AVEC 2017: real-life depression, and affect recognition workshop and challenge

Fabien Ringeval, Maja Pantic, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt

Angaben zur Veröffentlichung / Publication details:

Ringeval, Fabien, Maja Pantic, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, and Maximilian Schmitt. 2017. "AVEC 2017: real-life depression, and affect recognition workshop and challenge." In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge - AVEC '17, October 23 - 23, 2017, Mountain View, California, USA*, edited by Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, and Maja Pantic, 3–9. New York, NY: ACM Press. https://doi.org/10.1145/3133944.3133953.

AVEC 2017 – Real-life Depression, and Affect Recognition Workshop and Challenge

Fabien Ringeval
Univ. Grenoble Alpes, CNRS,
Grenoble INP, LIG
F-38000 Grenoble, France

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

Michel Valstar
University of Nottingham
Mixed Reality Lab
Nottingham, UK

Jonathan Gratch
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA

Roddy Cowie
Queen's University Belfast
Department of Psychology
Dublin, UK

Stefan Scherer
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA

Sharon Mozgai
University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA

Nicholas Cummins
University of Passau, Chair of
Complex & Intelligent Systems
Passau, Germany

Maximilian Schmitt
University of Passau, Chair of
Complex & Intelligent Systems
Passau, Germany

Maja Pantic†
Imperial College London
Department of Computing
London, UK

ABSTRACT

The Audio/Visual Emotion Challenge and Workshop (AVEC 2017) "Real-life depression, and affect" will be the seventh competition event aimed at comparison of multimedia processing and machine learning methods for automatic audiovisual 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 multimodal information processing and to bring together the depression and emotion recognition communities, as well as the audiovisual processing communities, to compare the relative merits of the various approaches to depression and emotion recognition from real-life data. This paper presents the novelties introduced this year, the challenge guidelines, the data used, and the performance of the baseline system on the two proposed tasks: dimensional emotion recognition (time and value-continuous), and dimensional depression estimation (value-continuous).

*The author is further affiliated with the Department of Computing, Imperial College London, London, UK and the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany.
†The author is further affiliated with Twente University, EEMCS, Twente, The Netherlands.

AVEC'17, October 23, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-5502-5/17/10…$15.00
DOI: https://doi.org/10.1145/3133944.3133953

CCS CONCEPTS

•General and reference → Performance; •Computing methodologies → Biometrics;

KEYWORDS

Affective Computing; Social Signal Processing; Automatic Emotion/Depression Recognition

1 INTRODUCTION

The 2017 Audio-Visual Emotion Challenge and Workshop (AVEC 2017) will be the seventh competition event aimed at comparison of multimedia processing and machine learning methods for automatic audiovisual analysis of emotion and depression, with all participants competing under strictly the same conditions [19, 26–28, 30, 31]. The goal of the Challenge is to compare the relative merits of the approaches for audiovisual 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. The main underlying motivation is the need to advance emotion recognition and depression estimation for multimedia retrieval to a level where behaviors expressed during human-human, or human-agent interactions, can be reliably sensed in real-life conditions, as this is exactly the type of data that applications would have to face in the real world.

AVEC 2017 shall help raise the bar for emotion and depression detection by challenging participants to estimate levels of depression and affect from audiovisual data captured in real-life conditions, and will continue to bridge the gap between research on emotion and depression recognition and low comparability of results.

1.1 Novelties and Challenge Guidelines

The Affect Sub-Challenge (ASC) is based on a novel database of human-human interactions recorded 'in-the-wild': SEWA[1] data set. Hence, audiovisual signals were not recorded with high-quality equipments and in dedicated laboratory rooms with ideal recording conditions, but in various places (e. g., home, work place) and with arbitrary personal equipments. In addition to those new challenging conditions tailored to real-life applications of affective computing technologies, we introduce the prediction of likability, along the usual (time- and value-continuous) emotional dimensions: arousal and valence. The Depression Sub-Challenge (DSC) is a refined re-run of the AVEC 2016 challenge [29], based on the DAIC-WOZ data set [10], and involving human-agent interactions; whereas the severity of depression was estimated as a binary task in AVEC 2016, we address this year the inference of the level of severity as a continuous-value.

