Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models

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

Abstract—Research on automatic emotion recognition from speech has recently focused on the prediction of time-continuous dimensions (e. g., arousal and valence) of spontaneous and realistic expressions of emotion, as found in real-life interactions. However, the automatic prediction of such emotions poses several challenges, such as the subjectivity found in the definition of a gold-standard from a pool of raters and the issue of data scarcity in training models. In this work, we introduce a novel emotion recognition system, based on ensembles of single-speaker-regression-models. The estimation of emotion is provided by combining a subset of the initial pool of single-speaker-regression-models selecting those that are most concordant among them. The proposed approach allows the addition or removal of speakers from the ensemble without the necessity to re-build the entire recognition system. The simplicity of this aggregation strategy, coupled with the flexibility assured by the modular architecture, and the promising results observed on the RECOLA database highlight the potential implications of the proposed method in a real-life scenario and in particular in web-based applications.

Index Terms—Speech emotion recognition, cooperative regression model, naturalistic emotional display

1 Introduction

Speech is one of, if not the, most natural way for humans to communicate. In everyday social interactions, humans express various complex feelings such as emotion and empathy. Despite the fact that the cognitive processes used to encode affective information during social interactions are relatively complex, humans can easily manage to decode such information in real time from multimodal cues. Conversely, the effort required of computer-based systems for a reliable and autonomous understanding of emotion is still challenging, even for the unimodal analysis of speech. Nonetheless, the development of such affective computing systems is promising for many distinct fields of research. Health care systems may offer a personalised treatment according to the measured emotional content, along with an auxiliary diagnostic tool of the psychological or developmental state of the patient, such as depression [1], [2] or autism spectrum conditions [3], [4]. Remote care assistance can benefit from the estimation of the affective state (e. g., stress or fear) in the voice of elder people [5]. Moreover, applications such as speech based advertising [6], remote teaching (e-learning) [7], job interview [8], and surveillance systems [9] may be incredibly enriched by customer-affect oriented services and monitoring, among many others.

Beyond the proven interests in the relatively new discipline of affective computing, until now numerous issues have limited the full development and use of speech emotion recognition (SER) systems in real-life applications [10]. Whereas the automatic recognition of acted emotion can provide useful insights in the process of affective behaviours encoding into speech and lead to very high recognition rates [11], [12], [13], it is widely acknowledged that such data cannot be a good representative of the emotions produced in real-life interactions [14]. Spontaneous emotions are indeed much more subtle and almost never appear as a "full-blown" expression [15]. As a result, the automatic recognition of spontaneous emotions is much more challenging in comparison to the automatic recognition of acted emotions. In such scenario, we aimed at developing a system able to continuously and automatically predict the perceived emotional condition of a subject expressed in any kind of naturalistic environment.

1.1 Related work

Recently, databases of emotion collected during natural interactions with time-continuous ratings (e. g., arousal and valence [16]) have emerged, such as the Sensitive Artificial Listener (SAL) set in the HUMAINE database [17], the SEMAINE database [18] and the RECOLA database [19]. Such databases have caused a shift in methods, first of all moving from

- *A. Mencattini, E. Martinelli, and C. Di Natale are with the Department of Electronic Engineering, University of Rome Tor Vergata, Rome, Italy.*
  E-mail: mencattini,martinelli@ing.uniroma2.it, dinatale@uniroma2.it
- *F. Ringeval is with the Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany.*
  E-mail: fabien.ringeval@uni-passau.de
- *B. Schuller is with the Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany and with the Department of Computing, Imperial College London, London, UK.*
  E-mail: bjoern.schuller@imperial.ac.uk

classification to regression to be able to model continuous affective dimensions [20], and next moving from utterance or segment level labels [21] to quasi time-continuous labels [18], [22]. Automatic recognition of naturalistic emotion from time-continuous labels presents, however, several challenges that are not yet solved [10], such as the definition of a reliable gold-standard from a pool of raters and the issue of data scarcity in training models.

In the light of the appraisal theory from the domain of emotion psychology [23], each annotator may have a subjective perception of the affective state expressed by an individual, motivated by his/her own past and present experience, memories, reasoning, etc. Additionally, humans have natural bias and inconsistencies in their judgement [24], which creates additional noise in the ratings. Further, the variability in emotion perception can also be observed in the time domain, since the evaluators may have different reaction lag (RL) during the procedure of time-continuous annotation [25]. However, the natural diversity found in emotion perception is usually merged when a machine learning model is trained, by averaging several evaluations from a pool of raters into a single gold-standard. Whereas the use of all annotation data can help at preserving diversity in emotion perception, e. g., by using multi-task learning of each annotator [26], [27], it has the main disadvantage to increase the overall complexity of the model according to the number of available raters. The issue of synchronisation of various individual ratings for defining a gold-standard has also been investigated with signal processing techniques. Models of RL have been estimated from the data, by maximising the correlation coefficient ([28], [29], [30]), or the mutual information [25] between audiovisual features and emotional ratings while shifting back in time the latter.

Regarding the issue of data scarcity, the main question to be solved is how to deal with the huge diversity found in a collection of spontaneous displays of emotion. The common approach in the literature is to use all the emotion variability found in the data as training material and tune the machine learning system in order to disregard the less relevant instances (e. g., by optimising the number of support vectors and the soft margin in Support Vector Regression (SVR)) for emotion prediction [20], [31], [32], [33]. Recent work has proposed to use cooperative learning as a means to select the most informative instances from a set of unlabelled acoustic utterances [34]. But the core underlying idea of this approach is to reduce the cost of the human annotation task, e. g., by selecting instances which are predicted with a low confidence level, not to consider consensus as a way to optimise the predictability of a given SER system. Attempts have already been made in developing cooperative strategies in supervised classification with ensemble models [35], or by considering multi-scaled sliding windows for binary classification [36]. Cooperative strategies have also been used to perform fusion of multimodal stimuli, by using either early (i. e., feature-level) or late (i. e., decision-level) fusion techniques [27], [37], [38], [39].

Taking inspiration from the cooperative strategy proposed in [40], here we introduce a system able to autonomously and temporarily change the composition of a restricted group of predictors provided by single-speaker-regression-models (SSRMs) in a cooperation task governed by a concordance paradigm.

1.2 Main contributions

Motivation of our work lies in the intention to produce a system that can predict the perceived level of emotion of a subject from speech analysis through the fusion of multiple independently trained systems. To this end, we propose a three-topics formulation of the problem of SER from time-continuous labels: (1) emotion subjectivity, (2) models concordance, and (3) dynamic settings.

As mentioned earlier, the use of annotated data of emotion has the immediate consequence of forcing the discrepancy between the emotion produced by the subject and that perceived by the evaluators [23]. Even though the latter may not match the actual affective state of the subject, the evaluators provide the unique available judgement about the emotion, transferring the natural subjectivity of the speaker into the subjectivity of a group of listeners. Hence, in this article, we propose a modular strategy based on cooperative models to perform emotion prediction from speech data. A consensus-based merging strategy is crucial for the cooperation of concordant responses, either of the evaluators (e. g., the Evaluator Weighted Estimator (EWE) [41]) or of the model developed for each speaker. The main goal here is not to consider emotion prediction as a fixed evaluation procedure, but rather as a dynamic cooperative task.

The first stage of our SER system consists of developing an SSRM for each speaker. Then, a second stage follows that consists of applying a cooperative strategy to merge the responses provided by the different SSRMs, while dynamically selecting the window of observation in which the concordance of the responses is estimated. The possibility to develop single-speaker-models merged through a cooperative strategy makes the proposed method easily applicable for real-time applications. Mobile devices and web-based applications require that the regression model can be continuously updated with new data, while avoiding the exponential increase of the learning time or the re-training of the whole model after the addition of new speakers to the system. The cooperative approach proposed in this article offers an elegant solution to this constraint, because it is able to embed new speakers' models independently trained on

separate speech sequences in a dynamic cooperative generation rule. Further, the dynamic adaptation of the SSRM along the observation window allows the system to automatically select the most concordant models and thus maximise the overall performance.

In line with the three-topics formulation named above, and in accordance with the article's organisation, the main contributions of this contribution can be listed as follow: (i) we propose to use a quadrant-based temporal division to estimate the RL of emotion annotation and perform feature selection (topic emotion subjectivity), (ii) we define a dynamic consensus-based cooperative strategy to predict emotion from several SSRMs (topics dynamic settings and models concordance), and (iii) we perform extensive evaluations on a fully naturalistic database of emotion (RECOLA) to compare the performance of our system with methods from the state-of-the-art.

The remainder of this article is structured as follows: first, Section 2 gives a detailed description of the proposed consensus-based SER system, and introduces the database used for the experiments; next Section 3 reports results. Final remarks and direction of future research are given in Section 4.

