Valstar, Michel F. and Gratch, Jonathan and Schuller, Björn and Ringeval, Fabien and Lalanne, Denis and Torres, Mercedes Torres and Scherer, Stefan and Stratou, Giota and Cowie, Roddy and Pantic, Maja (2016) AVEC 2016 – Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, 16 Oct 2016, Amsterdam, Netherlands.

AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge

Michel Valstar
University of Nottingham
School of Computer Science

Jonathan Gratch
University of Southern California
ICT

Björn Schuller*
University of Passau
Chair of Complex & Intelligent Systems

Fabien Ringeval†
Université Grenoble Alpes
Laboratoire d'Informatique de Grenoble

Denis Lalanne
University of Fribourg
Human-IST Research Center

Mercedes Torres Torres
University of Nottingham
School of Computer Science

Stefan Scherer
University of Southern California
ICT

Giota Stratou
University of Southern California
ICT

Roddy Cowie
Queen's University Belfast
Department of Psychology

Maja Pantic‡
Imperial College London
Intelligent Behaviour Understanding Group

ABSTRACT

The Audio/Visual Emotion Challenge and Workshop (AVEC 2016) "Depression, Mood and Emotion" will be the sixth competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and physiological depression and emotion analysis, with all participants competing under strictly the same conditions. The goal of the Challenge is to provide a common benchmark test set for multi-modal information processing and to bring together the depression and emotion recognition communities, as well as the audio, video and physiological processing communities, to compare the relative merits of the various approaches to depression and emotion recognition under well-defined and strictly comparable conditions and establish to what extent fusion of the approaches is possible and beneficial. This paper presents the challenge guidelines, the common data used, and the performance of the baseline system on the two tasks.

Keywords

Affective Computing, Emotion Recognition, Speech, Facial Expression, Physiological signals, Challenge

1. INTRODUCTION

The 2016 Audio-Visual Emotion Challenge and Workshop (AVEC 2016) will be the sixth competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, video, and physiological analysis of emotion and depression, with all participants competing under strictly the same conditions. The goal of the Challenge is to compare the relative merits of the approaches (audio, video, and/or physiologic) to emotion recognition and severity of depression estimation under well-defined and strictly comparable conditions, and establish to what extent fusion of the approaches is possible and beneficial. A second motivation is the need to advance emotion recognition for multimedia retrieval to a level where behaviomedical systems [38] are able to deal with large volumes of non-prototypical naturalistic behaviour in reaction to known stimuli, as this is exactly the type of data that diagnostic and in particular monitoring tools, as well as other applications, would have to face in the real world.

AVEC 2016 will address emotion and depression recognition. The emotion recognition sub-challenge is a refined re-run of the AVEC 2015 challenge [27], largely based on the same dataset. The depression severity estimation sub-challenge is based on a novel dataset of human-agent interactions, and sees the return of depression analysis, which

*The author is further affiliated with Imperial College London, Department of Computing, London, U.K.

†The author is further affiliated with University of Passau, Chair of Complex & Intelligent Systems

‡The author is further affiliated with Twente University, EEMCS, Twente, The Netherlands.

AVEC'16 16 October 2016, Amsterdam, NL

ACM 978-1-4503-4516-3/16/10…$15.00.
DOI: http://dx.doi.org/10.1145/2988257.2988258.

was a huge success in the AVEC 2013 [41] and 2014 [40] challenges.

- **Depression Classification Sub-Challenge** (DCC): participants are required to classify whether a person is classified as depressed or not, where the binary ground-truth is based on the severity of self-reported depression as indicated by the PHQ-8 score for every human-agent interaction. For the DCC, performance in the competition will be measured using the average **F1 score** for both classes *depressed* and *not_depressed*. Participants are encouraged to provide an estimate of the severity of depression, by calculating the root mean square error over all HCI experiment sessions between the predicted and ground-truth PHQ-8 score. In addition, participants are also encouraged to report on overall accuracy, average precision, and average recall to further analyse their results in the paper accompanying their submission.

- **Multimodal Affect Recognition Sub-Challenge** (MASC) participants are required to perform fully continuous affect recognition of two affective dimensions: Arousal, and Valence, where the level of affect has to be predicted for every moment of the recording. For the MASC, two regression problems need to be solved: prediction of the continuous dimensions Valence and Arousal. The MASC competition measure is the **Concordance Correlation Coefficient (CCC)**, which combines the Pearson's correlation coefficient (CC) with the square difference between the mean of the two compared time series, as shown in 1.

