AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition

Fabien Ringeval
Université Grenoble Alpes, CNRS
Grenoble, France
fabien.ringeval@imag.fr

Björn Schuller*
University of Augsburg
Augsburg, Germany
schuller@ieee.org

Michel Valstar
University of Nottingham
Nottingham, UK
michel.valstar@nottingham.ac.uk

Roddy Cowie
Queen's University Belfast
Belfast, UK

Heysem Kaya
Namık Kemal University
Çorlu, Turkey

Maximilian Schmitt
University of Augsburg
Augsburg, Germany

Shahin Amiriparian
University of Augsburg
Augsburg, Germany

Nicholas Cummins
University of Augsburg
Augsburg, Germany

Denis Lalanne
University of Fribourg
Fribourg, Switzerland

Adrien Michaud
Université Grenoble Alpes, CNRS
Grenoble, France

Elvan Çiftçi
University of Health Sciences
Istanbul, Turkey

Hüseyin Güleç
University of Health Sciences
Istanbul, Turkey

Albert Ali Salah†
Boğaziçi University
Istanbul, Turkey

Maja Pantic‡
Imperial College London
London, UK

ABSTRACT

The Audio/Visual Emotion Challenge and Workshop (AVEC 2018) "Bipolar disorder, and cross-cultural affect recognition" is the eighth competition event aimed at the comparison of multimedia processing and machine learning methods for automatic audiovisual health and emotion analysis, with all participants competing strictly under 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 health and emotion recognition communities, as well as the audiovisual processing communities, to compare the relative merits of various approaches to health and emotion recognition from real-life data. This paper presents the major novelties introduced this year, the challenge guidelines, the data used, and the performance of the baseline systems on the three proposed tasks: bipolar disorder classification, cross-cultural dimensional emotion recognition, and emotional label generation from individual ratings, respectively.

CCS CONCEPTS

• **General and reference → Performance**;

KEYWORDS

Affective Computing; Bipolar Disorder; Cross-Cultural Emotion

ACM Reference Format:
Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, Elvan Çiftçi, Hüseyin Güleç, Albert Ali Salah, and Maja Pantic. 2018. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In *2018 Audio/Visual Emotion Challenge and Workshop (AVEC'18), October 22, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3266302.3266316

1 INTRODUCTION

The Audio/Visual Emotion Challenge and Workshop (AVEC 2018) is the eighth competition aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual, and audiovisual health and emotion sensing, with all participants competing strictly under the same conditions [59, 61, 71, 72, 78, 80, 81].

One of the goals of the AVEC series is to bring together multiple communities from different disciplines, in particular the multimedia communities and those in the psychological and social sciences who study expressive behaviour. Another objective is to provide a common benchmark test set for multimodal information processing, to compare the relative merits of the approaches to automatic health and emotion analysis under well-defined conditions, i. e. , with large volumes of un-segmented, non-prototypical and non-preselected data of fully naturalistic behaviour.

*The author is further affiliated with Imperial College London, London, UK.
†The author is further affiliated with Nagoya University, Nagoya, Japan.
‡The author is further affiliated with University of Twente, Twente, The Netherlands.

AVEC'18, October 22, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5983-2/18/10…$15.00
https://doi.org/10.1145/3266302.3266316

AVEC 2018 is themed around two topics: bipolar disorder (for the first time in a challenge), and emotion recognition. Major novelties are introduced this year with three separated Sub-challenges focusing on health and emotion analysis: (i) Bipolar Disorder Sub-challenge (BDS), (ii) Cross-cultural Emotion Sub-challenge (CES), and (iii) Gold-standard Emotion Sub-challenge (GES). We describe in the following the novelties introduced in the Challenge and the Challenge guidelines.

The Bipolar Disorder Sub-challenge (BDS) is a new challenge, the first of its kind in the scope of mental health analysis. Whereas the topic of depression analysis was featured in the previous editions of AVEC [59, 78], we introduce this year the analysis of the manic episodes of bipolar disorder (BD), using the BD corpus [11]. In the BDS, participants have to classify patients suffering from BD and admitted to a hospital after a mania episode into three categories of mania, hypo-mania, and remission, following the Young Mania Rating Scale (YMRS) [85]. The participants estimate the classes from audiovisual recordings of structured interviews recorded periodically from the day of admittance to discharge, and the performance is measured by the Unweighted Average Recall (UAR), which is the average of the recall in percentage obtained on each of the three classes.

The Cross-cultural Emotion Sub-challenge (CES) is a major extension of the Emotion Sub-challenge previously run in AVEC 2017 [60], where dimensional emotion recognition was performed on data collected '*in-the-wild*' by the German participants of the SEWA dataset[1]; audiovisual signals were recorded in various places, e. g. , home, work place, and with arbitrary personal equipments, thus providing noisy but realistic data. For the AVEC 2018 CES, an extended version of the SEWA dataset, with new data collected in the same conditions from Hungarian participants, is used as a blind test set for the first ever cross-cultural emotion recognition competition task: participants have to predict the level of three emotional dimensions (*arousal*, *valence*, and *liking*) time-continuously from audiovisual recordings of dyadic interactions, and performance is the *total Concordance Correlation Coefficient (CCC)* [45, 83] averaged over the dimensions.

The Gold-standard Emotion Sub-challenge (GES) is a new task focusing on the generation of dimensional emotion labels. The task consists in the fusion of time-continuous annotations of dimensional emotions consistently provided by several annotators. The obtained labels are then used to train and evaluate a baseline multimodal emotion recognition system on the RECOLA dataset [63], which includes audiovisual and physiological recordings of dyadic interactions from French speaking subjects, and annotations data from six gender-balanced external annotators with the same mother tongue. The performance obtained by the baseline system on the test partition, for which the labels are generated by the algorithm provided by the participants in a blind manner, is used for ranking the contribution with the *averaged CCC* [83]. In order to avoid a simple system ignoring the annotations but performing well, visual inspection and statistical analysis is performed on the labels to ensure that the original variance found in the annotations is sufficiently well preserved.

All Sub-challenges allow contributors to find their own features to use with their own machine learning algorithm. In addition, standard feature sets are provided for audio and video data (cf. section 4), along with scripts available in a public repository[2], which participants are free to use for reproducing both the baseline features and recognition systems. 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 relies on the scoring metric of each respective Sub-challenge, i. e. , UAR for the BDS, *total CCC* for the CES, and *averaged CCC* for the GES. The UAR better reflects classification performance when the distribution of instances in the classes is not balanced, and is thus used as ranking metric for the BDS. Regarding the two regression tasks (CES and GES), the CCC is preferred over the Pearson's Correlation Coefficient (PCC), because it is sensitive to bias and scaling, and permit discriminative training when used as cost function [83]. The *averaged CCC* differs from the *total CCC* in that the former necessarily enforces accurate prediction of the target contour within each sequence, while the latter could assign a good score to over-smoothed regression outputs that only predict the average label right [76, 83]. Therefore, we used the *total CCC* as metric for the CES, as an overall accurate prediction is desired for the cross-cultural paradigm, and the *averaged CCC* for the GES, as the focus is on a detailed description of the labels.

