AVEC 2019 Workshop and Challenge: State-of-Mind, Depression with AI, and Cross-Cultural Affect Recognition

Fabien Ringeval
Université Grenoble Alpes, CNRS
Grenoble, France

Björn Schuller[*]
University of Augsburg
Augsburg, Germany

Michel Valstar
University of Nottingham
Nottingham, UK

Nicholas Cummins
University of Augsburg
Augsburg, Germany

Roddy Cowie
Queen's University Belfast
Belfast, UK

Mohammad Soleymani
University of Southern California
Los Angeles, USA

Maximilian Schmitt
University of Augsburg
Augsburg, Germany

Shahin Amiriparian
University of Augsburg
Augsburg, Germany

Eva-Maria Messner
University of Ulm
Ulm, Germany

Leili Tavabi
University of Southern California
Los Angeles, USA

Siyang Song
University of Nottingham
Nottingham, UK

Sina Alisamir
Université Grenoble Alpes, CNRS
Grenoble, France

Shuo Liu
Tianjin Normal University
Tianjin, China

Ziping Zhao
Tianjin Normal University
Tianjin, China

Adria Mallol-Ragolta
University of Augsburg
Augsburg, Germany

Zhao Ren
University of Augsburg
Augsburg, Germany

Maja Pantic[†]
Imperial College London
London, UK

ABSTRACT

The Audio/Visual Emotion Challenge and Workshop (AVEC 2019) "State-of-Mind, Depression with AI, and Cross-cultural Affect Recognition" is the ninth 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: state-of-mind recognition, depression assessment with AI, and cross-cultural affect sensing, respectively.

CCS CONCEPTS

• **General and reference → Performance**;

KEYWORDS

Affective Computing; State-of-Mind; Cross-Cultural Emotion

ACM Reference Format:
Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Mohammad Soleymani, Maximilian Schmitt, Shahin Amiriparian, Eva-Maria Messner, Leili Tavabi, Siyang Song, Sina Alisamir, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, and Maja Pantic. 2019. AVEC 2019 Workshop and Challenge: State-of-Mind, Depression with AI, and Cross-Cultural Affect Recognition. In *2019 Audio/Visual Emotion Challenge and Workshop (AVEC'19), October, 2019, Nice, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3266302.3266316

1 INTRODUCTION

The Audio/Visual Emotion Challenge and Workshop (AVEC 2019) is the ninth 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 [55–57, 68, 69, 79, 81, 82].

[*] The author is further affiliated with Imperial College London, London, UK.

[†] The author is further affiliated with University of Twente, Twente, The Netherlands.

AVEC'19, October, 2019, Nice, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5983-2/18/10.
https://doi.org/10.1145/3266302.3266316

One of the goals of the AVEC series is to bring together multiple communities from different disciplines, in particular, the audiovisual multimedia communities and those in the psychological and social sciences who study expressive behaviour. Another objective is to advance health and emotion recognition systems by providing a common benchmark test set for multimodal information processing, in order 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 wholly naturalistic behaviour. Because this is precisely the type of data that the new generation of affect-oriented multimedia and human-machine/human-robot communication interfaces have to face in the real world.

Major novelties are introduced for the AVEC 2019 with three separated Sub-challenges focusing on health and emotion analysis: (i) State-of-Mind Sub-challenge (SoMS), (ii) Detecting Depression with AI Sub-challenge (DDS), and (iii) Cross-cultural Emotion Sub-challenge (CES). Herein, we describe the novelties introduced in the Challenge and the guidelines for participating.

The State-of-Mind Sub-challenge (SoMS) is a new task focusing on the continuous adaptation of human state-of-mind (SOM), which is pivotal for mental functioning and behaviour regulation [32]. SOM is constantly shifting due to internal and external stimuli, and frequent use of either adaptive or maladaptive SOM influences our mental health. One key aspect of the human experience are our emotions, as they reflect our SOM [71, 75]. In the SoMS, self-reported mood (10-point Likert scale) after the narrative of personal stories (two positive and two negative), has to be predicted automatically from the audiovisual recordings of those stories; USoM corpus [53].

The Detecting Depression with AI Sub-challenge (DDS) is a major extension of the AVEC 2016 DSC [80], where the level of depression severity (PHQ-8 questionnaire) was assessed from audiovisual recordings of patients interacting with a virtual agent conducting a clinical interview and driven by a human as a Wizard-of-Oz (WoZ); DAIC-WOZ corpus [29]. The DAIC data set contains new recordings of the same population with the virtual agent being, this time, wholly driven by AI, i. e. , without any human intervention. Those new recordings are used as a test partition for the DDS, and will help to understand how the absence of a human for conducting the virtual agent impacts on automatic depression severity assessment.

The Cross-cultural Emotion Sub-challenge (CES) is a large extension of the AVEC 2018 CES [54], where dimensions of emotion were inferred from audiovisual recordings collected "*in-the-wild*", i. e. , with standard webcams and at home/work place. A cross-cultural setup was further exploited for inferring emotion: knowledge of German culture was leveraged to infer emotion on the Hungarian culture, using the SEWA corpus [38]. This dataset now includes data collected from new participants with Chinese culture, which is used as a test set for the CES, whose aim is, therefore, to investigate how emotion knowledge of Western European cultures (German, Hungarian) can be transferred to the Chinese culture.

