Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics

Heysem Kaya
Namık Kemal University
Çorlu, Turkey

Dmitrii Fedotov*
Ulm University
Ulm, Germany

Denis Dresvyanskiy
Ulm University
Ulm, Germany

Metehan Doyran
Utrecht University
Utrecht, The Netherlands

Danila Mamontov
Ulm University
Ulm, Germany

Maxim Markitantov
SPIIRAS
St. Petersburg, Russia

Alkim Almila Akdag Salah[†]
Utrecht University
Utrecht, The Netherlands

Evrim Kavcar
Mardin Artuklu University
Mardin, Turkey

Alexey Karpov
SPIIRAS
St. Petersburg, Russia

Albert Ali Salah[‡]
Utrecht University
Utrecht, The Netherlands

ABSTRACT

Cross-language, cross-cultural emotion recognition and accurate prediction of affective disorders are two of the major challenges in affective computing today. In this work, we compare several systems for Detecting Depression with AI Sub-challenge (DDS) and Cross-cultural Emotion Sub-challenge (CES) that are published as part of the Audio-Visual Emotion Challenge (AVEC) 2019. For both sub-challenges, we benefit from the baselines, while introducing our own features and regression models. For the DDS challenge, where ASR transcripts are provided by the organizers, we propose simple linguistic and word-duration features. These ASR transcript-based features are shown to outperform the state of the art audio visual features for this task, reaching a test set Concordance Correlation Coefficient (CCC) performance of 0.344 in comparison to a challenge baseline of 0.120. Our results show that non-verbal parts of the signal are important for detection of depression, and combining this with linguistic information produces the best results. For CES, the proposed systems using unsupervised feature adaptation outperform the challenge baselines on emotional primitives, reaching test set CCC performances of 0.466 and 0.499 for arousal and valence, respectively.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning approaches*; *Feature selection*; • **General and reference** → **Performance**;

KEYWORDS

Affective Computing; Depression Severity Prediction; PTSD; Cross-Cultural Emotion Recognition

ACM Reference Format:
Heysem Kaya, Dmitrii Fedotov, Denis Dresvyanskiy, Metehan Doyran, Danila Mamontov, Maxim Markitantov, Alkim Almila Akdag Salah, Evrim Kavcar, Alexey Karpov, and Albert Ali Salah. 2019. Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *9th International Audio/Visual Emotion Challenge and Workshop (AVEC '19), October 21, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3347320.3357691

1 INTRODUCTION

Multimodal affective computing is gaining momentum in both science and technological domains, while a large set of unsolved problems remain, including, but not limited to cross-corpus and cross-cultural emotion recognition and real life robustness. Predicting the severity level of affective disorders such as unipolar and bipolar depression also remains a challenge to cope with [41, 55, 57].

Thanks to the challenges organized in the field that introduce new data and tasks with a common protocol to the community, the sharing of resources and comparability/transparency of the works have been boosted. Challenge competitions help bridge the comparability and state-of-the-art (SoA) in the field, bringing together teams from multi-disciplinary backgrounds, such as the AVEC challenge series that were initiated in 2011 [46]. Enjoying the ninth competition in the series, AVEC 2019 presents three sub-challenges, namely, predicting i) State of Mind (SoMS), ii) Depression with AI (DDS) and iii) Cross-cultural Emotion (CES) [17]. In this paper, we tackle the latter two, namely DDS and CES, presenting our proposed methods and results.

[*] The author is further affiliated with ITMO University, St. Petersburg, Russia.
[†] The author is further affiliated with Sehir University, Istanbul, Turkey.
[‡] The author is further affiliated with Boğaziçi University, Istanbul, Turkey.

AVEC '19, October 21, 2019, Nice, France
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6913-8/19/10…$15.00
https://doi.org/10.1145/3347320.3357691

The contribution of this paper is threefold. First, we introduce a simple but very effective set of Automatic Speech Recognizer (ASR) transcription based features, which are shown to generalize better than state of the art audio and video modality features for depression severity level prediction. Second, we annotate a small portion of the training data for DDS and experiment automatic segmentation, in scope of an effort to extract useful information from silences and non-linguistic vocalizations. Finally, we present our own systems for DDS and CES challenges.

The remainder of the paper is organized as follows. In the next section we provide a brief review of literature on depression detection/severity estimation and cross-corpus emotion recognition. In Section 3, we cover the corpora and baseline features used in our experiments, including the additional corpora used for cross-cultural emotion recognition. Sections 4 and 5 present our proposed methods for DDS and CES sub-challenges, respectively. The experimental results are given in Section 6. Finally, we conclude with Section 7.

2 RELATED WORK

We summarize below the state-of-the-art in the automatic analysis of affect with a focus on: (i) depression detection, and (ii) cross-cultural emotion recognition.

