



HAL
open science

An array of physical sensors and an adaptive regression strategy for emotion recognition in a noisy scenario

Francesco Mosciano, Arianna Mencattini, Fabien Ringeval, Björn Schuller,
Eugenio Martinelli, Corrado Di Natale

► To cite this version:

Francesco Mosciano, Arianna Mencattini, Fabien Ringeval, Björn Schuller, Eugenio Martinelli, et al.. An array of physical sensors and an adaptive regression strategy for emotion recognition in a noisy scenario. *Sensors and Actuators A: Physical*, 2017, 267, pp.48-59. 10.1016/j.sna.2017.09.056 . hal-01993393

HAL Id: hal-01993393

<https://hal.science/hal-01993393v1>

Submitted on 24 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MANUSCRIPT

Title: An array of physical sensors and an adaptive regression strategy for emotion recognition in a noisy scenario

*Authors: Francesco Mosciano¹, Arianna Mencattini¹, Fabien Ringeval²,
Björn Schuller^{3,4}, Eugenio Martinelli¹, Corrado Di Natale¹*

- 1. Dept. of Electronic Engineering, University of Rome Tor Vergata, Rome, Italy*
- 2. Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, Grenoble, France*
- 3. Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany*
- 4. Department of Computing, Imperial College London, London, UK.*

Contacting author:

Prof. Eugenio Martinelli

Department of Electronic Engineering, University of Rome Tor Vergata

Via del Politecnico 1; 00133 Roma

Phone: +39 06 72597259

Fax: +39 06 2020519

Email: martinelli@ing.uniroma2.it

30 **Abstract**

31 Several studies demonstrate that since emotions are spontaneously manifested through different
32 measurable quantities (e.g. vocal and facial expressions), this makes possible a sort of automatic estimation
33 of emotion from objective measurements. However, the reliability of such estimations is strongly influenced
34 by the availability of the different sensor modalities used to monitor the affective status of a subject, and
35 furthermore the extraction of objective parameters is sometime thwarted in a noisy and disturbed
36 environment. This paper introduces a personalized emotion estimation based on a heterogeneous array of
37 physical sensors for the measurement of vocal, facial, and physiological (electro-cardiogram and electro-
38 dermal) activities. As a proof of concept, changes in the levels of both emotion reactivity and pleasantness
39 are estimated under critical operative conditions. The estimator model takes advantage from the time-
40 varying selection of the most relevant non-spurious sensors features and the adaptation of the k-nearest
41 neighbour paradigm to the continuous identification of the most affine model templates. The model, once
42 trained, demonstrated to autonomously embed new sensorial input and adapt to unwanted/unpredicted
43 sensor noise or emotion alteration. The proposed approach has been successfully tested on the RECOLA
44 database, a multi-sensorial corpus of spontaneous emotional interactions in French.

45 **Keywords:** Sensor Array; Adaptive regression strategy; Emotion recognition;

46

48 Several studies show that quantitative measurements of human expression can be used to estimate
49 psychological and physical conditions in humans [1,2,3]. Artificial empathic systems are expected to be of
50 benefit in many diverse domains such a precision medicine, personalized care and therapy, customer
51 satisfaction studies, or web profiling, to mention a few [4-6]. According to a commonly accepted model [7],
52 the complexity of emotions may be simplified using two features describing the level of reactivity to
53 stimuli, named *arousal*, and the level of attractiveness/averseness of an emotion, named *valence*.

54 A subject may manifest his/her affective condition using facial mimic, voice alteration, electro-cardiogram
55 deviation from normal status and electro-dermal activities modification due for example to sweating. The
56 increasing interest in modalities that complement to audio/visual is motivated by the growing availability of
57 wearable devices that include physiological sensors, such as electro-cardiogram and electro-dermal activity,
58 at an affordable cost. We indicate such plethora of verbal/nonverbal messages as a multi-sensorial
59 representation of the emotion. Previous studies [8-11] indicate that there can be a strong correlation
60 between alterations in sensor acquisition and in the level of arousal and valence observed over small time
61 intervals of about half a second. If correctly estimated, such relevance can be exploited in order to construct
62 a reliable model able to predict the affective status in the future. The most important implication can be to
63 optimize the therapeutic plans, advertising strategies, or not less fascinating, a cybernetic retelling of the
64 home-sweet-home paradigm. With respect to subject independent emotion estimation, customized emotion
65 prediction adds the simplification of not requiring a large amount of labelled data for training the model and
66 a greatest robustness to an affective content baseline. In fact, personalized estimation can be achieved
67 without the need to normalize the descriptors but rather exploiting data variability within the session in
68 favour of increased prediction performance. At the same time, it may suffer from reduced information about
69 past affective content that can negatively impact on the capability to accurately anticipate the subject's
70 affective status and consequently design prompt actions. Unfortunately, the complexity of the scenario to
71 model, and the diverse sources of variability present, make it very difficult to develop a robust emotion
72 prediction system even in case of customised affective estimation, as graphically illustrated in Fig. 1, where
73 the inter-relations between problems and sources of variability are outlined.

74 In analogy to what happens in very complex sensor systems, this scenario poses many problems originally
75 solved in different engineering fields. Hence, in the following sections, we present and discuss the main
76 sources of variability that should be accounted for in multi-sensorial emotion prediction, providing, at the
77 same time, the original contextualization in sensor networks management. As outlined in [12], no sensor
78 fusion technique is superior to others in all contexts, but there can be an optimal configuration for a specific
79 application. With this in mind, in this work, we intend to implement the C3 paradigm of multi-sensor fusion
80 theory [13], i.e., complementary, competitive, and cooperative. A complementary configuration allows the
81 sensors to not directly depend on each other, but to be combinable in order to give a more complete image
82 of the phenomenon under observation. Different sensors may acquire different cues of the same affective
83 condition. In a competitive architecture, each sensor delivers independent measurements of the same
84 property. Examples of competitive sensors are redundant configuration and fault tolerance that allow the
85 possibility to capture the same general affective condition using a different modality (e.g., negative valence,
86 positive arousal, etc.). A cooperative sensor network uses the information provided by each sensor modality
87 to increase the degree of knowledge of a given property, overcoming individual inaccuracy and uncertainties.
88 To demonstrate the effectiveness of the cooperative fusion approach, we will compare performance of
89 individual sensor modalities in emotion recognition with those obtained by a multi-sensor configuration.

90 The success of such an architecture lies in the hybrid fusion of characteristics that, if properly accounted,
91 ensure robustness, reliability, and effectiveness to the solution. In emotion prediction, complementary
92 requires that different aspects of the same affective condition are measured, and this is achieved through
93 the different sensor modalities. Competitiveness asks for the independence of the different sensor
94 acquisitions, obtained by different hardware, physical principles, as well as acquisition settings. Finally,
95 cooperativeness allows the information acquired from different sensors to be synergistically aggregated to
96 construct a unified model of prediction, as will be demonstrated by a dynamic adaptive procedure. The
97 novelty of our approach lies in the strategy used to implement the C3 paradigm, and in how it is applied to
98 personalized emotion prediction. First of all, each modality of the affective manifestation is acquired through
99 a dedicated sensor device, later indicated as AUDIO, VIDEO, ECG, and EDA. The four sensors acquire a one-
100 dimensional audio speech signal, the video sequence of the upper body of the subject, the ECG signal, and
101 the EDA signal. Sensor spurious data (blue subtree node in Fig. 1) may occasionally or permanently occur due
102 to ill positioning of the sensors, subject movement, power blackout, or transmission interruption, for
103 example. Such situations may cause a modality unavailable or unreliable, and if this issue is not properly
104 accounted for, it may dramatically alter the prediction effectiveness [14-16]. Second, when all the sensor
105 modalities are correctly acquired and all sensors work properly, some specific descriptors extracted from a
106 single modality may result unreliable or missing (green subtree node in Fig. 1). This may be due to unexpected
107 values of the subject's physiological state, or to the spectral content of the speech, or to particular conditions
108 of video acquisition that cause incorrect working of the face landmarking procedure and produce spurious
109 feature values. Such a situation, sometimes addressed using sensor redundancy [17,18] or algorithmic
110 corrections [19], such as cooperative integration strategies [15], may alter the goodness of prediction and
111 bring the module to produce out-of-range predictions.

112 When multiple sensors are embedded into a unique device with the aim to estimate the affective status of
113 the subject, the descriptors extracted can be highly heterogeneous either in distribution over the training
114 population or in the range and average values (headed by the pink subtree node in Fig.1). Sensor fusion
115 techniques then allow to embed in a unique sensorial system all the information related to the subjective
116 externalization of emotions using cooperative strategies able to combine descriptors with different range
117 and relevance levels [1,20].

118 One of the most crucial aspect is whether or not leaving the possibility to the model to integrate the new
119 data along with the corresponding prediction in the training knowledge-base for re-calibration of the entire
120 sensor networks (yellow subtree node in Fig.1). Abandoning the old paradigm of a static configuration, even
121 if it may assure a low computation effort, is a delicate aspect to consider. Reliability of the novel acquired
122 predictions should be accurately accounted for to avoid the risk of confounding the retrained model and
123 producing erroneous future predictions and instability of the dynamically configured system [8].

124 In order to account for all of the outlined problems, in this work, we propose a novel dynamic prediction
125 approach. First, the configuration allows to face descriptors' unreliability, and data missing occurring in new
126 tested instances. Correlation based feature selection methodologies are applied to all the descriptors in the
127 training thus equating the relevance of each modality by the objective correlation metric and selecting the
128 optimal descriptors to build the model. At each prediction, a new model is re-trained by selecting the
129 templates from the training set that are in the nearest neighbourhood of the data in test. Finally, the new
130 data along with the respective output predictions are included in the training set when the output level is
131 comparable with those already acquired to avoid a training confounding effect.