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[1], which participants are free to use for reproducing both the baseline features and recognition systems (cf. Section 5). 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 of the labels relies on the *Concordance Correlation Coefficient (CCC)* [43] for all Sub-challenges; the Root Mean Squared Error (*RMSE*) is additionally reported. Whereas many others metrics of performance could be exploited for ranking the contributions, such as the Spearman's *CC*, or the coefficient of determination (r^2), we believe that the indice of reproducibility *CCC* is the most suitable metric to use, as it is not biased by changes in scale and location, and elegantly includes information on both precision and accuracy in a single statistical measure [43]. Moreover, its theoretical definition and properties are well rooted in the literature [48], and it can be easily exploited as a loss function for training neural networks [84].

To be eligible to participate in the Challenge, every entry has to be accompanied by a paper submitted to the AVEC 2019 Data Challenge and Workshop, describing the results and the methods that created them. These papers undergo peer-review by the technical program committee. Only contributions with a relevant accepted paper and at least a submission of test results are eligible for participation. The organisers do not participate in the Challenge themselves, but re-evaluate 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

This section is a summary of the current state-of-the-art in the automatic analysis of affect with a focus on: (i) human state-of-mind, (ii) depression assessment in the context of AI-driven virtual agents, and (iii) dimensional analysis in cross-cultural paradigms.

2.1 State-of-Mind

The concept of a human SOM describes the phenomenon that our consciousness and emotions are constantly fluctuating over time; this is due to internal and external biological, psychological, and social demands [32, 53]. One key aspect of SOM is our emotions. They provide valuable information that influences our basic human processes in a bidirectional manner [74, 75]. Such processes include attention, perception, cognition, memory retrieval, memory storage, and behaviour regulation. In fact, depending on our actual SOM, some emotions, cognitions, and behaviours are more likely to occur, while others may be suppressed. This effect is the underlying principle of mood congruence [59, 71].

Despite the major impact of SOM on health and social functioning, the quantification of current emotional states, with therapy contexts, has its pitfalls. The major of these being that it relies

[1] https://github.com/AudioVisualEmotionChallenge/AVEC2019

heavily on self-reports of emotional states, which are inherently biased [86]. As humans are structurally determined closed systems, the assumptions that the quantification of one person's actual SOM might be matchable to another person's actual SOM, is not always observable, even if they rate the same score [45]. Moreover, even within a person, the current rating of the emotional state is rooted in previous experiences, known as the adaption level, and therefore is not really accurate in an absolute way [60].

One method to overcome this is to treat emotional state values as ordinal variables [86]; according to Russel's theory of core affect, every current emotion can be quantified on the orthogonal axes *arousal* (from sleepy to hyper-aroused) and *valence* (with the poles negative and positive) without having to limit the quantification to a given language [59]. Another is to complement self-ratings with expert ratings or physiological recordings. Each of these methods has its limitations; the mismatch between different emotion assessments is still very much a matter of scientific discourse [11, 72, 73].

Despite the given limitations of the scientific assessment of emotional states, humans constantly monitor their own and others' emotions and organise themselves within social systems [16, 62]. Given the need for humans to socially interact and the increased occurrence of human-machine-interactions, the development of a real-time SOM data-driven recognition system has the potential to enhance user experience, user satisfaction, and subsequently to foster user adherence [7, 52, 53]. Such a system could assist the society in various aspects; i) decreasing bias in the monitoring of SOM; ii) collecting more objective data to aid the diagnostics of affective disorders; iii) delivering tailored interventions fostering treatment of disease; iv) reducing the time spent in the evaluation of treatment outcome, and in e-treatment by presenting SOM related content resulting in reduced individual and societal burden [52, 53, 70, 77].

2.2 Depression Detection with AI

Depression, otherwise known as major depressive disorder (MDD), is a common mental health disorder that negatively impacts the way one thinks, feels, and acts [1]. It can lead to a variety of emotional and physical problems and can decrease a person's ability to carry out daily professional and personal activities. The World Health Organisation (WHO) declared depression as the leading cause of ill health and disability worldwide in 2015, with more than 300 million people living with depression [47]. Given the high prevalence of depression and its suicidal risk, finding new methods for diagnosis and treatment of depression becomes more and more critical.

There is growing interest in using automatic human behaviour analysis for computer-aided depression diagnosis based on behavioural cues such as facial expressions and speech prosody, because of convincing evidences that depression and related mental health disorders are associated with changes in patterns of behaviour [10, 15, 35, 65, 85]. Facial activity, gesturing, head movements and expressivity are among behavioural signals that are strongly correlated with depression.

Early paralinguistic investigations into depressed speech found that patients consistently demonstrated prosodic speech abnormalities such as reduced pitch, reduced pitch range, slower speaking rate, and higher articulation errors [15]. Facial expression and head gestures that can be tracked by computer vision are also good predictors of depression; e. g., a more downward angle of the gaze, less intense smiles, and shorter average duration of smiles have been reported as the most salient facial cues of depression [64]. Further, body expressions, gestures, head movements, and linguistic cues have also been reported to provide relevant cues for depression detection [2, 46, 50, 51].

Taking all those evidences together, it has been proposed to integrate affective computing technology into a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illnesses [18]. Data collected with subjects suffering from post-traumatic stress disorder showed that the automatic evaluation of their level of depression severity (PHQ-8 questionnaire) can achieve a *RMSE* less than 5 when the agent is driven by a human acting as a WoZ [28]; PHQ-8's range $\in [0, 24]$ and cutpoints are defined at $[5, 10, 15, 20]$ for mild, moderate, moderately severe, and severe depression, respectively. Those results need to be investigated further, with the agent being wholly driven by AI, as the wizard might drive the virtual agent to a situation that eases the observation of patterns associated with depression, or the autonomous agent might have issues in conducting the interview appropriately.

