of the human skin in handling electricity that can be measured by wearable sensors. EDA sensors are usually placed on the fingers to measure skin conductance after applying a fixed small voltage to that area [13]. This is currently an important topic for researchers, the correlation between such signals and the affective state is constantly investigated and many studies are already available in the literature [2]. In particular, it is known that this signal embeds human-centered information: it is highly correlated with the user's affective arousal [9] and might be used to provide a continuous and implicit feedback about the level of excitement of the user [6]. This topic is further examined in Section 3.

Although it is quite easy to measure EDA signals, the interpretation of the gathered data implies the implementation of an algorithm that processes raw data to yield meaningful information. First of all, EDA variations are quite subject-specific, which means that the baseline for the EDA readings is different from person to person and depends on several factors, such as the person's temperature and diet as well as how the sensor is attached to the hand. For this reason, it is not possible to work with absolute values and the algorithm should keep these factors into account, and extract meaningful quantities from the raw data. Applying signal processing techniques to the EDA signal, such as filtering and thresholding, may not be robust, since as stated above EDA variations are subject-specific [5]. Moreover, the algorithm has to deal with the noise of the signals, in particular it should be able to handle missing and corrupted data.

1.1 Research question

The main objective of this research is to plan and implement an algorithm which takes EDA data as input, processes it in real-time and gives the estimation of the level of users' arousal as output. Estimating the affective state in real-time and directly from the physiological data is still a difficult task [5].

In order to propose a new real-time affect detector based on EDA as the only source of physiological information, we especially focus on changes of the affective state in terms of arousal, rather than in the identification of a precise emotion. Therefore, the proposed detector will aim at automatically analyse the intensity of the emotion felt by a group of people, without a qualitative judgement relative to valence (positive or negative) or the kind of the emotion (sadness, happiness, interest, etc.).

In particular, the context of our research is about measuring the engagement of the audience during a performance. However, there is not a unique definition of engagement, which implies there are several ways to measure it and to interpret this information. Therefore, a primary question is about the meaning of physiological signals and their relation with the concept of engagement. The first question we want to answer is the following:

- What does 'engagement' mean? What is the relation between engagement and EDA signals?

In Section 3.1 we provide an overview on previous literature to answer this question from a theoretical point of view. Besides that, the main research question that we attempt to address is:

- To what extent can we measure the level of engagement of the audience during a performance using a machine-learning based algorithm?

To answer this question we collected EDA sensor data from 40 participants attending a live jazz concert. Afterwards, we processed the raw data and used machine-learning methods to train a classifier. The output is a representation of the audience's engagement, divided into three levels of intensity.
Specifically the contribution of this research is to provide creative people with a tool of real-time analysis of audience reaction during a live show or performance. While there is large literature on mapping EDA to arousal, there is only a small amount of work about measuring the audience engagement in real-time during a performance [10]. However, research in this direction can have a direct impact on many applications [18].

1.2 Structure of the paper

The paper is structured as follows: the next section briefly introduces the research context of our project, and provides some examples of possible applications. In Section 3 we present a conceptual exploration of audience engagement, how we define it and how it is related to physiological sensors. Moreover, a review of the relevant work is presented. Section 4 describes our method. We initially collected EDA data from 40 subjects attending a live jazz concert. We implemented an algorithm based on a machine-learning approach to learn the level of engagement of the audience in such a way that it can be possible to measure the audience feedback minute by minute. Afterwards, we present our results in Section 4 and we discuss strengths and limitations of our project in Section 5. We finally outline our conclusion in Section 6.

2. MOTIVATION

Emotion is a fundamental part of human experience, however only few decades ago researchers started to investigate how machines can help in understanding emotions, and to develop new technologies aimed at understanding affective states [4]. Nowadays, including emotion to improve the experience of users with computers has become a main concern in the area of Human Computer Interaction [13].
Affective Computing is a scientific field focused on the study and development of intelligent systems that recognize and react to human affective states. Researchers working in this field believe in the importance of emotion in communication, and are interested in exploring the application of emotion recognition by machines. As stated by Picard, emotions have a major impact on essential cognitive processes and if one wants to improve the interaction with computers, it is important to "teach" them to recognize affect [12]. One main step of this process consists in detecting affect.
Affect detection is critical since it aspires at understanding the user's affective state, however, it is still a challenging issue because emotions are concepts that can not be measured directly [4]. Moreover, the interpretation of the gathered data implies the application of an accurate algorithm that processes raw data to yield meaningful information. To detect user's emotions, a technological system can get input data from different sources: such as facial expression, speech, gesture movements, etc. However, these behaviours can be easily masked by intentional control. Instead, to avoid possible artefacts, physiological signals have recently become a more reliable emotional channel for human emotion recognition, especially thanks to the unbiased nature of the signals that originate autonomously from the central nervous system [13]. Generally, these signals can be collected from the cardiovascular system, respiratory system, muscular system, or via brain activities and electrodermal activity, [13]. Measuring these parameters means obtaining streams of numbers, apparently meaningless in itself. However, it is possible to interpret this data and link it to emotions, through understanding their relation and implementing mathematical tools that facilitate its interpretation. For this reason, the field of physiological computing has become an active area of research [16].