2 Data and Methods

2.1 Database

A new multimodal corpus of spontaneous interactions in French called RECOLA, for REmote COLlaborative and Affective interactions, was recently introduced by Ringeval et al. [19]. Spontaneous interactions were collected during the resolving of a collaborative task ("Winter survival task") that was performed in dyads (i. e., interaction of two speakers at a time) and remotely by video conference. The RECOLA database includes 9.5 h of multimodal recordings, i. e., audio, video, electro-cardiogram (ECG) and electrodermal activity (EDA), that were continuously and synchronously recorded from 46 participants. Ratings of emotion were performed by six French-speaking assistants (three male, three female) via the ANNEMO web-based annotation tool [19], i. e., time- and value-continuous, for the first five minutes of all recorded sequences. The dataset for which participants gave their consent to share their data is reduced to a set of 34 participants for an overall duration of seven hours, from which the annotation of 23 participants (10 male, 13 female; age: $\mu = 21.3$ years and $\sigma = 4.1$ years) were made publicly available[1]. Even though all participants were French speakers, they had different mother tongue: 17 subjects were French, three German and three Italian. Note that the nonconsecutive numeric speaker labels displayed in this article – e. g., P16, P17, P21, and so on – originate from the RECOLA dataset.

1. https://diuf.unifr.ch/diva/recola/

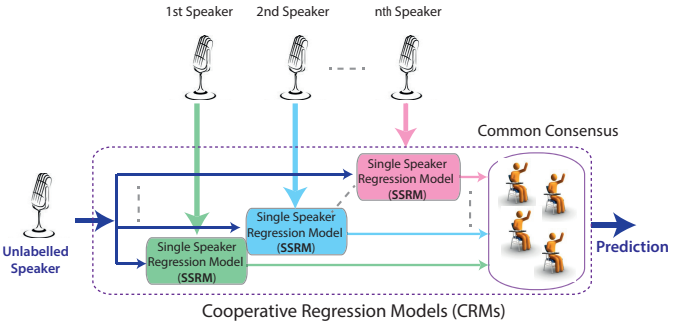

Fig. 1. Schematic description of the consensus based speech emotion recognition system.

Algorithm 1 Construction of each Single Speaker Regression Model (SSRM)

1: acoustic features extraction
2: gold-standard estimation
3: Quadrant-Based Temporal Division (QBTD)
4: **for all** $q = \{a^-, a^+, v^-, v^+\}$ **do**
5: $L_q \leftarrow$ length of each segment
6: **for all** $RL = 0$ **to** $8\,s$ step $0.04\,s$ **do**
7: shift gold-standard of RL
8: **return** $CFS(RL)$ for feature selection
9: **end for**
10: $RL^q_{opt} \leftarrow argmax_{RL}(CFS)$
11: save selected features according to RL^q_{opt}
12: **end for**
13: **return** $RL^a \leftarrow \frac{1}{L_{a^-}+L_{a^+}}\left(L_{a^-} \cdot RL^{a^-}_{opt} + L_{a^+} \cdot RL^{a^+}_{opt}\right)$
14: **return** $RL^v \leftarrow \frac{1}{L_{v^-}+L_{v^+}}\left(L_{v^-} \cdot RL^{v^-}_{opt} + L_{v^+} \cdot RL^{v^+}_{opt}\right)$
15: gold standard synchronisation by RL^a and RL^v
16: concatenate selected features for each dimension
17: features normalisation by $Z-$score
18: linear regression by Partial Least Square (PLS)

2.2 Single Speaker Regression Model (SSRM)

Fig. 1 shows a schematic description of the whole method. Coloured blocks identify each SSRM receiving as input the speech of a speaker as well as the corresponding annotations in terms of arousal and valence. The cooperative regression model (CRM) used for the prediction of an emotional dimension (e. g., arousal or valence) from an unlabelled speech sequence involves to average the responses of each SSRM exhibiting a common consensus, as illustrated by the stylised men with raised hand. The steps needed for the construction of SSRM and CRM are listed in Algorithm 1 and 2, respectively, and are detailed in the following sections.

2.2.1 Acoustic features extraction

According to previous work [27], we consider the 65 acoustic low level descriptors (LLDs) and their first order derivatives (producing 130 LLDs in total) that

were used for the INTERPSEECH Computational Paralinguistic challengE since its 2013 edition [42]. The COMPARE feature set has been computed with the open source extractor OPENSMILE (release 2.0) [43]. This feature set includes a group of 4 energy related LLDs, 55 spectral related LLDs, and 6 voicing related LLDs, cf. Table 1 and step 1 in Algorithm 1. For more details on the COMPARE feature set, the reader is referred to [44]. In what follows, we denote with N_t the temporal length of each speech sequence, with N_f the total number of acoustic features, with N_e the number of evaluators for each speech sequence, and with N_{sp} the number of speakers for which data and annotations are available as training material.

2.2.2 Gold-standard estimation

Learning the acoustic model of an emotional dimension requires the computation of a gold-standard from the annotated data of each speaker, cf. step 2 in Algorithm 1. This is often achieved by averaging the traces provided by each rater. The EWE [41] procedure can be used to centre the ratings to a value that maximises the inter-rater agreement [27]. Assuming that individual mean centring of each annotation may alter the original rating by resetting the natural bias of each annotator, i. e., the subjective perception of each rater, here we propose a new weighted averaging strategy that maintains the original dynamic of the annotations similarly to the one used in [27].

Formally, indicating with d each dimension, i. e., $d = \{a, v\}$, and starting from the evaluation provided by each rater, e_i, $y_d^{e_i}(t)$, $i = 1, \ldots, N_e$, the six evaluations are shifted by the same quantity $\bar{y}_d$ that is obtained by applying Eqs. (1) – (3).

$$\bar{\rho}_d(i) = \frac{1}{N_e - 1} \sum_{\substack{j=1 \\ (j \neq i, \rho_d(i,j) > 0)}}^{N_e} \tilde{\rho}_d(i,j) \tag{1}$$

$$\bar{y}_d = \frac{1}{\sum_{i=1}^{N_e} \bar{\rho}_d(i)} \sum_{i=1}^{N_e} \frac{1}{T} \sum_t y_d^{e_i}(t) \bar{\rho}_d(i) \tag{2}$$

$$y_d(t) = \frac{1}{N_e} \sum_{i=1}^{N_e} (y_d^{e_i}(t) - \bar{y}_d), \tag{3}$$

with $\bar{\rho}_d(i)$ the mean pair-wise Pearson's correlation coefficient of the annotation provided by the evaluator e_i with the remaining $N_e - 1$, and $\tilde{\rho}_d(i,j) = \max(0, \rho_d(i,j))$ the positive Pearson's correlation coefficient of the ratings provided by the evaluators e_i and e_j.

Such procedure gives thus priority to the raters that agree more with the pool when averaging their respective annotation. If all raters perfectly agree with each other, then all pair-wise correlation coefficients are equal to one and our procedure corresponds to a simple average of the annotations after mean centring.

TABLE 1
COMPARE acoustic feature set: 65 low-level descriptors (LLDs).

4 energy related LLDs	**Group**
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
RMS energy, Zero-crossing rate	Prosodic
55 spectral LLDs	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	Spectral
MFCC 1–14	Cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral roll-off pt. 0.25, 0.5, 0.75, 0.9	Spectral
Spectral flux, Centroid, Entropy, Slope	Spectral
Psychoacoustic sharpness, Harmonicity	Spectral
Spectral variance, Skewness, Kurtosis	Spectral
6 voicing related LLDs	**Group**
F_0 (SHS & Viterbi smoothing)	Prosodic
Probability of voicing	Voice quality
log. HNR, Jitter (local & δ), Shimmer (local)	Voice qual.

Note that we do not consider in the computation of the gold-standard the annotations that exhibit negative correlation coefficients to avoid unwanted compensation effects in the normalisation procedure.

2.2.3 Quadrant-based temporal division (QBTD)

According to Russell's two dimensional representation of emotions [16], each quadrant of the diagram conveys specific characteristics of emotion. Further, all emotions are not conveyed by a unique acoustic feature set [45], and such associations can also vary according to the age and the gender of the speaker, among many other paralinguistic traits and states [46].