2.3 Cross-cultural Emotion Recognition

Cross-cultural emotion recognition has long been highlighted as an open research question within the affective computing community [19, 21, 23, 49], and was introduced as an AVEC Sub-challenge in 2018 [54]. Whereas the AVEC 2018 CES focused on detecting *arousal*, *valence*, and *liking* from Hungarian speakers using only German speakers for training and development of the models [54], in this year's AVEC CES the test cohort is Chinese speakers with speakers from the two cultures mentioned earlier being available for training, development, and additional testing.

A common idiom in facial expression recognition is that emotional expressions have a large degree of universality across cultures [13, 20]. This statement was on the whole supported by both baseline results and works submitted to the AVEC 2018 CES, with either vision-only or multimodal systems achieving higher cross-culture accuracies than speech-only approaches [34, 56, 83, 89]. These results were insightful, as previously, there were only a few works in the affective computing literature which supported this claim [12, 19].

Interestingly, approaches in the AVEC 2018 CES did not employ approaches such as transfer learning [87, 88] or domain adaptation techniques [36, 61] typically seen in cross-cultural testing. In [83], the authors proposed a model based on emotional salient detection to identify emotion markers invariant to socio-cultural context. The other two entrants employed data driven approaches based on long short-term memory recurrent neural networks (LSTM-RNN) [34, 89]. Matching with similar results in the literature [26, 63], all entrants in the AVEC 2018 CES observed a drop in system performance when testing on the Hungarian data [34, 83, 89].

3 CHALLENGE CORPORA

The AVEC 2019 Challenge relies on three corpora: (i) the USoM corpus [53] for the SoMS, (ii) the Extended-DAIC corpus [29] for the DDS, and (iii) the SEWA dataset [38] for the CES. We provide below a short overview of each dataset and refer the reader to the original work for a more complete description.

3.1 Ulm State-of-Mind Corpus

The Ulm state of mind database was collected to assess the association between personal story telling and current SOM, operationalised by affective state according to Russel's theory [53, 59, 70]. Parts of this dataset have been released for the Interspeech 2018 Computational Paralinguistics (ComParE) challenge [70].

Participants of the USoM corpus were instructed to first tell two negative personal narratives $NN_{1,2}$ and subsequently two positive personal narratives $PN_{1,2}$, each for five minutes in front of a camera. They were also asked to rate their current affect (CA) on a 10-point likert scale for the dimensions *arousal* and *valence* before and after telling each narrative, resulting in the following protocol: (t_0), CA_0, NN_1, (t_1), CA_1, NN_2, (t_2), CA_2, PN_1, (t_3), CA_3, PN_2, and (t_4), CA_4. For the purpose of the Challenge, the USoM dataset was partitioned into training, development and test sets while preserving the overall speaker diversity – in terms of age, gender distribution, and core affect evaluations – within the partitions. Table 1 shows the number of subjects and duration for each partition.

As the interest of the SoMS is on the change in mood, rather than just it's static observation, the initial self-reports made before the storytelling are included in the data package given to participants for all partitions, including the test set. Exploiting such contextual information in an automatic system predicting the level of mood is a realistic scenario in the real-world, because a therapist would always ask a person baseline emotion at the start of a session. It is thus essential to provide machine learning algorithms with the same prior information as a therapist would have.

3.2 Distress Analysis Interview Corpus

The Extended Distress Analysis Interview Corpus (E-DAIC) [18] is an extended version of WOZ-DAIC [29] that contains semi-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 large effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illnesses [29].

Data collected include audio and video recordings, automatically transcribed text using Google Cloud's speech recognition service, and extensive questionnaire responses. The interviews are conducted by an animated virtual interviewer called Ellie. In the WoZ interviews, the virtual agent is controlled by a human interviewer (wizard) in another room, whereas in the AI interviews, the agent acts in a fully autonomous way using different automated perception and behavior generation modules.

For the purpose of the Challenge, the E-DAIC dataset was partitioned into training, development and test sets while preserving the overall speaker diversity – in terms of age, gender distribution, and the eight-item Patient Health Questionnaire (PHQ-8) scores – within the partitions. Whereas the training and development sets include a mix of WoZ and AI scenarios, the test set is solely constituted from the data collected by the autonomous AI. Details regarding the speaker distribution over the partitions are given in Table 2.

Table 1: Number of subjects and duration of the storytellings contained in the USoM database [53].

Partition	# Subjects	Duration [h:min:s]
Training	45	13:49:38
Development	33	10:46:57
Test	33	9:46:14
All	**111**	**34:22:49**

Table 2: Number of subjects and duration of the interviews included in the Extended-DAIC database [29].

Partition	# Subjects	Duration [h:min:s]
Training	163	43:30:20
Development	56	14:47:31
Test	56	14:52:42
All	**275**	**73:10:33**

Table 3: Number of subjects and duration of the video chats contained in the SEWA database [38].

Culture	Partition	# Subjects	Duration [h:min:s]
German	Training	34	1:33:12
German	Devel.	14	0:37:46
German	Test	16	0:46:38
Hungarian	Training	34	1:08:24
Hungarian	Devel.	14	0:28:42
Hungarian	Test	18	0:36:06
Chinese	Test	70	3:17:52
All		**200**	**8:28:40**

3.3 Cross-cultural Emotion Database (SEWA)

The SEWA database consists of audiovisual recordings of spontaneous behaviour of participants captured using an *in-the-wild* recording paradigm [38]. Pairs of friends or relatives from German, Hungarian, and Chinese cultures 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.