2.1 Applications

In general, Affective Computing can be used in a broad range of applications**,** from improving human-computer interaction according to the user's emotional state, to assess the audience feedback during a show.
Getting the affective state of an audience member attending a performance or watching a film may have many potential applications in the creation and distribution of an artistic product [6]. It can be useful to select the best target, know what emotional effect the show has on the audience, or identify which are the most interesting parts according to the audience's reaction. Another example consists of recommendation systems, where users' ratings are used to outline their preferences, predict the success of a new product or other marketing strategies.
The real-time property is fundamental, since the main difference between an offline algorithm and a real-time method is that the former can be more useful for research purposes, especially for the analysis and understanding of physiological data, in relation to human emotions. The latter, on the other hand, can be used also for more artistic applications. For example, it could be possible to enhance performances with the audience feedback by showing a real-time visualization of the physiological data obtained as output of our algorithm.
There is a small amount of work in which the audience's engagement, measured through biometrics, is used as real-time input to performance. The main challenge of this project consists in the interpretation of the EDA signal in real-time, because the processing of the raw data should be simple and fast so that meaningful interpretations can be used by artists or performers.

3. RELATED WORK

This project combines computer science with psychophysiology, since it deals with the relation between emotions and physiological signals and aims at creating a bridge through the implementation of an algorithm. The

main focus is the detection of affect states by using techniques that identify patterns in physiological activity, which are known to be related to the emotional state of the subject [5].

3.1 Measuring the arousal of emotions

Emotion is a subjective concept, hard to define and measure. For our research, we referred to the dimensional theory developed by Lang [9], that is derived from the scientific view of emotion by Russell [17]. He asserts that all emotions can be located in a two-dimensional space, as coordinates of affective valence and arousal. Arousal denotes the level of intensity, with passive emotions having a low arousal, and energetic emotions having a high arousal, while valence represents the positive and negative scale of the emotion (see Fig. 1).

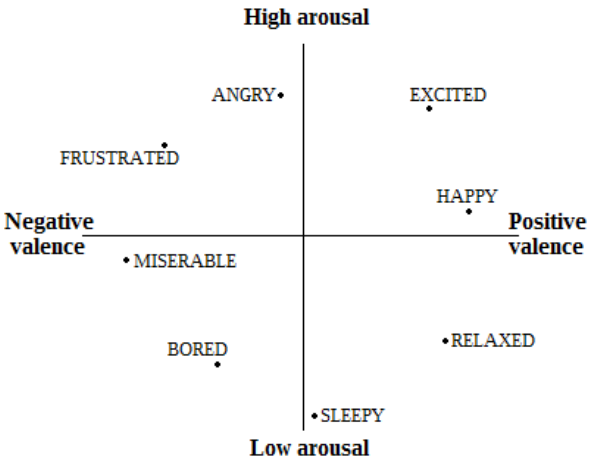

Fig. 1: The circumplex model by Russell [17], eight affect states plotted in the two-dimensional space

EDA sensors give an indication of the user's reaction in terms of the arousal. However, through the analysis of this simple biometric data, we are not able to assign a positive or negative quality to the emotion recorded. EDA is thus linearly correlated to arousal and reflects emotional intensity changes.

As a consequence, we can identify emotional events analysing the EDA signal, and we are interested in the detection of relevant changes in the emotion's intensity, rather than in the identification of the precise emotion. We aim at automatically identifying those events in real-time and assess the intensity of the audience arousal, which we will refer to as the engagement of the audience. Assuming that, we define audience engagement in terms of affective states of arousal. Normally, we think of engagement as a synonym of attention and interest with a positive valence, but one can be attentive and interested with negative valence too. The aim of a performance is to provoke an emotional reaction on the audience, either with a positive (eg. funny) or negative (eg. dramatic) valence. In other words, arousal is more significant in measuring engagement: in these terms, we can measure how sleepy or how active the audience is. Intuitively, it seems unlikely that a performer would want the audience to feel sleepy or bored.

Latulipe et al. point out that measuring valence can also be misleading, since the main risks would be in the differentiation between how the performance makes a person feel and how much they like a performance. Given this, measuring valence is considerably more difficult that measuring arousal. Moreover, investigating the importance of valence data, they found that performers are less interested in it, since positive and negative emotions are equally valued in art manifestations [10]. In conclusion, the dimension of experience that we want to measure deals with the ability of the performers to capture and maintain the audience's attention.

After defining the concept of engagement, another important choice to take is related to the methodology used to measure it. Hernandez et al. list three different approaches: self-reports (interviews or surveys), external ratings, and physiological information [7]. While collecting explicit feedback (surveys, focus groups with test audiences) is still very common since it is fast and direct, EDA signals provide a source of implicit feedback which can be used to infer users' reaction at a fine granularity, enabling us to get for example the mean reaction of a whole audience at different instants of the performance. On the contrary, explicit feedback usually provide only a single rating for the entire show, moreover survey forms are limited by their reliance on viewer memory, and focus groups are constrained by participation costs and time limitations [18]. An alternative method consists of asking experts to annotate participants' engagement state during the presentation of stimuli. This approach is based on the theory of facial expression recognition. Although facial movements can be detected at a distance, they can be voluntarily controlled and in general not completely representative of the actual affective state [7]. Although each one of them has its advantages and drawbacks, physiological signals seem to represent the less disruptive method to collect data and they perform as an effective indicator of the engagement level of people in different settings. In addition, among the different physiological signals, EDA is easier to measure, especially if compared to functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) [11].