132 The remainder of this article is organized as follows. In Section 2, we will describe the dataset used for the
133 experiments and provide a sketch of the descriptors extracted. Then, in Section 3, we will describe the diverse
134 modules composing the system and illustrate the way they cooperate. Next, in Section 4, we will detail out
135 the experiments run, and in Section 5, we will provide some hints for discussion. Finally, in Section 6, we will
136 draw conclusions.

137 **2. Physical multimodal sensors and data recording**

138 In order to provide experimental results on the novel method proposed for the personalized emotion
139 estimation using multi-sensorial acquisition, we consider as benchmarking database the RECOLA (REmote
140 COLlaborative and Affective interaction) corpus [20], a recently developed multimodal corpus of
141 spontaneous interactions in French. It has been used as benchmarking dataset for the multimodal affect
142 recognition sub-challenge in the last two editions of the Audio/Visual Emotion recognition Challenge
143 (AVEC'15, AVEC'16) [9, 10]. The RECOLA corpus, freely available at <https://diuf.unifr.ch/diva/recola/>, was
144 recorded to study socio-affective behaviours from multimodal data in the context of remote collaborative
145 work, for the development of computer-mediated communication tools. Spontaneous and naturalistic
146 interactions were collected during the course of a collaborative task that was performed remotely in dyads
147 through video conference. Physical multimodal sensor acquisition, i.e., audio (AUDIO), video (VIDEO),
148 electro-cardiogram (ECG) and electro-dermal activity (EDA), were simultaneously recorded from 27 French-
149 speaking subjects. Audio data were captured by unidirectional headset microphones (AKG C520L) and
150 recorded using the *Audacity software* at 44.1 kHz with 16 bits. An external sound card (Lexicon Omega) was
151 used to split the audio data in order to be simultaneously processed by Skype and Audacity. As this solution
152 was not applicable to the video sensors without compromising the frame rate, two HD webcams (Logitech
153 C270) were used for each participant. The first webcam only captured the video data to be used for the Skype
154 video-conference, whereas the second webcam was used to record both audio, from the built-in
155 omnidirectional microphone, and video with the software provided by the manufacturer; audio was recorded
156 at 48 kHz, 16 bits, and brightness auto adjustments were turned off for video recording. Regarding
157 physiological data, the Biopac MP36 unit and the Biopac Student Lab software were used to record both EDA
158 and ECG signals with a sampling frequency of 1 kHz.

159 Even though all subjects speak French fluently, they have different nationalities (i.e., French, Italian or
160 German, cf. below), which thus provides some diversity in the encoding of affect especially in emotion
161 externalization through speech and facial mimic. Out of the 27 speakers, only 18 gave their consent for
162 publication and sharing their data. In order to train and validate a prediction model, time-continuous ratings
163 (40 ms binned frames) of emotional arousal and valence were also acquired by six gender balanced (three
164 women and three men) French-speaking assistants. Since participants showed emotions mostly at the
165 beginning of their interaction, the annotations were collected during the first five minutes of all recordings.
166 In order to collect a unified gold standard from the six annotations available for each speaker, a normalization
167 technique based on the Evaluator Weighted Estimator introduced in [22], that demonstrated a significant
168 improvement ($p < 0.001$ for correlation coefficient using Student t-test) in the inter-rater reliability for both
169 arousal and valence, was applied. For the task of demonstrating the effectiveness of personalized emotion
170 estimation in real-life condition, and according to previous studies [23, 31], we reduced the sampling rate of
171 the signal acquired to 400 ms, leading to sequences with a duration of 750 samples.

172 *Eligibility criteria.* Personalized emotion estimation requires an initial phase of acquisition with provided
173 annotations in both the arousal and valence dimensions, then followed by a session of prediction. To do this,
174 at most an initial 30% of the sequence should be used for training the model and the remaining 70% part for
175 testing. In light of this, we eliminated conversations in which emotions abruptly changed after about the first

176 third of the sequence. Such sequences would cause the inability to achieve reliable performance in testing,
177 since the model would have been tested on affective contents far different from the ones in training,
178 occurring in an unrealizable task of predicting the unknowable instead of the unknown. At the same time,
179 we selected subjects balancing for gender. The present method will be then verified on the conversations of
180 10 speakers whose metadata are reported in Table 1. Out of the speakers, seven ones are French, two ones
181 are Italian and only one is German. The average age is 21.4 years ($\sigma = 2.5$ years). Five speakers are female
182 and five ones are male. In addition, the fifth and sixth columns in Table 1 list the average and the standard
183 deviation values of arousal and valence for each speaker. *Data Descriptors*. Each of the four (AUDIO, VIDEO,
184 ECG, and EDA) provided a signal that was further processed using well-founded feature extractors: for the
185 *acoustic features*, we used the well standardised and broadly used ComParE set of low-level-descriptors,
186 which includes 65 acoustic descriptors with their first order derivate (130 acoustic descriptors) extracted
187 through the openSMILE (Release 2.0) [24] open source software. For the facial action features, 49 landmarks
188 of the face are returned using the Supervised Descent Method (SDM) [25] applied to each frame. According
189 to the Facial Action Coding System (FACS) [26], human facial movements may be taxonomized by their
190 appearance on the face. Firstly, developed by the Swedish anatomist Carl-Herman Hjortsjö, this standard was
191 later adopted by P. Ekman and W.V. Friesen, and published in 1978 [26]. Movements of individual facial
192 muscles are encoded by FACS from slightly different instant changes in facial appearance. Based on the Action
193 Units (AUs) of the subject, which quantify the activity of groups of facial muscles according to the FACS
194 lexicon, it is common to systematically categorize the facial expression of emotions, and it has proven useful
195 to psychologists and to animators. The AUs of the subject were detected starting from high level processing
196 applied to the 49 face landmarks [27,28]. Post processing performed on the AUs led to a set of 40 visual
197 descriptors: numerical details can be found in [29]. Regarding electro-cardiogram descriptors, 28 spectral
198 features are extracted from the ECG signal along with their first order derivatives additionally computed on
199 all features except for two (heart rate and heart rate variability) providing 54 features in total; finally,
200 concerning the EDA signal that reflects a rapid, transient response called skin conductance response (SCR),
201 and a slower, basal drift called skin conductance level (SCL) [21,30,31], 30 features are computed along with
202 the first order derivative for all, providing 60 features in total. The four modalities collected led to a total of
203 284 descriptors for each subject. Beyond the largely proven relevance of audio/visual descriptors for emotion
204 estimation, also physiological signals have demonstrated to correlate with emotion [32,33] – in particular to
205 arousal, despite not being directly perceptible as the way audio-visual data are to humans. Despite
206 controversies that arose about the relation between peripheral physiology and emotions [34,35], autonomic
207 measures have the strong advantage to be easily and continuously monitored using wearable sensors [36-
208 38]. To verify such controversial relevance in our context, Fig. 2 illustrates the average absolute values of the
209 Pearson's Correlation Coefficient (CC) computed over each modality with respect to the arousal (left) and
210 valence (right) dimensions for the 10 speakers. Noteworthy is the fact that there is a large variability over the
211 10 subjects of how each modality impacts on the expected output, thus giving sense to the personalized
212 strategy and to the need of implementing a dynamic selection of the descriptors in the prediction. Moreover,
213 it can be noted that if the AUDIO modality is dominantly correlated with the arousal dimension (see for
214 example speakers P28 and P56), in case of valence all the four modalities balance their importance and
215 support the need for a multimodal strategy.

216

217 3. Dynamic Feature selection and regression strategy

218 In this section, we will describe the innovative architecture presented for the task of multi-sensorial emotion
 219 recognition. The proposed sensor network consists of the five distinct cooperating modules represented in
 220 Fig. 3 each dealing with a specific predictive property. Interactions of the modules are also indicated.

221 1. The Missing Measure Protection (MMP) module aims to protect against missing or unreliable feature
 222 values that can occur in one of the sensors. As a frequent condition, a subset of features acquired by
 223 a single sensor may be missing due to the diverse situations depicted by the scheme in Fig. 1. When
 224 this situation occurs, the system dynamically switches off features that are missing or unreliable
 225 during model construction and prediction. More specifically, the MMP module checks whether
 226 features acquired by a single sensor modality are out of range with respect to the corresponding
 227 training reference values or totally missing. By indicating with x_i the i -th feature in a test sample, and
 228 with y_{i1}, \dots, y_{in} the corresponding n values in the training set, then if $|x_i - \mu_i| > 6\sigma_i$ (condition 1),
 229 where

$$230 \quad \mu_i = \frac{1}{n} \sum_{j=1}^n y_{ij}, \quad (1)$$

$$231 \quad \sigma_i^2 = \frac{1}{n-1} \sum_{j=1}^n (y_{ij} - \mu_i)^2, \quad (2)$$

232 Feature y_i is totally eliminated from the model computation for the prediction of that test sample.
 233

234 The MMP module represents an unsupervised processing block required to maintain sufficiently high
 235 accuracy also in accidental situations such as a power blackout, or a general inability to acquire data (e.g.,
 236 a temporary subject out of focus event). MMP may have the consequence to reduce the number of
 237 features used in the construction of the regression model. The MMP module provides the test input to
 238 the regression model (RM).
 239

240 2. The Dynamic Training Template Recruitment (DTTR) module aims to select, for each data in test, the
 241 nearest measured values in the training set for the construction of the optimal prediction model. The
 242 DTTR module is based on the well-known concept of training templates. When a test data is
 243 processed, the Euclidean distance between its feature vector and those of the training data are
 244 computed, after min-max normalization. The smallest K distances are identified and the
 245 corresponding K training samples are taken as model templates for further model construction. The
 246 module DTTR is an unsupervised block that mimics the K -nearest neighbours (K-NN) pattern
 247 recognition paradigm. It consists in identifying and extracting the K nearest K templates in training
 248 that exhibit the highest affinity with the data in test (provided by the MMP module) in terms of a
 249 given distance metric. Here, K is a fixed parameter selected with the requirement to store at most K
 250 elements at each time provided that the starting training set contains 200 samples that enlarges as
 251 time goes on. In our work, we selected $K=200$ as a trade off between computation burden and
 252 template representativeness.
 253