The video chats of the three cultures have been annotated w. r. t. the emotional dimensions *arousal* and *valence*, and a third dimension describing *liking* (or sentiment), independently by several native speakers; German and Chinese: six annotators, Hungarian: five annotators. The annotation contours (traces) are combined into a single gold-standard using the same *evaluator weighted estimator (EWE)*-based approach that was used in the last two editions of AVEC [54**?**]. Table 3 shows the number of subjects and the duration of the recordings for each partition.

4 BASELINE FEATURES

Emotion recognition from audiovisual signals usually relies on feature sets whose extraction is based on expertise gained over several decades of research in the domains of speech processing, e. g., Mel Frequency Cepstral Coefficients (MFCCs), and vision computing, e. g., Facial Action Units (FAUs). However, recent advances in the field of representation learning, whose objective is to learn representations of data that are best suited for the recognition task [8], have shown that efficient representations of audiovisual signals can be learnt in the context of emotion [4, 14, 27, 66, 78].

Audiovisual representations can be learnt from expert-driven information extracted from the raw signals [27, 66], or directly from the raw signals [78]. They can also be generated using adversarial networks [17], or using convolutional neural networks trained on out-of-domain data and for a different task, e. g. , audio representations extracted by a model trained for objects classification in images [4, 14].

4.1 Expert-knowledge

The traditional approach in affect sensing consists in summarising low-level descriptors (LLDs) of audioviusal signals over time with a set of statistical measures computed over a fixed-duration sliding analysis window. Those descriptors usually include spectral, cepstral, prosodic, and voice quality information for the audio channel, and appearance, geometric, and FAUs information for the video channel.

As audio features, we compute the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [24], which contains 88 measures covering the aforementioned acoustic dimensions, and used here as baseline. In addition, MFCCs 1-13, including their 1st- and 2nd-order derivatives (deltas and double-deltas) are computed as a set of acoustic LLDs, using the openSMILE[2] [25] toolkit. As visual features, we extract the intensities of 17 FAUs for each video frame, along with a confidence measure, using the toolkit openFace[3] [6]. Descriptors of pose and gaze are additionnaly extracted.

Audiovisual LLDs are summarised over time by computing their mean and standard-deviation using a sliding window of 4 s length, and a hop size of 1 s for the USoM and E-DAIC datasets, and 100 ms for the SEWA dataset, excepted for the eGeMAPS set, which is computed on each window.

4.2 Bags-of-Words

The technique of bags-of-words (BoW), which originates from text processing, represents the distribution of LLDs according to a dictionary learned from them. As a front-end of the BoW, we use the MFCCs and the eGeMAPS set for the acoustic data, and the intensities of the FAUs for the video data; MFCCs and eGeMAPS LLDs 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 BoW representations, both the acoustic and the visual features are processed and summarised over a block of a 4 s length duration, for each step of 100 ms for the SEWA dataset, and 1 s for the USoM and E-DAIC datasets. The codebook size is 100. 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[4] [67].

4.3 Deep Representations

As in last year's challenge [54], we have included Deep Spectrum[5] features as a deep learning based audio baseline feature representation [4]. Deep Spectrum features are inspired by deep representation learning paradigms common in image processing: spectral images of speech instances are fed into pre-trained image recognition CNNs and a set of the resulting activations are extracted as feature vectors. Such representations have since been used in a wide array of speech and audio processing tasks [3, 5, 31].

For this year's challenge, we extracted Deep Spectrum features from four robust pre-trained CNNs using VGG-16 [76], AlexNet [41], DenseNet-121, and DenseNet-201 [33]; AlexNet was used solely for the AVEC 2019 CES for being consistent with the previous AVEC 2018 CES. The speech files are first transformed into mel-spectrogram images with 128 mel-frequency bands, a window width of 4 s for all challenge corpora and a hop size of 1 s for the USoM and E-DAIC datasets, and 100 ms for the SEWA dataset. Afterwards, the spectral-based images are forwarded through the pre-trained networks. A 4 096-dimensional feature vector is then formed from the activations of the second fully connected layer in VGG-16 and AlexNet, and a 1 024 and a 1 920-dimensional feature vector is obtained from the activations of the last average pooling layer of the DenseNet-121 and DenseNet-201 networks, respectively.

We also provide two baseline deep visual representations. For these, we employed a VGG-16 [76] network and a ResNet-50 network [30] that are pre-trained with the Affwild dataset [37]. The pipeline starts with applying the openFace toolkit [6] to detect the face region and perform then face alignment. Then, we froze the weights of two pre-trained models and fed the aligned face images to both CNNs individually. To obtain the deep representations for each frame, we extract the output of the first fully-connected layer from the pre-trained VGG-16 network, and the output of the global average pooling layer from the pre-trained ResNet-50 network, respectively. As a result, a 4 096-dimensional deep feature from VGG and a 2 048-dimensional deep feature from ResNet are provided for each frame.

5 BASELINE SYSTEMS

All baseline systems rely exclusively on existing open-source machine learning toolkits to ensure the reproducibility of the results. In this section, we describe the systems developed for each Subchallenge, and present the obtained results. For evaluation on the test set, we retained the two audio representations with the best performance, and the two video representations with the best performance, in addition to the fusion of all audiovisual representations.

[2] http://audeering.com/technology/opensmile/

[3] https://github.com/TadasBaltrusaitis/OpenFace/

[4] https://github.com/openXBOW/openXBOW

[5] https://github.com/DeepSpectrum/DeepSpectrum

Table 4: Baseline results evaluated with *CCC* for the AVEC 2019 SoMS; USoM data set [53]; BoAW-M/e: bags-of-audio-words with MFCCs/eGeMAPS; DS-DNet: Deep Spectrum using DenseNet-121; DS-VGG: Deep Spectrum using VGG-16; best result on the test partition is highlighted in bold.