To be eligible to participate in the challenge, every entry had to be accompanied by a paper submitted to the AVEC 2018 Data Challenge and Workshop, describing the results and the methods that created them. These papers underwent peer-review by the technical program committee. Only contributions with a relevant accepted paper and at least a submission of test results were eligible for challenge participation. The organisers did not participate in the Challenge themselves, but re-evaluated the findings of the best performing system of each Sub-challenge.

The remainder of this article is organised as follows. We summarise relevant related work in Section 2, introduce the challenge corpora in Section 3, the common audiovisual baseline feature sets in Section 4 and the developed baseline recognition systems with the obtained results in Section 5, before concluding in Section 6.

2 RELATED WORK

We summarise below the state-of-the-art in the automatic analysis of affect with a focus on: (i) the population of BD, (ii) dimensional analysis in cross-cultural paradigms, and (iii) generation of emotion labels from individual annotations.

2.1 Bipolar Disorder

BD is a serious, persistent (possibly lifelong) mental health disorder, with subjects experiencing intermittent episodes of mania and depression that can last from days to months, and cause unusual shifts in mood, energy, activity levels, and the ability to carry out day-to-day tasks. While one in four people in the world will be affected by mental or neurological disorder at some point in their

[1] https://db.sewaproject.eu/

[2] https://github.com/AudioVisualEmotionChallenge/AVEC2018

lives [54], BD lifelong prevalence is about 2.1 % worldwide [49], and suicidal risk amongst BD patients is around 15 %. Consequently, World Health Organisation has ranked BD among the top ten diseases for young adults according to the Disability-Adjusted Life Year (DALY) indicator [55].

Access to treatment and treatment resistance have been identified as one of the big challenges for medical care in BD [8]. In addition, many patients with BD are seen exclusively in primary care [43], due to a lack of access to specialised mental health care services, and the stigma associated with receiving care in a mental health setting [42, 57]. Recent advances in the automatic recognition of human behaviours from multimodal data [7, 31] constitute thus promising avenues for facilitating the monitoring of patients suffering from BD. Some studies have investigated the use of automated methods for the analysis of BD.

Multimodal data passively collected from smartphones, such as location, distance travelled, conversation frequency, and non-stationary duration, by seven BD patients for four weeks, were used to assess the Social Rhythm Metric (SRM) in [2], which is a measure of stability and periodicity for individuals with BD [29]. Authors reported that the SRM score (0-7) could be predicted by Support Vector Machines (SVMs) with a Root Mean Square Error (RMSE) of 1.40, and further improved to 0.92 when using personalised models.

Another study exploited data collected from smartphones for twelve weeks to classify 28 patients into manic or mixed states [25], following the YMRS [85]. In addition to metadata, such as the number of text messages and phone calls emitted per day, they used speech features (6.5k acoustic measures computed with the openSMILE toolkit [22] – *emolarge* feature set) collected during phone calls. Random forests were exploited to perform the classification task and speech features provided the best performance with an area under the curve (AUC) of 0.89.

In [39], long-term monitoring (6-12 months) of mood of 43 subjects suffering from BD was investigated from speech data collected with cellular phones. The authors further developed a program termed PRIORI (Predicting Individual Outcomes for Rapid Intervention), to analyse acoustics of speech as predictors of mood state from mobile smartphone data [41]. Subject-specific mood variation was captured in an i-vector space and then used with SVMs as features to classify euthymia vs. depression state of BD patients. The system performed best with a soft decision fusion and obtained an AUC of 0.78.

2.2 Cross-cultural Emotion Recognition

The other theme of AVEC 2018 – obviously – relies on emotion, but on two different facets never explored in a challenge: (i) cross-cultural inference of emotional dimensions *in-the-wild* and (ii) generation of emotional labels from individual ratings. The generalisability of recognising emotion expression across different cultures has continually been highlighted as an open research challenge in affective computing [19, 20, 56]. Indeed, it was observed in a recent meta-review into multimodal affect detection systems [17], that only minimal research efforts have been made into assessing the performance and robustness of state-of-the-art affective computing approaches across cultures.

A common idiom in facial expression recognition is that, due in part to consistencies across all humans in terms of the make-up of our facial muscles [12], emotional expressions – especially within the basic emotions of happiness, sadness, fear, anger, disgust, and surprise – have a large degree of universality across cultures [12, 18]. However, it is very difficult to find works in the affective computing recognition literature, particularly for dimensional affect representations, which support this claim [17]. Furthermore, it has also been argued that the results of facial expression perception studies can be easily biased by the manner in which the answers are elicited [65].

Despite the apparent complexity, some efforts have been made for the empirical evaluation of emotion recognition systems' performance in cross-cultural settings. Accuracies achieved in such settings are, in general, less than those achieved within-culture, and training with cultures from similar language families has been shown to improve the accuracy [28, 67]. Moreover, as cross-culture testing is more often performed in a cross-corpus scenario, transfer learning [86, 87] and domain adaptation techniques [40, 66] can be leveraged to deal with aspects such as channel effects and covariate shift [87].

2.3 Generation of Emotion Labels

Emotion recognition, as many other supervised machine learning tasks, requires a large amount of labels of sufficient quality to train systems learning the appropriate mapping between input data – usually collected from sensors like web-cameras – and the labels describing the emotion. Because those labels are at the core of the machine learning, they need a careful attention in their definition, especially when they consist in human judgements of human behaviours, which are by nature, highly variable and subjective [77]. Whereas humans seem to be more efficient at discriminating among options than assigning absolute values to subjective variables [50, 84], the dominant approach in affect modelling relies on absolute values of dimensional attributes, such as *arousal* or *valence* [64], that are annotated time-continuously over the recordings by a pool of annotators using a tool like Feeltrace [13].

Even though individual ratings of emotion can be modelled all together in a multi-task learning framework [58], the dominant practice is to summarise the annotations for each recording into a single time-series, referred as *gold-standard* in the literature[3], that can be then easily processed by any machine learning algorithm. However, many difficulties arise during the fusion of the individual annotations, as inconsistencies appear between values reported by the annotators [77], and a delay is present between the emotional event expressed in the data and the corresponding annotation value, which is due to the reaction time and other related cognitive factors.