We therefore propose to consider such peculiarities to select relevant acoustic feature subsets and estimate RL of the raters. For the purpose of optimising the feature selection as well as the reaction lag estimation procedures, we decide to segment the gold-standards $y_d(t)$ and the corresponding acoustic features $x^k(t)$, $k = 1, \ldots, N_f$ into segments of positive and negative arousal or valence. Denoting with $q = \{a^+, a^-, v^+, v^-\}$ each possible quadrant of the 2D arousal-valence space, cf. step 3 in Algorithm 1, the corresponding segments of the gold-standard are indicated by $y_{a^+}(t)$, $y_{a^-}(t)$ and $y_{v^+}(t)$, $y_{v^-}(t)$, and the corresponding segments of acoustic features by $x^k_{a^+}(t)$, $x^k_{a^-}(t)$ and $x^k_{v^+}(t)$, $x^k_{v^-}(t)$, where $y_{a^+}(t) = \{y_a | y_a \geq 0\}$, $y_{a^-}(t) = \{y_a | y_a < 0\}$ and $y_{v^+}(t) = \{y_v | y_v \geq 0\}$, $y_{v^-}(t) = \{y_v | y_v < 0\}$. With reference to the Russell representation, we call this segmentation the quadrant-based temporal division (QBTD). Segmentation is performed by simply concatenating all the segments of a single quadrant. Such procedure adds the benefit to avoid that feature selection is mostly guided by the most populated quadrant.

2.2.4 Reaction lag estimation and feature selection

It is known that, evaluators need some time to evaluate the cues observable in an audiovisual sequence and then report the corresponding emotion. This is especially observable on time-continuous ratings used on dimensional models of emotion, where a delay occurs between the observable cues and the reported emotional value. According to the evaluations performed in [25], we assume here a RL distinct for each speaker and emotional dimension with a negligible variation among the six ratings of the same speaker, compensating this effect with the correlation-based estimation of the gold-standard. However, we relate the estimation of the optimal RL to a feature selection procedure that is performed independently on each quadrant of the 2D arousal-valence emotional space, to consider the peculiarities of the acoustic features according to the emotions.

The importance of such kind of analysis has been demonstrated by the results obtained in preliminary comparative simulations performed without the RL-based synchronisation of features and gold-standard. In this regard, in Section 3.5 we will discuss results of the related experiments run to reinforce our assumption.

All gold-standard segments $y_{a^+}(t)$, $y_{a^-}(t)$, $y_{v^+}(t)$, and $y_{v^-}(t)$ extracted by the QBTD decomposition are thus used separately for each quadrant to perform synchronisation with the corresponding acoustic features. For each quadrant q and a variable RL value in the range $[0,\,8]\,s$ with a step of $0.04\,s$, the corresponding gold-standard segment is shifted back in time with a lag equal to RL and the correlation-based feature selection (CFS) measure is computed [47], [48] (steps 7-8 in Algorithm 1). The optimal reaction lag RL^q_{opt} is then defined as the RL that maximises the CFS measure (step 10 in Algorithm 1). Given the two optimal values RL^q_{opt} for a given dimension (i. e., arousal or valence), the final reaction lag is estimated by weighting the two values obtained on each side of the considered dimension with the length of the corresponding segments (step 13 for arousal and step 14 for valence in Algorithm 1). Compensation of the annotation delay is finally obtained by shifting back in time the gold-standard with the corresponding RL (step 15 in Algorithm 1).

The results show that, an average RL of $3.89\,s$ is obtained for arousal ($\sigma = 1.16\,s$) and $4.52\,s$ ($\sigma = 2.15\,s$) for valence, in total agreement with the experimental results reported in the literature [25], [27]. Arousal is indeed a less subjective emotion dimension than valence is, and thus requires less time for being evaluated. Concerning the results of feature selection, we list in Appendix A the most frequently selected features in each quadrant along with the related description. Note that, the list of features that are selected in each quadrant are saved (step 11 in Algorithm 1) and concatenated (step 16 in Algorithm 1) for each affective dimension in order to be used for the prediction of an unknown speaker's emotion.

2.2.5 Feature normalisation and linear regression

The features selected using the QBTD procedure are normalised by a Z-score (step 17 in Algorithm 1), i. e., the mean is removed from the features and the values are further divided by the standard-deviation, and the normalisation parameters $\mu_{\tilde{x}^k_q}$ and $\sigma_{\tilde{x}^k_q}$ (mean and standard deviation) are stored in the SSRM's parameters for being used later in the cooperative regression. Concerning the regression part of the SSRM, we trained Partial Least Square regression (PLS) on the selected features (step 18 in Algorithm 1). The SIMPLS algorithm is used for this purpose [49]. The optimal numbers of latent variables LV_a and LV_v (for arousal and valence, respectively) are extracted through contiguous block splitting cross-validation (10 splits) performed on the entire speech of the speaker.

2.3 Cooperative Regression Model (CRM)

The principle of the cooperative regression model (CRM) is illustrated in Fig. 1. The CRM receives as inputs the predictions provided by each SSRM. Only the predictions that exhibit a common consensus (indicated by the men with raised hand) are averaged, and a final prediction is produced. The cooperation principle is based on a two-fold strategy. First, each SSRM is applied on the speech of a new speaker sp_x which produces an individual response. Then, only the most concordant responses among the N_{sp} available ones are retained and merged to produce the final prediction. In order to select the most concordant predictions, we used the mutual concordance correlation coefficient (CCC), ρ_c [50]. It is a measure of agreement between two time-continuous predictions that non-linearly combines in a unique parameter the Pearson correlation coefficient (CC), ρ, and the mean square error. The parameter CCC computed on two time-series $y_1(t)$ and $y_2(t)$ on a given observation time-interval T is defined as follows:

$$\rho_c(y_1, y_2) = \frac{2\rho(y_1, y_2)\,\sigma_{y_1}\,\sigma_{y_2}}{\sigma^2_{y_1} + \sigma^2_{y_2} + \left(\mu_{y_1} - \mu_{y_2}\right)^2}, \qquad (4)$$

where the CC (ρ), the mean (μ), and the standard deviation (σ) are meant to be computed under the assumption of stationarity of the two time-series $y_1(t)$ and $y_2(t)$ on the observation time-interval T. The underlying idea of using the CCC is to measure the consensus of the predictions provided by the speakers in the cooperation observed on a given time period T. The steps used in the CRM are listed in Algorithm 2 and detailed below.

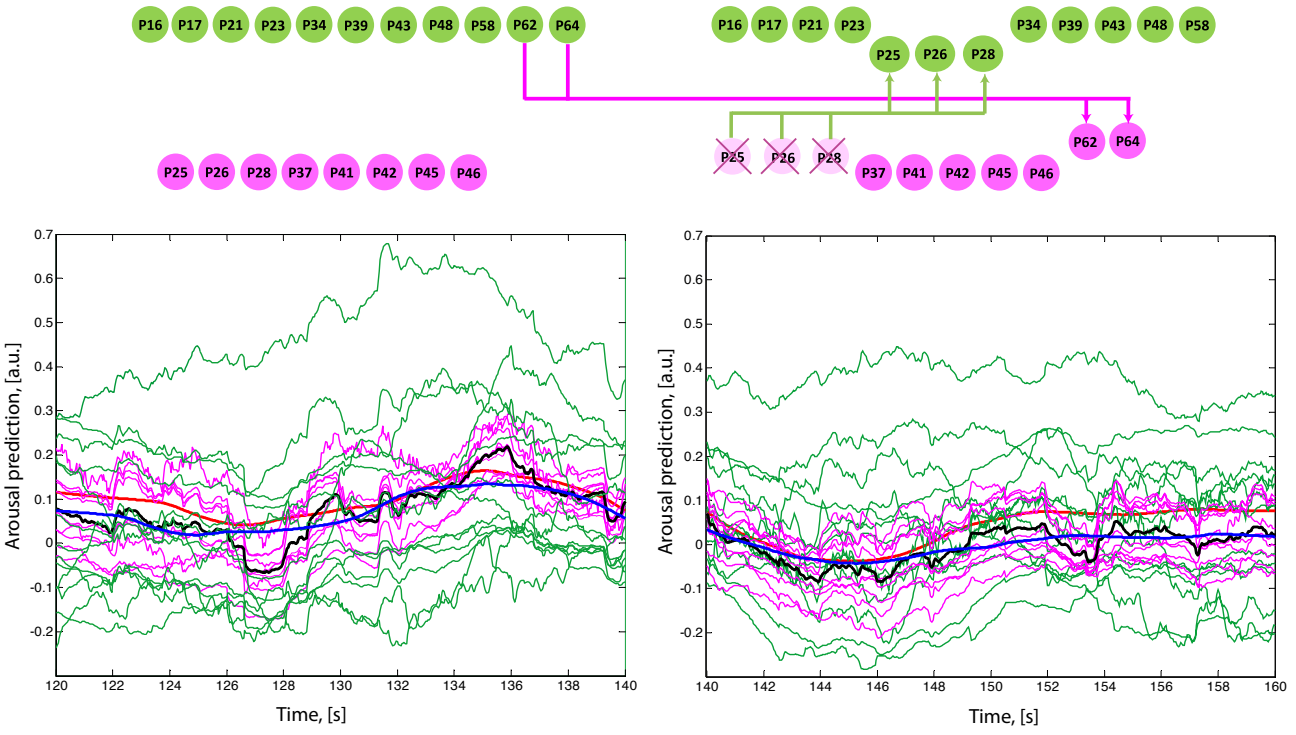

Fig. 2. Illustration of the dynamic consensus-based cooperative merging rule for emotion prediction on arousal. The left and right plots represent two consecutive segments each of $20\,s$ of duration. The red curve represents the gold-standard (speaker P19), the black curve is the unsmoothed prediction obtained after averaging the most concordant predictions (the magenta curves, i. e., those that fall below the $60th$-percentile) and excluding the less concordant ones (the green curves). The blue curve is the final prediction obtained after applying a moving average sliding window for smoothing purpose. On the top of the figure, the magenta circles indicate the speakers that were included in the cooperation process for each of the two segments whereas green circles indicate those that were excluded. The magenta arrows indicate the process of inclusion of the SSRMs (speakers P62 and P64) and the green ones the process of exclusion (speakers P25, P26 and P28) when in the second segment.