3.2 EDA signal

Electrodermal signals are composed of two main components: skin conductance level (SCL) and skin conductance responses (SCRs). The first is the tonic component that represents slow changes while the second one, the phasic component, is characterized by rapid peaks. SCL is, in fact, the product of the general activation of the sympathetic nervous system [11], and it could be measured considering the absolute level of conductance of the skin in absence of any particular external stimuli. The values of this component change slowly over time, according to many different factors, such as personal psychological state, temperature, or hydration, so it also varies between individuals. Instead, SCR reflects the effect of unexpected or relevant events. Discrete environmental stimuli are associated with the shape of electrodermal signals, being represented by abrupt increases in skin conductance. Fig. 2 shows an example of an electrodermal signal over time: SCL is modulated by temporary peaks (SCRs). Usually, in experimental settings, the aim is to prove that distinct and isolated stimuli are associated with peaks in the signal.

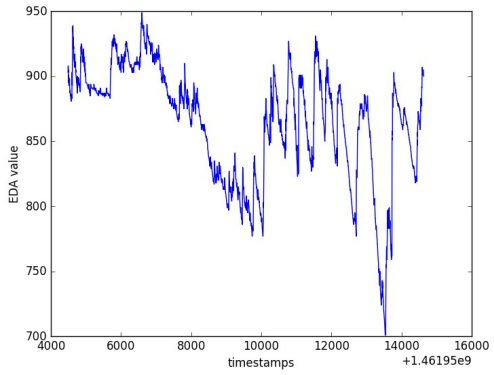

Fig. 2: An example of EDA signal plotted over time

For this reason, it is necessary to analyse the two components separately. The illustration of SCR is presented in Fig. 3, where the most common features are outlined. However, these measurements are not always easy to compute since multiple skin conductance responses can overlap, for example when subsequent reactions occur before the first has ended [11]. The features selected for the aim of this work are described in Section 4.3.

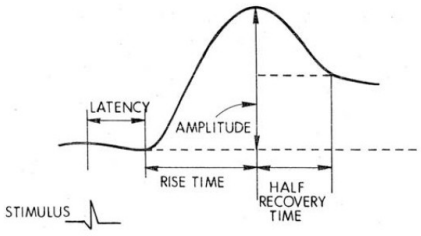

Fig. 3: Graphical representation of principal SCR components [3]

4. METHOD

We propose a novel EDA signal analysis that aims at detecting and quantifying the level of engagement of the audience minute by minute. To this end, a machine-learning framework has been developed, each component is described following the main steps in the pipeline illustrated by Ping et al. [13]. (see Fig.4)

4.1 Data collection

We conducted an experiment during a live jazz concert in order to collect data from EDA sensors. The experiment took place at the Goethe-Institut Niederlande in Amsterdam on the 29th of April 2016. Thanks to this acquisition stage we were able to gather input data from three different sources and organise it accordingly in three datasets:

- EDA signals
- Annotations
- Questionnaires

The objective is to combine these pieces of information in order to find patterns in the data, and these observations will help us to train a model that measures audience engagement. It has to be noticed that data collection is the only step that took place during the experiment. The data was successively analysed in laboratory.

4.1.1 EDA signals

Wearable sensors specifically assembled by the Distributed and Interactive Systems group [19] at the Centrum Wiskunde and Informatica (CWI) in Amsterdam were used. Thanks to the network setting used, it was possible to collect data from several users at the same time exploiting a wireless connection. In this way, we were able to receive packets of data simultaneously, which is very important for real-time experiments.

Our sample consists of 40 participants, 15 male and 25 female with an average age of 32 years. No specific criteria were adopted for the recruitment. Participants were given short instructions about the aim of the experiment and about the meaning of the sensor data. They were asked to be on location 30 minutes before the start of the concert, in order to put the sensors on and fill in the pre-questionnaire. They were asked to wear the sensors for the entire event and they were given free entrance. The concert lasted approximately three hours and was divided as follows: performance of first band, break, performance of second band. The entire recording process was synchronized and controlled and the EDA measurements were sent to a receiver through wireless transmission. For technical reasons, we set up two receivers in the concert room. Participants were divided into two abstract groups, according to the sensors ID number. Each receiver recorded data from 20 sensors (i.e. one group). The outcome of this step concretely consists of two CSV files containing all data related to the two groups. For each file (respectively length: 121328 and 114327 lines), every line is composed of a timestamp, a string message, a sensor ID number, a EDA value (from 0 to 1023).

```
1461954491.286,,1625,835
1461954491.250,"crcerror 0xB202",,,
1461954491.411,"next loop",,,
```

Tab. 1: An example of input EDA data

Some fields can be empty. An example is shown in Tab. 1. The first line represents a complete data packet from sensor number 1625, with EDA value equal to 835. The string message is present only in case of an error which implies missing data (see line 2 in Tab.1) or in case of "next loop" (see line 3 in Tab.1). This last message appears each time after the receiver communicated with each sensor in the group singularly. Each loop has no a fixed duration, but in general it lasts around 1/1.5 seconds. This information has been used to organise the

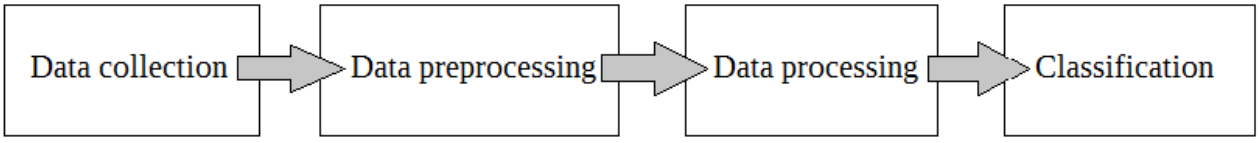

Fig. 4: Pipeline adopted for the implementation of the algorithm

data collected in this stage. Therefore, the final outcome is represented by a matrix where the number of rows corresponds to the number of total loops recorded and the number of columns is the number of sensors, in this case 20 for each group. Each item in the matrix is an integer number that ranges from 0 to 1023 and represents the EDA value detected by the sensors.