2.1 Predicting Depression Severity Level

Depression, affecting more than 300 million people as the World Health Organisation (WHO) declared in 2015 [36], is a common mental health disorder which alters one's actions, thoughts, and feelings negatively [4] and affects the social life of the patients as well as the society as whole. Often, depression patients suffer from multiple comorbidities, and the medical costs of the disease is estimated to be very high, economically [21].

Automatic depression detection has received some interest in the literature in the last ten years. Language usage [3], facial actions [10], vocal prosody [10], and speech [11] have been known to be associated with mental health, and were used as indicators for automatic analysis. These indicators can be applied not only to identify people with depression, but also for quantifying the severity of depression. Public challenges and datasets focus on different modalities and help researchers to extend the collective knowledge on automatic depression detection. Morales et al. [35] recently compared different public depression detection datasets, as well as performances of some approaches for automatic depression detection. While the ReachOut Triage Shared Task [34], and SemEval-2014 Task 7 [37] datasets focused only on conversation transcriptions (i.e. text data) of people with depression, AVEC 2013 [58], and AVEC 2014 [57] datasets had audiovisual modalities. The DementiaBank [6] and DAIC-WOZ [20] datasets had all three modalities (audio, visual, and text).

In this problem, multimodal approaches are gaining popularity, since they seem to be outperforming unimodal methods in depression detection. Various levels of fusion methods are tried to combine bag of audio features with bag of video features for diagnosing depression [27]. Semantic features are combined with audiovisual features to create context-aware methods [19].

Transfer learning is being used by many researchers to extract visual features using deep neural networks [26, 45]. These networks are pre-trained on different datasets, depending on the task. The main idea is to use very large, annotated datasets to learn good low-level and mid-level representations, and fine-tune these networks with smaller amounts of task-specific datasets.

Deep learning training is also applied to automatic depression detection in an end-to-end fashion [62]. Zhu *et al.* [63] used a two-stream deep convolutional neural network architecture [53], where one network is used for appearance (takes regular images), and the other for dynamics (takes optical flow images), respectively. Furthermore, researchers use recurrent neural networks and more specifically, long short-term memory modules [7] to capture longitudinal features.

2.2 Cross-cultural Emotion Recognition

Cross-cultural emotion recognition was first introduced as a sub-challenge in AVEC 2018, Cross-cultural Emotion Sub-challenge (CES) [41]. AVEC 2019 CES includes an additional Chinese corpus on top of Hungarian and German corpora introduced previously in AVEC 2018 CES [17].

There were few researchers focusing on the universality of emotional expressions across cultures before AVEC 2018 CES [14, 54]. The results of AVEC 2018 CES support the idea that facial expressions are much more universal than speech-based emotion expression, as vision-only approaches outperform speech-only methods [8, 17, 25, 40, 60]. Linguistic variation can be a major issue when learning from one corpus and testing it on another.

Researchers employ hidden Markov models (HMMs) [33], long short-term memory recurrent neural networks (LSTM-RNNs) [28, 59], Bidirectional LSTMs (BLSTMs) [33] to extract longitudinal features. In cross-cultural settings, domain adaptation techniques are shown to be vital [1]. Normalization-based and machine learning-based adaptation techniques are often used. Corpus-level and speaker-level normalization are two straightforward domain adaptation techniques that are used widely [18, 47–49]. Cascaded speaker-level normalization combines feature-, value- and instance-level normalization [29]. Transfer learning methods [22, 65], denoising auto-encoders (DAEs) [12, 13], Principal Component Analysis (PCA), and Canonical Correlation Analysis (CCA) [43] are proposed to solve the domain adaptation problem.

3 CORPORA AND BASELINE FEATURES

The AVEC 2019 Detecting Depression with AI Sub-Challenge (DDS) Task [17] extends the previously conducted AVEC 2016 Depression Severity Challenge [56], where US Army veterans were (clinically) interviewed with a virtual agent in a Wizard-of-Oz (WoZ) setup. The veterans' depression severity levels were assessed with a PHQ-8 questionnaire. The DAIC-WOZ corpus [20] used in the 2016 challenge is extended with new recordings in the test partition, where the WoZ setup is replaced by a fully autonomous conversational system. Performance is evaluated with the Concordance Correlation Coefficient (CCC) [31].

For cross-cultural emotion recognition, a number of corpora are used. The SEWA database consists of audiovisual recordings of spontaneous emotional behaviour in-the-wild [30]. Participants watch a set of commercials and discuss the last one with an interlocutor. This year, subjects represent German, Hungarian and Chinese cultures. The dataset is divided into three partitions: training (169 minutes), development (67 minutes) and test (281 minutes). Chinese data are neither presented in the training, nor in the development partitions, and make approximately 70% of all the test data.