Algorithm 2 Implementation of the CRM

1: **for** $t = 0$ **to** N_t step $t_0 = 200\,ms$ **do**
2: **for** $w = 0$ **to** $80\,s$ step $2\,s$ **do**
3: application of each SSRM at time t
4: $y(t, sp_x, sp_p) \leftarrow$ prediction of sp_x provided by the pth-SSRM
5: $\rho_c(w,p) \leftarrow$ average pair-wise ρ_c of pth-prediction in each w
6: $\rho_c(w) \leftarrow$ average over $\rho_c(w,p)$ in the $60th$-percentile
7: **end for**
8: $w_{opt} \leftarrow argmax_w(\rho_c(w))$
9: average prediction values in the optimal window w_{opt}
10: **end for**
11: **return** $y(t) \leftarrow$ average predictions collected for each time step
12: **return** output smoothing by moving average with time lag of $8\,s$

At each time t (step 1 in Algorithm 2) and for a given temporal window w (step 2 in Algorithm 2), the pth-SSRM is first applied to the unlabelled speech sequence sp_x producing a response $y(t, sp_x, sp_p)$ (step 4 in Algorithm 2). We emphasise that the range of $[0 - 80]s$ where to select the most concordant responses has been chosen to let the approach have a wide range of possibilities to choose the optimum interval of concordance from. Then, for each SSRM, the average pair-wise CCC is computed considering its prediction with the others $\rho_c(w,p)$ (step 5 in Algorithm 2). A global concordance factor $\rho_c(w)$, for the duration w, is obtained by averaging only the $\rho_c(w,p)$ that fall into the $60th$-percentile (step 6 in Algorithm 2). This value has been selected after running experiments using values in the range $[50 - 70]th$-percentile and selecting the optimal trade-off between the number of predictions merged on average and the performance in the prediction. The optimal window duration w_{opt} and the most concordant predictions are defined by the arguments that maximise the value of ρ_c, (step 8 in Algorithm 2). The most concordant responses are then averaged which produces the final prediction in w_{opt} (step 9 in Algorithm 2). Continuous monitoring can

be achieved by implementing a sliding windowing procedure with a time lag $t_0 = 200\ ms$. Due to the optimal duration selection (w_{opt}) and to the used sliding window, there can be overlapping predictions that are finally averaged time by time (step 11 in Algorithm 2). Finally, a moving average procedure over $8\,s$ is applied to produce a smoothed response in the final prediction (step 12 in Algorithm 2).

The described procedure illustrates how the most concordant predictions are selected according to the average pair-wise CCC computed on a dynamically changing window. This implementation choice is motivated by the fact that it is not *a priori* known which is the duration of consensus or of disagreement of each predictor with the majority. As a consequence, the composition of the cooperation changes dynamically over time as it is shown in Fig. 2.

3 Results and Discussion

The proposed method has been tested using the RECOLA database which contains 23 publicly available emotion speech sequences of five minutes length each that were annotated in terms of arousal and valence. To assess the performance of the CRM we implemented a leave-one-speaker-out (LOSO) cross-validation strategy to ensure speaker independence in testing the system.

In the following, we describe each test that has been performed to evaluate system performance.

3.1 Training and optimisation of SSRM

We first evaluated the performance obtained during the training and the optimisation of the SSRM. The CCC, CC, and root mean square error (RMSE) between the gold-standard and the prediction, as well as the average CFS (i. e., averaged over the two CFS carried out in the two quadrants of the same dimension) computed during the optimisation step are given in Fig. 3 and 4 for arousal and valence, respectively. Results show that, arousal is significantly better recognised from the acoustic features than valence. This result is in agreement with the literature, where acoustic features have always been shown to present a stronger correlation with the arousal dimension in comparison to valence [22], [26], [27], [31], [38]. The values of CCC and CC are most of the time almost identical, as the RMSE is quite low; we obtained an average RMSE of 0.068 for arousal and of 0.128 for valence over a range of 2.

3.2 Overall performance of the CRM

We tested our system on the RECOLA database by applying the CRM on the predictions provided by each SSRM with a LOSO evaluation framework. The performance obtained for each speaker is combined in the box-plot in Fig. 5 for CCC (top) and CC (bottom)

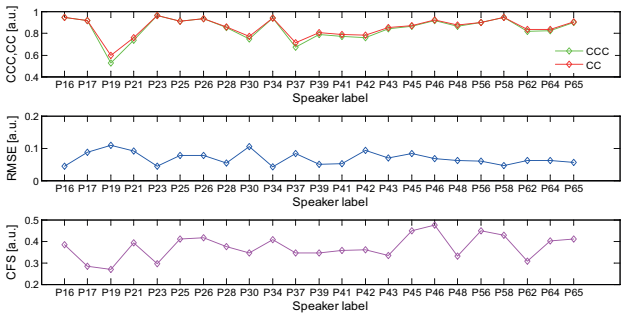

Fig. 3. Performance obtained during the training of each SSRM for the **arousal** dimension (from top to bottom): concordance correlation coefficient (CCC), Pearson's correlation coefficient (CC), root mean square error (RMSE), and correlation-based feature selection (CFS).

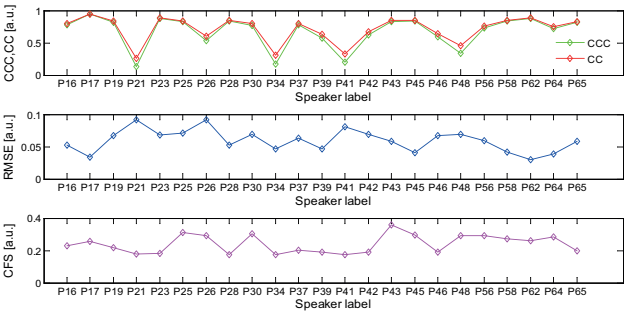

Fig. 4. Performance obtained during the training of each SSRM for the **valence** dimension (from top to bottom): concordance correlation coefficient (CCC), Pearson's correlation coefficient (CC), root mean square error (RMSE), and correlation-based feature selection (CFS).

and for arousal (left) and valence (right) dimensions. Results confirm that the prediction of arousal from acoustic features provides significantly better results than for valence. The combination of weak predictors (PLS) in the CRM, which is similar to a boosting strategy [51], provides a performance that is comparable with the one obtained with more complex machine learning methods that are trained on a full set of speakers [27], [38].

3.3 Inclusion of the SSRM in the CRM

Since our system dynamically adapts the ensemble of SSRM used in the cooperation strategy to perform emotion prediction, we have analysed the frequency of inclusion (i. e., the number of times the SSRM of a speaker is included in the cooperation over the number of observation windows) of each speaker in the model. Fig. 6 illustrates two bar diagrams (the upper for arousal and the lower for valence), representing the frequency with which each speaker is included in the cooperation. The x-axis reports the speaker labels. Results highlight that some speakers such as

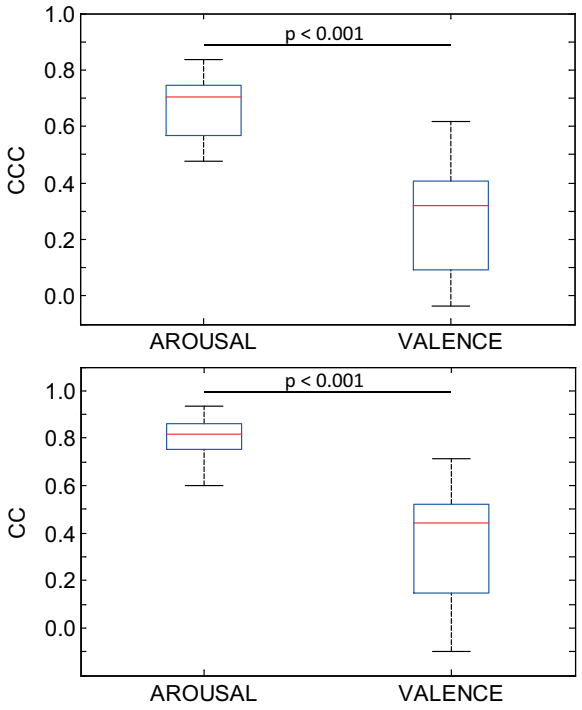

Fig. 5. Box-plots of CCC (top) and CC (bottom) values of the cooperative regression model applied to each speaker in testing phase for arousal and valence.