4.1.2 Annotations

To evaluate the algorithm, specific moments during the concert were reported: through an application, three people from the team were in charge to annotate when the level of engagement of the audience was perceived to be either very high or very low. This step allowed us to have a ground-truth in order to assess the final accuracy of the algorithm. In fact, at the end of the concert, we obtained three files. Each file contains a list of timestamp-annotation pairs created by each annotator. The annotations are extremely simple and represented by a -1 or a +1, respectively the atmosphere perceived in the room is very calm or very excited, according to the definition of engagement given in Section 3.

Another common method to define a ground truth is trough a direct self-assessment [10], [15]. It consists of asking each participant to annotate explicitly his own mood, minute by minute. However, this approach has some drawbacks: first of all it is quite subjective and too intrusive, it may distract the subject from the performance.

Instead, according to the method we adopted, the outcome of this step consists of a unique file where all the annotations were put together and ordered by timestamps. Finally, values were arithmetically summed minute by minute and separated in two vectors. The first vector has length 75, which is the number of minutes of the first band's session. Similarly, the second vector contains 59 values and it is related to the second band (see Fig.5 and Fig. 6). Data collected during the break has not been further analysed. In conclusion, these vectors represent the ground truth and are used to train and evaluate the algorithm (see Section 4.4).

4.1.3 Questionnaires

Before and after the concert, a questionnaire was provided to the participants in order to assess mood, interest in the concert, quantity of alcohol consumed during the course of the night. Particularly, we needed this information to further analyse our results. In Section 5, we describe different results, obtained by comparing different subsets of our sample.

4.2 Data preprocessing

We implemented the next steps using Python as programming language, since it offers useful libraries for data manipulation and analysis (Pandas), and for general scientific computing (NumPy, matplotlib).

In Section 4.1.1 the structure of the actual input data is described. The input matrix has been divided in two smaller datasets: “df1” and “df2”, containing the EDA values collected during the first and the second band respectively.

The first task is to deal with missing values. During the experiment almost 30% of data was lost, which means it has not been received correctly by the receiver. In programming, this is translated by a not defined value, that causes “holes” within the sequence of integer numbers accurately recorded. To solve this, an interpolation method has been implemented. Missing data are filled using the *interpolate()* function available in SciPy. Values are estimated with a linear interpolation method. Moreover, at this step, the data of five sensors have been removed from the analysis, since the signal was too corrupted or people left before the end of the concert. Therefore, the actual sample consists of 35 participants.

The other important factor to consider in the preprocessing stage is the subjectivity. As stated above, EDA signals are affected by individual differences. This challenge is relevant especially for algorithms that rely on a machine-learning approach, because basically the idea is to learn from a set of data (i.e. a group of people) and then generalize to a broader sample that includes new data. Usually, this problem is solved taking into account normalized values or individual baselines. However, in our case, the implementation of both methods is not straightforward. Normalization is computed knowing all the values over the period of time studied, not feasible when processing data in real-time, or minute by minute as it is our case. Concerning the baseline, it is useful to detect a general trend in the signal but it can be recorded asking people to relax for 10/15 minutes, which was not possible during our experiment, for logistic reasons.

However, since our target is the entire audience and we take into account average values, we can assume that the subjectivity factor will affect each minute in the same way. This assumption is further analysed in Section 5, where a comparison between collected values and

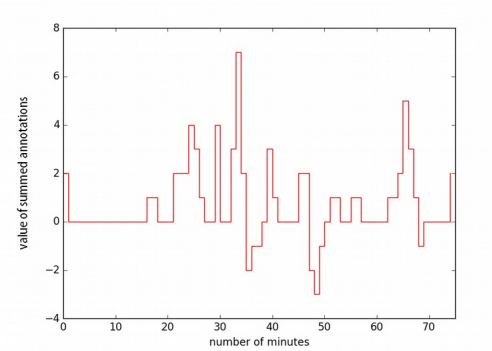

Fig. 5: Annotations summed during the first part of the concert

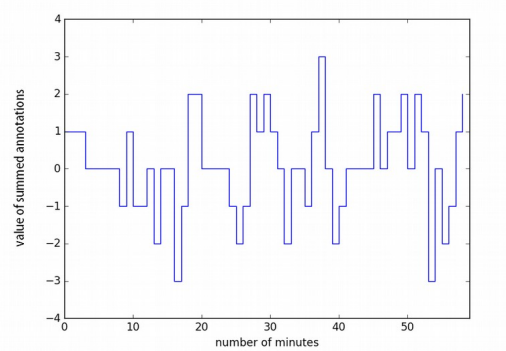

Fig. 6: Annotations summed during the second part of the concert

normalized values is given.
Finally, we implemented a function to reduce the effect of outliers. In our case, the main source of artefacts is due to quick sensor movements, which are impossible to control during a concert [7]. High frequency motion artefacts have been detected and attenuated using linear interpolation.