	Audio						**Video**				**Fusion**
Partition	MFCCs	eGeMAPS	BoAW-M	BoAW-e	DS-DNet	DS-VGG	FAUs	BoVW	ResNet	VGG	*All*
					Random sampling of training instances						
Development	.282	.412	.336	.295	.280	.384	.372	.317	.261	.318	.417
Test	–	.276	–	–	–	**.289**	.119	–	–	.191	.278
					Curriculum sampling of training instances						
Development	.299	.378	.334	.288	.326	.437	.419	.313	.300	.318	.464
Test	–	**.294**	–	–	–	.208	.151	–	-	.160	.236

5.1 State-of-Mind Sub-challenge

We use a gated recurrent unit (GRU) network with two layers, each having 64 nodes for their hidden layers, for each audiovisual representation. As a pre-processing step, all input features are normalised to have zero mean and unit variance. Dropout, at a rate of 10 %, is employed during training. The GRU is then followed by a fully connected neural network that has one hidden layer with 32 nodes, followed by a single linear layer to map to the desired output size of one. Note that a middle-fusion of the audiovisual representations is performed by concatenating their respective GRU outputs.

The model is implemented using a PYTORCH framework and is trained with an ADAM optimiser. As previous studies have shown the benefits of training a network following a curriculum [9, 22, 44], where instances are gradually presented in increasing level of difficulty, we implemented this approach using the following strategy: a uniform distribution of valence labels is first obtained by duplicating training instances, then, a sub-set of the training set with only the data instances with $CA \in [2-3] \cup [9-10]$, i. e., the most positive and negative storytellings, is firstly used for training, followed by a larger sub-set with data instances with $CA \in [2-4] \cup [8-10]$, each for 32 epochs. We then exploited the whole training set until *early stopping* occurs; once 60 epochs have passed, training is stopped if there is no improvement within the last 25 epochs.

Because the interest of the SoMS is in the prediction of a change in human SOM, the network is trained to model the difference between the self-reported core affect after each story and before the first story: $CA_i - CA_0$, $i = 1, 2, 3, 4$. Results are reported for each audiovisual representation, and for the two training approaches, i. e., with or without curriculum, in Table 4. Whereas the mid-fusion of all audiovisual representations provides the best result on the development set for the two learning approaches, audio descriptors achieve higher performance on the test set, with the expert-based eGeMAPS set performing best with curriculum learning.

A summary of the results obtained with either a static (CA_i) or a dynamic ($CA_i - CA_0$) view of the self-reported mood used for training or testing the system is additionally provided in Table 5. Interestingly, results show that, the automatic inference of the self-reported mood performs much better in a mixed scenario with either a training on the static view (CA_i) and an evaluation on the change ($CA_i - CA_0$), or *vice-versa*, i. e., training on the change and testing on the static label, compared to a 'consistent' approach with both training and testing performed on the same view, i. e., either static or dynamic.

Table 5: Comparison of the approaches – training or testing on a static or dynamic measure of mood – used for the AVEC 2019 SoMS; averaged *CCC* results are reported; $[\mu(\sigma)]$.

Partition	Static training	Dynamic training
	Static evaluation	
Development	.149 (.108)	.335 (.050)
Test	.037 (.063)	.219 (.068)
	Dynamic evaluation	
Development	.368 (.150)	.102 (.066)
Test	**.325 (.052)**	.040 (.094)

This result might stem from the fact that emotion data is hierarchically organised. As such, each self-reported emotion is nested within a person over a period of time [39]. Due to the incapability of humans to assess their own emotions as an absolute value, self-reported emotion can only be interpreted as a current assessment of emotional differences in relation to the nearest past. Furthermore, there is also variance in emotion dynamics between people and not only within a person [40]. The inter-individual and intra-individual variance in emotion dynamics are strongly related to one another, but add both new information to predictions. While the variance between persons might be best captured in a scenario where machine learning is applied to raw values, the intra-individual auto-correlation of emotion, the so called emotional inertia, is portrayed in the dynamic evaluation [42]. Therefore training on static data and evaluating on dynamic data, such as emotional inertia, might be the *state-of-the-art* approach to handle this phenomenon.

5.2 Detecting Depression Sub-challenge

For the depression detection baseline, we employ a single-layer 64-d GRU as our recurrent network with a dropout regularization of rate 0.2, followed by a 64-d fully-connected layer to obtain a single-value regression score. To handle bias, prior to training we convert the

Table 6: Baseline results evaluated with *CCC* for the AVEC 2019 DDS; *RMSE* is additionally reported; BoAW-M/e: bags-of-audio-words with MFCCs/eGeMAPS; DS-DNet: Deep Spectrum using DenseNet-121; DS-VGG: Deep Spectrum using VGG-16; best result on the test partition is highlighted in bold; * note that results with the eGeMAPS set are pending update.