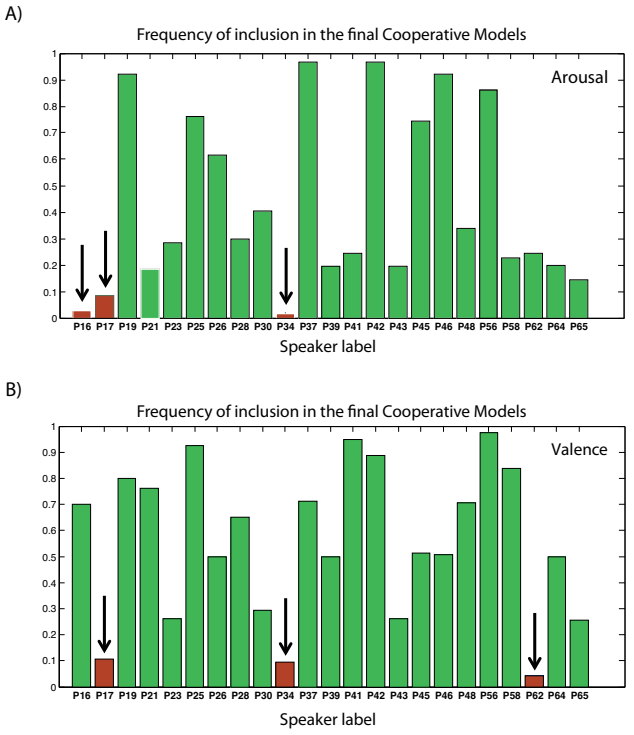

Fig. 6. Bar diagrams (top: arousal, bottom: valence) showing the frequency of inclusion of the SSRM in the CRM. Bars corresponding to the SSRMs that are rarely involved in the CRM are coloured in red and indicated with a descending black arrow.

P16, P17, and P21 (for arousal), P17, P34, and P62 (for valence), marked by the black arrows and represented by red bars, are rarely selected in the cooperative rule. Indeed, if one speaker generally produces emotion in such a specific way that her/his data cannot be used to predict efficiently another speaker's affective behaviour, then these data are not included in the cooperation rule. In addition, we observed that the gold-standard annotation of these speakers (in terms of arousal, valence, or both) exhibit a very small total variation (quantified by the sum of the absolute first derivative over the entire period), meaning that the annotations remain almost stable except for a few small time intervals. This strong heterogeneity in terms of depicted emotions is another possible explanation for the exclusion of the corresponding SSRM from the consensus rule. Therefore, the system autonomously solves this aspect by the dynamic selection of the members of the cooperation, assuring that speakers with low generalisation capabilities do not deteriorate the overall prediction performance.

3.4 Comparison with standard approaches

To further quantify the performance of the proposed method (i. e., SSRM combined with CRM) with respect to standard regression approaches, we also implemented two other emotion recognition strategies.

The first one, labelled as AVERAGE, consists in averaging predictions from all the SSRMs without using the cooperation rule. Such test allows to verify the improvement achieved by the proposed adaptive merging procedure.

The second comparative approach, labelled as GLOBAL, is based on a global training of a unique PLS model performed on the entire training dataset. This comparison allows to highlight the advantage of using an ensemble of SSRM in a modular architecture without taking into account the benefits of the CRM for adaptive merging. Note that, the learning of the global model is computationally much more demanding than the other two approaches, because all speakers are used to compute the PLS model. Moreover, such approach is not flexible to the on-line addition of new speech sequences.

Performance is quantified through the median CCC value and the corresponding inter-quartile range (i. e., the distance between the $75th$ and the $25th$ percentiles), and is given in Fig. 7 for each of the three comparative methods, i. e., CRM, AVERAGE and GLOBAL.

The results show that, the performance obtained with the CRM approach is significantly higher than the two other strategies (i. e., AVERAGE and GLOBAL) for both arousal ($p < 0.001$) and valence ($p < 0.05$, paired t-test). Although the performance is slightly higher for AVERAGE in comparison to GLOBAL, for both arousal and valence, the differ-

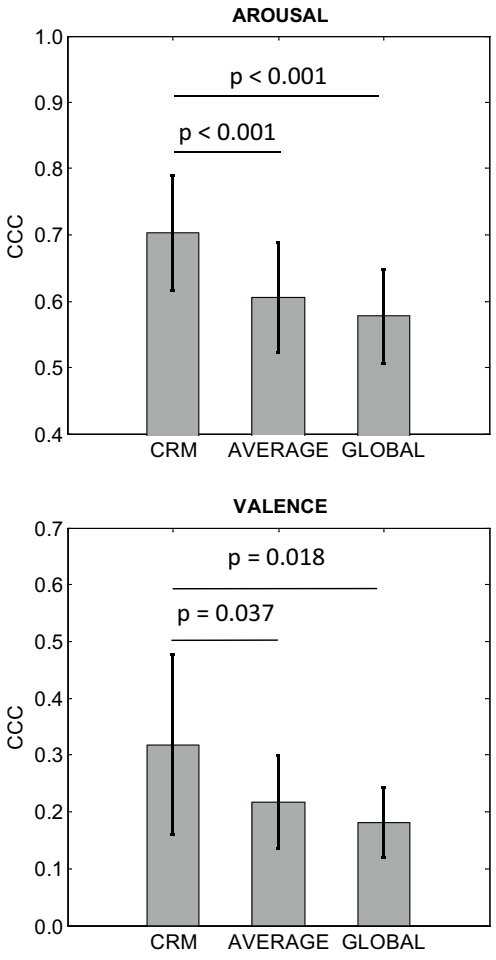

Fig. 7. Comparison of the performance (median and inter-quartile range of CCC) obtained with the proposed CRM, the average of all SSRM (AVERAGE), and a single PLS model learnt on all the training data (GLOBAL) for arousal (top) and valence (bottom). The p-values obtained by a t-test on the CCC values between CRM and the two other methods are also indicated.

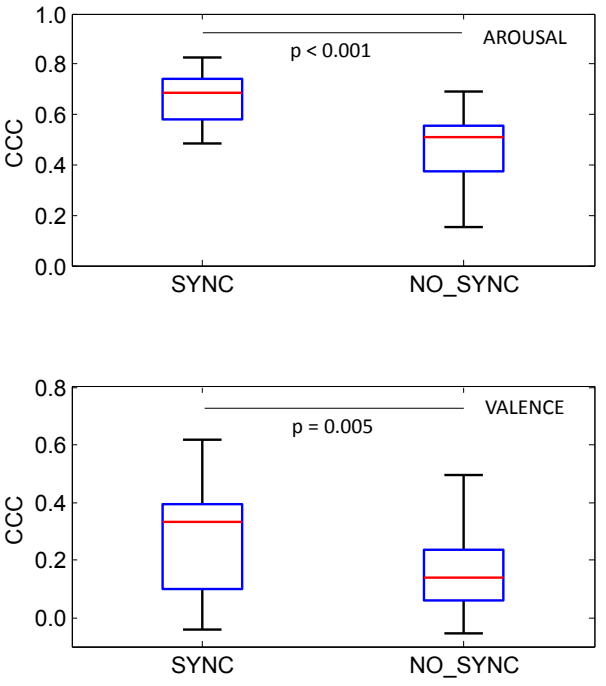

Fig. 8. Box-plots of the performance in terms of CCC values obtained with (left – labelled as SYNC) or without (right – labelled as NO-SYNC) the shifting back of the gold-standard by the estimated reaction lag for each dimension, for arousal (top) and valence (bottom). The p-values of a paired t-test between the CCC values obtained on those two approaches are reported.

ences are not statistically significant (i. e., ($p > 0.05$)).

3.5 Gold-standard and features synchronisation by the estimated RL

Another novelty proposed in this article is the synchronisation of the gold-standard with acoustic features for the construction of each SSRM, performed using the reaction lag estimated separately for arousal and valence. To prove the importance of such procedure, we compare the CCC values computed on the predictions achieved by the proposed approach with those obtained without the synchronisation and the RL estimation procedures. In the latter case, features are selected without shifting back the gold-standard of a quantity equal to the estimated reaction lag.

Fig. 8 shows the box-plots of the CCC values obtained in the two experiments for arousal (top) and valence (bottom). The statistical significance of the improvements obtained with the inclusion of the synchronisation procedure is verified by a paired t-test for both arousal and valence; we obtained $p < 0.001$ for those two dimensions, demonstrating the importance of the synchronisation procedure for constructing the SSRMs that cooperate in the CRM.