- **Affect Sub-Challenge** (ASC) participants are required to perform fully continuous affect recognition of three affective dimensions: Arousal, Valence, and Likability, where the level of affect has to be predicted for every moment of the recording. The competition measure is the *concordance correlation coefficient* (*CCC*) [13], as previously used in the last two editions of AVEC [19, 29]; CCC evaluates the agreement between two time series by scaling their correlation coefficient with their mean square difference (1). Therefore, predictions that are well correlated with the gold-standard but shifted in value are penalised in proportion to the deviation [32]. Moreover, the intra-class correlation coefficient usually needs ANOVA assumptions while CCC does not [3].

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

  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.
- **Depression Sub-Challenge** (DSC): participants are required to assess the depression severity of the interviewed subject, where the target depression severity is based on the self-report PHQ-8 scores recorded prior to every human-agent interaction. For the DSC, performance in the competition will be measured using the *root mean square error* (*RMSE*). Participants in the competition, however, are also encouraged to provide classification output whether the participant also scored as *depressed* or *not depressed* according to the PHQ-8 score, i. e., score >= 10. In addition, participants are also encouraged to report on overall accuracy, correlation with the PHQ-8 score, average precision, and average recall to further analyse their results. As an additional novelty over the AVEC 2016 DSC, we encourage participants to provide symptom predictions, i. e., values of 0-3 for each of the eight questions on the PHQ-8 depression inventory[2].

Both Sub-Challenges allow contributors to find their own features to use with their regression algorithm. In addition, standard feature sets are provided for audio, video, and text separately, which participants are free to use. The labels of the test partition remain unknown to the participants, and participants have to stick to the definition of training, development, and test partition. They may freely report on results obtained on the development partition, but are limited to five trials per Sub-Challenge in submitting their results on the test partition. Ranking will rely on the scoring metric of each respective Sub-Challenge, i. e., *RMSE* for the DSC, and *CCC* for the ASC.

To be eligible to participate in the challenge, every entry has to be accompanied by a paper submitted to the AVEC 2017 Data Challenge and Workshop, and presenting the results and the methods that created them, which will undergo peer-review by the technical program committee. Only contributions with a relevant accepted paper will be eligible for challenge participation. The organisers will not participate in the Challenge themselves, but will re-evaluate the findings of the two best performing systems of each Sub-Challenge.

The remainder of this article is organised as follow: we introduce the corpus and baseline features for the ASC and the DSC in Section 2 and Section 3, respectively, baseline methods and results obtained for the two Sub-Challenges are then presented in Section 4, before concluding in Section 5.

2 AFFECT ANALYSIS CORPUS

The corpus used in the AVEC 2017 ASC is a subset of the Sentiment Analysis in the Wild (SEWA) database[1]. This data set consists of audiovisual recordings of subjects showing spontaneous and natural behaviours. All recordings were collected 'in-the-wild', i. e., using standard webcams and microphones from the computers in the subjects' offices or homes. The subset of the SEWA database exploited for the ASC is the video chat recording of German subjects.

Subjects participated in pairs and were given the task of discussing a commercial they have just viewed. The commercial was a 90 seconds long video clip advertising a tap. The participants were allowed to discuss arbitrary aspects of the commercial, e. g., if it was produced well, if it was too long, or the usefulness of the product itself. The maximum duration of the dyadic conversation was 3 minutes, but participants were free to stop the video chat at any time. Each conversational partner was required to know their chat partner beforehand (relatives, friends, or colleagues), in order to ensure an unreserved discussion. The data set was recorded using an online platform through the OpenTok API[3].