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

  where ρ is the Pearson correlation coefficient between two time series (e. g., prediction and gold-standard), σ_x^2 and σ_y^2 is the variance of each time series, and μ_x and μ_y are the mean value of each. Therefore, predictions that are well correlated with the gold standard but shifted in value are penalised in proportion to the deviation.

To be eligible to participate in the challenge, every entry has to be accompanied by a paper presenting the results and the methods that created them, which will undergo peer-review. Only contributions with a relevant accepted paper will be eligible for challenge participation. The organisers reserve the right to re-evaluate the findings, but will not participate in the Challenge themselves.

2. DEPRESSION ANALYSIS CORPUS

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) database is part of a larger corpus, the Distress Analysis Interview Corpus (DAIC) [13], that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illness [8]. Data collected include audio and video recordings and extensive questionnaire responses; this part of the corpus includes the Wizard-of-Oz interviews, conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room. Data has been transcribed and annotated for a variety of verbal and non-verbal features.

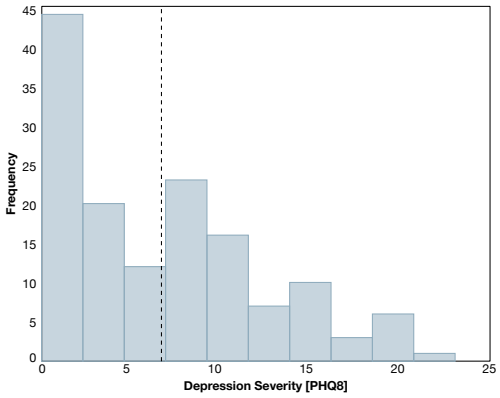

Figure 1: Histogram of depression severity scores for DESC challenge. Data of training and development set are provided here.

Information on how to obtain shared data can be found in this location: http://dcapswoz.ict.usc.edu. Data is freely available for research purposes.

2.1 Depression Analysis Labels

The level of depression is labelled with a single value per recording using a standardised self-assessed subjective depression questionnaire, the PHQ-8 [21]. This is similar to the PHQ-9 questionnaire, but with the suicidal ideation question removed for ethical reasons. The average depression severity on the training and development set of the challenge is $M = 6.67$ ($SD = 5.75$). The distribution of the depression severity scores based on the challenge training and development set is provided in Figure 1. A baseline classifier that constantly predicts the mean score of depression provides an $RMSE = 5.73$ and an $MAE = 4.74$.

2.2 Depression Analysis Baseline Features

In the following sections we describe how the publicly available baseline feature sets are computed for either the audio or the video data. Participants can use these feature sets exclusively or in addition to their own features. For ethical reasons, no raw video is made available.

2.2.1 Video Features

Based on the *OpenFace* [2] framework[1], we provide different types of video features:

- facial landmarks: 2D and 3D coordinates of 68 points on the face, estimated from video
- HOG (histogram of oriented gradients) features on the aligned 112x112 area of the face
- gaze direction estimate for both eyes
- head pose: 3D position and orientation of the head

[1] `https://github.com/TadasBaltrusaitis/CLM-framework`

In addition to that, we provide emotion and facial action unit continuous measures based on *FACET* software[23]. Specifically, we provide the following measures:

- emotion: {Anger, Contempt, Disgust, Joy, Fear, Neutral, Sadness, Surprise, Confusion, Frustration}
- AUs: {AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, AU28, AU43}

2.2.2 Audio Features

For the audio features we utilized COVAREP(v1.3.2), a freely available open source Matlab and Octave toolbox for speech analyses [7][2]. The toolbox comprises well validated and tested feature extraction methods that aim to capture both voice quality as well as prosodic characteristics of the speaker. These methods have been successfully shown to be correlated with psychological distress and depression in particular [32, 33]. In particular, we extracted the following features:

- **Prosodic:** Fundamental frequency (F0) and voicing (VUV)
- **Voice Quality:** Normalized amplitude quotient (NAQ), Quasi open quotient (QOQ), the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), spectral tilt/slope of wavelet responses (peakSlope), and shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (Rd)
- **Spectral:** Mel cepstral coefficients (MCEP0-24), Harmonic Model and Phase Distortion mean (HMPDM0-24) and deviations (HMPDD0-12).