Some studies have investigated methods to process noisy time-continuous labels reported by humans on dimensional attributes of emotion, in order to learn more reliable predictive models of human behaviours. A notable work is the winning contribution of the AVEC 2012 Challenge [71], where authors proposed to maximise the PCC between audiovisual features and the gold-standard to

[3]The term *gold-standard* is preferred over the more common terminology of *ground-truth*, as there is not a direct and unique mean to measure the expressed emotions in a natural interaction; only subjective evaluations of those emotions are available, hence there is no *truth* in the definition of emotion labels.

estimate the delay used to compensate the reaction time of annotators [53]. This approach was further investigated by maximising the Mutual Information (MI) instead of the PCC, which was reported to be less reliable and more sensitive to noise present in the data in comparison with MI [47, 73]. Some participants of the AVEC 2015 Challenge [61] proposed to estimate an overall reaction time for each emotional dimension by maximising the recognition performance while varying the delay in a grid search [35, 36]; this method has been used since then in the baseline system of the AVEC Challenge as reaction time compensation.

Dynamic Time Warping (DTW) [51], which is a popular method to perform monotonic time-alignment between two time-series, has also been successfully employed to compensate reaction time using different variants of the method, such as Canonical Time Warping [26, 52], Generalised Time Warping [27], and Deep Canonical Time Warping [75]. Additional comparative rank-based information with triplet embedding [82] was also proposed for warping dimensional annotations in [10].

Alternative strategies explored in the literature consist in minimising the distortions amongst the annotators using the expectation-maximisation algorithm [33], smoothing the annotation values [74], or exploiting measures of inter-rater agreement as joint predictive information [34]. In addition, it is worth to note that, as emotions are dynamic by nature, the prediction of a change in emotion has been proven to perform better than predicting the value itself in some cases [37, 46, 48].

3 CHALLENGE CORPORA

The AVEC 2018 Challenge relies on three corpora: (i) the BD corpus [11] for the BDS, (ii) SEWA dataset[1] for the CES, and (iii) RECOLA dataset[4] [63] for the GES. We provide below a short overview of each dataset and refer the reader to the original work for a more complete description.

3.1 Bipolar Disorder Corpus

The BD corpus [11] used for the AVEC 2018 BDS includes audiovisual recordings of structured interviews performed by 46 Turkish speaking subjects. All the subjects suffered from BD and were recruited from a mental health service hospital, where they were diagnosed by clinicians following DSM-5's inclusion criteria [1]. Only bipolar patients at mania episode were included in the study, while being in depressive episode was part of the exclusion criteria. Other exclusion criteria included being younger than 18 years or older than 60 years, showing low mental capacity during interview, expression of hallucinations and disruptive behaviours during interview, presence of severe organic disease and diagnosis of substance or alcohol abuse in the last three months. Participants of the BD corpus were asked to complete seven tasks guided by a presentation, inspired by the collection of the Audio-Visual Depressive Language Corpus (AVDLC) used in former AVEC challenges [80, 81]. The tasks included explaining the reason to participate in the activity, describing happy and sad memories, counting up to thirty, and explaining two emotion eliciting pictures. The video recordings are given in their entirety, without task-based segmentation, but a tone is played when the task is switched, produced by the system that presented the instructions on the screen.

Table 1: Number and duration (minutes : seconds) of video clips of the BD corpus used for the AVEC 2018 BDS; details are blinded on the test partition.

Category	Training	Development	Test
Remission	25 – 64:52	18 – 42:47	–
Hypomania	38 – 167:42	21 – 62:24	–
Mania	41 – 189:29	21 – 71:01	–
All	**104 – 422:04**	**60 – 176:13**	**54 – 207:07**

Table 2: Number of subjects and duration (minutes : seconds) of the video chats of the SEWA database used for the AVEC 2018 CES.

Culture	Partition	# Subjects	Duration
German	Training	34	93:12
German	Development	14	37:46
German	Test	16	46:38
Hungarian	Test	66	133:12
All		**130**	**310:48**

The full corpus was collected from over 100 subjects, half of which forming the healthy control group, and under ethical committee approval. In the scope of the challenge, we focus on the portion collected from 35 male and 16 female bipolar subjects from the psychiatry inpatient service. Out of 51 patients, five subjects did not give consent for sharing their data publicly, thus the number of subjects present in this challenge is 46 (30 M, 16 F). Video recordings and session level annotations were carried out during hospitalization, in every follow up day (0th- 3rd- 7th- 14th- 28th day) and after discharge on the 3rd month. Presence of manic features were evaluated after each session using the Young Mania Rating Scale (YMRS) [85]; scores obtained at time t are grouped into three levels: Remission ($YMRS_t <= 7$), hypomania ($7 < YMRS_t < 20$), and mania ($YMRS_t >= 20$).

For the purpose of the AVEC 2018 BDS, the BD corpus was segmented into three subject-independent partitions (training – 22 subjects development – 12 subjects, and test – 12 subjects, respectively), while preserving the first two statistical moments of gender, age, and level of mania over the partitions. Details such as number and duration of video clips (except test) are given in Table 1.

3.2 SEWA database

The SEWA database consists of audiovisual recordings of spontaneous behaviour of participants captured using an *in-the-wild* recording paradigm. Subjects from German and Hungarian cultures (pairs of friends or relatives) were recorded through a dedicated video chat platform which utilised participants' own – standard – web-cameras and microphones. After watching a set of commercials, pairs of participants were given the task to discuss the last advert watched (a video clip advertising a water tap) for up to three minutes. The aim of this discussion was to elicit further reactions and opinions about the advert and the advertised product.

In addition to the audio and video modalities, manual transcription of the speech is available for the whole dataset. The transcription was done by a native speaker, providing timestamps for each utterance. The video chats of both cultures have been annotated w. r. t. the emotional dimensions *arousal* and *valence*, and a third dimension describing *liking* (or sentiment), independently by six (German) or 5 (Hungarian) native speakers. The annotation contours are combined into a single gold-standard using the same *evaluator weighted estimator (EWE)*-based approach that was used in AVEC 2017 [60]. Table 2 shows the number of subjects and the duration of the recordings for each partition.

3.3 RECOLA database

The Remote Collaborative and Affective Interactions (RECOLA) database [63], which is available on-line[4], was recorded to study socio-affective behaviours from multimodal data in the context of computer-supported collaborative work [62]. 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, who did not know each other well in most cases. The database was annotated time-continuously for the dimensions of *arousal* and *valence* by six gender-balanced French speaking annotators. The gold-standard is generated with the same *evaluator weighted estimator*-based [32] approach as used in the previous editions of AVEC [61, 79]. Partitioning of the dataset into training, development, and test sets (with exactly nine subjects and an overall duration of 45 minutes per partition) was performed while preserving the distribution of mother tongue, age and gender, and is exactly the same as used in the preceding editions of the AVEC Challenge [61, 79].