	Audio						**Video**				**Fusion**
Partition	MFCCs	eGeMAPS	BoAW-M	BoAW-e	DS-DNet	DS-VGG	FAUs	BoVW	ResNet	VGG	*All*
					Regression of PHQ-8 score (CCC)						
Development	.198	.076*	.102	.272	.165	.305	.198	.107	.269	.108	.043
Test	–	–	–	.045	–	.108	-.098	–	.120	–	**.121**
					Regression of PHQ-8 score (RMSE)						
Development	7.77	9.32*	9.61	7.67	8.24	8.56	7.51	6.71	8.81	7.69	6.01
Test	–	–	–	8.83	–	8.78	6.94	–	7.65	–	**6.31**

PHQ score labels to floating point numbers by downscaling with a factor of 25. The network is trained and evaluated using a *CCC* loss function and evaluation score, and the results are reported using the original PHQ scale. A batch size of 15 is used consistently and the learning rate is optimized across different feature sets. In order for the data to fit on GPU memory, a maximum sequence length has been assigned for the sessions. For the MFCCs and eGeMAPS LLDs, and the high dimensional deep representations like DeepSpectrum, ResNet and VGG, a maximum sequence length of 20 minutes has been used. Additionally for ResNet, VGG and deep spectrum representations the frames have been dropped keeping one out of two, or one out of four frames depending on the dimensionality, so the data can be loaded on memory. Fusion of the different audiovisual representations is achieved by simply averaging their scores.

Baseline results of the DDS are given in Table 6. They show that, on the development set, the best *CCC* score from audio features was achieved with Deep spectrum (DS-VGG) features, and the model with ResNet features achieved the best result for visual features. This shows the power of representations learned by deep neural networks with large amount of data that can be used in a different context than the one they have been originally designed for. Whereas fusion of the different representations performs poorly on the development set, it achieves the best result on the test set, with yet a relatively low value of *CCC*, but a corresponding *RMSE* that is slighlty better than the one obtained on the DAIC-WoZ dataset with the AVEC 2017 baseline system [58]; *RMSE* = 6.31 for AVEC 2019 compared to *RMSE* = 6.97 for AVEC 2017. However, the baseline system developed for this year's Challenge is more complex – simple linear regression models *vs.* GRU-RNNs for this year –, and the corresponding scores should be therefore best regarded in light of the best results of the AVEC 2017 Depression Sub-challenge [28], which was *RMSE* = 4.99. Results obtained in the automatic sensing of the level of depression from interactions with the virtual agent thus suggest that, a setting where the agent is wholly driven by AI makes the recognition more challenging compared to a setting with a human driving the agent as a WoZ.

5.3 Cross-cultural Emotion Sub-challenge

For the baseline system of the CES, we employ a 2-layer LSTM-RNN (64 / 32 units) as a time-dependent regressor of the three targets (learned together) 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. The model is implemented using the KERAS framework. The network is trained for 50 epochs with the RMSPROP optimiser using a dropout of 10%, and the model providing the highest CCC on the development set of the German and Hungarian culture is used to generate the predictions for the test sets (German, Hungarian, and all clips of the Chinese culture). Even though the model has three outputs modelling each dimension, the optimum model for each dimension is selected separately. The predictions of all test sequences from each culture are concatenated prior to computing the *CCC*, whose opposite is used as loss function for training the networks [78, 84].

In order to perform time-continuous prediction of the emotional dimensions, audiovisual signals were 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. We utilised frame-stacking for the SVM-based late fusion of the audiovisual representations with either past, or future context.

Baseline results of the CES are given in Table 7. They show improvements over the performance reported in the previous edition of the AVEC CES; relative improvement for German is 7.25% and 8.25% for arousal and valence, respectively, and for Hungarian, 17.3% and 13.3%, respectively. The inclusion of instances of the Hungarian culture as training and development material, in addition to those of the German culture, might explain the large increase in performance for both cultures, as only instances of the German culture were available for training and development in AVEC 2018 CES. In addition, a more recent version of the OPENFACE toolkit [6] was exploited, which provided the best results on the test set for both arousal and valence with FAUs based features. Those results confirm the common idiom that facial expressions of emotion have a large degree of universality across cultures, compared to the vocal expressions, where the acoustic and prosodic dimensions already play a key role in the oral communication by serving many grammatical and pragmatic functionalities, e. g. , tonal languages like Mandarin associate different meanings to a same syllable depending on it's pitch contour, with language dependent peculiarities that

Table 7: Baseline results evaluated with *CCC* for the AVEC 2019 CES; SEWA dataset [38]; DeepSpec: Deep Spectrum; best result on the test partition is highlighted in bold.

		Audio					Video				Fusion
Culture	**Partition**	MFCCs	eGeMAPS	BoAW-M	BoAW-e	DS	FAUs	BoVW	ResNet	VGG	*All*
Arousal											
German	Dev.	.389	.396	.323	.434	.380	.606	.556	.475	.561	.629
German	Test	–	.293	–	.276	–	.562	–	–	.505	.517
Hungarian	Dev.	.236	.305	.237	.291	.156	.425	.321	.460	.367	.583
Hungarian	Test	–	.272	–	.250	–	.527	–	–	.396	.525
Ger. + Hun.	Dev.	.326	.371	.298	.398	.312	.531	.467	.473	.493	.614
Chinese	Test	–	.100	–	.107	–	**.355**	–	–	.297	.238
Valence											
German	Dev.	.344	.405	.190	.455	.317	.639	.594	.552	.595	.684
German	Test	–	.309	–	.325	–	.627	–	–	.646	.622
Hungarian	Dev.	.017	.073	.042	.135	.084	.463	.421	.373	.363	.508
Hungarian	Test	–	.166	–	.151	.173	.459	–	–	.548	.397
Ger. + Hun.	Dev.	.187	.286	.134	.352	.233	.565	.523	.487	.505	.615
Chinese	Test	.–	.267	–	.281	–	**.468**	–	–	.398	.423
Liking											
German	Dev.	.159	.136	.140	.003	.164	.056	.073	.057	.244	.048
German	Test	–	.012	–	.074	–	-.042	–	–	-.052	-.019
Hungarian	Dev.	.115	.192	-.027	.253	.121	.104	.041	.028	.028	.260
Hungarian	Test	–	.051	–	.089	–	-.062	–	–	-.069	-.22
Ger. + Hun.	Dev.	.144	.159	.074	.138	.142	.083	.057	.040	.037	.222
Chinese	Test	–	.007	–	**.041**	–	.006	–	–	-.006	-.012

make cross-cultural settings highly challenging, especially when noise comes into play because of the ecological conditions of study.