In addition to the feature set above, raw audio and transcripts of the interview are being provided, allowing the participants to compute additional features on their own. For more details on the shared features and the format of the files participants should also review the DAIC-WOZ documentation [3].

3. EMOTION ANALYSIS CORPUS

The Remote Collaborative and Affective Interactions (RECOLA) database [29] was recorded to study socio-affective behaviours from multimodal data in the context of computer supported collaborative work [28]. Spontaneous and naturalistic interactions were collected during the resolution of a collaborative task that was performed in dyads and remotely through video conference. Multimodal signals, i. e., audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA), were synchronously recorded from 27 French-speaking subjects. Even though all subjects speak French fluently, they have different nationalities (i. e., French, Italian or German), which thus provide some diversity in the expression of emotion.

Data is freely available for research purposes, information on how to obtain the RECOLA database can be found on this location: `http://diuf.unifr.ch/diva/recola`.

[2] `http://covarep.github.io/covarep/`

[3] `http://dcapswoz.ict.usc.edu/wwwutil_files/DAICWOZDepression_Documentation.pdf`

Table 1: Inter-rater reliability on arousal and valence for the 6 raters and the 27 subjects of the RECOLA database; raw or normalised ratings [26].

	RMSE	CC	CCC	ICC	α
Raw					
Arousal	.344	.400	.277	.775	.800
Valence	.218	.446	.370	.811	.802
Normalised					
Arousal	.263	.496	.431	.827	.856
Valence	.174	.492	.478	.844	.829

Table 2: Partitioning of the RECOLA database into train, development, and test sets.

#	train	dev	test
female	6	5	5
male	3	4	4
French	6	7	7
Italian	2	1	2
German	1	1	0
age μ (σ)	21.2 (1.9)	21.8 (2.5)	21.2 (1.9)

3.1 Emotion Analysis Labels

Regarding the annotation of the dataset, time-continuous ratings (40 ms binned frames) of emotional arousal and valence were created by six gender balanced French-speaking assistants for the first five minutes of all recordings, because participants discussed more about their strategy – hence showing emotions – at the beginning of their interaction.

To assess inter-rater reliability, we computed the intra-class correlation coefficient (ICC(3,1)) [36], and Cronbach's α [5]; ratings are concatenated over all subjects. Additionally, we computed the root-mean-square error (RMSE), Pearson's CC and the CCC [22]; values are averaged over the C_2^6 pairs of raters. Results indicate a very strong inter-rater reliability for both arousal and valence, cf. Table 1. A normalisation technique based on the Evaluator Weighted Estimator [14], is used prior to the computation of the gold-standard, i. e., the average of all ratings for each subject [26]. This technique has significantly ($p < 0.001$ for CC) improved the inter-rater reliability for both arousal and valence; the Fisher Z-transform is used to perform statistical comparisons between CC in this study.

The dataset was divided into speaker disjoint subsets for training, development (validation) and testing, by stratifying (balancing) on gender and mother tongue, cf. Table 2.

3.2 Emotion Analysis Baseline Features

In the followings we describe how the baseline feature sets are computed for video, audio, and physiological data.

3.2.1 Video Features

Facial expressions play an important role in the communication of emotion [9]. Features are usually grouped in two types of facial descriptors: appearance and geometric based [39]. For the video baseline features set, we computed

both, using Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [1] for appearance and facial landmarks [42] for geometric.

The LGBP-TOP are computed by splitting the video into spatio-temporal video volumes. Each slice of the video volume extracted along 3 orthogonal planes (x-y, x-t and y-t) is first convolved with a bank of 2D Gabor filters. The resulting Gabor pictures in the direction of x-y plane are divided into 4x4 blocks. In the x-t and y-t directions they are divided into 4x1 blocks. The LBP operator is then applied to each of these resulting blocks followed by the concatenation of the resulting LBP histograms from all the blocks. A feature reduction is then performed by applying a Principal Component Analysis (PCA) from a low-rank (up to rank 500) approximation [15]. We obtained 84 features representing 98 % of the variance.