3.6 QBTD-optimisation of the SSRM

We also propose in this article the use of a QBTD-optimisation for the construction of each SSRM. To demonstrate the importance of the QBTD procedure, we performed a global optimisation of the SSRM by using all the quadrants of a given emotional dimension, i. e., passive and active for arousal and negative and positive for valence. This global optimisation is labelled as ALL in the following. Related results, comparing the QBTD and the ALL procedures in terms of CCC values obtained for arousal (top) and valence (bottom), are collected in the box-plots shown in Fig. 9.

The statistical significance of the improvements obtained with the QBTD procedure over the global optimisation (ALL), is verified with a paired t-test

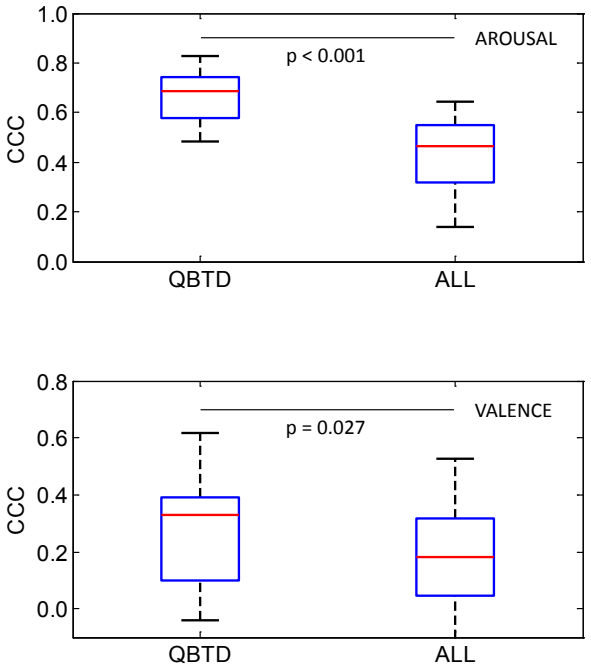

Fig. 9. Box-plots of the performance in terms of CCC values obtained with (left – labelled as QBTD) or without (right – labelled as ALL) the use of the QBTD procedure for the construction of the SSRM for arousal (top) and valence (bottom). The p-values of a paired t-test between the CCC values obtained on those two approaches are reported.

for both arousal and valence; we obtained $p < 0.001$ and $p = 0.027$ for arousal and valence, respectively, demonstrating the importance of the QBTD procedure for constructing the SSRM. Indeed, the QBTD allows the selection of acoustic features that are well correlated with each quadrant of the 2D arousal-valence space. Further, the analysis of the selected acoustic feature sets shows that they strongly depend on the quadrant, especially for valence, cf. Appendix A.

3.7 Correlation between inter-rater agreement and prediction performance

According to our preliminary statements on the importance given to the perceived emotions, we also show that on average, the prediction performance in terms of CC is positively correlated with the mean inter-rater agreement (evaluated through the average pair-wise CC of the ratings for each speaker), cf. Fig. 10. This fact demonstrates how concordance can be considered as a very promising merging principle, both for the design of the cooperation of the models, and for the collection of the gold-standard. Note that, there is a good linear correlation (with ρ equal to 0.75 and 0.61 for arousal and valence, respectively) among the two metrics, especially for arousal, that also presents higher average inter-rater agreement as expected. Moreover, we did not find any statistically

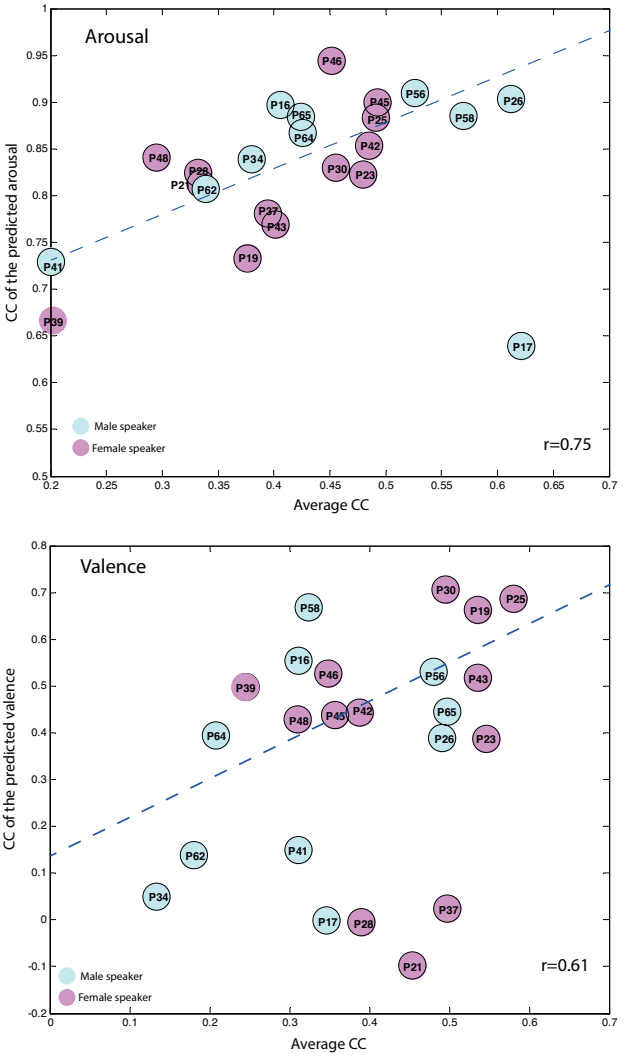

Fig. 10. CC values of the prediction for each speaker during testing versus the average CC value of the evaluators: (top) arousal and (bottom) valence. Colours identify female subjects (magenta) and male subjects (cyan).

significant difference on the CC values grouped according to the gender of speakers, proving that the system is both gender and speaker independent.

3.8 Comparison between PLS and SVR

We investigate here the benefit of using a PLS regression approach to perform adaptive boosting as proposed with the CRM. The generalisation capability of the CRM system based on PLS regression is compared with the use of a predictor based on SVR, with default settings, i. e., a complexity value of $C = 1$, and a Gaussian kernel with $\sigma = 1/f_{sel}$ [52], being f_{sel} the number of features selected in each SSRM. The results reported in Fig. 11 illustrate two kinds of experiments. The first two columns, labelled as SSRM-PLS and SSRM-SVR, respectively, are the box-plots of the CCC values obtained by subject-dependent cross-

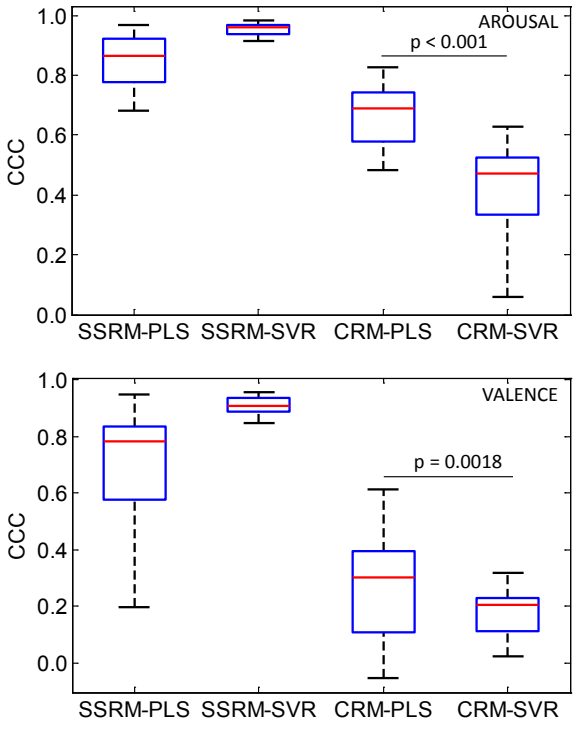

Fig. 11. Box-plots of the CCC values for the CRM applied using PLS regression compared with the CRM based on SVR. The first and the second column report the CCC values obtained during subject-dependent validation of each SSRM (SSRM-PLS and SSRM-SVR), the third and the fourth column indicate the CCC values during testing (CRM-PLS and CRM-SVR): p-values obtained by running paired t-test on the CCC values obtained for the CRM-PLS and the CRM-SVR are also indicated. Results are presented separately for arousal (top) and valence (bottom).

validation of each SSRM in the corresponding speaker speech sequence, comparing PLS and SVR regression methods. In addition, the third and fourth columns, labelled as CRM-PLS and CRM-SVR, respectively, represent the box-plots of the CCC values obtained by merging the responses of all the SSRMs in the training set and estimating the response in the test set, using a LOSO subject-independent cross-validation technique. Results are presented separately for arousal (top) and valence (bottom). One can observe that, even though the SVR provides the best performance in the validation of each SSRM for both arousal and valence, the PLS algorithm is more robust to overfitting and thus produces significantly improved performance on the test set. Our conclusion is that, weak predictors are indeed more suitable to perform boosting than more sophisticated algorithms [51].