254 3. The Maximum Relevance Modality Selection (MRMS) module aims to select in the training set the
 255 modalities that best correlate with the expected output to build the predictive model. The MRMS
 256 method selects features that mostly correlate with the output level of arousal and valence
 257 independently. Denoting with y_i the i -th feature in the training set, $i=1, \dots, Q$, being Q the total
 258 number of features provided by the DTTR module, and with z being the expected level of the output
 259 property at a given time (provided by the expert evaluation), then the MRMS criterion finds a subset
 260 of D features $\tilde{y}_i, i = 1, \dots, D \leq Q$, to maximize the following quantity

261
$$MRMS \equiv \max \left(\frac{1}{Q} \sum_{y_i} I(y_i, z) \right), \quad (3)$$

262 where $I(y_i, z)$ represents the mutual information between feature y_i and z defined as

263
$$I(y_i, z) = \sum_{y_i} \sum_{y_k} p(y_i, z) \log \left(\frac{p(y_i, z)}{p(y_i)p(z)} \right), \quad (4)$$

264 where $p(y_i, z)$ is the discrete joint probability distribution (or probability mass function) of variables
 265 y_i and output z , and $p(y_i)$ represents the marginal probability mass function of the variable y_i . The
 266 mutual information between two variables tells us how much one of the two variables can be
 267 statistically explained by the other, hence, in our case, to which extent the output can be explained
 268 by a feature. Note that, the term in eq. (3) maximizes the mutual information between each feature
 269 and the expected output, and it is also called *Relevance Term*. Further details can be found in [1,40].
 270 Feature selection is performed modality by modality and therefore we named it as *maximum*
 271 *relevance modality selection*. As outlined in the Introduction, feature redundancy is not solved at this
 272 step as a strategy to increase system robustness. However, to assure the system works even in
 273 presence of highly correlated features, we decided to implement a regression module based on
 274 partial least square regression. It is well known that this kind of technique solves collinearity problem
 275 since the input space is preliminary projected into a smaller domain with the immediate effect of
 276 reducing such phenomenon.

- 277
- 278 4. The Regression Module (RM) aims at predicting the level of arousal and valence of the new data (test
 279 input from the MMP module). It is built using the template (DTTR) and the features selected in the
 280 training set (MRMS). The Partial Least Square (PLS) regression approach is applied to arousal and
 281 valence prediction, iteratively and adaptively trained according to the selected and reduced set of
 282 features from the training set [42,43]. In our work, we built a separate PLS model for arousal and
 283 valence expected output. The optimal number of latent variables for each regression model is chosen
 284 as the value for which the root mean square error of the validation (RMSECV) computed in the
 285 validation dataset (20 splits, 10 iterations, and mean centring in both input and output variables)
 286 reaches the minimum value [45].45 The corresponding test output values are sent to the dynamic
 287 measure integration (DMI) module.

288

289 The MRMS and RM modules constitute the supervised prediction blocks that use the information about
 290 the expected levels of arousal and valence in the training set to construct the optimal predictive model.
 291 PLS regression has been preferred over other available regression approaches as it accounts for data
 292 collinearity. Feature redundancy, related to the competitive property of sensor fusion in C3 paradigm, is
 293 a crucial concept here with positive effects. In fact, fault tolerance can be supported also by a certain
 294 level of redundancy that may protect from the accidental fault of a single measurement in a given sensor
 295 modality. For this reason, methods that preliminarily reduce redundancy (such as stepwise or sequential
 296 floating forward feature selection), or regression approaches that generally suffer from excessive feature
 297 redundancy (such as ANN) may not have a chance of success in respect of robustness.

- 298 5. The Dynamic Measures Integration (DMI) module aims to selectively enlarge the training set with the
 299 predicted data (test output) to increase the sensor network knowledge-base and improve its
 300 capability to deal with future unknown affective conditions. This semi-supervised module provides a
 301 feedback input to the approach (to the DTTR and the MRMS modules) by dynamically increasing the
 302 training set with the selective inclusion of the predicted test data, aiming to maintain a high accuracy

also if unknown affective conditions are presented as input to the system, especially when the training set did not embrace all the plethora of possible emotions. Of course, to simultaneously maintain the robustness of the system to prediction errors, only data whose level of predicted emotions (arousal and valence separately) is within a certain range with respect to the already predicted content, are included. In particular, indicating with σ_z and μ_z the standard deviation and the mean values of the predicted output z computed up to time t , then the new predicted value \hat{z} at time $t+1$ will be added to the training labels only if

$$|\hat{z} - \mu_z| < 3 \sigma_z . \quad (5)$$

Otherwise, the sample is not included in the knowledge-base (here, three times the standard deviation has been considered as a reliable range). The DTTR module will then consider the additional data for selecting the K-NN templates in case of high affinity with the new test data. The selected constrain is needed to avoid the system has to predict output values out of the range (i.e., the need to extrapolate data). For this reason, the training set has to be selected in order to cover a wide range of output values for arousal and valence. In practical applications, it means that it is impossible to ask the system to predict strongly positive valence values if it has been trained only on negative and low positive values.

4. Experiments

In order to demonstrate the effectiveness of the proposed strategy, we performed the following extensive experiments.

First, in Test 1, the dynamic approach was applied considering all the sensor modalities in a whole and compared with a static configuration (sensor fusion at feature level). Such an experiment allows us to predict the potential improvement of the proposed novel configuration. The prediction capability was assessed through the Concordance Correlation Coefficient (CCC) [44], which combines the Pearson's correlation coefficient (CC) with the square difference between the mean of the two compared time series as follows:

$$CCC(x, y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \quad (6)$$

where ρ is the Pearson correlation coefficient between two time series (e. g., prediction and gold standard) x and y , σ_x^2 , σ_y^2 , μ_x and μ_y are the variances and the mean values of the two time series. Such metric, used as official scoring metric for the AVEC'15 and AVEC'16 challenges, outperforms other kind of standard performance measures (e.g., CC or RMSE) on this task in terms of suitability and meaningfulness, since predictions that are well correlated with the gold standard but shifted in value are penalised in proportion to the deviation while predictions that averagely approach each other but do not correlate, still exhibit low CCC values with respect for example to RMSE [45].

Second, in Test 2, we evaluated the robustness of the proposed system to noisy or unreliable feature values, assuming that also under a sensor *operation status*, some of the measured values cannot be used for the prediction, as highlighted in Fig.1. Single feature unreliability was artificially injected by randomly setting the

343 10% of the descriptors of each modality in test in turn to spurious or missing values for the entire duration,
344 and applying the multimodal prediction approaches (dynamic and static configurations) as in Test 1.

345 All the experiments were conducted on the 10 subjects selected. Each feature sequence was originally
346 reduced at the beginning and at the end of an amount of samples needed to eliminate N/A descriptors in the
347 video modality probably arisen from excessive movements of the subject at the beginning and at end of the
348 task. The resulting sequences included 671 samples and had a duration of 268.4 s. Then, for each subject, we
349 considered a training set represented by the initial 200 samples of each modality with the corresponding
350 annotations (80 s), and the remaining 471 samples (188.4 s) for testing.

351 5. Results

352 In this section, we provide numerical results of the two tests performed.

353
354 *Test 1.* The *dynamic* approach was applied considering all the sensor modalities in a whole and compared
355 with a *static configuration*, i.e., a PLS regression module trained on the same original training set of the
356 *dynamic* configuration with the addition of the MRMS feature selection module. Test 1 demonstrated that,
357 the multimodal configuration allows to obtain a more reliable emotion prediction, being more robust to the
358 intra-subject variability occurring during the predictions of the emotion thanks to a higher level of adaptation.
359 Median (std) CCC values of 0.62 (0.12) and 0.60 (0.20) for the arousal (dynamic and static configuration) and
360 of 0.30 (0.26) and 0.05 (0.25) for the valence (dynamic and static), respectively, have been obtained. In
361 addition, Fig. 4 illustrates the boxplots of the CCC values obtained. A paired t-test was run to estimate the
362 statistical significance of the achieved improvement. It can be noted that, the best improvement was
363 obtained in the valence estimation which is the most difficult affective dimension to estimate according to
364 the literature [1,2,3,9,10,24,45]. In Fig. 5, we also showed the result of the arousal (top) and valence (bottom)
365 prediction obtained using the dynamic configuration for the speaker P16. The blue line identifies the
366 prediction achieved, the red line represents the expected affective dimension, and the green squares locate
367 the training session. The corresponding CCC values obtained are also indicated.

368 In order to compare the performance of the multimodal configuration with those of each single modality, in
369 Table 2 we listed the median CCC values along with the corresponding standard deviation values of the single
370 modalities. Note that the multimodal dynamic configuration outperforms each of the single modalities using
371 the same dynamic configuration in case of arousal prediction, and strongly improves the results in case of
372 valence prediction.

373 *Test 2.* Single feature unreliability has been simulated by randomly setting 10 % of the descriptors of each
374 modality to spurious or missing values for the entire duration and applying the multimodal prediction
375 approaches (dynamic and static configurations) as before in Test 1. Test 2 aims at proving the effectiveness
376 of the proposed approach to protect the system against random (occasional) descriptor unreliability due, for
377 example, to errors in the post-processing, in the storage, or in the transmission of data. Of course, if
378 correlated features are simultaneously missing and those features are all significant for the prediction, the
379 system performance can heavily degrade. On the other hand, we injected missing data only on the 10% of
380 descriptors for each modality hence avoiding, on average, to fall in such critical scenario.

381

382 Figure 6 shows the boxplots of the CCC values of the proposed strategy for arousal (upper) and valence
383 (lower) prediction. The proposed dynamic multimodal configuration (Dyn Mul) is compared against the static
384 multimodal configuration (Stat Mult) as in Test 1.