In order to extract geometric features, we tracked 49 facial landmarks with the Supervised Descent Method (SDM) [42] and aligned them with a mean shape from stable points (located on the eye corners and on the nose region). As features, we computed the difference between the coordinates of the aligned landmarks and those from the mean shape, and also between the aligned landmark locations in the previous and the current frame; this procedure provided 196 features in total. We then split the facial landmarks into groups according to three different regions: i) the left eye and left eyebrow, ii) the right eye and right eyebrow and iii) the mouth. For each of these groups, the Euclidean distances (L2-norm) and the angles (in radians) between the points are computed, providing 71 features. We also computed the Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame. In total the geometric set includes 316 features.

Both appearance and geometric feature sets are interpolated by a piecewise cubic Hermite polynomial to cope with dropped frames. Finally, the arithmetic mean and the standard-deviation are computed on all features using a sliding window, which is shifted forward at a rate of 40 ms.

3.2.2 Audio Features

In contrast to large scale feature sets, which have been successfully applied to many speech classification tasks [34, 35], smaller, expert-knowledge based feature sets have also shown high robustness for the modelling of emotion from speech [25, 3]. Some recommendations for the definition of a minimalistic acoustic standard parameter set have been recently investigated, and have led to the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), and to an extended version (eGeMAPS) [10], which is used here as baseline. The acoustic low-level descriptors (LLD) cover spectral, cepstral, prosodic and voice quality information and are extracted with the openSMILE toolkit [11].

As the data in the RECOLA database contains long continuous recordings, we used overlapping fixed length segments, which are shifted forward at a rate of 40 ms, to extract functionals; the arithmetic mean and the coefficient of variation are computed on all 42 LLD. To pitch and loudness the following functionals are additionally applied: percentiles 20, 50 and 80, the range of percentiles 20 – 80 and the mean and standard deviation of the slope of rising/falling signal parts. Functionals applied to the pitch, jitter, shimmer, and all formant related LLDs, are applied to voiced regions only. Additionally, the average RMS energy is computed and 6 temporal features are included: the rate of loudness peaks per second, mean length and standard deviation of continuous voiced and unvoiced segments and the rate of voiced segments per second, approximating the pseudo syllable rate. Overall, the acoustic baseline features set contains 88 features.

3.2.3 Physiological Features

Physiological signals are known to be well correlated with emotion [20, 19], despite not being directly perceptible the way audio-visual are. Although there are some controversies about peripheral physiology and emotion [31, 18], we believe that autonomic measures should be considered along with audio-visual data in the realm of affective computing, as they do not only provide complementary descriptions of affect, but can also be easily and continuously monitored with wearable sensors [30, 24, 4].

As baseline features, we extracted features from both ECG and EDA signals with overlapping (step of 40 ms) windows. The ECG signal was firstly band-pass filtered ($[3 - 27]$ Hz) with a zero-delay 6th order Butterworth filter [26], and 19 features were then computed: the zero-crossing rate, the four first statistical moments, the normalised length density, the non-stationary index, the spectral entropy, slope, mean frequency plus 6 spectral coefficients, the power in low frequency (LF, 0.04-0.15 Hz), high frequency (HF, 0.15-0.4 Hz) and the LF/HF power ratio. Additionally, we extracted the heart rate (HR) and its measure of variability (HRV) from the filtered ECG signal [26]. For each of those two descriptors, we computed the two first statistical moments, the arithmetic mean of rising and falling slope, and the percentage of rising values, which provided 10 features in total.

EDA reflects a rapid, transient response called skin conductance response (SCR), as well as a slower, basal drift called skin conductance level (SCL) [6]. Both, SCL (0–0.5 Hz) and SCR (0.5–1 Hz) are estimated using a 3rd order Butterworth filter, 8 features are then computed for each of those three low-level descriptors: the four first statistical moments from the original time-series and its first order derivate w.r.t. time.

4. CHALLENGE BASELINES

For transparency and reproducibility, we use standard and open-source algorithms for both sub-challenges. We describe below how the baseline system was defined and the results we obtained for each modality separately, as well as on the fusion of all modalities.