3.9 Dynamic evaluation of the prediction performance

As a final consideration, and due to the large duration of the recorded speech signal (5 minutes for each sequence), it is interesting to quantify the tightness of the prediction. To this regard, after we collect the prediction for each speaker, we apply a sliding windowing with observation time frame w_o in the range $[5, 300]s$ on each prediction in testing, and computed the corresponding CCC and the CC values achieved in that segment with respect to the corresponding gold-standard. Given a w_o, the maximum CCC and the maximum CC values computed over all segments of the same length w_o are extracted. Then, by collecting these values for all the 23 speakers, we have a single box-plot related to a given w_o. By repeating for each w_o, we derive the graph in Fig. 12. Such further test allows us to emphasise the fact that for each window length there is at least a segment for each speaker exhibiting very high CCC and CC values in both dimensions. The results indicate that, as long as the window length w_o decreases, the performance metrics increase. This result can be explained by the fact that it is more probable for the prediction to reach a high concordance level with the gold-standard in a small interval than in very long ones. However, from a preliminary analysis, we also noted that, the significance of the metrics CCC and CC decreased on very short segments (i. e., less than $4\,$s), since the reliability of the computation of CCC and of CC values depends on the size of the data used for the calculus. For this reason, we decided to consider the observation windows of duration less than $4\,$s not as meaningful.

4 Conclusion

In this article, we presented a new strategy for continuous speech emotion estimation. New paradigms have been presented concerning single speaker and cooperative regression models. Those novel strategies allow a system to dynamically select the most concordant models over time, which provide an elegant solution to the issues of data scarcity and inconsistencies in the definition of emotion, by fostering the paradigm of perception of an unknown speaker's emotion. A novel quadrant-based decomposition of a speech sequence is used for model optimisation to achieve emotion-related feature selection. Concepts like evaluator's reaction lag and concordance for aggregation have also been addressed and embedded in the whole method. As demonstrated with extensive experiments on a database featuring spontaneous and natural emotions, our approach confers robustness to inter-rater agreement variability, but also to variations in both gender and age of the speaker.

The proposed system presents important potential implications. First of all, new speakers can be added to the cooperative system simply by training a new

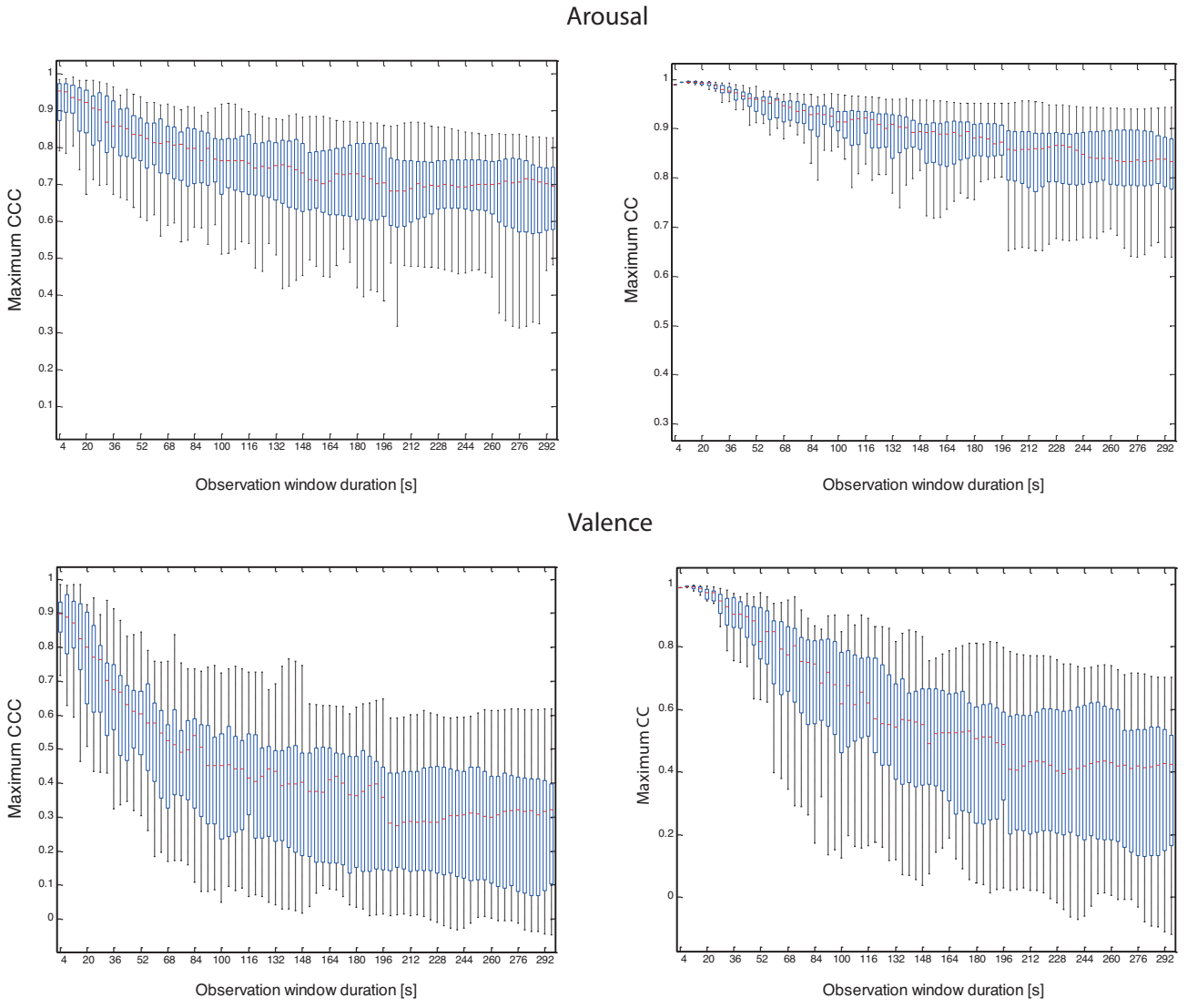

Fig. 12. Box-plots of the maximum CCC and CC values computed over all the possible segments of the same length w_o and distributed over the 23 speakers: (top) arousal and (bottom) valence. The graph is obtained by varying the window length w_o in the range [5, 300] s.

SSRM using the speech sequence along with the corresponding annotations for the new speaker. Second, new affective contents of a speaker already present in the system may be included in the cooperation simply by performing re-learning of the SSRM of that speaker, adding a new speech sequence with a strong reduction of the required learning time. Consequently, system updating can be seen as a parallel procedure that does not influence the normal functioning and, in addition, it does not require time consuming re-learning of the whole prediction system. For this reason, the proposed architecture is perfectly suitable for mobile applications, thanks to the easiness and flexibility to develop single models separately trained on distinct speech sequences with different emotional contents. Web-based applications could offer the possibility to everyone to upload to the cloud his/her speech sequence along with the corresponding annotation.

Finally, the introduction of the QBTD paradigm suggests future developments based on modular architecture in which each SSRM is trained and optimised on each quadrant and then merged using a cooperative rule based on different machine learning scenarios and other databases of emotional speech. This strategy could also be applied for multimodal emotion recognition, to ensure that only relevant cues are effectively used over time [53], [54].

ACKNOWLEDGMENTS

The research leading to these results has been partially funded by the European Union's ERC Starting Grant No. 338164 (iHEARu), and Horizon 2020 Programme through the Innovation Action (IA) #644632 (MixedEmotions), and #645094 (SEWA), and the Research IA #645378 (ARIA-VALUSPA).

REFERENCES

[1] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "From joyous to clinically depressed: Mood detection using spontaneous speech," in *Proceedings of the 25th International Florida Artificial Intelligence Research Society Conference (FLAIRS)*. Marco Island, (FL) USA: AAAI, 2012, pp. 141–146.

[2] K. E. B. Ooi, M. Lech, and N. B. Allen, "Prediction of major depression in adolescents using an optimized multi-channel weighted speech classification system," *Biomedical Signal Processing and Control*, vol. 14, pp. 228–239, November 2014.

[3] E. Marchi, F. Ringeval, and B. Schuller, "Voice-enabled assistive robots for handling autism spectrum conditions: an examination of the role of prosody," in *Speech and Automata in Health Care (Speech Technology and Text Mining in Medicine and Healthcare)*, A. Neustein, Ed. Boston/Berlin/Munich: De Gruyter, 2014, pp. 207–236, invited contribution.

[4] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for prosodic assessment of language impaired children," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1328–1342, July 2011.