The subset of the SEWA database used for the ASC consists of 32 pairs in total, i. e., 64 subjects. The data is provided in three partitions (Training, Development, and Test), where both partners of one video chat appear in the same partition. Different combinations of gender are included, cf. Table 1. All subjects are between 18 and 60 years old. All video chats have been manually transcribed. Speaker turn timings have been further derived to know which subject is speaking when. Information on how to obtain shared data can be found in this location: http://sewaproject.eu. Data is freely available for research purposes.

[1] http://sewaproject.eu
[2] http://patienteducation.stanford.edu/research/phq.pdf
[3] https://tokbox.com

2.1 Emotion Analysis Labels

The video chat recordings were annotated time-continuously in terms of the emotional dimensions *arousal*, *valence*, and *liking*, i. e., how much a subject expresses a positive or a negative attitude while speaking, either with respect to the commercial, the advertised product, or any other matters discussed. The annotation process was conducted by 6 annotators (3 female, 3 male) aged between 20 and 24. No annotator is present in the database and all were German native speakers. Each dimension was annotated separately; the video chat recordings of each subject were shown in random order to each annotator, who was asked to rate the current expressed emotional dimension using a joystick on a continuous scale.

In order to create one unique *gold-standard* from the annotations, the six single annotations for each dimension were processed in the following manner. First, as the ratings from the joysticks are non-uniform, a Hermitian resampling to the final annotation sample rate (100 ms, 10 fps) was performed. The resulting contours are then normalised to the range of −1 to +1, based on the peak amplitude of the joysticks and median filtered (with a width of 3 samples). Then, in order to attenuate the effect of a different interpretation of the scale, the normalised and filtered ratings are standardised to the average standard deviation of all annotators.

One single gold-standard y_{EWE} for each audio-visual sequence n and dimension d is then formed exploiting the *evaluator weighted estimator* (*EWE*) approach based on the inter-rater agreement, similar to the basic approach described by Schuller [25]. The gold-standard is given by:

$$\mathbf{y}_{\text{EWE}_{n,d}} = \sum_{k=1}^{K} r_{n,d,k}\ \mathbf{y}_{n,d,k}, \tag{2}$$

where the index k denotes the annotator $k = 1, 2, \ldots, K$ and $\mathbf{y}_{n,d,k}$ denotes the pre-processed annotation values.

The annotator-specific weight $r_{n,d,k}$ is a different one for each sequence and dimension and computed as follows. First, the pairwise linear correlation coefficients (*CC*) $r'_{n,d,(k,k_i)}$ between the annotations of all raters k are computed, as well as the autocorrelation (obtained when $k = k_i$). Then, the mean pairwise correlation for each annotator k is computed using:

$$r'_{n,d,k} = \frac{1}{K} \sum_{k_i=1}^{K} r'_{n,d,(k,k_i)}. \tag{3}$$

The weight of the unreliable annotators, i. e., where $r'_{n,d,k} < 0$ is then set to zero ($r'_{n,d,k} = 0$). This ensures that negatively correlated annotations are not taken into account in the gold-standard [15]. The resulting coefficients are finally normalised with respect to their sum to obtain an annotator-specific weight:

$$r_{n,d,k} = \frac{1}{\sum_{k_i=1}^{K} r'_{n,d,k_i}} r'_{n,d,k}. \tag{4}$$

The average inter-rater agreement $\frac{1}{N} \sum_{n=1}^{N} r'_{n,d,k}$ for each annotator k and dimension d is given in Table 2.

2.2 Emotion Analysis Baseline Features

Below we describe the features that were extracted for the Affect Analysis sub-challenge.

Table 1: Pair distributions for the training and development partitions of the ASC dataset (subset of the SEWA database); FF: Female-Female; FM: Female-Male; MM: Male-Male.

Pair	Train	Devel
FF	4	2
FM	5	2
MM	8	3

Table 2: Average inter-rater agreement $r'_{n,d,k}$ of each annotator of the ASC dataset (subset of the SEWA database) for each emotional dimension (arousal, valence, and liking); values were computed before sorting out negatively correlated annotators.