385 From the observation of the results, it is evident that the static configuration is critically influenced by
386 spurious values injected in the 10 % of the feature values of a single modality. This is due to the fact that the
387 static configuration selects the optimal features only once at the beginning of the procedure. Hence, if a
388 descriptor is selected and then it assumes spurious values during test, all the predicted output sequence
389 results in failure. On the contrary, the proposed approach, thanks to the dynamic selection of the features
390 and the outlier elimination procedure forwarded by the data acquired in test, eliminates the features that
391 exhibit spurious values in the test and protects the prediction from failures.

392

393 **6. Discussion**

394 In this study, we presented a novel multi-sensor platform and fault-tolerance processing strategies able to
395 integrate multimodal emotion data collection, and robustness to occasional data unreliability during test.
396 Moreover, the system is able to incorporate the new input data along with the corresponding estimated
397 affective content, thanks to a dedicated architecture able to protect the system from drift and instability of
398 the future predictions. The proposed architecture satisfies the C3 paradigm being complementary,
399 competitive, and cooperative as demonstrated in the Tests 1 and 2.

400 An interesting part of this study was to investigate the distribution over the four sensor modalities of the
401 features selected at each frame for the prediction of arousal and valence. Figure 7 shows a stacked bar graph
402 in which for each speaker the length of each coloured bar represents the relative frequency of selection of
403 the modality. Modalities are identified by colours. Noteworthy is the fact that, if for arousal the audio
404 modality is almost always the most selected one, for the valence dimension the frequency of selection is
405 much more distributed over the modalities. This fact demonstrates once more the usefulness of the
406 multimodal approach to account for the inter-subject variability. Moreover, the diverse distribution over the
407 speakers in the valence prediction allows to assign equal importance to the four modalities and in general
408 reduces the superiority of the audio modality that is manifested instead in the arousal prediction.

409 One of the most interesting properties of the system presented here is the ability to cope with the
410 unexpected lack of data (an entire sensor modality or a subset of feature values). To further demonstrate
411 this ability, we ran a dedicated experiment with the aim to simulate a progressive, not instant, sensor missing.
412 We artificially set to missing values an increasing percentage of feature values for the same modality (video
413 in the experiment), until one obtains the entire fault of that sensor modality. The simulated scenario is
414 illustrated in Fig. 8 for the valence dimension. The progressive degradation of sensor performance is
415 graphically illustrated by the insets shown over the prediction at three time instants (80 s, 160 s, and 268 s).
416 Coloured bars represent the four modalities as indicated in the legend. The bottom bars represent the
417 percentage of non-spurious features for each modality and shows that features in the video modality
418 progressively move entirely towards fault. The top bars indicate the corresponding percentage of features
419 selected in each modality. Note that, a subset of features from the video modality is still chosen by the
420 algorithm at the beginning of the prediction, but the percentage of selection goes down to zero as long as
421 the percentage of spurious video features increases (268 s).

422 As a final evidence of the effectiveness of the dynamic multimodal configuration proposed in this work, we
423 staged a more critical scenario, in which an entire modality in turn is missed by assuming that one of the four
424 input signals is not available during the entire duration of test. The comparative approaches are again the
425 dynamic and the static multimodal architectures. A total of four different simulation results are collected,
426 one for each missing modality. Such a test aims at verifying the effectiveness of the proposed strategy in a
427 fault risk scenario, where one modality of the test sequence can be totally off-line. Figure 9 shows the CCC
428 values of arousal (upper), and valence (lower) predictions, when the dynamic configuration is compared to
429 the static solution. Note that, in this case, especially when the audio modality is totally missing, prediction
430 performance drastically degrades, demonstrating once more the crucial role of the audio modality in emotion
431 recognition, especially for arousal.

432 As a further important aspect, the proposed system has been specifically developed for real time applications
433 when, after an initial training phase, the system is able to predict one sample at a time, in less than 400ms
434 that is the sampling time chosen in this study. All the techniques implemented in the approach have been
435 selected since they require very small computational effort and can take advantage of matrix calculus (PLS,
436 Euclidean distance among vectors, MRMS, etc.). Simulations were performed using Matlab 2017®, on an
437 Intel Core i7 machine. Of course, for longer prediction session, a restricted number of training samples have
438 to be chosen by definitely eliminating oldest training samples hence avoiding to calculate Euclidean distance
439 over an even larger training set.

440 To conclude, it is important to highlight the potential application of the proposed strategy to diversified
441 scenarios, such as chemical sensors, or signal processing in general. The approach presents an architecture
442 that can be easily adapted to alternative frameworks with the aim to protect systems from failure and out-
443 of-range predictions. Moreover, even under the general assumption that there is not any superior
444 configuration for sensor fusion, the C3 paradigm assumption is a fundamental step toward better
445 performance in a large plethora of scenarios.

446

447 **7. Conclusion**

448 Despite the increasing interest, human emotion remains a complex, dynamically changing scenario that
449 expresses itself in diversified human spheres. Therefore, a smart sensor network architecture is required for
450 reliable prediction. In this paper, we presented a novel dynamic sensor network configuration that receives
451 the emotion events communicated through the speech, the facial mimic, and physiological signals of a
452 subject acquired while he/she is involved in natural conversation in a controlled environment. The shown
453 system is able to continuously predict the two dimensions of affect (arousal and valence) under normal and
454 simulated critical working conditions as demonstrated by specific experiments run. Although adequate
455 performance was obtained in the experimental tests, more investigations are required in order to improve
456 robustness, especially with the prediction of the emotional valence, that still appears as the most critical
457 affective dimension. This is crucial especially in those applications where positive, neutral, and negative
458 valence levels are used to calibrate the results of behavioural disorder studies conducted such as on children.
459 To further verify the robustness of the prediction system, future work will also account for the presence of
460 noise sources (wind noise, phone noise, car and train noises, etc.) investigating strategies for protecting the
461 system from both soft and hard failures in real-life operation modes.

462

463 **Acknowledgment**

464 The research leading to these results has been partially funded by the PainTCare project (University of Rome
465 Tor Vergata, Uncovering Excellence program), the European Commission's Seventh Framework Programme
466 through the ERC Starting Grant No. 338164 (iHEARu), and the European Union's Horizon 2020 Programme
467 through the Innovative Action No. 645094 (SEWA) and the Research Innovative Action No. 645378 (ARIA-
468 VALUSPA).

469

470 **References**

- 471 [1] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, C. Di Natale, Continuous Estimation of Emotions
472 in Speech by Dynamic Cooperative Speaker Models, *IEEE Transactions on Affective Computing*, pages
473 14,2016, doi: 10.1109/TAFFC.2016.2531664.
- 474 [2] E. Martinelli, A. Mencattini, E. Daprat, C. Di Natale, Strength Is in Numbers: Can Concordant Artificial
475 Listeners Improve Prediction of Emotion from Speech?, *PLoS one*, 11, 8, pp. e0161752, 2016.
- 476 [3] F. Wengler, M. Wöllmer, B. Schuller, Emotion Recognition in Naturalistic Speech and Language—A
477 Survey. *Emotion Recognition: A Pattern Analysis Approach* 237-267, 2015.
- 478 [4] M. Mansoorizadeh and N. M. Charkari, *Multimodal information fusion application to human emotion*
479 *recognition from face and speech*, *Multimedia Tools and Applications* 49,2, pp. 277-297, 2010.
- 480 [5] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their
481 applications, *Affective Computing, IEEE Transactions on*, 1, 1, pp.18-37, 2010.
- 482 [6] N. Sebe, I. Cohen. T. Gevers, T.S. Huang, Multimodal approaches for emotion recognition: a survey.
483 *In Electronic Imaging*, pp. 56-67, International Society for Optics and Photonics 2005.
- 484 [7] J. Posner, J. Russell, and B. Peterson, "The circumplex model of affect: an integrative approach to
485 affective neuroscience, cognitive development, and psychopathology," *Develop. Psychopathol.*, 17,
486 3, pp. 715-734, 2005.
- 487 [8] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, Enhanced semi-supervised learning for
488 multimodal emotion recognition. In *2016 IEEE International Conference on Acoustics, Speech and*
489 *Signal Processing (ICASSP)* IEEE, pp. 5185-5189.
- 490 [9] F. Ringeval et al. AV+ EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video,
491 and Physiological Data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion*
492 *Challenge*, ACM, pp. 3-8, 2015.
- 493 [10] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie,
494 and M. Pantic. AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge.
495 *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC'16*, co-
496 *located with the 24th ACM International Conference on Multimedia, MM 2016*, pages 3-10,
497 Amsterdam, The Netherlands, October 2016.
- 498 [11] A. Mencattini, F. Ringeval, B. Schuller, E. Martinelli, C. Di Natale, Continuous Monitoring of Emotions
499 by a Multimodal Cooperative Sensor System. *Procedia Engineering*, 120, pp. 556-559, 2015.
- 500 [12] K.M. Tsang, W.L. Chan, Data validation of intelligent sensor using predictive filters and fuzzy logic,
501 *Sensors and Actuators A: Physical*, 159, 2, pp. 149-156, 2010.
- 502 [13] W. Elmenreich, An introduction to sensor fusion. Vienna University of Technology, Austria (2002).
- 503 [14] E. Martinelli, G. Magna, S. De Vito, R. Di Fuccio, G. Di Francia, A. Vergara, C. Di Natale, An adaptive
504 classification model based on the Artificial Immune System for chemical sensor drift mitigation.
505 *Sensors and Actuators B: Chemical* 177, pp. 1017-1026, 2013.
- 506 [15] E. Martinelli, G. Magna, A. Vergara, C. Di Natale, Cooperative classifiers for reconfigurable sensor
507 arrays. *Sensors and Actuators B: Chemical*, 199, pp. 83-92, 2014.
- 508 [16] G. Magna, F. Mosciano, E. Martinelli, C. Di Natale, An on-line reconfigurable classification algorithm
509 improves the long-term stability of gas sensor arrays in case of faulty and drifting sensors. *Procedia*
510 *Engineering*, 120, pp. 249-252, 2015.
- 511 [17] J. Yao, H. Zhang, X. Xiang, H. Bai, Y. Zhao, A 3-D printed redundant six-component force sensor with
512 eight parallel limbs. *Sensors and Actuators A: Physical*, 247, pp. 90-97, 2016.
- 513 [18] L. Fernandez, S. Marco, A. Gutierrez-Galvez, Robustness to sensor damage of a highly redundant gas
514 sensor array, *Sensors and Actuators, B: Chemical*, 218, pp. 296-302.