4 BASELINE FEATURES

Emotion recognition from audiovisual signals usually relies on feature sets whose extraction is based on knowledge gained over several decades of research in the domains of speech processing and vision computing. Along with the recent trend of representation learning, whose objective is to learn representations of data that are best suited for the recognition task [9], there has been some noticeable efforts in the field of affective computing to learn representations of audio/visual data in the context of emotion [14, 15, 30, 68, 76]. In order to better reflect those advances, we introduce for the AVEC 2018 Challenge an ensemble of three methods that use a different level of supervision in the way expert knowledge is exploited at the feature extraction step: (i) supervised: features rely directly on expert-knowledge based representations, i. e. , the usual approach [6, 21], (ii) semi-supervised: features are learned from expert-knowledge based representations [30, 68], and (iii) unsupervised: features are directly learned from the raw signals [76], or generated [16], with eventual use of out-of-domain data [14, 15].

[4]http://diuf.unifr.ch/diva/recola

4.1 Supervised: Expert-knowledge

The traditional approach in time-continuous emotion recognition consists in summarising low-level descriptors (LLDs) of speech and video data over time with a set of statistical measures computed over a fixed-duration sliding window. Those descriptors usually include spectral, cepstral, prosodic, and voice quality information for the audio channel, appearance and geometric information for the video channel. Features can be either brute-forced with a large ensemble of LLDs that are all combined with a large set of statistical measures, e. g. , the ComParE acoustic feature set [70], or they can be reduced to smaller, expert-knowledge based information. As visual features, we extract the intensities of 17 Facial Action Units (FAUs) for each video frame, along with a confidence measure, using the toolkit openFace[5] [6], as the FAUs have proven to perform well for facial emotion recognition. Recommendations for the definition of a minimalistic acoustic standard parameter set have led to the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [21], and to an extended version (eGeMAPS), which contains 88 measures covering the aforementioned acoustic dimensions, and used here as baseline. In addition, Mel-frequency cepstral coefficients (MFCCs) 1-13, including their 1^{st}- and 2^{nd}-order derivatives (deltas and double-deltas) are computed as a set of acoustic LLDs. The open-source toolkit openSMILE[6] [22] is used to extract the acoustic features.

4.2 Semi-supervised: Bags-of-X-Words

The technique of bags-of-words (BoW), which originates from text processing, can be seen as a semi-supervised representation learning, because it represents the distribution of LLDs according to a dictionary learned from them. As a front-end of the bags-of-words, we use the MFCCs for the acoustic data, and the intensities of the FAUs for the video data; MFCCs are standardised (zero mean, unit variance) in an on-line approach prior to vector quantisation, while this step is not required for the FAU intensities. To generate the XBoW-representations, both the acoustic and the visual features are processed and summarised over a block of a fixed length duration, for each step of 100 ms or 400 ms, in order to match the frequency of the gold-standard of the CES and GES, respectively. Instances are sampled at random to build the dictionary, and the logarithm is taken from resulting term frequencies in order to compress their range. The whole XBoW processing chain is executed using the open-source toolkit openXBOW[7] [69].

4.3 Unsupervised: Deep Spectrum

As unsupervised audio baseline feature representation in this year's challenge we have included Deep Spectrum[8] features, which were first introduced for snore sound classification [4], and are extracted using deep representation learning paradigm heavily inspired by image processing. Deep Spectrum features have been shown to be effective in tasks highly related to the presented Sub-challenges, including emotion recognition [14], sentiment classification [3, 15] and autism severity recognition [5].

[5]https://github.com/TadasBaltrusaitis/OpenFace/
[6]http://audeering.com/technology/opensmile/
[7]https://github.com/openXBOW/openXBOW
[8]https://github.com/DeepSpectrum/DeepSpectrum

To generate Deep Spectrum features, the speech files are first transformed into mel-spectrogram images using Hanning windows (default configuration: 1 *s* duration and hop-size of 100 *ms*) and a power spectral density computed on the dB power scale. For this and the generation of the plots with a *viridis* colour mapping, the *matplotlib* python package [38] is used. The plots are then scaled and cropped to square images of size 227×227 pixels without axes and margins to comply with the input needs of AlexNet [44] – a deep CNN pre-trained for image classification. Afterwards, the spectral-based images are forwarded through AlexNet. Finally, 4 096-dimensional feature vectors are extracted from the mel-spectrogram images using the activations from the second fully-connected layer (*fc7*) of AlexNet.

5 BASELINE SYSTEMS

All baseline scripts are provided in the GitHub repository[2] of the AVEC Challenge, enabling participants to reproduce both the feature extraction and machine learning from the raw audiovisual files. Baseline systems rely exclusively on existing open-source machine learning toolkits to ease the reproducibility of the results. We describe in the following the systems developed for each Sub-challenge, and then present the obtained results.

5.1 Bipolar Disorder Sub-challenge

The baseline recognition system of the BDS consists of a late fusion of the best performing audio and video representations using linear SVMs with the liblinear toolkit [24]; training instances of the minority classes are duplicated to be balanced with the majority class, and the type of solver and the value of complexity C are optimised by a grid search, using a logarithmic scale for the latter, with $C \in [1.10^{-5}, 2.10^{-5}, 5.10^{-5}, 1.10^{-4}, \dots, 10^{0}]$.

For audio data, the MFCCs are computed at the frame level, and the eGeMAPS set at the speaker turn level. The turns are estimated automatically by a voice activity detection based on Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) [23], and timings are further post-processed to ignore pauses shorter than 500 *ms* and speech segments shorter than 1 *s*. In addition, the tone played when switching the task is detected automatically by template matching, but is only utilised for ignoring the corresponding acoustic LLDs, and not for topic-based segmentation, because the number of clearly audible tones present in the data varied across sessions. Bags-of-audio-words (BoAW) are generated with a window size of 2 *s* (best duration obtained by a grid search) and a hop-size of 1 *s*, and 20 soft assignments are performed on a codebook size of 1000 instances. The Deep spectrum features are computed with the default configuration.

For video data, frames are filtered in order to keep the estimations of the FAUs' intensity that were obtained with a minimum confidence level of 95 %, which removed 2.71% of the frames. The intensity of the FAUs are then summarised over each recording session by using the mean, standard-deviation, and the relative position of the maximum as statistical measures. Bag-of-video-words (BoVW) are generated with a window size of 11 *s* (best duration found in a grid search) and a hop-size of 1 *s*, and the same parameters for soft assignments and codebook size as defined for the audio data are used.