Annotator	Arousal	Valence	Liking
Female 1	.405	.491	.548
Female 2	.381	.385	.447
Female 3	.481	.524	.478
Male 1	.353	.406	.429
Male 2	.392	.497	.516
Male 3	.406	.486	.554

2.2.1 Audio Features. Different acoustic feature sets are included in the challenge package, all of them based on the extended version of the *Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) feature set [7]. eGeMAPS is an expert-knowledge based feature set consisting of 23 acoustic *low-level descriptors* (LLDs) extracted every 10 ms over a short-term frame. The LLDs set consists of energy, spectral and cepstral features, pitch, voice quality, and micro-prosodic features. This feature set has been successfully used in many affect related prediction tasks [17, 18, 22].

As the LLDs capture only very local information in time, a segment-level representation of the features is required, especially when static predictors are used, such as *Support Vector Machine* (SVMs). In the ASC, two different types of segment-level features are provided, using the eGeMAPS LLDs: *functionals* (as defined in the eGeMAPS feature set) and *bag-of-audio-words* (BoAW). The latter feature type was introduced for text features originally, but has been successfully applied also to others modalities, such as the audio and the video domain [11, 23]. In the BoAW framework, the LLDs over a certain segment are first quantised using predefined templates of a codebook of 'audio words' and then, an histogram of the audio words occurring in the corresponding segment is created.

Different methods can be employed to create a codebook of audio words [24]. For the provided BoAW features, a random sampling of the LLDs present in the training partition was employed. One important parameter to define is the codebook size, i. e., the number of audio words, which is set to 1 000 for the baseline BoAW acoustic features. As a pre-processing step, the LLDs are standardised to zero mean and unit variance prior to vector quantisation. As a post-precessing step, the term frequencies in the BoAW are logarithmised in order to compress their numerical range.

The extraction of the LLDs and the functionals was done using the openSMILE[4] toolkit [8]; for the BoAW generation, the openXBOW[5] toolkit [24] was employed. Both segment-level acoustic feature types were computed over segments of 6 seconds. The given timestamps correspond to the centres of each segment. Overall, the acoustic baseline feature sets with functionals contain 88 features, the BoAW features contain 1 000 features.

2.2.2 Video Features. The video feature set consists of three types of features related to the position and expression of the subjects' face:

(1) Face orientation (pitch, yaw, and roll) in degrees (3 features)
(2) Pixel coordinates for 10 eye points (x and y coordinates, i. e., 20 features in total)
(3) Pixel coordinates for 49 facial landmarks (x and y coordinates, i. e., 98 features in total)

The facial features have been extracted for each video frame (frame step 20 ms) using the Chehra face tracker [1]. In addition to the raw features, a normalised version of them is provided where all coordinates, x and y respectively, are standardised to zero mean and unit variance on frame level. This step removes the influence of the position of the face within the video image on the final features.

To obtain a segment-level representation, *bag-of-video-words* (BoVW) were computed from the normalised facial features using the same segment lengths as for the BoAW (6 seconds). The process to generate the BoVW is the same as described above, however, separate codebooks and histograms were created for each three facial feature type, with a codebook size of 1 000 each, resulting in a final segment-level feature vector of length 3 000.

2.2.3 Text Features. In addition to audio and video features, a bag-of-words feature representation based on the transcription of the speech are generated with openXBOW and used as additional features. The dictionary for these textual features is learnt from the training partition taking only the terms with at least two occurrences into account. This results in a dictionary of 521 words, where only unigrams are considered. As for the acoustic and facial features, the histograms are created over a segment of 6 *s* in time and the logarithm is taken from the term frequencies. In total, the *bag-of-text-words* (BoTW) features contain 521 features.

3 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) [10], 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 non-verbal indicators of mental illness [6]. 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.