- 515 [19] J. Fonollosa, A. Vergara, R. Huerta, Algorithmic mitigation of sensor failure: Is sensor replacement
516 really necessary? *Sensors and Actuators, B: Chemical*, 183, pp. 211-221, 2013.
- 517 [20] W.-T. Sung, and K.-Y. Chang, Evidence-based multi-sensor information fusion for remote health care
518 systems, *Sensors and Actuators A: Physical*, 204, pp. 1-19, 2013.
- 519 [21] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the RECOLA multimodal corpus of
520 remote collaborative and affective interactions. Automatic Face and Gesture Recognition (FG), 2013
521 *10th IEEE International Conference and Workshops on. IEEE*, pp. 1-8, 2013.
- 522 [22] M. Grimm, K. Kroschel, S. Narayanan, The Vera am Mittag German audio-visual emotional speech
523 database, in: *Proc. of the 8th Inter. Conf. on Multimedia & Expo (ICME)*, Hannover, Germany. pp. 865–
524 868, 2008.
- 525 [23] W.J. Yan, Q. Wu, J. Liang, Y.H. Chen, W. Fu, How fast are the leaked facial expressions: The duration
526 of micro-expressions. *Journal of Nonverbal Behavior* 37, pp. 217–230, 2013.
- 527 [24] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F.
528 Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim. The INTERSPEECH
529 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In
530 Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech
531 Communication Association, pages 148–152, Lyon, France, August 2013. ISCA. Xiong and F. De la
532 Torre. Supervised descent method and its applications to face alignment. *In Proc. of CVPR*, pp. 532–
533 539, Portland (OR), USA, 2013. IEEE.
- 534 [25] P. Ekman, W.V. Friesen, *Facial action coding system: A technique for the measurement of facial*
535 *movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.
- 536 [26] D. Lowe, Distinctive image features from scale-invariant keypoints, *Intern. Journal of Computer*
537 *Vision*, 60, pp. 91–110, 2004.
- 538 [27] J. Xiao, T. Moriyama, T. Kanade, J.F. Cohn, Robust full-motion recovery of head by dynamic templates
539 and re-registration techniques. *Inter. Journal of Imaging Systems and Technology*, 13, pp. 85–94,
540 2003.
- 541 [28] G. Farneback, Two-frame motion estimation based on polynomial expansion, in: *Image Analysis,*
542 *Special Issue, 13th Scandinavian Conf. (SCIA). Springer*, Halmstad, Sweden. volume 2749, pp. 363–
543 370, 2003.
- 544 [29] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.P. Thiran, T. Ebrahimi, D. Lalanne, B. Schuller Prediction of
545 asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern*
546 *Recognition Letters*, 66, pp. 22-30, 2015.
- 547 [30] M. Dawson, A. Schell, D. Filion, *The electrodermal system*, in: Cacioppo, J.T., Tassinari, L.G.,
548 Berntson, G.G. (Eds.), *Handbook of psychophysiology*. Cambridge: Cambridge University Press.
549 volume 2, pp. 200–223, 2007.
- 550 [31] R. B. Knapp, J. Kim, and E. André. *Physiological signals and their use in augmenting emotion*
551 *recognition for human-machine interaction*. In *Emotion-Oriented Systems – The Humaine Handbook*,
552 pp. 133–159. Springer Berlin Heidelberg, 2011.
- 553 [32] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, and A. N. I. Patras. DEAP:
554 A database for emotion analysis using physiological signals. *IEEE Transactions on Affective*
555 *Computing*, 3, pp. 18–31, 2012.
- 556 [33] S. Schachter. *Cognition and peripheralist-centralist controversies in motivation and emotion*. In M. S.
557 Gazzaniga, editor, *Handbook of Psychobiology*, pp. 529–564. Academic Press Inc., 2012.
- 558 [34] D. Keltner and J. S. Lerner. *Emotion*. In S. Fiske, D. Gilbert, and G. Lindzey, editors, *Handbook of Social*
559 *Psychology*, 1, pp. 317–331. John Wiley & Sons Inc., 5th edition, 2010.

- 560 [35] A. Sanoa, R. W. Picard, and R. Stickgold. Quantitative analysis of wrist electrodermal activity during
561 sleep, *International Journal of Psychophysiology*, 94, 3, pp. 382–389, 2014.
- 562 [36] R. Picard. Affective media and wearables: surprising findings. *In Proc. of ACM MM*, pp. 3–4, Orlando
563 (FL), USA, 2014. ACM.
- 564 [37] M. Chen, Y. Zhang, M. M. H. Yong Li, and A. Alarmi. AIWAC: Affective interaction through wearable
565 computing and cloud technology, *IEEE Mobile Wearable Communications*, 22, 1, pp. 20–27, 2015.
- 566 [38] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency,
567 max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (8) (2005) 1226–1238.
- 568 [39] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemometr. Intell.*
569 *Lab. Syst.* 58 (2) (2001) 109–130.
- 570 [40] V. Vinzi, L. Trinchera, S. Amato, *PLS path modeling: from foundations to recent developments and*
571 *open issues for model assessment and improvement*, in: *Handbook of Partial Least Squares*, Springer,
572 2010, pp. 47–82.
- 573 [41] X.-S. Gan, J.-S. Duanmu, J.-F. Wang, W. Cong, Anomaly intrusion detection based on PLS feature
574 extraction and core vector machine, *Knowl.-Based Syst.* 40 (2013) 1–6.
- 575 [42] S.-S. Chen, Y.-W. Chuang, P.-Y. Chen, Behavioral intention formation in knowledge sharing: examining
576 the roles of KMS quality, KMS self-efficacy, and organizational climate, *Knowl.-Based Syst.* 31 (2012)
577 106–118.
- 578 [43] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, *Chemometr. Intell.*
579 *Lab. Syst.* 18 (3) (1993) 251–263.
- 580 [44] I.-K. L. Lin, A concordance correlation coefficient to evaluate reproducibility, *Biometrics*, 45, 1, pp.
581 255–268, 1989.
- 582 [45] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller. Discriminatively Trained Recurrent Neural
583 Networks for Continuous Dimensional Emotion Recognition from Audio. In *Proceedings of the 25th*
584 *International Joint Conference on Artificial Intelligence (IJCAI 2016)*, pp. 2196–2202, New York City
585 (NY), USA, July 2016. IJCAI/AAAI.
- 586

587

588

589

590

591

592

593

594

595

596

597

598 **Tables Captions:**

599 **Table 1.** Metadata of the 10 speakers used in this work.

600 **Table 2.** CCC median and median absolute values for the arousal and valence prediction using the dynamic
601 single modality configuration: multi (multimodal).

602

603

604

605

606

607

608

Table 1.

Speaker Label	Age	Sex	Mother tongue	Arousal ($\mu \pm \sigma$)	Valence ($\mu \pm \sigma$)
P16	21	M	French	0.03 \pm 0.17	0.13 \pm 0.10
P19	20	F	French	0.09 \pm 0.15	0.12 \pm 0.13
P21	19	F	Italian	0.03 \pm 0.15	0.11 \pm 0.11
P28	18	F	French	-0.09 \pm 0.14	0.09 \pm 0.12
P30	22	F	Italian	-0.15 \pm 0.19	0.03 \pm 0.14
P34	25	M	French	-0.19 \pm 0.17	-0.02 \pm 0.05
P41	23	M	German	0.15 \pm 0.10	0.11 \pm 0.09
P45	19	F	French	-0.02 \pm 0.19	0.08 \pm 0.09
P56	22	M	French	0.07 \pm 0.16	0.12 \pm 0.11
P64	25	M	French	0.01 \pm 0.14	0.08 \pm 0.07

609

610

611

Table 2.

	Arousal					Valence				
	multi	audio	video	ecg	eda	multi	audio	Video	ecg	eda
<i>median CCC</i>	0.622	0.587	0.117	0.098	-0.005	0.303	0.093	-0.022	0.028	0.050
<i>std. dev.</i>	0.121	0.111	0.132	0.119	0.063	0.266	0.166	0.259	0.092	0.145

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637 **Figure Captions:**

638 **Figure 1.** Graphical representation of the problems to account for in a personalized multimodal mood estimation
639 framework.

640 **Figure 2.** Distribution over the 10 speakers (denoted with P16 - P64) of the average absolute correlation coefficient (CC)
641 of the features in each modality (distinguished by colours) with respect to the arousal (top) and valence dimension
642 (bottom).

643 **Figure 3.** A schematic illustration of the proposed method. The missing measure protection (MMP) module identifies
644 features/sensor modality that are spurious or missing in the test input and eliminate them from the training set. The
645 dynamic training template recruitment (DTTR) module extracts from the reduced training set the templates most affine
646 to the reduced test input and sends them to the feature selection module based on maximum relevance modality
647 selection (MRMS) criterion. Selected training templates are then used to construct the regression module (RM) that
648 assigns a predicted output to the test input (test output). Finally, the dynamic measure integration (DMI) module aims
649 at including the predicted output into the training set to enlarge the knowledge-base.

650 **Figure 4.** Boxplot of the concordance correlation coefficients (CCC) computed for the dynamic multimodal configuration
651 of the arousal (first column) and valence (third column) prediction against the static multimodal configuration for
652 arousal (second column) and valence (fourth column) (Test 1). The p-values of the improvement quantified by running
653 paired t-test on the CCC values are also reported. ***Dyn Mult Arousal*** and ***Dyn Mult Valence*** indicate the proposed
654 dynamic multimodal prediction of arousal and valence, respectively. ***Stat Mult Arousal*** and ***Stat Mult Valence*** indicate
655 the standard static multimodal prediction of arousal and valence, respectively.