4.3 Data processing

Before training the classifier, signals have to be translated into vectors. Each element in the vector is a number that describes a feature. In order to continuously measure arousal from the analysis of EDA signals, a large number of features can be extracted [8]. According to previous research ([21], [7]), we selected the most important and common features used to describe EDA signals.
SCR events are generally sparse and vary considerably in their intensities. Furthermore, due to the subjective differences across people, the SCR events may not be temporally aligned and could also consist of some events that may not be related to the same stimulus [18]. To mitigate this effect in the algorithm, we considered EDA values by aggregating them into intervals of one minute. We focus on statistical measures in the time domain, since they are easy and fast to compute. Therefore, for each minute, we compute the average value of the following variables considering all sensors:

- minimum EDA value;
- maximum EDA value;
- mean EDA value;
- median EDA value;
- negative slope: it captures an overall decrease of the response;
- positive slope: it captures an overall increase of the response;
- number of peaks: peaks are detected analysing changes in the derivatives.

These selected features are used to train the classifier. In practice the output data of this step is represented by a list of 7-unit length vector. Each vector describes one minute during the concert according to the features selected, therefore we have 75 vectors for the first part and 59 vectors for the second part.

4.4 Classification

The first parameter to define is the number of classes. In Section 4.1.2 we describe how we created the ground-truth through the annotations process. Vectors' values range from -3 to +7 for the first band and from -3 to +3 for the second band. In order to reduce the number of classes we mapped each value into three superclasses. We use {0, 1, 2} as labels for the new classes. Class 0 is associated to those minutes characterized by a low engagement by the audience, this is when the total value in the annotation is less than -2 (i.e. at least two negative annotations were reported in that minute). Class 2, on the contrary, represents intervals of time characterized by a high engagement, and consequently the total value in the annotation is more than +2 (i.e. at least two positive annotations were reported in that minute). In order to make this representation more objective, we check if the annotations have been reported by different annotators. Lastly, for intermediate values, it is hard to define the level of engagement perceived, therefore we map them in class 1. The input data for the machine-learning part is presented as follow:

- 75 vectors with 7-unit length (first part of the concert);
- 75 respective labels;
- 59 vectors with 7-unit length (second part of the concert);
- 59 respective labels.

To proceed, a training set that represents the data to learn from, and a test set that is used to make prediction have to be defined. Since each element in the test set has a label associated (validation set) that represents which class it belongs to, we are able to evaluate the algorithm comparing predictions and real labels.
To accomplish this task, we implement k-Nearest Neighbour (k-NN). By using k-NN, we first store data and then elaborate it, associating classes through a comparison between the training set and the test element we want to classify.
We also implemented a Support Vector Machine model which requires a training phase to learn the classes using the function *fit()* before we can apply it to the validation set with *predict()*. The SVM approach aims at finding a hyperplane that separates classes, more precisely the one that best generalise the classification. It means that the objective is to find the separating hyperplane that has the largest distance to the nearest training elements of different classes. At the same time, this hyperplane has to correctly separate as many instances as possible. These two objectives can be in opposition to each other. The C parameter gives us the opportunity to manage this problem. A low value of C gives a large 'minimum margin', even if that hyperplane misclassifies more points (so the number of training errors increases). A high value of C allows in particular situations to correctly classify elements but with a smaller minimum margin between different classes (so a loss in generalization properties of the classifier). So the parameter C can control the trade-off between errors of the SVM on training data and margin maximization.
In addition to performing linear classification, with SVM we can apply a non-linear classification using the so called 'kernel trick'. Through its use, the input is mapped into high-dimensional features spaces. Solving the algorithm with a linear kernel is faster, but typically the predictive performance is better with a non-linear kernel.
Data has been analysed in two different ways. First we consider the two bands together. Therefore the input data includes 134 vectors with 7-unit length. Using 10-fold cross validation, original data is partitioned into 10 sub samples: 9 are used as training set while the other one is the test set. This process is then repeated 10 times, and the 10 results are averaged to obtain a final accuracy estimation. This approach based on repeated sub-sampling guarantees that all vectors are used both for training and validation, especially each vector is used for validation only once.
With the same procedure, we analysed three more different scenarios, in order to investigate possibly noisy factors. Therefore, in the second scenario we use normalized factors. In the third scenario, four participants

have been removed, those who drank more than three drinks before and during the concert (according to the questionnaires). In the last scenario, six participants have been removed, those who were not really interested in the concert (according to the questionnaire).
Finally, in order to explore how the music has an influence in our experiment, we analysed the data collected during the first band and the second band separately. Similarly as with the previous scenarios, a 5-fold cross validation is used. In the next section accuracy results are presented.