[5] R. Looije, M. A. Neerincx, and F. Cnossen, "Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors," *International Journal of Human-Computer Studies*, vol. 68, no. 6, pp. 386–397, June 2010.

[6] A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth, "A taxonomy of applications that utilize emotional awareness," in *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference (IS-LTC)*. Ljubljana, Slovenia: Informacijska Druzba (Information Society), 2006, pp. 246–250.

[7] Q. Luo, "Research on e-learning system using speech emotion recognition," *Advanced Science Letters*, vol. 5, no. 1, pp. 363–366, January 2012.

[8] H. Lu, M. Rabbi, G. T. Chittaranjan, D. Frauendorfer, M. S. Mast, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the ACM 14th International Conference on Ubiquitous Computing (UBICOMP)*. Pittsburgh, (PA) USA: ACM, 2012, pp. 351–360.

[9] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, June 2008.

[10] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing: Affect Analysis in Continuous Input*, vol. 31, no. 2, pp. 120–136, February 2013.

[11] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, September 2006.

[12] M. E. Ayadi, M. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, March 2011.

[13] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, and C. D. Natale, "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure," *Knowledge-Based Systems*, vol. 63, pp. 68–81, June 2014.

[14] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, "The Automatic Recognition of Emotions in Speech," in *Emotion-Oriented Systems: The HUMAINE Handbook, Cognitive Technologies*, R. Cowie, P. Petta, and C. Pelachaud, Eds. Springer, 2010, pp. 71–99.

[15] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.

[16] J. Posner, J. Russell, and B. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology," *Developmental Psychopathology*, vol. 17, no. 3, pp. 715–734, September 2005.

[17] E. Douglas-Cowie *et al.*, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *2nd International Conference on Affective Computing and Intelligent Interaction (ACII), LNCS*, A. Paiva, R. Prada, and R. Picard, Eds. Lisbon, Portugal: Springer-Verlag Berlin Heidelberg, 2007, vol. 4738, pp. 488–500.

[18] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wöllmer, "Building Autonomous Sensitive Artificial Listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, April–June 2012.

[19] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proceedings of Face and Gestures 2013, 2nd IEEE International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*. Shanghai, China: IEEE, April 2013, 8 pages.

[20] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Proceedings of the 32th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. Honolulu, HI, USA: IEEE, 2007, pp. 1085–1088.

[21] M. Chetouani, A. Mahdhaoui, and F. Ringeval, "Time-scale feature extractions for emotional speech characterisation," *Cognitive Computation*, vol. 1, no. 2, pp. 194–201, June 2009.

[22] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1–2, pp. 7–12, March 2010.

[23] K. R. Scherer, A. Schorr, and T. Johnstone, "Appraisal processes in emotion: Theory, methods, research," in *Series in Affective Science*, K. R. Scherer, A. Schorr, and T. Johnstone, Eds. Oxford University Press, New York and Oxford, 2001.

[24] A. Tversky, "Intransitivity of preferences," *Psychological Review*, vol. 76, pp. 31–48, January 1969.

[25] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, pp. 97–108, April-June 2015.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "A multi-task approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems (TiiS) - Special Issue on Affective Interaction in Natural Environments*, vol. 2, no. 1, March 2012, Article No. 6, 29 pages.

[27] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, November 2015.

[28] M. Nicolaou, H. Gunes, and M. Pantic, "Automatic segmentation of spontaneous data using dimensional labels from multiple coders," in *Proceedings of the LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing*. Istambul, Turkey: ELRA, 2010, pp. 43–48.

[29] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI)*. Istambul, Turkey: ACM, 2012, pp. 501–508.

[30] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Proceedings of the 5th Humaine Association Conference on Affective Computing and Intelligent Interactions (ACII)*. Geneva, Switzerland: IEEE, September 2013, pp. 85–90.

[31] D. Wu, T. Parsons, E. Mower, and S. Narayanan, "Speech Emotion Estimation in 3D space," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. Suntec City, Singapore: IEEE, 2010, pp. 737–742.

[32] M. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM Regression for Dimensional and Continuous Emotion Prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, March 2012.

[33] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K.R., "On the acoustics of emotion in audio: What speech, music and sound have in common," *Frontiers in Psychology, Emotion Science, Special Issue on Expression of emotion in music and vocal communication*, vol. 4, May 2013, Article ID 292, 12 pages.

[34] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Shanghai, China: IEEE, April 2016, to appear.
[35] J. Mendes-Moreira, C. Soares, A. Jorge, and J. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, November 2012, Article No. 10.
[36] Y. Fan, M. Xu, Z. Wu, and L. Cai, "Automatic emotion variation detection using multi-scaled sliding window," in *Proceedings of the IEEE International Conference on Orange Technologies (ICOT)*. Xian, China: IEEE, 2014, pp. 232–236.
[37] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage, Special issue on Multimodal Data Fusion*, vol. 102, no. 1, pp. 162–172, November 2013.
[38] F. Ringeval *et al.*, "AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in *Proceedings of the 23rd ACM International Conference on Multimedia (ACM MM), 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*. Brisbane, Australia: ACM, October 2015, pp. 3–8.
[39] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys*, vol. 47, no. 3, April 2015.
[40] E. Martinelli, G. Magna, A. Vergara, and C. D. Natale, "Cooperative classifiers for reconfigurable sensor arrays," *Sensors and Actuators B: Chemical*, vol. 199, pp. 83–92, August 2014.
[41] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proceedings of the 8th International Conference on Multimedia and Expo (ICME)*. Hannover, Germany: IEEE, 2008, pp. 865–868.
[42] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association (ISCA)*. Lyon, France: ISCA, 2013, pp. 148–152.
[43] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of ACM Multimedia (MM)*. Barcelona, Spain: ACM, 2013, pp. 835–838.
[44] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Springer, 2016.
[45] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion," *Journal of Personality and Social Pscychology*, vol. 70, no. 3, pp. 614–636, March 2009.
[46] B. Schuller and A. Batliner, *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Wiley, November 2013.
[47] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, August 2007.
[48] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, August 2005.
[49] S. de Jong, "Simpls: an alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, March 1993.
[50] I.-K. L. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, March 1989.
[51] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, "Emotional audiovisual speech synthesis based on PAD," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 570–582, March 2011.
[52] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, April 2011, Article No. 27.
[53] F. Ringeval, E. Marchi, M. Méhu, K. Scherer, and B. Schuller, "Face Reading from Speech – Predicting Facial Action Units from Audio Cues," in *Proceedings of INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association (ISCA)*. Dresden, Germany: ISCA, September 2015, pp. 1977–1981.
[54] A. Mencattini, F. Ringeval, B. Schuller, E. Martinelli, and C. D. Natale, "Continuous monitoring of emotions by a multimodal cooperative sensor system," *Procedia Engineering, Special Issue Eurosensors 2015*, vol. 120, pp. 556–559, July 2015.

APPENDIX

In this section, we provide additional results concerning feature selection based on the QBTD procedure. Table 2 lists for each quadrant the most frequently selected features along with the corresponding LLD name; the reader is referred to [44] for more information on the computation of the features. These results clearly show that, the sets of features selected for the two partitions of arousal and of valence are almost entirely disjoint—especially for valence—, underlining the importance of a quadrant-based selection. Additionally, spectral based acoustic features appear to be the most robust ones for emotion prediction of both arousal and valence.

TABLE 2
Most selected acoustic LLD in each quadrant; R-PLP stands for RASTA-PLP psychoacoustic filtering; for the purpose of readability, only the minimum and maximum value of frequency band are given for consecutive spectral related features (this case is indicated by a parenthesis including the number of consecutive features).

Negative valence
Energy in R-PLP spectrum $[547-801]Hz$
Energy in R-PLP spectrum $[945-1279]Hz$
Energy in R-PLP spectrum $[1469-1911]Hz$
Positive valence
Zero crossing rate
Energy in R-PLP spectrum $[5865-7203]Hz$
Spectral roll off point at 90%
Negative arousal
Loudness (sum of all R-PLP coefficients)
Root mean square energy
Energy in R-PLP spectrum $[799-3077]Hz$ (9)
Energy in R-PLP spectrum $[3074-4280]Hz$ (2)
Energy in R-PLP spectrum $[4277-5870]Hz$ (3)
Energy in spectrum $[250-650]Hz$
Energy in spectrum $[1000-4000]Hz$
Spectral flux
Spectral slope
Positive arousal
Probability of voicing
Loudness (sum of all R-PLP coefficients)
Root mean square energy
Energy in R-PLP spectrum $[4277-5291]Hz$
Energy in spectrum $[250-650]Hz$
Energy in spectrum $[1000-4000]Hz$
Spectral flux
Spectral variance
Energy in 1st MFCC $[17-163]Hz$