6 CONCLUSIONS

We introduced AVEC 2019 – the sixth combined open Audio/Visual Emotion and Health assessment challenge. It comprises three Sub-challenges: i) SoMS, where the level of mood has to predicted from positive and negative personal stories, ii) DDS, where the level of depression (PHQ-8 score) has to be predicted from structured interviews conducted by a virtual agent wholly driven by AI, and iii) CES, where the level of affective dimensions of *arousal*, *valence*, and *liking* has to be inferred in a cross-cultural *in-the-wild* paradigm with German and Hungarian cultures as training and testing material, and Chinese culture as solely testing material.

By intention, we opted to use exclusively open-source software and the highest possible transparency and realism for the baselines, by using the same number of trials as given to participants for reporting results on the test partition, and sharing all the developed scripts for both features extraction and machine learning on a public platform. Results showed that: i) in the SoMS, the level of mood was best predicted when the system was trained on the static scores and evaluated on their dynamic view, i. e. , between the label provided after the storytellings, and before the first story, which can be explained by inertial emotion theories, ii) in the DDS, prediction of the level of depression (PHQ-8) is reported to be more challenging when the virtual agent conducting the interview is wholly driven by AI, compared to a WoZ setup, and iii) in the CES, dimensional emotions are more challenging to sense in a cross-cultural setting for audio descriptors compared to video descriptors, which confirm on one hand the universality of facial expressions for Asian (Chinese) and Western European cultures (German and Hungarian), and show on the other the challenge of using audio descriptors for paralinguistics analysis in languages presenting dissimilarities in their acoustic, in particular when data are collected in an ecological (noisy) environment.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the Horizon 2020 Programme through the Innovation Action No. 645094 (SEWA), and the Research Innovation Action No. 826506 (sustAGE), and No. 688835 (DE-ENIGMA). The work on the depression challenge was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. The authors further thank the sponsor of the challenge – audEERING GmbH.

REFERENCES

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

[2] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to

mental health. *Transactions of the Association for Computational Linguistics* 4 (2016), 463–476.

[3] Shahin Amiriparian and Nicholas Cummins. [n. d.]. ([n. d.]).

[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] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Sergey Pugachevskiy, and Björn Schuller. 2018. Bag-of-Deep-Features: Noise-Robust Deep Feature Representations for Audio Analysis. In *Proceedings 31st International Joint Conference on Neural Networks (IJCNN)*. INNS/IEEE, IEEE, Rio de Janeiro, Brazil, 1–7.

[6] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[7] Amit Baumel and Elad Yom-Tov. 2018. Predicting user adherence to behavioral eHealth interventions in the real world: examining which aspects of intervention design matter most. *Translational behavioral medicine* 8, 5 (2018), 793–798.

[8] 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.

[9] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. Montreal, QC, Canada, 41–48.

[10] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. 2009. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, Amsterdam, Netherlands, 1–7.

[11] Tamlin S Conner and Matthias R Mehl. 2015. Ambulatory assessment: Methods for studying everyday life. *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* (2015), 1–15.

[12] Daniel T Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. 2018. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion* 18 (2018), 75–93.

[13] 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.

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

[15] 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.

[16] Kerstin Dautenhahn. 2002. The origins of narrative: In search of the transactional format of narratives in humans and other social animals. *International Journal of Cognition and Technology* 1, 1 (2002), 97–123.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Jeffrey L. Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48, 1 (Juky 1993), 71–79.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Yuan Gong and Christian Poellabauer. 2017. Topic Modeling Based Multi-modal Depression Detection. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, AVEC 2017*. ACM, Mountain View (CA), USA, 69–76.

[29] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*. ELRA, Reykjavik, Iceland, 3123–3128.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[31] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *Proc. 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[32] Marlies Houben, Wim Van Den Noortgate, and Peter Kuppens. 2015. The relation between short term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin* 141, 4 (July 2015), 901–930.

[33] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, HW, 4700–4708.

[34] Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Minghao Yang. 2018. Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks. In *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*. ACM, Seoul, South Korea, 57–64.

[35] Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear. 2013. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces* 7, 3 (2013), 217–228.

[36] Heysem Kaya and Alexey A. Karpov. 2018. Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing* 275 (January 2018), 1028–034.

[37] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision* (2019), 1–23.

[38] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Bjorn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. 2019. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *arxiv.org* 1901.02839 (January 2019). 17 pages.

[39] Peter Koval, Peter Kuppens, Nicholas B Allen, and Lisa Sheeber. 2012. Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition & emotion* 26, 8 (2012), 1412–1427.

[40] Peter Koval, Madeline L Pe, Kristof Meers, and Peter Kuppens. 2013. Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion* 13, 6 (2013), 1132.

[41] 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.

[42] Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. 2010. Emotional inertia and psychological maladjustment. *Psychological science* 21, 7 (2010), 984–991.

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

[44] Reza Lotfian and Carlos Busso. 2019. Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels. *IEEE Transactions on Audio, Speech &*

Language Processing 27, 4 (2019), 815–826.