5. RESULTS

The last step is the evaluation. Comparing associated labels with the predicted ones, we know how many correct matches we obtain from our algorithm and we can compute accuracy, which is the fraction of predictions that are correct over the number of all predictions and gives an understanding of how efficient our algorithm is. The following tables show the accuracy obtained.
Tab. 2, Tab. 3, and Tab.4 present the result for the four scenarios previously explained, using k-NN and SVM (with linear and rbf kernel). In particular, Scenario 1 includes 35 participants (only corrupted data has been removed from the original sample), Scenario 2 is represented by the same input data that has been normalized during the preprocessing step, considering for each sensor data its minimum and maximum collected during the concert. Comparing Scenario 1 and 2 it is possible to analyse potential differences caused by the subjectivity factor that is inherent to physiological signals. Scenario 3 includes 31 participants: from the sample in the previous situations, four participants that asserted to having consumed a certain amount of alcoholic beverages before and during the concert have been further removed from the analysis. Finally, in Scenario 4 we want to investigate the effect of removing those participants who confirmed not to be interested in the concert itself.
Tab. 2 present the results obtained in the different scenarios using k-NN method and tuning the parameter k from 1 to 10. Accuracy values range between 0.48 and 0.7 (underlined values in Tab.2). Although with a low value for k (k = 1, 2, 3) accuracy is lower than 0.6, in the other cases values oscillate around 0.65. Higher values are obtained using SVM approach. In Tab. 3 we present results obtained with the linear kernel and in Tab. 4 with rbf kernel (with parameter C = {0.001, 0.01, 0.1, 1, 10, 100, 1000}). The highest accuracy obtained by this experiment is 0.71 (underlined value in Tab. 3) which is a positive result. It has to be noticed that the aim of this work is to explore the extent to which we can predict the level of engagement in an audience using machine learning methods. This approach represents the novelty of our work: the interpretation of results should take this into consideration. However, there is room for improvements. An interesting output, other than accuracy values, arises from comparisons between the different methods adopted and the different scenarios analysed. Firstly, the SVM method gives better result than k-NN, while there are no substantial differences between linear and rbf kernel. Tuning the C parameter also gives only slightly different results, however the best results are obtained when C is set to 0.1. Interestingly, similar results are obtained in the four different scenarios. The reason is likely to be "hidden" in the algorithm itself: in fact, all the features selected are computed as an average over all the sensors data. Assuming there is a noisy factor (due for example to the alcohol, or to the limited interest in the event), this affects all the signal in the same way. Although potential differences might even each other out, carefully comparing the results in the four scenarios, we can notice that in scenario 2 (normalized data) and in scenario 4 ("not interested" participants removed) slightly higher results are obtained. This means that further research in understanding how different moods, habits, interests can affect EDA data is needed, although for different applications.

	k= 1	**k= 2**	**k= 3**	**k= 4**	**k= 5**	**k= 6**	**k= 7**	**k= 8**	**k= 9**	**k= 10**
Scenario 1	0.6	0.63	0.58	0.63	0.62	0.64	0.65	0.65	0.65	0.67
Scenario 2	0.54	0.57	0.56	0.62	0.64	0.68	0.68	0.69	0.70	0.69
Scenario 3	0.48	0.61	0.63	0.65	0.6	0.66	0.65	0.65	0.64	0.65
Scenario 4	0.62	0.64	0.6	0.68	0.67	0.63	0.66	0.65	0.66	0.65

Tab. 2: Accuracy results using k-NN

	C= 0.001	**C= 0.01**	**C= 0.1**	**C= 1**	**C= 10**	**C= 100**	**C= 1000**
Scenario 1	0.69	0.69	0.69	0.68	0.65	0.66	0.68
Scenario 2	0.69	0.69	0.71	0.68	0.63	0.6	0.6
Scenario 3	0.69	0.69	0.7	0.68	0.66	0.64	0.65
Scenario 4	0.69	0.69	0.69	0.67	0.67	0.68	0.66

Tab. 3: Accuracy results using SVM with linear kernel

	C= 0.001	**C= 0.01**	**C= 0.1**	**C= 1**	**C= 10**	**C= 100**	**C= 1000**
Scenario 1	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Scenario 2	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Scenario 3	0.69	0.69	0.69	0.69	0.65	0.65	0.65
Scenario 4	0.69	0.69	0.69	0.69	0.69	0.69	0.69

Tab. 4: Accuracy results using SVM with rbf kernel

	k= 1	k= 2	k= 3	k= 4	k= 5	k= 6	k= 7	k= 8	k= 9	k= 10
Band 1	0.49	0.52	0.48	0.6	0.57	0.56	0.6	0.6	0.59	0.61
Band 2	0.58	0.55	0.45	0.65	0.65	0.64	0.71	0.67	0.67	0.67

Tab. 5: Accuracy results using k-NN, data divided into Band 1 and Band 2

	C= 0.001	C= 0.01	C= 0.1	C= 1	C= 10	C= 100	C= 1000
Band 1 Linear kernel	0.65	0.65	0.61	0.63	0.56	0.53	0.52
Band 2 Linear kernel	0.75	0.75	0.75	0.75	0.71	0.71	0.69
Band 1 Rbf kernel	0.65	0.65	0.65	0.61	0.57	0.57	0.57
Band 2 Rbf kernel	0.75	0.75	0.75	0.75	0.75	0.75	0.75

Tab. 6: Accuracy results using SVM (linear and rbf kernel), data divided into Band 1 and Band 2

Tab. 5 and Tab. 6 represent the accuracy obtained analysing data collected during the first part of the concert and the second part separately. Tab. 5 presents results obtained by the k-NN method and Tab. 6 by SVM. For this analysis we only consider the sample from scenario 1 (35 participants, which means the original sample with corrupted data removed). Surprisingly, significant differences are obtained: in particular, considerably higher results are recorded while processing data from the second part of the concert. It is hard to assess that the music being played has a direct influence in this result. It is rather possible that during the second band it was easier for the annotators to decide whether the audience engagement was high or low. In fact, while the first band performed acoustic jazz and played a soft music driven by the sound of a saxophone, the second trio played a more modern jazz, with electric guitars and faster rhythm.

6. DISCUSSION

The aim of this study is to explore the implementation of a machine-learning based algorithm to detect the level of engagement in an audience during a live performance. As described in Section 4, the implementation of a EDA signals-based system involves various steps, from data collection to the final classification. The precision of each stage is interdependent and besides that, each stage can be completed using different techniques. This means that several choices have to be made during the process, and each of them affect the final accuracy.