[4]http://audeering.com/technology/opensmile/
[5]https://github.com/openXBOW/openXBOW

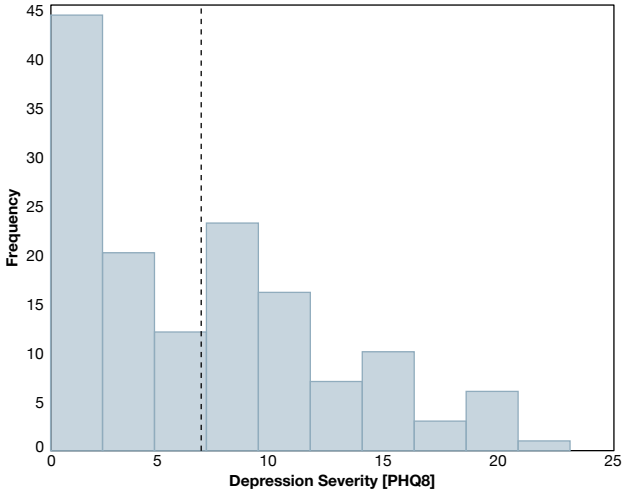

Figure 1: Histogram of depression severity scores for DSC challenge (training and development sets). Mean score of depression is shown as a vertical dashed line.

Further, we provide the scores that every individual provided on the PHQ-8 depression inventory to allow participants to better model the varied characteristics and symptoms of depression [4]. 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.

3.1 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 [12]. The average depression severity on the training and development set of the challenge is $M = 6.67$ ($SD = 5.75$) out of a maximum score of 24. The distribution of the depression severity scores based on the challenge training and development set is provided in Figure 1. A baseline regression model that constantly predicts the mean score of depression provides an $RMSE = 5.73$ and an $MAE = 4.74$.

3.2 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. For ethical reasons, no raw video is made available.

3.2.1 Audio Features. For the audio features we utilised COVAREP (v1.3.2), a freely available open source Matlab and Octave toolbox for speech analyses [5]. The toolbox[6] 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 [20, 21]. In particular, we extracted the following features:

[6]http://covarep.github.io/covarep/

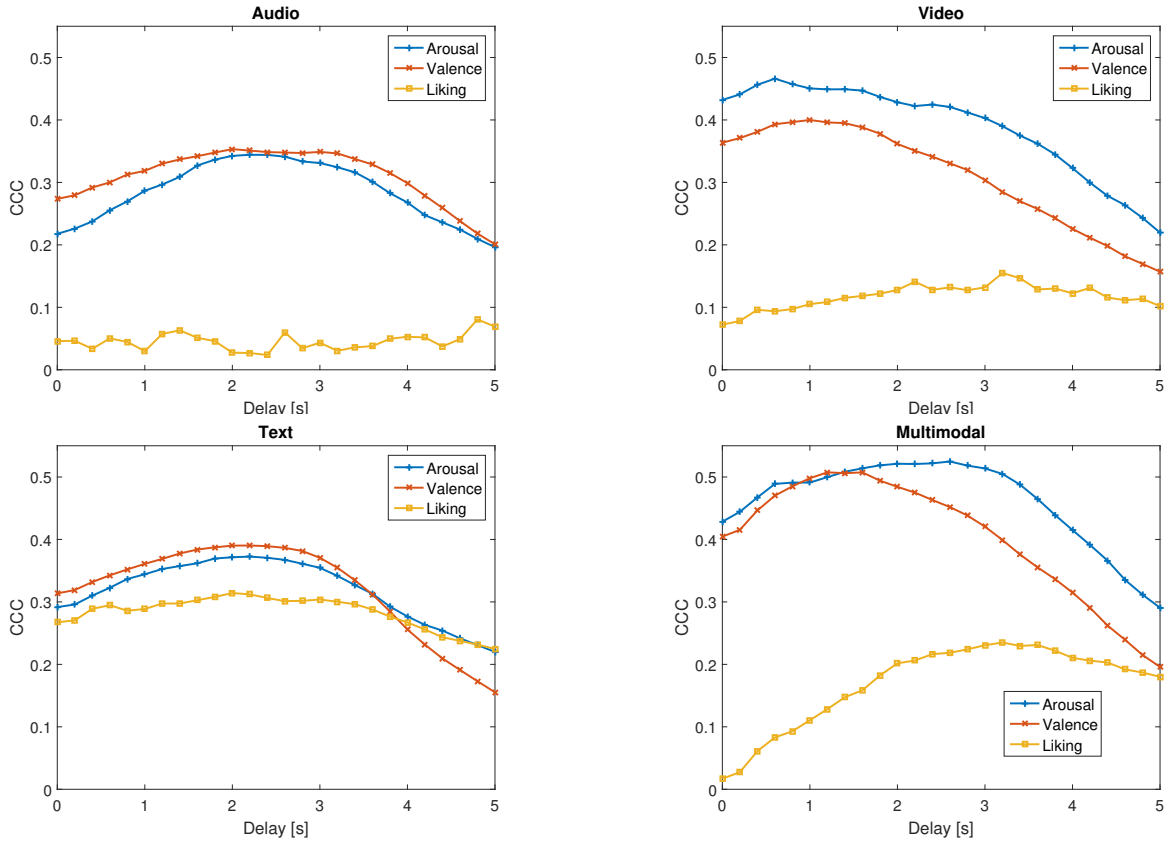

Figure 2: Results for the Development partition of the ASC corpus. Results are displayed over different delays for each modality (i. e., audio, video, text, and all) and dimension (i. e., arousal, valence and liking)

.

- Prosodic: Fundamental frequency (F0) and voicing (VUV)
- Voice quality: Normalised 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 (peak-slope), 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.

3.2.2 Video Features. Based on the *OpenFace*[7] framework [2], we provide different types of video features[8]:

- Facial landmarks: 2D and 3D coordinates of 68 points on the face, estimated from video
- Histogram of oriented gradients (HOG) features on the aligned 112x112 area of the face
- Gaze direction estimate for both eyes
- Head pose: 3D position and orientation of the head
- Action units (AUs): {AU01, AU02, AU04, AU05, AU06, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU25, AU26}

4 CHALLENGE BASELINES

For transparency and reproducibility, we use standard and open-source algorithms for both Sub-Challenges; SCIKIT-LEARN toolbox[9]. We describe below how the baseline system was defined and the results we obtained for each modality, as well as on the fusion of all modalities.

4.1 Emotion

An emotion recognition baseline system is obtained using the BoAW, BoVW, and BoTW features with a segment length of 6 seconds described above and a Support Vector Regression (SVR).

[7] https://github.com/TadasBaltrusaitis/OpenFace
[8] For additional information consult the challenge manual provided after entering the challenge.
[9] http://scikit-learn.org/

Generally, time-continuous annotations obtained in real-time suffer from a certain delay as the annotators cannot react on changes in the shown emotion immediately [14, 16]. In order to compensate for the reaction time of the annotators, the features are shifted ahead for a variable delay of $[0, 5]$ seconds with stride $0.2s$. A corresponding number of feature vectors at the end of each recording is dropped, while the 'missing' feature vectors at the beginning of each recording are filled with copies of the first feature vector.

To train the SVR, the LIBLINEAR library [9] through the Python machine learning framework is used with default options. The complexity parameter of SVM, as well as the delay, is tuned on the development set in the range $c \in \{2^{-15}, 2^{-14}, \ldots, 2^{}\}$.

The results of this analysis, in terms of *CCC* for the features of each single domain and multimodal (cross-modal) bag-of-words (early fusion), are presented in Table 3. Complexity and delay have been optimised on the Development partition separately for each dimension. Figure 2 shows the *CCC* on the Development partition for different delays. To define a baseline, the SVR is trained on the fusion of Training and Development partition with the complexity and delay that is optimum on the Development partition for the corresponding case. For each emotional dimension, we picked the modality (audio, video, text, or multimodal) that provides the highest *CCC* on the Test partition. This required the organisers to complete four trials under challenge conditions, one less than the five trials the participants of the challenge have.