[45] Humberto R Maturana and Francisco J Varela. 1987. *The tree of knowledge: The biological roots of human understanding.* New Science Library/Shambhala Publications.

[46] Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A cross-modal review of indicators for depression detection systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical PsychologyÃćÂĂÂŤFrom Linguistic Signal to Clinical Reality*. 1–12.

[47] World Health Organization et al. 2017. *Depression and other common mental disorders: global health estimates.* Technical Report. World Health Organization.

[48] Vedhas Pandit and Björn Schuller. 2019. On Many-to-Many Mapping Between Concordance Correlation Coefficient and Mean Square Error. *arxiv.org* 1902.05180 (February 2019). 23 pages.

[49] 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.

[50] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[51] Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The Psychology of Word Use in Depression Forums in English and in Spanish: Texting Two Text Analytic Approaches.. In *ICWSM*.

[52] Eva-Maria Rathner, Julia Djamali, Yannik Terhorst, Björn Schuller, Nicholas Cummins, Gudrun Salamon, Christina Hunger-Schoppe, and Harald Baumeister. 2018. How Did You like 2017? Detection of Language Markers of Depression and Narcissism in Personal Narratives. *Future* 1, 2.58 (2018), 0.

[53] Eva-Maria Rathner, Yannik Terhorst, Nicholas Cummins, Björn Schuller, and Harald Baumeister. 2018. State of Mind: Classification through Self-reported Affect and Word Use in Speech. In *Proceedings of Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*. ISCA, Hyderabad, India, 267–271.

[54] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Dennis Lalanne, Adrien Michaud, Elvan Ciftci, Hüseyin Gülec, Albert Ali Salah, and Maja Pantic. 2018. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*. ACM, Seoul, South Korea, 3–13.

[55] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2015. AVEC 2015 – The 5th International Audio/Visual Emotion Challenge and Workshop. In *Proceedings of the 23rd ACM International Conference on Multimedia, MM 2015*. ACM, Brisbane, Australia, 1335–1336.

[56] 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.

[57] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2018. Summary for AVEC 2018: Bipolar Disorder and Cross-Cultural Affect Recognition. In *Proceedings of the 26th ACM International Conference on Multimedia, MM 2018*. ACM, Seoul, South Korea, 2111–2112.

[58] 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.

[59] James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (2003), 145.

[60] James A Russell and Ulrich F Lanius. 1984. Adaptation level and the affective appraisal of environments. *Journal of Environmental Psychology* 4, 2 (1984), 119–135.

[61] 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 *Proc. 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* . IEEE, Shanghai, P. R. China, 5800–5804.

[62] Robert M Sapolsky. 2004. Social status and health in humans and other animals. *Annu. Rev. Anthropol.* 33 (2004), 393–418.

[63] 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.

[64] Stefan Scherer, Giota Stratou, Jonathan Gratch, Jill Boberg, Marwa Mahmoud, Albert (Skip) Rizzo, and Louis-Philippe Morency. 2013. Automatic Behavior Descriptors for Psychological Disorder Analysis. In *Proceedings of FG*. IEEE, Shanghai, China.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] Björn W Schuller, Stefan Steidl, Anton Batliner, Peter B Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian Pokorny, Eva-Maria Rathner, Katrin D Bartl-Pokorny, et al. 2018. The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. *Proceedings of INTERSPEECH, Hyderabad, India* 5 (2018).

[71] Norbert Schwarz and Gerard L. Clore. 1983. Mood, misattribution, and judgements of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology* 45, 3 (September 1983), 512–523.

[72] Andreas Schwerdtfeger. 2004. Predicting autonomic reactivity to public speaking: don't get fixed on self-report data! *International Journal of Psychophysiology* 52, 3 (2004), 217–224.

[73] Andreas R Schwerdtfeger and Eva-Maria Rathner. 2016. The ecological validity of the autonomic-subjective response dissociation in repressive coping. *Anxiety, Stress, & Coping* 29, 3 (2016), 241–258.

[74] Caifeng Shan, Shaogang Gong, and Peter W Mcowan. 2009. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image and Vision Computing* 27, 6 (2009), 803–816.

[75] Stewart Shapiro and Deborah J. MacInnis. 2002. Understanding program-induced mood effects: Decoupling arousal from valence. *Journal of Advertising* 31, 4 (May 2002), 15–26.

[76] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[77] Lukas Stappen, Nicholas Cummins, Eva Messner, Harald Baumeister, Judith Dineley, and Björn W. Schuller. 2019. Context Modelling Using Hierarchical Attention Networks for Sentiment and Self-assessed Emotion Detection in Spoken Narratives. In *Proc. 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom, 6680–6684.

[78] 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 *Proc. 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Shanghai, China, 5200–5204.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] Kalani Wataraka Gamage, Ting Dang, Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. 2018. Speech-based Continuous Emotion Prediction by Learning Perception Responses Related to Salient Events: A Study Based on Vocal Affect Bursts and Cross-Cultural Affect in AVEC 2018. In *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*. ACM, Seoul, South Korea, 47–55.

[84] 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.

[85] James R. Williamson, Thomas F. Quatieri, Brian S. Helfer, Rachelle Horwitz, Bea Yu, and Daryush D. Mehta. 2013. Vocal Biomarkers of Depression Based on Motor Incoordination. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC’13)*. ACM, New York, NY, USA, 41–48. https://doi.org/10.1145/2512530.2512531

[86] 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.

[87] 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.

[88] 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.

[89] Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. 2018. Multi-modal Multi-cultural Dimensional Continues Emotion Recognition in Dyadic Interactions. In *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC’18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*. ACM, Seoul, South Korea, 65–72.