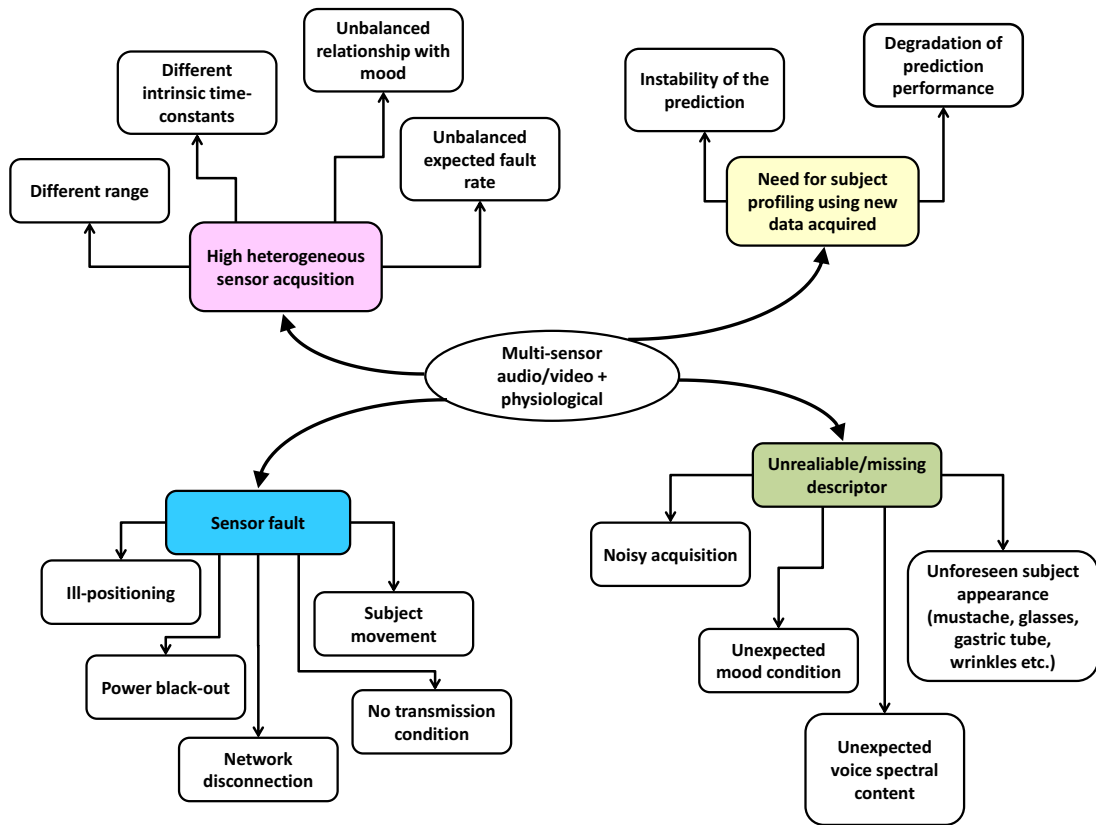
656 **Figure 5.** Two examples of arousal (top) and valence (bottom) prediction for subject P16. The blue line indicates the
657 achieved prediction, the red line represents the expected output, and the green squares locate the training session. The
658 corresponding CCC values are also reported.

659 **Figure 6.** Boxplot of the concordance correlation coefficients (CCC) computed for dynamic multimodal configuration of
660 the arousal (upper) and valence (lower) prediction against the static multimodal configuration in case of 10% of features
661 for each modality in turn randomly set to missing or spurious values (Test 2). ***Dyn Mult Arousal*** and ***Dyn Mult Valence***
662 indicate the proposed dynamic multimodal prediction of arousal and valence, respectively. ***Stat Mult Arousal*** and ***Stat***
663 ***Mult Valence*** indicate the standard static multimodal prediction of arousal and valence, respectively.

664 **Figure 7.** Relative frequency of selection of features of each modality during prediction of arousal (left) and valence
665 (right) in the dynamic configuration.

666 **Figure 8.** Progressive spurious feature injection during valence prediction in the video modality. Three time instants are
667 shown (80s, 160s, 268s). Bar colours indicate the four modalities as explained in the legend. Bottom bars represent the
668 percentage of non-spurious features for each modality. Top bars indicate the percentage of selected features for each
669 modality.

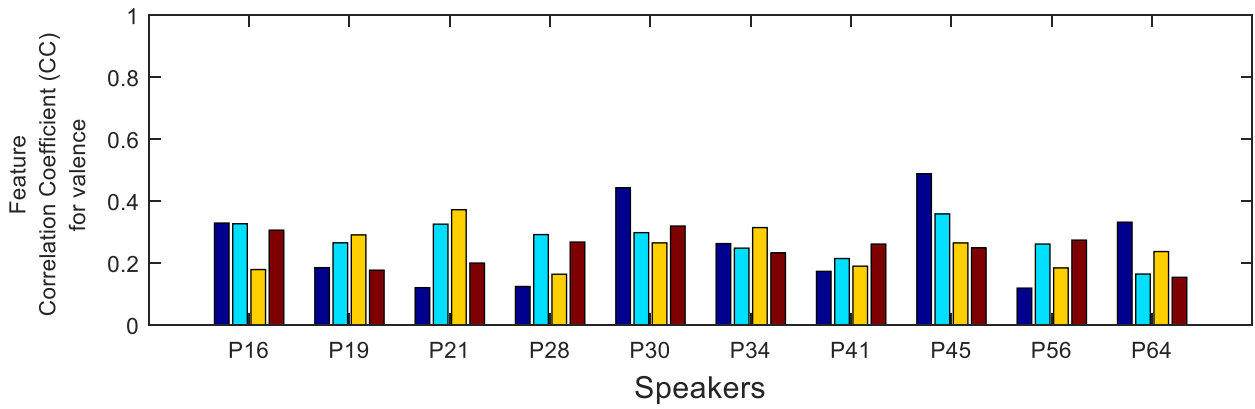
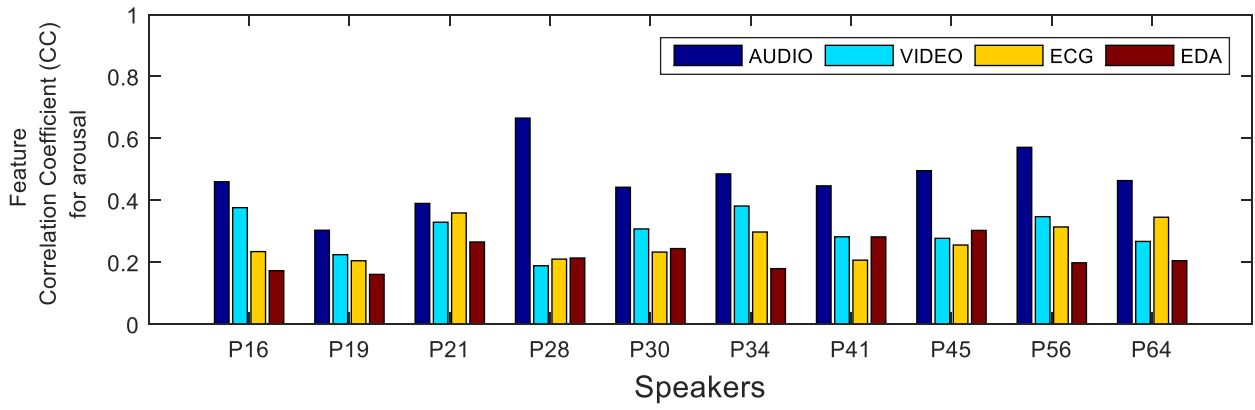
670 **Figure 9.** Boxplot of the concordance correlation coefficients (CCC) computed for dynamic multimodal configuration of
671 the arousal (upper) and valence (lower) prediction against the static multimodal configuration in case of modality
672 missing in turn. ***Dyn Mult Arousal*** and ***Dyn Mult Valence*** indicate the proposed dynamic multimodal prediction of
673 arousal and valence, respectively. ***Stat Mult Arousal*** and ***Stat Mult Valence*** indicate the standard static multimodal
674 prediction of arousal and valence, respectively.