4.2 Depression

We computed the depression severity baseline using random forest regression. The only hyper-parameter in this experiment was the number of trees $\in \{10, 20, 50, 100, 200\}$. For both audio and video the best performing random forest has 10 trees. 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 3: Baseline results for emotion recognition on the Development (D) and Test (T) partitions from audio, video, and text feature sets, and their early fusion (multimodal). Performance is measured in terms of the *Concordance Correlation Coefficient* (*CCC*). Best performance obtained on the test partition for each dimension, i. e., the ASC baseline performance, is highlighted in bold format.

Modality	Arousal	Valence	Liking
D-Audio	.344	.351	.081
D-Video	.466	.400	.155
D-Text	.373	.390	.314
D-Multimodal	.525	.507	.235
T-Audio	.225	.244	-.020
T-Video	.308	.455	.002
T-Text	**.375**	.425	**.246**
T-Multimodal	.306	**.466**	.048

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. Best performance obtained on the test partition, i. e., the DSC baseline performance, is highlighted in bold format.

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

5 CONCLUSION

We introduced AVEC 2017 – the fourth combined open Audio/Visual Emotion and Depression Severity Assessment Challenge. It comprises two Sub-Challenges: (i) ASC, where the level of affective dimensions of arousal, valence, and - for the first time - likability has to be infered from audiovisual data collected 'in-the-wild' during human-human interactions, and (ii) DSC, where a self-reported level of depression needs to be estimated from audiovisual data collected during human-machine interactions. This manuscript described AVEC 2017'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, using less number of trials as given to participants for reporting results on the test partition. In addition, baseline scripts have been made available in the data repositories, which should help improving the reproducibility of the baseline results.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), the Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA), and the Research Innovative Action No. 645378 (ARIA-VALUSPA), and No. 688835 (DE-ENIGMA). The data provided for the depression severity Sub-Challenge was sponsored by DARPA and the Army Research Laboratory under contracts W911NF-04-D-0005 and W911NF-14-D-0005. The authors further thank the sponsors of the challenge – audEERING GmbH and the Association for the Advancement of Affective Computing (AAAC).

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Government, and no official endorsement should be inferred.

REFERENCES

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. 2014. Incremental face alignment in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE, 1859–1866.

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: an open source facial behavior analysis toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV*. IEEE, Lake Placid, NY, USA.

[3] Chia-Cheng Chen and Huiman X. Barnhart. 2013. Assessing agreement with intraclass correlation coefficient and concordance correlation coefficient for data with repeated measures. *Journal of Computational Statistics & Data Analysis* 60 (April 2013), 132–145.

[4] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication* 71 (July 2015), 10–49.

[5] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP – A collaborative voice analysis repository for speech technologies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. IEEE, Florence, Italy, 960–964.

[6] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jonathan Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Alberto Rizzo, and Louis-Philippe Morency. 2014. SimSensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS'14*. ACM, Paris, France, 1061–1068.

[7] Florian Eyben, Klaus R. Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (April-June 2016), 190–202.

[8] Florian Eyben, Felix Weninger, Florian Groß, and Björn Schuller. 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the ACM International Conference on Multimedia, MM'13*. Barcelona, Spain, 835–838.

[9] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research* 9 (June 2008), 1871–1874.

[10] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Albert Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC*. ELRA, Reykjavik, Iceland, 3123–3128.

[11] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. 2014. Predicting emotions in user-generated videos. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI'14*. Association for the Advancement of Artificial Intelligence, Québec, Canada, 73–79.

[12] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders* 114, 1-3 (April 2009), 163–173.

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

[14] Soroosh Mariooryad and Carlos Busso. 2015. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing* 6 (April-June 2015), 97–108.