4.1 Depression

The challenge baseline for the depression classification sub-challenge is computed using the scikit-learn toolbox[4]. In particular, we fit a linear support vector machine with stochastic gradient descent, i. e. the loss is computed one sample at a time and the model is sequentially updated. We validated the model on the development set and conducted a grid search for optimal hyper-parameters on the development set of both the audio data and video data separately. Features of both modalities are taken from the provided challenge baseline features. Classification and training was performed on a frame-wise basis (i.e., at 100Hz for audio

[4] `http://scikit-learn.org/`

Table 3: Baseline results for depression classification. Performance is measured in F1 score for *depressed* and *not depressed* classes as reported through the PHQ-8. In addition, precision and recall are provided. Values for class *not depressed* are reported in brackets.

Partition	Modality	F1 score	Precision	Recall
Development	Audio	.462 (.682)	.316 (.938)	.857 (0.54)
Development	Video	.500 (.896)	.600 (.867)	.428 (.928)
Development	Ensemble	.500 (.896)	.600 (.867)	.428 (.928)
Test	Audio	.410 (.582)	.267 (.941)	.889 (.421)
Test	Video	.583 (.851)	.467 (.938)	.778 (.790)
Test	Ensemble	.583 (.857)	.467 (.938)	.778 (.790)

Table 4: Baseline results for depression severity estimation. Performance is measured in mean absolute error (MAE) and root mean square error (RMSE) between the predicted and reported PHQ-8 scores, averaged over all sequences.

Partition	Modality	RMSE	MAE
Development	Audio	6.74	5.36
Development	Video	7.13	5.88
Development	Audio-Video	6.62	5.52
Test	Audio	7.78	5.72
Test	Video	6.97	6.12
Test	Audio-Video	7.05	5.66

and 30Hz for video); temporal fusion was conducted through simple majority voting of all the frames within an entire screening interview. For both modalities we conducted a grid search for the following parameters: loss function $\in$ {logarithmic, hinge loss}, regularization $\in$ {L1, L2}, and $\alpha \in \{1e1, 1e0, \ldots, 1e-5\}$. For the audio data the optimal identified hyper-parameters are loss function = hinge loss, regularization = L1, and $\alpha = 1e-3$. For the video data the optimal identified hyper-parameters are loss function = logarithmic, regularization = L1, and $\alpha = 1e0$. The ensemble of audio and video was computed through a simple binary fusion of a logical AND. The test performance was computed on a classifier trained using the found optimal parameters from the grid search. Since the positive outputs of the video modality are a subset of those of the audio the ensemble classifier's performance is exactly the same as the video modality for both the development and test sets. Results are summarized in Table 3.

In addition to classification baseline, we also computed a regression baseline using random forest regressor. The only hyper-parameter in this experiment was the number of trees $\in 10, 20, 50, 100, 200$ in the random forest. For both audio and video the best performing random forest has trees = 10. Regression was performed on a frame-wise basis as the classification and temporal fusion over the interview was conduced by averaging of outputs over the entire screening interview. Fusion of audio and video modalities was performed by averaging the regression outputs of the unimodal random forest regressors. The performance for both root mean square error (RMSE) and mean absolute error (MAE) for development and test sets is provided in Table 4.

Table 5: Size of the window W in seconds used to extract features on the different modalities, and delay D in seconds applied to the gold-standard, according to the emotional dimension, i. e., arousal (A), and valence (V); parameters were obtained as the result of an optimisation of the performance measured as CCC on the development partition.

	Arousal		Valence	
Modality	W_A	D_A	W_V	D_V
Audio	4	2.8	6	3.6
Video-appearance	6	2.8	4	2.4
Video-geometric	4	2.4	8	2.8
ECG	4	0.4	10	2.0
HRHRV	8	0.0	8	0.0
EDA	8	0.0	10	0.4
SCL	4	0.0	14	2.4
SCR	4	0.8	14	0.8

4.2 Affect

Mono-modal emotion recognition was first investigated separately for each modality. Baseline features were extracted as previously described, with a window size W ranging from four to 14 seconds, and a step of two seconds. The window was centred, i. e., the first feature vector was assigned to the center of the window ($W/2$), and duplicated for the previous frames; the same procedure was applied for the last frames. For video data, frames for which the face was not detected were ignored. For EDA, SCL, and SCR, test data from the subject #7 was not used, due to issue during the recording of this subject (sensor was partially detached from the skin). Two different techniques were investigated to standardise the features: (i) online (standardisation parameters μ and σ are computed on the training partition and used on all partitions), and (ii) speaker dependent (μ and σ are computed and applied on features of each subject). In order to compensate time reaction of the raters, a time delay D is applied to the gold-standard, by shifting back in time the values of the time-series (last value was duplicated), with a delay ranging from zero to eight seconds, and a step of 400ms.