675

676 **Figure 1.**

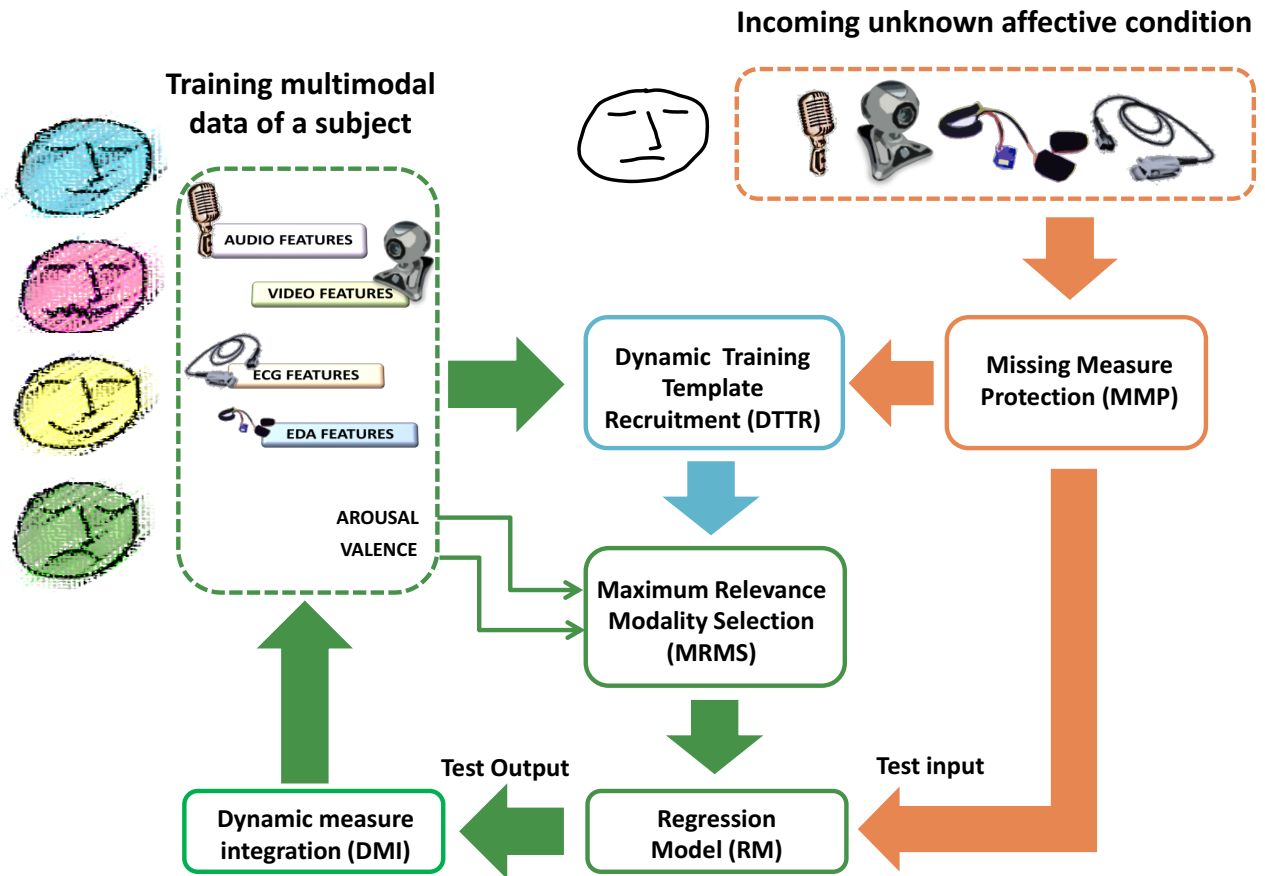
677



678

679 **Figure 2.**

680

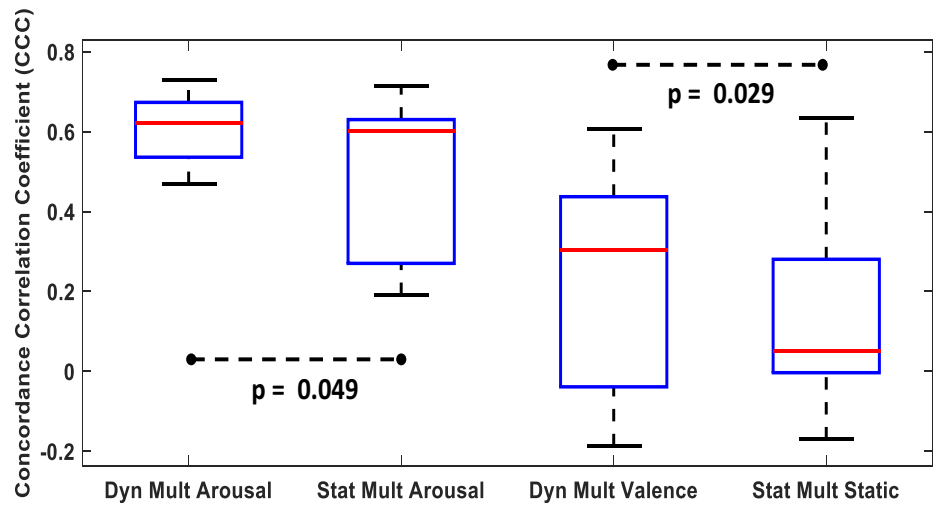


681

682

683 Figure 3.

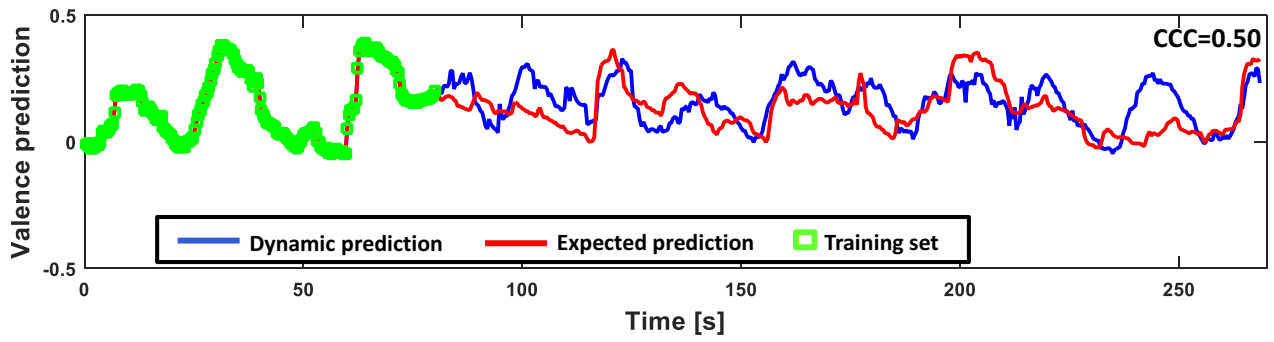
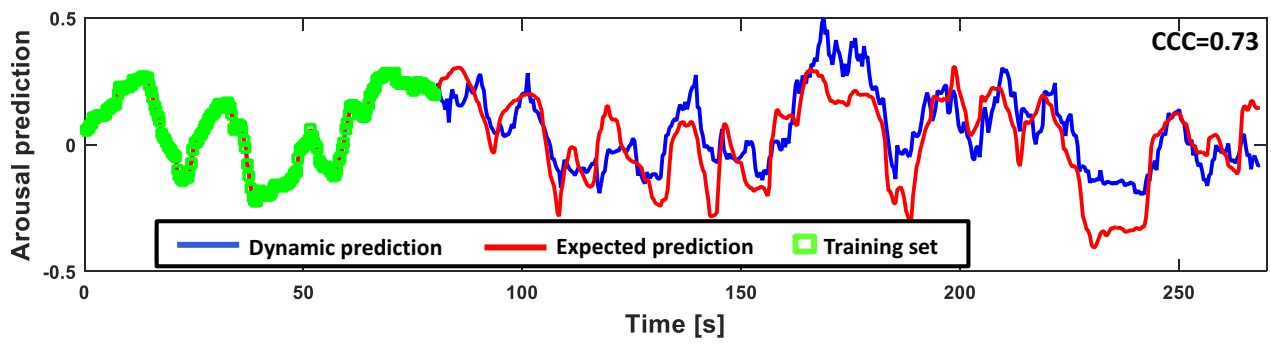
684