On-line standardisation of the features is performed only for the supervised representations, i. e. , eGeMAPS for audio and the FAUs for video; values of BoW and Deep Spectrum representations are processed directly. Final decision is based on majority voting for all audiovisual representations, except for the statistics of FAU intensity, which are computed at the session level.

5.2 Cross-cultural Emotion Sub-challenge

For the baseline system of the CES, we employ a 2-layer LSTM-RNN as a time-dependent regressor for each representation of the audiovisual signals, and SVMs (liblinear with L2-L2 dual form of the objective function) for the late fusion of the predictions. To limit the complexity of the LSTM-RNN based approach, only two recurrent layers consisting of 64 / 32 LSTM units each have been employed. The three targets are learned together. The model is implemented using the Keras framework. The network is trained for 50 epochs with the RMSprop optimiser, and the model providing the highest CCC on the development set (German culture) is used to generate the predictions for the test sets (Hungarian culture). The predictions of all test sequences from each culture are concatenated prior to computing the performance, i. e. , *total CCC* whose the opposite is used as loss function for training the networks [76, 83].

In order to perform time-continuous prediction of the emotional dimensions, audiovisual signals are processed with a sliding window of 4 s length, which is a compromise to capture enough information to be used with both static regressors, such as SVMs, and context-aware regressors, such as RNNs; even though we utilised frame-stacking for the SVM-based late fusion of the audiovisual representations with either past, future, or past and future context. As supervised representations, the MFCCs and FAUs are summarised with mean and standard-deviation, and the eGeMAPS set is computed on each window. BoW are extracted from the three sets of LLDs with one hard assignment on a codebook size of 100, to reduce the complexity. Deep Spectrum features are extracted with default parameters.

Note that transcriptions have not been taken into account for the baseline system of this year for several reasons. First, for the cross-cultural evaluation, transcriptions would need to be translated and there is a multitude of options to fuse the modalities. This leads to a complex model with many hyper-parameters, whose tuning would already offer a lot of room for improvement above the baseline results. However, the organisers wanted to offer the participants the option to propose a system handling the translation problem, even though it is not required to exploit those transcriptions.

5.3 Gold-standard Emotion Sub-challenge

Baseline emotion recognition systems were previously developed in the context of the AVEC Challenge [61, 79] for the RECOLA dataset [63]. They were based on SVMs used as static regressors of multimodal feature sets including audio, video, and physiological data. Because the objective of the GES is to generate emotion labels that maximise the recognition performance, while minimising the unexplained variance from the individual ratings, we re-designed the baseline system by incorporating a hierarchical fusion of the different representations of the modalities, which allows for more flexibility in the learning of the gold-standard emotion labels.

Table 3: Baseline results for the AVEC 2018 BDS. Unweighted Average Recall (% UAR) of the three classes of BD (remission, hypo-mania, and mania) is used as scoring metric; DeepSpec: Deep Spectrum; best result on the test partition is highlighted in bold.

	Audio				Video		Late fusion	
Partition	MFCCs	eGeMAPS	BoAW-MFCCs	DeepSpec	FAUs	BoVW	eGeMAPS + FAUs	DeepSpec + FAUs
Development	49.47	55.03	55.03	58.20	55.82	55.82	60.32	63.49
Test	–	50.00	–	44.44	46.30	–	**57.41**	44.44

Table 4: Baseline results for the AVEC 2018 CES. Concordance correlation coefficient (*total CCC*) is used as scoring metric; Devel: development; BoAW-M/e: bags-of-audio-words with MFCCs/eGeMAPS; DeepSpec: Deep Spectrum; best result on the test partition is highlighted in bold.

		Audio					Video		Contextualised fusion		
Culture	**Partition**	MFCCs	eGeMAPS	BoAW-M	BoAW-e	DeepSpec	FAUs	BoVW	*past*	*future*	*past+future*
					Arousal						
German	Development	.253	.124	.282	.421	.332	.486	.500	.577	.565	.581
German	Test	–	–	–	.247	.101	.524	.450	–	–	.470
Hungarian	Test	–	–	–	.226	.238	**.436**	.426	–	–	.418
					Valence						
German	Development	.217	.112	.306	.398	.276	.549	.536	.649	.625	.646
German	Test	–	–	.229	.268	–	.577	.507	.563	–	–
Hungarian	Test	–	–	.098	.166	–	**.405**	.363	.343	–	–
					Liking						
German	Development	.136	.001	.143	.003	.150	.212	.188	.288	.244	.271
German	Test	–	–	.060	–	.004	.038	.041	.035	–	–
Hungarian	Test	–	–	-.006	–	**.023**	-.036	-.031	-.003	–	–

Table 5: Baseline results achieved by the individual models for the AVEC 2018 GES. Concordance correlation coefficient (*averaged CCC*) is used as scoring metric; DeepSpec: Deep Spectrum; best regression model (E: *Elastic Net*, L: *Lasso*, ML: *Multi-task Lasso*, R: *Ridge* or S: *SVMs*) is given in superscript; best result on the test partition is highlighted in bold.

	Audio			Video				Physiology				
Part.	eGeMAPS	BoAW	DeepSpec	Appearance	Geometric	FAUs	BoVW	ECG	HRHRV	EDA	SCL	SCR
					Arousal							
Devel	.749 ^{S}	.760 ^{S}	.621 ^{L}	.483 ^{ML}	.344 ^{L}	.309 ^{L}	.197 ^{R}	.118 ^{S}	.193 ^{S}	.064 ^{S}	.083 ^{S}	.109 ^{S}
Test	.628 ^{S}	**.651** ^{S}	.495 ^{L}	.312 ^{ML}	.241 ^{L}	.233 ^{L}	.230 ^{R}	.065 ^{S}	.153 ^{S}	.029 ^{S}	.038 ^{S}	.056 ^{S}
					Valence							
Devel	.319 ^{S}	.364 ^{L}	.220 ^{E}	.416 ^{S}	.506 ^{S}	.482 ^{S}	.395 ^{ML}	.085 ^{S}	.177 ^{S}	.083 ^{S}	.129 ^{S}	.090 ^{S}
Test	.195 ^{S}	.346 ^{L}	.158 ^{E}	.382 ^{S}	**.438** ^{S}	.373 ^{S}	.433 ^{ML}	.043 ^{S}	.108 ^{S}	.058 ^{S}	.099 ^{S}	.096 ^{S}

As dimensional regressors, we explored various linear models using SVMs from the *liblinear* toolkit [24], and Generalised Linear Models (GLMs) such as *Ridge regression*, *Lasso*, *Elastic Net* from the scikit-learn toolbox[9]; multi-task formulation of the *Lasso* and *Elastic Net* algorithms are also used to take benefits of correlations between the dimensions. Optimisation is performed on the development partition by a grid-search over the following parameters: window size, time delay uniformly applied to the gold-standard, regularisation coefficients (C for SVMs, α for the GLMs), and post-processing parameters (bias and scaling factor), respectively. Two different late fusion strategies are then exploited: one that fuses all predictions obtained by the different modalities (audio, video,

[9] http://scikit-learn.org/

Table 6: Baseline results achieved by the hierarchical fusion for the AVEC 2018 GES. Concordance correlation coefficient (*averaged CCC*) is used as scoring metric; MT-Lasso: *Multi-task Lasso,* MT-E. Net: *Multi-task Elastic Net*; best result on the test partition is highlighted in bold.