One main complication encountered in the data collection step is about the annotation process. The output of the algorithm gives a general representation of the average engagement of the entire audience. We decided to evaluate the audience as a whole instead of analysing individual signals. This decision is motivated also by practical reasons. In fact, to evaluate an algorithm that processes signals separately, individual annotations would have been necessary. However, asking participants to report their emotion continually during the performance is very intrusive, and the risk is to collect subjective data affected by distraction.

The annotation method we adopted aims at avoiding this issue. However, it is hard to understand whether the audience is actually engaged or not. This has also been reported by Webb et al. as a result highlighted during interviews with artists [20]. People clapping or dancing are examples of the audience being engaged, but a viewer can also be absorbed in the performance just with his eyes closed. Moreover, the biometric response collected throughout a performance may be completely unaffected by the performance being viewed [10]. The annotators had to pay attention to the general mood perceived in the room. Whether participants were paying attention to the concert or not, does not represent the main focus. Within the 40 participants, it often happened during the concert that they were divided into smaller groups: some of them were interested in the music, others were chatting close to the bar, others were maybe more relaxed. This approach of annotations consists of reporting massive levels of engagement of an audience and might not be fully representative and objective. For example, we noticed that more +1 were annotated, rather than -1. This fact can be explained in different ways: for example, it might be easier to identify when the audience is highly engaged, rather than annotate relaxed or boring moments. Moreover, other noisy factors are the room not particularly illuminated and the fact that people wearing sensors were just a fraction of the entire audience.

Also the processing of physiological signals is not straightforward and needs to address several important challenges. Physiological signals are noisy and vary considerably according to the type of stimulus. Additionally, they also depend on the individual user's physiological and psychological state [18]. There are many personal characteristics that affect the response to a stimulus (e.g. people become habituated to stimuli: repeated stimuli cause a decreasing reaction intensity, past experiences lead people to have different reactions to the same stimulus). Poh et al. also point out other smaller details that can affect the comparison between multiple signals: for example, there are differences in the measurements according to the hand the sensors are placed on (left or right), and there are different sweating mechanisms according to the type of stimulus (physical, cognitive or emotional) [14]. These factors make information extraction from EDA data difficult.

Moreover, in Section 3.2 we outlined the two distinct components of the signal. While the phasic component shows quick changes, the tonic component changes

slowly and has little correlation with the user reactions to stimuli [18]. Various signal decomposition approaches have been proposed to separate the two components, such as the linear convolution model proposed by Bach et al. [1] or the method based on nonnegative deconvolution by Benedek et al. [2]. However, these techniques are limited by computational complexity, for this reason we consider the original one-dimensional signal.

Each computation in the proposed algorithm has been implemented trying to find a trade-off between simplicity and efficiency. For example the number of peaks is obtained by the number of downward concavities analysing the derivatives. A more efficient approach would include overlapping peaks detection, the count of their duration and amplitude.

Therefore, using EDA signals to measure the arousal, some limitations must be considered. Finding the right trade-off between high accuracy and computational feasibility is the core issue.

We tried to deal with the presence of large artefacts and quantization problems, these challenges are even amplified in an uncontrolled setting. In fact, the majority of the experiments in this area is carried out in controlled laboratory settings [14]. On the contrary, our observational measurements were performed during a real public event. We believe that, on one hand, this provides more realistic data. As stated by Wang et al., being part of an audience is a group experience and this important factor is neglected in a laboratory setting, where physiological data is recorded for each participant individually [19]. On the other hand, a real settings involves some complications. First of all, during a controlled experiment it is possible to ask participants to sit quietly for few minutes. In this way the researcher is able to obtain a baseline signal that can be used to reduce the subjectivity factor [14]. Moreover, it is not possible to control the stimuli received by the participants, which makes it harder to analyse their physiological signals. Another difference consists in the type of stimuli. While it is common to analyse significant responses to distinct stimuli, our study deals with continuous input data. During the experiments there were many types of stimuli, some related to the bands playing, others caused by interactions between audience members. This causes unpredictable flows and overlapping peaks in each EDA signal.

In conclusion, although EDA signals are robust physiological signals in emotion recognition, since they reflect emotional arousal directly from the central nervous system. However, EDA signal measurement may be affected by environmental conditions, different types of stimulus sources and placement of the sensors. Although controlled experimental conditions can be applied, there is still room for improvement to reduce these artefacts in real settings [13].

Using more than one physiological measure at a time can lead to improvements [15]. Jerritta et al. show that the classification accuracy seems to depend on the number of physiological signals being measured: for the same dataset, classification accuracy is higher when the features from all the physiological signals are used. In contrast, when only one physiological signal is used, the classification accuracy is lower [8].

Regarding the classification step, the choice of which class best represents each minute during the concert was the main challenge. When the value was higher than +2 or lower than -2 with different annotators reporting the same level of engagement, the choice was actually straightforward. However, ambiguous situations arise otherwise. The task is to associate the most representative label to each minute during the concert, but it is hard to give an interpretation when annotations are contrasting or when during a minute no annotations are reported. This leads to a bad characterization of class 1, because different situations are described by the same label. Improvements should be applied in particular to this step, starting from a more detailed description of the classes, both in terms of quantity and quality.

Moreover, we noticed that the most predicted label is 1, especially for wrong predictions. This might be due to the fact that during the live event the level of engagement of the audience stayed moderate in general. However, it can also be that the annotations create an unbalanced distribution over the classes, leading the classifier to model the classes not appropriately.