As machine learning, we used a linear Support Vector Machine (SVM) to perform the regression task with the liblinear library [12]; the L2-regularised L2-loss dual solver was chosen (option -s 12) and a unit bias was added to the feature vector (option -B 1), all others parameters were kept

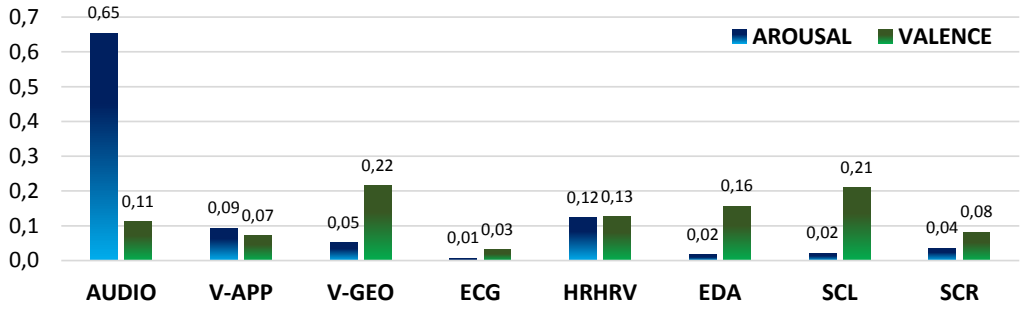

Figure 2: Percentage of contribution of each modality in the prediction of emotion; values are derived from the multimodal fusion model; V-APP: video appearance; V-GEO: video geometric; ECG: electrocardiogram; HRHV: heart rate and heart rate variability; EDA: electrodermal activity; SCL: skin conductance level; SCR: skin conductance resistance.

Table 6: Baseline results for affect recognition on the development (D) and test (T) partitions from audio, video (appearance and geometric), and physiologic (ECG, HRHRV, EDA, SCL, and SCR) feature sets, and their late fusion (multimodal). Performance is measured in Concordance correlation coefficient.

Modality	Arousal	Valence
D-Audio	.796	.455
D-Video-appearance	.483	.474
D-Video-geometric	.379	.612
D-ECG	.271	.153
D-HRHRV	.379	.293
D-EDA	.073	.194
D-SCL	.068	.166
D-SCR	.073	.085
D-Multimodal	**.821**	**.683**
T-Audio	.648	.375
T-Video-appearance	.343	.486
T-Video-geometric	.272	.507
T-ECG	.158	.121
T-HRHRV	.334	.198
T-EDA	.075	.228
T-SCL	.066	.216
T-SCR	.065	.145
T-Multimodal	**.683**	**.639**

to default. The complexity of the SVM was optimised in the range $[10^{-5} - 10^{0}]$. In order to compensate for scaling and bias issues in the predictions, but also noise in the data, we used the same post-processing chain as employed in [37]. The window size W and the time delay D were optimised by a grid search with an early stopping strategy, i. e., evaluations were stopped if no improvement was observed over the best score after two iterations. Experiments were always performed for both standardisation strategies, i. e., online and per speaker.

The best value of complexity, window size, time delay, and standardisation method were obtained by maximising the performance - measured as CCC - on the development partition with the model, learned on the training partition. Table 5 lists the best parameters for W and D, for each modality and emotional dimension, and shows that, the valence generally requires longer window size (to extract features) and time delay (to compensate for reaction time) than for arousal; $\bar{W}_A = 5.3$, $\bar{W}_V = 9.3$, $\bar{D}_A = 1.2$, $\bar{D}_V = 1.8$. Moreover, the results show that, a separate processing of features related to ECG, i. e., HRHV, and those related to EDA, i. e., SCL, and SCR, is justified as the best parameters obtained those signals differ from the ones obtained on their original signal. Regarding the standardisation technique, the online approach worked best for audio on both dimensions, and video data on valence, whereas standardisation of the features per subject worked best for all physiological features.

Mono-modal performance is reported in Table 6. Results show that, a significant improvement has been made for all modalities compared to the AVEC 2015 baseline [27], excepted for the EDA features. In agreement with the state-of-the-art, audio features perform significantly better than any other modality on arousal, and video features on valence. Interestingly, emotion prediction from the HRHRV signal performs significantly better than with the original ECG signal, and it is ranked as the most relevant physiological descriptor for arousal, when taken alone.