Part.	Ridge	Lasso	MT-Lasso	Elastic Net	MT-E. Net
			Arousal		
Devel	.770	.775	.588	.588	.587
Test	.647	**.657**	.494	.496	.494
			Valence		
Devel	.493	.492	.570	.555	.568
Test	.335	.333	**.515**	.513	.513

and physiological) and their corresponding representation(s)[10], and another that fuses the representations of each modality in a first stage, and then the fused predictions from the different modalities in a second final stage. Optimisation of the regression models used for the fusion includes only the regularisation parameter, and the best performing series of predictions is replicated in case the fusion deteriorates the performance. As we use static regressors, we introduce context by staking frames from either past, future, or past and future, with different sizes of context; type and size of context are optimised each as hyper-parameters.

5.4 Baseline results

The official baseline results for the AVEC 2018 BDS and CES are displayed in Table 3 and 4, respectively, and results of the AVEC 2018 GES are given in Table 5 for the individual models, and in Table 6 for the hierarchical fusion. Details are not given for the the fusion of all representations of all modalities and the fusion of all representations for each modality, as no improvement was observed over the individual models.

All the approaches investigated for the BDS perform above the chance score, which is 33 %. Interestingly, the unsupervised representation of audio data based on Deep Spectrum performs best on the development set (58.2%), followed by the two representations of video data (FAUs and BoVW) that perform equally, which shows the interest of unsupervised representation of speech data based on deep learning in the context of BD. As trials for test results evaluation, we use the best two representations of audio data, and the best representation of video data; the supervised representations are preferred over the semi-supervised representations because of their reduced complexity. Evaluations show that the eGeMAPS acoustic set performs best on the test set as individual descriptor. The last two trials amongst the allowed five consist in the fusion of the best two models obtained on the audio data, with the best model of video data; results show that fusion of the two supervised representations provides the best results on the test partition, cf. Table 3.

[10]Only the supervised representation of the physiological signals is used, and the sets of video features based on appearance and geometric descriptors, as previously defined in AVEC 2015 [61], are also utilised.

In the CES, the LSTM-based system outperforms the baseline results of the SVMs-based baseline of AVEC 2017 [60] for *arousal* and *valence* in the German culture. The worse performance for *liking* is mainly due to the disuse of the linguistic information from the transcriptions, which turned to be the most suitable information, despite a bias due to the absence of noise in the transcriptions, unlike audio and video data [60]. Concerning the acoustic feature representations, the BoAW representations outperform the set of statistics for both sets of LLDs, while similar results are achieved with Deep Spectrum features in most cases. For the video domain, BoVW is not superior to using only statistics of the FAUs. Fusion of the different representations of the modalities only improved the performance on the German training partition. Overall, results show that *arousal* and *valence* can be well predicted in the Hungarian culture from audiovisual data captured *in-the-wild*, by using only knowledge from the German culture as training and development material.

Best results obtained for the GES are slightly lower than those reported in AVEC 2016 [79] for both *arousal* and *valence*, which is partly due to the reduced range of hyper-parameters used to optimise the system in order to reduce computation time, as the system needs to be fully evaluated for each gold-standard submitted by participants, and the use of a more challenging scoring metric, i. e. , *averaged CCC* vs. *total CCC* [83]. The introduced novelties yet showed some interest, such as the hierarchical fusion and the contextualised fusion, which performed best for *valence*.

6 CONCLUSIONS

We introduced AVEC 2018 – the fifth combined open Audio/Visual Emotion and Health assessment challenge. It comprises three Sub-challenges: (i) BDS, where the level of mania of patients suffering from bipolar disorder has to be classified into three classes from audiovisual recordings of structured interviews, (ii) CES, where the level of affective dimensions of *arousal*, *valence*, and *liking* has to be inferred – for the first time – in a cross-cultural paradigm from audiovisual data collected *in-the-wild*, and (ii) GES, where dimensional labels of emotion have to be generated from individual annotations. This manuscript described AVEC 2018'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 using less or the same number of trials as given to participants for reporting results on the test partition. In addition, baseline scripts have been made available in a public data repository[2], which should help the reproducibility of the baseline results.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the EU 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), the Horizon 2020 Programme through the Innovation Action No. 645094 (SEWA), and the Research Innovation Action No. 645378 (ARIA-VALUSPA), and No. 688835 (DE-ENIGMA), and from the University of Fribourg, Switzerland, through the commercial licenses of the RECOLA project. The authors further thank the sponsors of the challenge – audEERING GmbH and the Association for the Advancement of Affective Computing (AAAC).

REFERENCES

[1] 2013. Diagnostic and Statistical Manual of mental disorders (5th Ed.). American Psychiatric Association, Washington, DC.

[2] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury. 2016. Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association* 23, 3,1 (March 2016), 538–543.

[3] Shahin Amiriparian, Nicholas Cummins, Sandra Ottl, Maurice Gerczuk, and Björn Schuller. [n. d.]. Sentiment analysis using image-based deep spectrum features. In *Proceedings of the 2nd International Workshop on Automatic Sentiment Analysis in the Wild (WASA), held in conjunction with the 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017), year = 2017, address = San Antonio, TX, USA, month = October, publisher = IEEE, note = 4 pages,*.

[4] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore sound classification using image-based deep spectrum features. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 3512–3516.

[5] Alice Baird, Shahin Amiriparian, Nicholas Cummins, Alyssa M. Alcorn, Anton Batliner, Sergey Pugachevskiy, Michael Freitag, Mauric Gerczuk, and Björn Schuller. 2017. Automatic classification of autistic child vocalisations: A novel database and results. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 849–853.

[6] Tadis 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. 10 pages.

[7] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (January 2018). Early Access, 20 pages.

[8] Isabelle E. Bauer, Jair C. Soares, Salih Selek, and Thomas D. Meyer. 2017. The link between refractoriness and neuroprogression in treatment-resistant bipolar disorder. In *Neuroprogression in Psychiatric Disorders. Mod Trends Pharmacopsychiatry*, A. Halaris and B.E. Leonard (Eds.). Vol. 31. Karger Publishers, 10–26.