The last point we want to outline is the temporal resolution. Ideally the method should give a second-by-second accounting of engagement. However, this configuration might depend on the application. While a continuous result could be useful for television, it might not be the most suitable for artistic performances. In fact, Latulipe et al. outline that live performances (concerts, ballet, theatre, etc.), typically have a narrative structure and the stimulus-response interpretation ignore this [10], since the audience would be labelled as highly engaged only during the fastest movements or “coup de theatre”. Further research could investigate how these results can be interpreted and implemented in a powerful tool for artists.

Only a small number of studies reported in the literature have investigated the potential of physiological signals as implicit measurement for the engagement of multiple users [15]. Our results outline the opportunity to employ a machine-learning approach to differentiate three classes according the level of engagement in the audience. It is not possible to directly compare similar studies, since the physiological signal measured and the number and type of classes are different. The real-time emotion recognition using physiological signals is still in its early stages [15].

7. CONCLUSION

In this paper, the implementation of a physiological signals-based system for measuring the engagement of the audience during a performance has been discussed. Engagement recognition from EDA signals still involves a number of challenges, especially for real-time applications.

Assuming one can collect implicit audience feedback, questions arise around parsing and making sense of the information. The main issues are sensor noise, environmental effects, and the choices in the implementation of a signal processing algorithm. Physiological measurements have the potential to play an important role in the investigation of audience response:

they represent an implicit and objective source of information and the devices to record such data are becoming more and more compact and non-obtrusive [6]. The availability of modern wearable physiological sensors represents the opportunity to investigate physiological signals in order to measure audience engagement during live scenarios [7].

Acknowledgement
Thanks to all the members of the Distributed and Interactive Systems group at CWI for the support given along all the steps of this research project. In particular, thanks to Pablo César for his hospitality and Thomas Röggla for his supervision.

REFERENCES

[1] Bach, D. R., Flandin, G., Friston, K. J., & Dolan, R. J. (2009). Time-series analysis for rapid event-related skin conductance responses. *Journal of neuroscience methods*, *184*(2), 224-234.

[2] Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, *47*(4), 647-658.

[3] Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (Eds.). (2007). *Handbook of psychophysiology*. Cambridge University Press.

[4] Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, *1*(1), 18-37.

[5] Fleureau, J., Guillotel, P., & Huynh-Thu, Q. (2012). Physiological-based affect event detector for entertainment video applications. *Affective Computing, IEEE Transactions on*, *3*(3), 379-385.

[6] Fleureau, J., Guillotel, P., & Orlac, I. (2013, September). Affective benchmarking of movies based on the physiological responses of a real audience. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on,* 73-78. IEEE.

[7] Hernandez, J., Riobo, I., Rozga, A., Abowd, G. D., & Picard, R. W. (2014, September). Using electrodermal activity to recognize ease of engagement in children during social interactions. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 307-317). ACM.

[8] Jerritta, S., Murugappan, M., Nagarajan, R., & Wan, K. (2011, March). Physiological signals based human emotion recognition: a review. In *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on* (pp. 410-415). IEEE.

[9] Lang, P. J. (1995). The emotion probe: studies of motivation and attention. *American psychologist*, *50*(5), 372.

[10] Latulipe, C., Carroll, E. A., & Lottridge, D. (2011, May). Love, hate, arousal and engagement: exploring audience responses to performing arts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1845-1854). ACM.

[11] Leiner, D., Fahr, A., & Früh, H. (2012). EDA positive change: A simple algorithm for electrodermal activity to measure general audience arousal during media exposure. *Communication Methods and Measures*, *6*(4), 237-250.

[12] Picard, R. W., & Picard, R. (1997). *Affective computing,* 252. Cambridge: MIT press.

[13] Ping, H. Y., Abdullah, L. N., Halin, A. A., & Sulaiman, P. S. (2013). A study of physiological signals-based emotion recognition systems. *Int J Comput & Technol*, *11*, 2189-2196.

[14] Poh, M. Z., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *Biomedical Engineering, IEEE Transactions on*, *57*(5), 1243-1252.

[15] Rigas, G., Katsis, C. D., Ganiatsas, G., & Fotiadis, D. I. (2007). A user independent, biosignal based, emotion recognition method. In *User Modeling 2007* (pp. 314-318). Springer Berlin Heidelberg.

[16] Röggla, T., Wang, C., & César, P. S. (2015, October). Analysing Audience Response to Performing Events: A Web Platform for Interactive Exploration of Physiological Sensor Data. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference* (pp. 749-750). ACM.

[17] Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology,* 39, 1161–1178

[18] Silveira, F., Eriksson, B., Sheth, A., & Sheppard, A. (2013, September). Predicting audience responses to movie content from electro-dermal activity signals. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (pp. 707-716). ACM.

[19] Wang, C., Geelhoed, E. N., Stenton, P. P., & Cesar, P. (2014). Sensing a live audience. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems,* 1909-1912. ACM.

[20] Webb, A. M., Wang, C., Kerne, A., & Cesar, P. (2016, February). Distributed Liveness: Understanding How New Technologies Transform Performance Experiences. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 432-437). ACM.

[21] Zong, C., & Chetouani, M. (2009, December). Hilbert-Huang transform based physiological signals analysis for emotion recognition. In *Signal processing and information technology (isspit), 2009 ieee international symposium on* (pp. 334-339). IEEE.