Multimodal emotion recognition is performed with three different late-fusion models, because frames might be missing on the video, and EDA related features; (i) audio-ECG, used for missing video and EDA; (ii) audio-ECG-EDA, used for missing video; (iii) audio-ECG-EDA-video, used otherwise. In order to keep the complexity low, and estimate the contribution of each modality in the fusion process, we build the fusion model by a simple linear regression of the predictions obtained on the development partition, using Weka 3.7 with default parameters [16]. Obtained predictions were then post-processed with the same approach used for the mono-modal predictions.

$$Pred_m = \epsilon_m + \sum_{i=1}^{N} \gamma_i * Pred_u(i), \tag{2}$$

where $Pred_u(i)$ is the mono-modal prediction of the modality i from the N available ones (ranging from two to eight), γ_i and ϵ_m are regression coefficients estimated on the development partition, and $Pred_m$ is the fused prediction.

Performance is reported in Table 6. Results show that, the baseline for the AVEC 2016 MASC is highly competitive, with the performance obtained on valence for the test partition being slightly better than the top-performer of AVEC

2015 [17]. In order to depict the contribution of each modality in the prediction of emotion, we normalised the linear regression coefficients that were learned for the multimodal fusion model (iv) into a percentage:

$$C_i = 100 * \frac{|\gamma_i|}{\sum_{k=1}^{N} |\gamma_k|}, \quad (3)$$

where C_i is the contribution of the modality i in percentage, and γ_k are the regression coefficients of the multimodal fusion model; $N = 8$.

Results show that, even if the mono-modal performance can be low for a given modality and emotion, e. g., EDA for arousal or SCR for valence, cf. Table 6, all modalities contribute, to a certain extent, to the prediction of arousal and valence in the fusion scenario, cf. Figure 2. This is especially the case for SCR features on arousal and for SCL features on valence, which did not perform well when used in isolation, but contribute even outperformed appearance features in the fusion model.

5. CONCLUSION

We introduced AVEC 2016 – the third combined open Audio/Visual Emotion and Depression recognition Challenge. It comprises two sub-challenges: the detection of the affective dimensions of arousal and valence, and the estimation of a self-reported level of depression. This manuscript describes AVEC 2016's challenge conditions, data, baseline features and results. By intention, we opted to use open-source software and the highest possible transparency and realism for the baselines by refraining from feature space optimisation and optimising on test data. This should improve the reproducibility of the baseline results.

Acknowledgements

The research leading to these results has received funding from the EC's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and the European Union's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA), and the Research Innovative Action No. 645378 (ARIA-VALUSPA), and No. 688835 (DE-ENIGMA).

6. REFERENCES

[1] T. R. Almaev and M. F. Valstar. Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition. In *Proc. of ACII*, pages 356–361, Geneva, Switzerland, September 2013. IEEE Computer Society.

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *Proc. of WACV*, Lake Placid (NY), USA, March 2016. IEEE.

[3] D. Bone, C.-C. Lee, and S. S. Narayanan. Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features. *IEEE Transactions on Affective Computing*, 5(2):201–213, April-June 2014.

[4] M. Chen, Y. Zhang, M. M. H. Yong Li, and A. Alarmi. AIWAC: Affective interaction through wearable computing and cloud technology. *IEEE Wireless Communications*, 22(1):20–27, February 2015.

[5] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[6] M. Dawson, A. Schell, and D. Filion. The electrodermal system. In J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, editors, *Handbook of psychophysiology*, volume 2, pages 200–223. Cambridge: Cambridge University Press, 2007.

[7] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. COVAREP – A collaborative voice analysis repository for speech technologies. In *Proc. of ICASSP*, pages 960–964, Florence, Italy, May 2014. IEEE.

[8] D. DeVault et al. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In *Proc. of AAMAS*, pages 1061–1068, Paris, France, May 2014. ACM.

[9] P. Ekman, W. V. Friesen, and J. Hager. *Facial action coding system*. Salt Lake City, UT: Research Nexus, 2002.

[10] F. Eyben et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 2015. to appear.