[9] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 4 (August 2013), 1798–1828.

[10] Brandon M. Booth, Karel Mundnicha, and Shrikanth S. Narayanana. 2018. A novel method for human bias correction of continuous-time annotations. In *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Calgary, Canada.

[11] Elvan Çiftçi, Heysem Kaya, Hüseyin Güleç, and Albert Ali Salah. 2018. The Turkish audio-visual bipolar disorder corpus. In *Proceedings of the 1st Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. AAAC, Beijing, China. 6 pages.

[12] Ciprian A. Corneanu, Marc O. Simón, Jeffrey F. Cohn, and Sergio E. Guerrero. 2016. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (August 2016), 1548–1568.

[13] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. 2000. Feeltrace: An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Belfast, UK, 19–24.

[14] Nicholas Cummins, Shahin Amiriparian, Gerhard Hagerer, Anton Batliner, Stefan Steidl, and Björn Schuller. 2017. An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 25th ACM International Conference on Multimedia (ACM MM)*. ACM, Mountain View, CA, USA, 478–484.

[15] Nicholas Cummins, Shahin Amiriparian, Sandra Ottl, Maurice Gerczuk, Maximilian Schmitt, and Björn Schuller. 2018. Multimodal Bag-of-Words for cross domains sentiment analysis. In *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Calgary, Canada. 5 pages, to appear.

[16] Jun Deng, Nicholas Cummins, Maximilian Schmitt, Kun Qian, Fabien Ringeval, and Björn Schuller. 2017. Speech-based diagnosis of autism spectrum condition by generative adversarial network representations. In *Proceedings of the 7th International Conference on Digital Health (DH)*. ACM, London, UK, 53–57.

[17] Sidney K. D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys* 47, 3 (February 2015). Article 43, 36 pages.

[18] Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on motivation*, Vol. 19. University of Nebraska Press, 207–283.

[19] Hillary Anger Elfenbein and Nalini Ambady. 2002. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin* 128, 2 (2002), 203–235.

[20] Anna Esposito, Antonietta M. Esposito, and Carl Vogel. 2015. Needs and challenges in human computer interaction for processing social emotional information. *Pattern Recognition Letters* 66 (November 2015), 41–51. Issue C.

[21] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. 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 (July 2016), 190–202.

[22] 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 21st ACM International Conference on Multimedia (ACM MM)*. ACM, Barcelona, Spain, 835–838.

[23] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller. 2013. Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Vancouver, Canada. 5 pages.

[24] 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.

[25] Maria Faurholt-Jepsen, Jonas Busk, Mads Frost, Maj Vinberg, Ellen M. Christensen, Ole Winther, Jakob E. Bardram, and Lars V. Kessing. 2016. Voice analysis as an objective state marker in bipolar disorder. *Transactional Psychiatry* 6 (July 2016), e856.

[26] Feng Zhou and Fernando de la Torre. 2009. Canonical time warping for alignment of human behavior. In *Proceedings of the 23rd Annual Conference on Advances in Neural Information Processing Systems (NIPS)*. Neural Information Processing Systems Foundation, Inc., Vancouver, Canada. Paper 3728. 9 pages.

[27] Feng Zhou and Fernando de la Torre. 2012. Generalized time warping for multi-modal alignment of human motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Providence, RI, USA, 1282–1289.

[28] Silvia Monica Feraru, Dagmar Schuller, and Björn Schuller. 2015. Cross-language acoustic emotion recognition: An overview and some tendencies. In *Proceedings of the 6th biannual Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Xi'an, P. R. China, 125–131.

[29] Ellen Frank, Isabella Soreca, Holly A. Swartz, Andrea M. Fagiolini, Alan G Mallinger, Michael E. Thase, Victoria J. Grochocinski, Patricia R. Houck, and David J. Kupfer. 2008. The role of interpersonal and social rhythm therapy in improving occupational functioning in patients with bipolar I disorder. *The American Journal of Psychiatry* 165, 12 (December 2008), 1559–1565.

[30] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation learning for speech emotion recognition. In *Proceedings of INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association*. ISCA, San Francisco, CA, USA, 3603–3607.

[31] Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. 2017. Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion* 35 (May 2017), 68–80.

[32] Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, San Juan, Puerto Rico, 381–385.

[33] Rahul Gupta, Kartik Audhkhasi, Zach Jacokes, Agata Rozga, and Shrikanth Narayanan. 2018. Modeling multiple time series annotations as noisy distortions of the ground truth: An Expectation-Maximization approach. *IEEE Transactions on Affective Computing* 9, 1 (January-March 2018), 76–89.

[34] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 25th ACM International Conference on Multimedia (ACM MM)*. ACM, Mountain View, CA, USA, 890–897.

[35] Lang He, Ercheng Pei, Dongmei Jiang, Peng Wu, Le Yang, and Hichem Sahli. 2015. Multimodal affective dimension prediction using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC), co-located with the 23rd ACM International Conference on Multimedia (ACM MM)*. ACM, Brisbane, Australia, 73–80.

[36] Zhaocheng Huang, Nicholas Cummins, Ting Dang, Brian Stasak, Phu Le, and Julien Epps. 2015. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC), co-located with the 23rd ACM International Conference on Multimedia (ACM MM)*. ACM, Brisbane, Australia, 41–48.

[37] Zhaocheng Huang and Julien Epps. 2018. Prediction of emotion change from speech. *Frontiers in ICT* 5 (June 2018). 11 pages.

[38] John D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (May-June 2007), 90–95.

[39] Zahi N. Karam, Emily Mower Provost, Satinder Singh, Jennifer Montgomery, Christopher Archer, Gloria Harrington, and Melvin G. Mcinnis. 2014. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using