[15] Arianna Mencattini, Eugenio Martinelli, Fabien Ringeval, Björn Schuller, and Corrado Di Natale. 2016. Continuous estimation of emotions in speech by dynamic cooperative speaker models. *IEEE Transactions on Affective Computing* (2016). 14 pages, to appear.

[16] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. 2012. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI'12*. ACM, Santa Monica, CA, USA, 501–508.

[17] Fabien Ringeval, Erik Marchi, Charline Grossard, Jean Xavier, Mohamed Chetouani, David Cohen, and Björn Schuller. 2016. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In *Proceedings of INTERSPEECH*. ISCA, San Francisco, CA, USA, 1210–1214.

[18] Fabien Ringeval, Erik Marchi, Marc Méhu, Klaus Scherer, and Björn Schuller. 2015. Face reading from speech – Predicting facial action units from audio cues. In *Proceedings of INTERSPEECH*. ISCA, Dresden, Germany, 1977–1981.

[19] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. AV+EC 2015 – The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC'15, ACM International Conference on Multimedia, MM'15*. ACM, Brisbane, Australia, 3–8.

[20] Stefan Scherer, Gale Lucas, Jonathan Gratch, Alberto Rizzo, and Louis-Philippe Morency. 2015. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing* 7, 1 (January-March 2015), 59–73.

[21] Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert (Skip) Rizzo, and Louis-Philippe Morency. 2014. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing* 32, 10 (October 2014), 648–658.

[22] Maximilian Schmitt, Erik Marchi, Fabien Ringeval, and Björn Schuller. 2016. Towards cross-lingual automatic diagnosis of autism spectrum condition in childrenfis voices. In *Proceedings of the 14th ITG Conference on Speech Communication, volume 267 of ITG-Fachbericht*. ITG/VDE, IEEE, Paderborn, Germany, 264–268.

[23] Maximilian Schmitt, Fabien Ringeval, and Björn Schuller. 2016. At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech. In *Proceedings of INTERSPEECH*. ISCA, San Francisco, CA, USA, 495–499.

[24] Maximilian Schmitt and Björn W. Schuller. 2016. openXBOW – Introducing the Passau open-source crossmodal Bag-of-Words toolkit. *preprint arXiv:1605.06778* (2016).

[25] Björn Schuller. 2013. *Intelligent Audio Analysis*. Springer.

[26] Björn Schuller, Michel Valstar, Florian Eyben, Roddy Cowie, and Maja Pantic. 2012. AVEC 2012 – The continuous Audio/Visual Emotion Challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI'12*. ACM, Santa Monica, CA, USA, 449–456.

[27] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. 2011. AVEC 2011 – The First International Audio/Visual Emotion Challenge. In *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction, ACII 2011*, Vol. II. Springer, Memphis, TN, USA, 415–424.

[28] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Roddy Cowie, and Maja Pantic. 2016. Summary for AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 24th ACM International Conference on Multimedia, MM'16*. ACM, Amsterdam, The Netherlands, 1483–1484.

[29] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016 – Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC'16, ACM International Conference on Multimedia, MM'16*. ACM, Amsterdam, The Netherlands, 3–10.

[30] Michel Valstar, Björn Schuller, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2013. Workshop summary for the 3rd international Audio/Visual Emotion Challenge and workshop (AVEC'13). In *Proceedings of the 21st ACM International Conference on Multimedia, MM'13*. ACM, Barcelona, Spain, 1085–1086.

[31] Michel Valstar, Björn Schuller, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. AVEC 2014: The 4th international Audio/Visual Emotion Challenge and workshop. In *Proceedings of the 22nd ACM International Conference on Multimedia, MM 2014*. ACM, Orlando, FL, USA, 1243–1244.

[32] Felix Weninger, Fabien Ringeval, Erik Marchi, and Björn Schuller. 2016. Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI'16*. IJCAI/AAAI, New York City, NY, USA, 2196–2202.