685

686 **Figure 4.**

687



688

689 Figure 5.

690

691

692

693

694

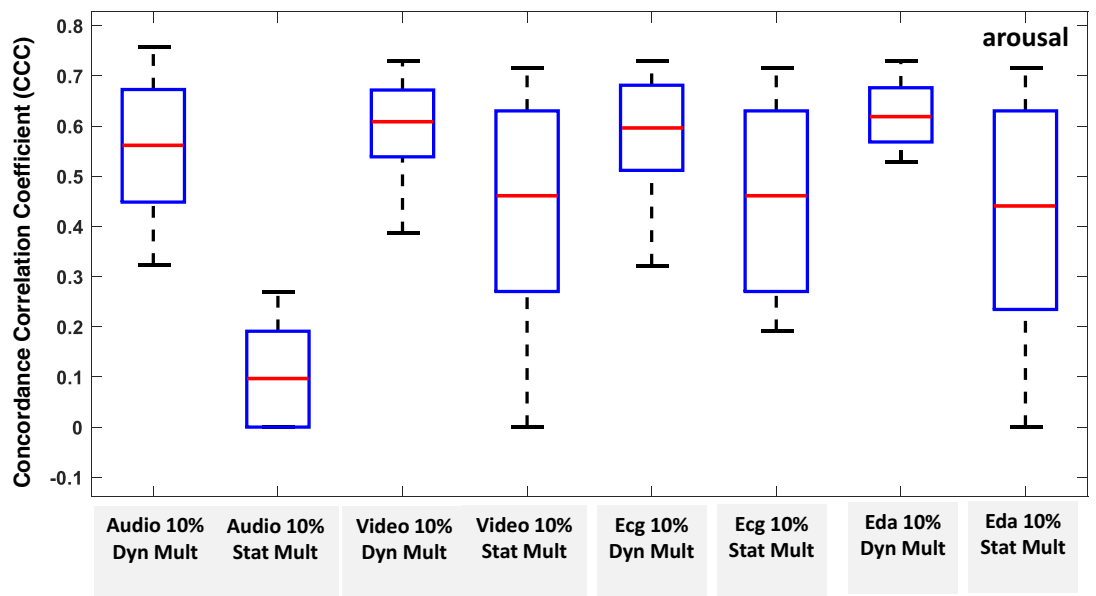
695

696

697

698

699



700

701

702

703

704

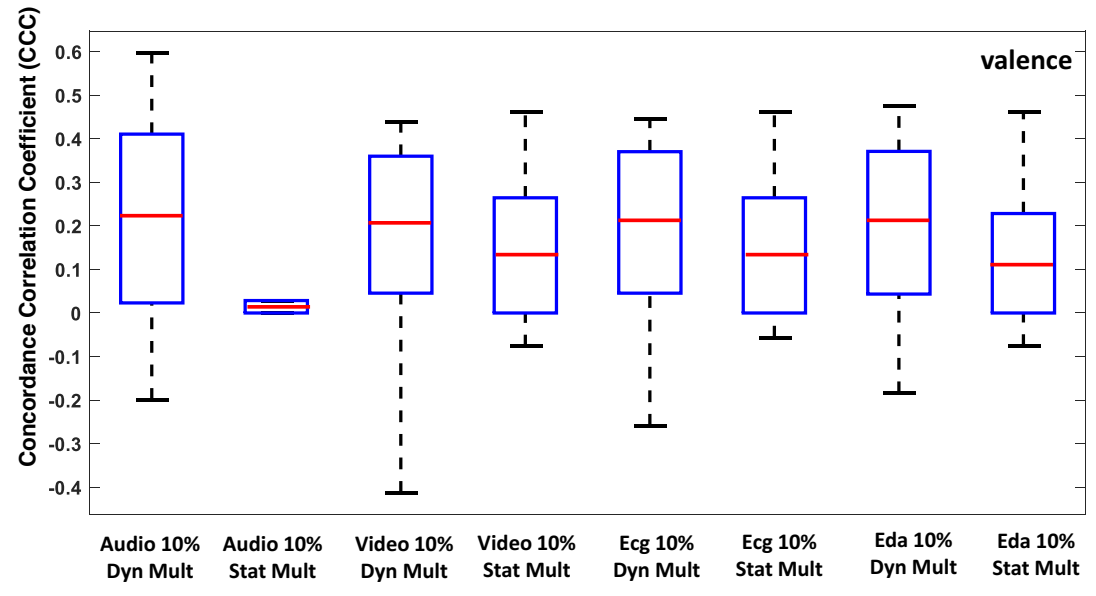
705

706

707

708

709



710

711

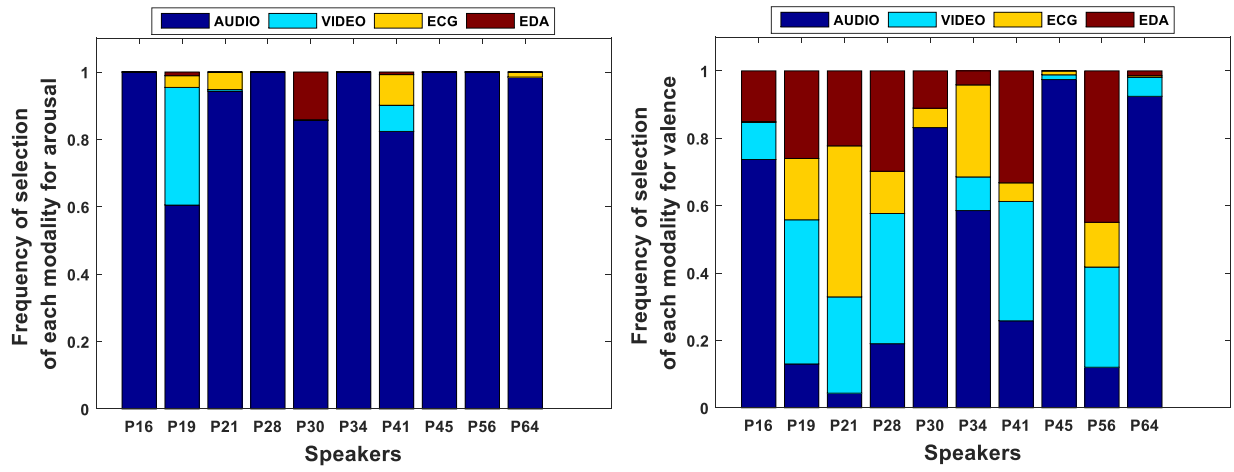
712 **Figure 6.**

713

714

715

716



717

718 **Figure 7.**

719

720

721

722

723

724

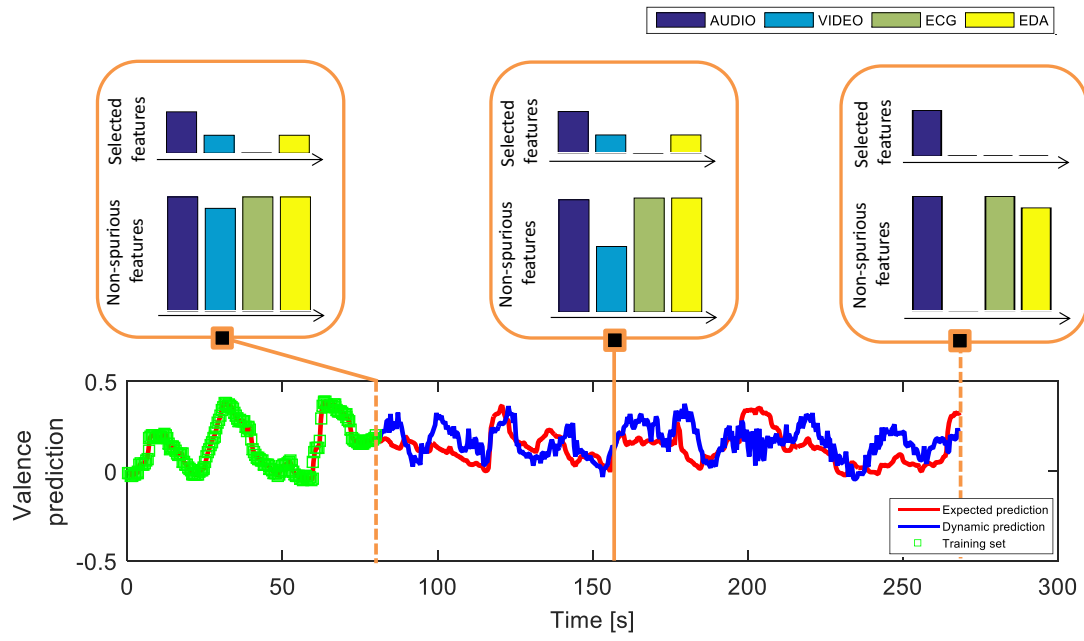
725

726

727

728

729



730 **Figure 8.**

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

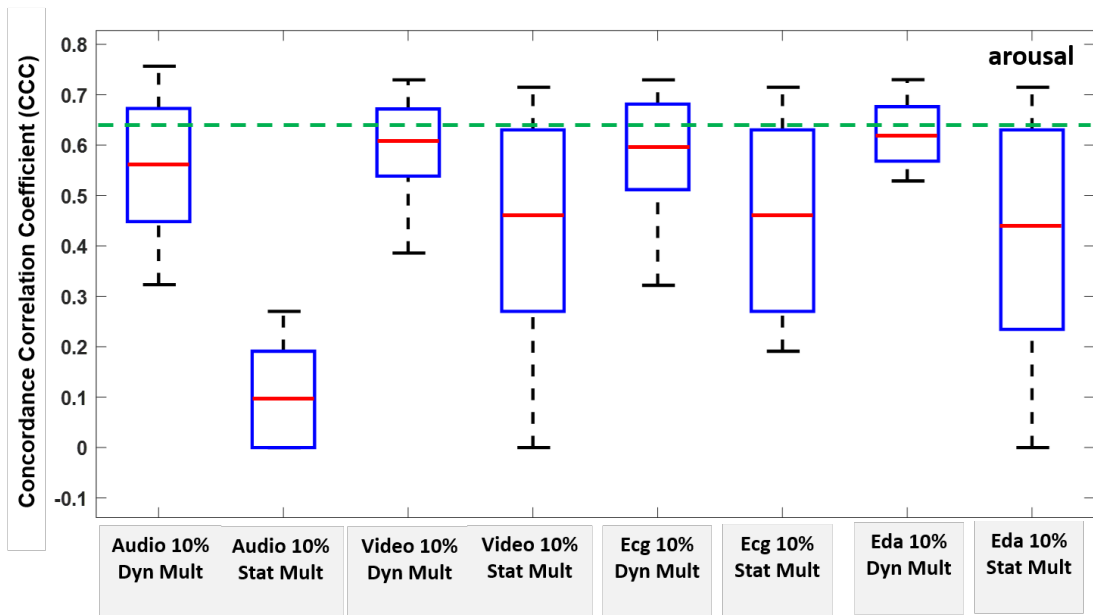
746

747

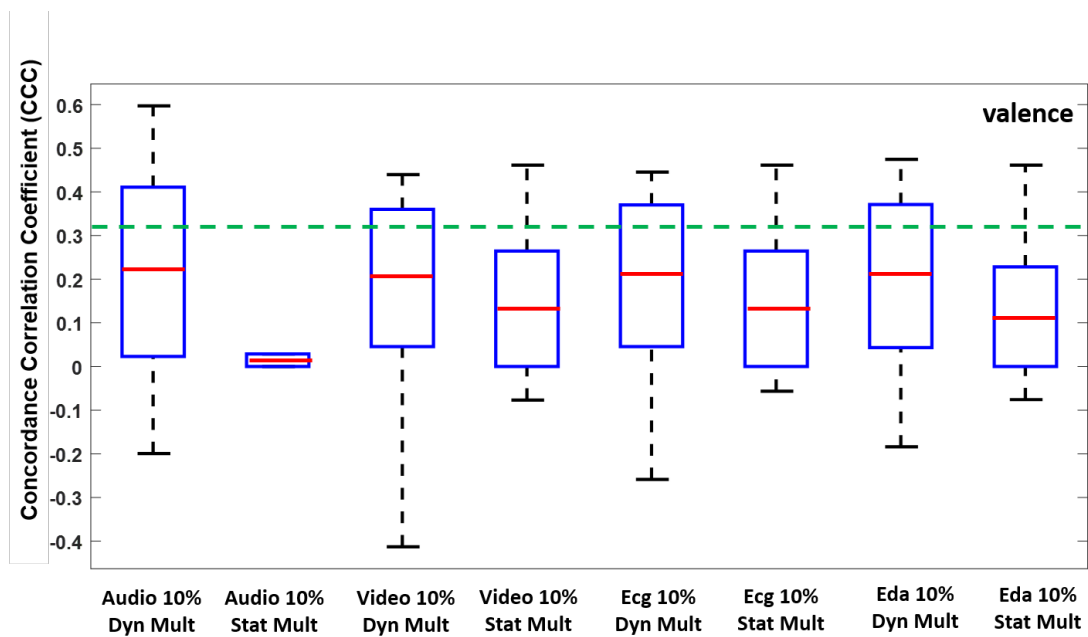
748

749

750



751



752

753

754

755 **Figure 9.**