[11] F. Eyben, F. Weninger, F. Groß, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proc. of ACM MM*, pages 835–838, Barcelona, Spain, October 2013.

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, June 2008.

[13] J. Gratch et al. The distress analysis interview corpus of human and computer interviews. In *Proc. of LREC*, pages 3123–3128, Reykjavik, Iceland, May 2014. ELRA.

[14] M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *Proc. of ASRU*, pages 381–385, San Juan, Puerto Rico, November-December 2005. IEEE.

[15] N. Halko, P.-G. Martinsson, Y. Shkolnisky, and M. Tygert. An algorithm for the principal component analysis of large data sets. *Journal on Scientific Computing*, 33(5):2580–2594, October 2011.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, June 2009.

[17] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In *Proc. of AVEC, ACM MM*, pages 73–80, Brisbane, Australia, October 2015. ACM.

[18] D. Keltner and J. S. Lerner. Emotion. In S. Fiske, D. Gilbert, and G. Lindzey, editors, *Handbook of Social Psychology*, volume 1, pages 317–331. John Wiley & Sons Inc., 5th edition, 2010.

[19] R. B. Knapp, J. Kim, and E. André. Physiological signals and their use in augmenting emotion recognition for human–machine interaction. In

Emotion-Oriented Systems – The Humaine Handbook, pages 133–159. Springer Berlin Heielberg, 2011.

[20] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, and A. N. I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, January-March 2012.

[21] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, April 2009.

[22] L. Li. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, March 1989.

[23] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In *Prof. of FG*, pages 298–305, Santa Barbara (CA), USA, March 2011. IEEE.

[24] R. Picard. Affective media and wearables: surprising findings. In *Proc. of ACM MM*, pages 3–4, Orlando (FL), USA, November 2014. ACM.

[25] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller. Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion. In *Proc. of EmotiW, ICMI*, pages 473–480, Istanbul, Turkey, November 2014.

[26] 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*, 66:22–30, November 2015.

[27] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. AVEC 2015–The first affect recognition challenge bridging across audio, video, and physiological Data. In *Proc. of AVEC, ACM MM*, pages 3–8, Brisbane, Australia, October 2015. ACM.

[28] F. Ringeval, A. Sonderegger, B. Noris, A. Billard, J. Sauer, and D. Lalanne. On the influence of emotional feedback on emotion awareness and gaze behavior. In *Proc. of ACII*, pages 448–453, Geneva, Switzerland, September 2013. IEEE Computer Society.

[29] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. of EmoSPACE, FG*, Shanghai, China, 2013.

[30] A. Sanoa, R. W. Picard, and R. Stickgold. Quantitative analysis of wrist electrodermal activity during sleep. *International Journal of Psychophysiology*, 94(3):382–389, December 2014.

[31] S. Schachter. Cognition and peripheralist-centralist controversies in motivation and emotion. In M. S. Gazzaniga, editor, *Handbook of Psychobioogy*, pages 529–564. Academic Press Inc., 2012.

[32] S. Scherer et al. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658, October 2014.

[33] S. Scherer, G. Lucas, J. Gratch, A. Rizzo, and L.-P. Morency. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1):59–73, January-March 2015.

[34] B. Schuller et al. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. INTERSPEECH*, pages 148–152, Lyon, France, August 2013. ISCA.

[35] B. Schuller et al. The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In *Proc. INTERSPEECH*, pages 427–431, Singapore, Singapore, September 2014. ISCA.

[36] P. Shrout and J. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, March 1979.

[37] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. of ICASSP*, pages 5200–5204, Shanghai, China, March 2016. IEEE.

[38] M. Valstar. Automatic behaviour understanding in medicine. In *Proc. of RFMIR, ICMI*, pages 57–60, Istambul, Turkey, November 2014. ACM.

[39] M. Valstar, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. In *Proc. of FG*, Ljublijana, Slovenia, May 2015. IEEE.

[40] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014 – The Three Dimensional Affect and Depression Challenge. In *Proc. of AVEC, ACM MM*, Orlando (FL), USA, November 2014. ACM.

[41] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013 – The Continuous Audio / Visual Emotion and Depression Recognition Challenge. In *Proc. of AVEC, ACM MM*, pages 3–10, Barcelona, Spain, October 2013.

[42] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. of CVPR*, pages 532–539, Portland (OR), USA, 2013. IEEE.