speech. In *Proceedings of the 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 4858–4862.
[40] Heysem Kaya and Alexey A. Karpov. 2018. Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing* 275 (January 2018), 1028–034.
[41] Soheil Khorram, John Gideon, Melvin McInnis, and Emily Mower Provost. 2016. Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge. In *Proceedings of INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association*. ISCA, San Francisco, CA, USA, 1215–1219.
[42] Amy M. Kilbourne, David Goodrich, David J. Miklowitz, Karen Austin, Edward P. Post, and Mark S. Bauer. 2010. Characteristics of patients with bipolar disorder managed in VA primary care or specialty mental health care settings. *Psychiatric Services* 61, 5 (May 2010), 500–507.
[43] Amy M. Kilbourne, David E. Goodrich, Allison N. O'Donnell, and Christopher J. Miller. 2012. Integrating bipolar disorder management in primary care. *Current Psychiatry Reports* 14, 6 (December 2012), 687–695.
[44] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep Convolutional Neural Networks. In *Proceedings of the 26th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Vol. 25. Curran Associates, Inc., 1097–1105.
[45] Lin Li. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 1 (March 1989), 255–268.
[46] Phil Lopes, Georgios N. Yannakakis, and Antonios Liapis. 2017. RankTrace: Relative and unbounded affect annotation. In *Proceedings of the 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE Computer Society, San Antonio, TX, USA, 158–163.
[47] 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.
[48] Arianna Mencattini, Francesco Mosciano, Maria Colomba Comes, Tania De Gregorio, Grazia Raguso, Elena Daprati, Fabien Ringeval, Björn Schuller, and Eugenio Martinelli. 2018. An emotional modulation model as signature for the identification of children developmental disorders. *Scientific Reports* (2018). 18 pages, to appear.
[49] Kathleen R. Merikangas, Minnie Ames, Lihong Cui, Paul E. Stang, T. Bedirhan Ustun, Michael Von Korff, and Ronald C. Kessler. 2007. The impact of comorbidity of mental and physical conditions on role disability in the US adult household population. *Archives of General Psychiatry* 64, 10 (October 2007), 1180–1188.
[50] George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63, 2 (March 1956), 81–97.
[51] Meinard Müller. 2007. Dynamic time warping. In *Information retrieval for music and motion*. Springer, 69–86.
[52] Mihalis A. Nicolaou, Vladimir Pavlovic, and Maja Pantic. 2014. Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (2014), 1299–1311.
[53] 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)*. ACM, Santa Monica, CA, USA, 501–508.
[54] World Health Organization. 2001. Mental disorders affect one in four people. http://www.who.int/whr/2001/media_centre/press_release/en/
[55] World Health Organization. 2008. The global burden of disease: 2004 update, Table A2: Burden of disease in DALYs by cause, sex and income group in WHO regions, estimates for 2004. WHO Press, Geneva.
[56] Maja Pantic, Nicu Sebe, Jeffrey F. Cohn, and Thomas Huang. 2005. Affective Multimodal Human-computer Interaction. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA)*. ACM, 669–676.
[57] Siobhan Reilly, Claire Planner, Mark Hann, David Reeves, Irwin Nazareth, and Helen Lester. 2012. The role of primary care in service provision for people with severe mental illness in the United kingdom. *PLoS One* (May 2012). Article e36468.
[58] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters* 66 (November 2015), 22–30.
[59] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2017. Summary for AVEC 2017 – Real-life depression and affect challenge and workshop. In *Proceedings of the 25th ACM International Conference on Multimedia (ACM MM)*. ACM, Mountain View, CA, USA, 1963–1964.
[60] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, and Maja Pantic. 2017. AVEC 2017 – Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge (AVEC), co-located with the 25th ACM International Conference on Multimedia (ACM MM)*. ACM, Mountain View, CA, USA, 3–9.
[61] 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), co-located with the ACM International Conference on Multimedia (ACM MM)*. ACM, Brisbane, Australia, 3–8.
[62] Fabien Ringeval, Andreas Sonderegger, Basilio Noris, Aude Billard, Jürgen Sauer, and Denis Lalanne. 2013. On the influence of emotional feedback on emotion awareness and gaze behavior. In *Proceedings of the 5th biannual Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE Computer Society, Geneva, Switzerland, 448–453.
[63] Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), held in conjunction with the 10th International IEEE Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, Shanghai, China. 8 pages.
[64] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (December 1980), 1161–1178.
[65] James A. Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin* 115, 1 (January 1994), 102–141.
[66] Hesam Sagha, Jun Deng, Maryna Gavryukova, Jing Han, and Björn Schuller. 2016. Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace. In *Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* . IEEE, Shanghai, P. R. China, 5800–5804.
[67] Klaus R. Scherer, Rainer Banse, and Harald G. Wallbott. 2001. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology* 32, 1 (January 2001), 76–92.
[68] 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 2016, 17th Annual Conference of the International Speech Communication Association*. ISCA, San Francisco, CA, USA, 495–499.
[69] Maximilian Schmitt and Björn Schuller. 2017. openXBOW – Introducing the Passau open-source crossmodal Bag-of-Words toolkit. *Journal of Machine Learning Research* 18 (2017), 1–5. Issue February - present.
[70] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. 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, 148–152.
[71] 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)*. ACM, Santa Monica, CA, USA, 449–456.
[72] 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 biannual International Conference on Affective Computing and Intelligent Interaction (ACII)*, Vol. II. Springer, Memphis, TN, USA, 415–424.
[73] Mariooryad Soroosh and Carlos Busso. 2013. Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In *Proceedings of the 5th biannual International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Geneva, Switzerland, 85–90.
[74] Nattapong Thammasan, Kenichi Fukui, and Masayuki Numao. 2016. An investigation of annotation smoothing for EEG-based continuous music-emotion Recognition. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, Budapest, Hungary.
[75] George Trigeorgis, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2018. Deep Canonical Time Warping for simultaneous alignment and representation learning of sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 5 (May 2018), 1128–1138.
[76] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep Convolutional Recurrent Network. In *Proceedings of the 41st International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, Shanghai, China, 5200–5204.
[77] Amos Tversky. 1969. Intransitivity of preferences. *Psychological Review* 76, 1 (January 1969), 31–48.
[78] 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 (ACM MM)*. ACM, Amsterdam, The Netherlands, 1483–1484.

[79] 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), co-located with the ACM International Conference on Multimedia (ACM MM)*. ACM, Amsterdam, The Netherlands, 3–10.

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

[81] 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 (ACM MM)*. ACM, Orlando, FL, USA, 1243–1244.

[82] Laurens van der Maaten and Kilian Weinberger. 2012. Stochastic triplet embedding. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, Santander, Spain.

[83] 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)*. IJCAI/AAAI, New York City, NY, USA, 2196–2202.

[84] Georgios N. Yannakakis, Roddy Cowie, and Carlos Busso. 2017. The ordinal nature of emotions. In *Proceedings of the 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, San Antonio, TX, USA. 8 pages.

[85] Robert C. Young, Jeffery T. Biggs, Veronika E. Ziegler, and Dolores A. Meyer. 1978. A rating scale for mania: Reliability, validity and sensitivity. *The British Journal of Psychiatry* 133, 5 (November 1978), 429–435.

[86] Biqiao Zhang, Emily Mower Provost, and Georg Essl. 2017. Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences. *IEEE Transactions on Affective Computing* (March 2017). Early Access, 14 pages.

[87] Zixing Zhang, Nicholas Cummins, and Björn Schuller. 2017. Advanced data exploitation in speech analysis – An overview. *IEEE Signal Processing Magazine* 34, 4 (July 2017), 107–129.