In addition to the SEWA corpus provided for CES, we use two external corpora with similar annotations and experimental procedure, namely, RECOLA [42] and SEMAINE [32]. They both have audio-visual data annotated time-continuously for arousal and valence. SEMAINE (Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression) database was designed to develop systems for machine-human interactions. It consists of three parts, representing interaction between a user and a sensitive artificial listener (SAL) on three levels. In our work, we use the solid SAL part, in which a human-operator simulates an agent. It consists of 95 audio-visual recordings with the total duration of 394 minutes. The language of interaction is English. RECOLA (Remote COLlaborative and Affective interactions) database was collected during spontaneous dyadic interactions between people while solving a cooperative problem. Recordings in French from 23 users are presented in the current version of the database, shared with the research community. Each recording has duration of five minutes, yielding 115 minutes of data in total.

Although additionally introduced corpora present European cultural background of participants, they expand original language span of the challenge corpus.

The baseline paper of AVEC 2019 [17] presents a rich set of audio and video feature sets. The baseline sets are categorized into three groups, namely, i) expert-knowledge based (such as Mel-frequency cepstral coefficients (MFCCs), extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [15] acoustic set extracted using openSMILE tool [16], and Facial Action Units (FAU) extracted via OpenFace [5]), ii) Bag-of-Words (BoW) representations of MFCC, eGeMAPS and FAUs, and iii) Deep representations of face and MFCC spectrograms. For further details on the baseline sets, the reader is referred to the paper on the challenge [17].

4 PROPOSED METHOD FOR DDS

The DDS sub-challenge is annotated both for depression severity (PHQ) level and Post Traumatic Stress Disorder (PTSD) level for each clip/participant. Only the audio modality signal is given, along with audio and video modality features. We use the baseline feature sets, which reflect the state-of-the-art in audio and video feature representation, and our own features to augment these. The suprasegmental feature vectors extracted from the audio, video and text modalities ultimately represent the whole clip and are used to predict the PHQ and PTSD levels.

4.1 Feature Extraction

In addition to the rich set of audio-visual challenge baseline features, we extracted two sets of features. The first makes use of the ASR transcripts provided by the organizers and the second is based on automatic segmentation of the speech signal to catch non-verbal aspects of depressed speech, such as silences and breathing. The proposed features are described in the subsequent sub-sections.

4.1.1 ASR Transcriptions based Features for DDS. Apart from the baseline feature sets, the organizers of AVEC 2019 provided Automatic Speech Recognizer (ASR) transcripts that include start time and end time of the subject, as well as the confidence of the automatically recognized words in the corresponding interval. While the ASR transcriptions are not perfect and do not provide sentence level information for high-level Natural Language Processing (NLP), they allow extraction of speech duration and analysis on the word level. Note that the baseline systems do not employ any ASR based features despite the provided transcripts. For the DDS challenge, we decided to investigate this modality with a hypothesis that depressed people's timing during conversations (duration of silences, word speaking rate, etc.) and their choice of words, or word repetition will show telling patterns.

For this purpose, we extracted four ASR-based low level descriptors (LLD) from each transcript record t; namely, word count (f_1^t), speech duration(f_2^t), words per second ($f_1^t = f_1^t / f_2^t$), and inter turn duration ($f_4^t = startTime^{t+1} - endTime^t$). These four ASR LLDs are then passed through ten functionals to obtain a high-level, fixed length representation over the whole transcript. The functionals employed are listed in Table 1. These $4 \times 10 = 40$ functional features are augmented with overall words per second and number of repetitions (total number of words subtracted from the number of unique words), which make up 42 ASR statistics based features for each transcript.

Functional	Description
Mean	Arithmetic mean
Std	Standard deviation
Curvature	Leading coefficient of the second order polynomial fit to the LLD contour
Slope and offset	Coefficients of the first order polynomial fit to the LLD contour
Min	Minimum value
Relative Min Location	Location index of min value divided by the length of LLD contour
Max	Maximum value
Relative Max Location	Location index of max value divided by the length of LLD contour
ZCR	Zero crossing rate of the LLD contour normalized into [-1,1] range

Table 1: List of statistical functionals applied to LLDs.

The words from ASR are also used for a simple bag-of-words representation, where stemming is not employed, only apostrophes (') and full stops (.) are removed from each word and the numeric words are removed from the word bag. Once the word bag is formed, the words from each ASR transcript document are represented with their corresponding term frequencies. Because there are only 163 clips/participants in the training set and over 8K words in the bag, Principal Component Analysis (PCA) is applied prior to regressor modeling, where the number of PCA eigenvectors are optimized on the development set.

4.1.2 Automatic Segmentation for Silences and Non-Linguistic Vocalizations.

We decided to use the inter turn duration from the ASR transcripts in order to estimate the total duration of silences and breathing for each participant. However, these episodes (where the participant's speech is not recognized) not only contain silences and breathing, but also the speech of the virtual agent. We aimed to discriminate between the segments corresponding to the subject's speech, subject's non-linguistic vocalizations (such as breathing, lip noise, laughter and fillers), silences, and the virtual agent's speech. To learn an automatic segmentation, a subset of the training set (17 audio files with varying PHQ levels, amounting to 138 minutes) are partially or fully annotated for 7 segment classes: AI (virtual agent's speech), B (breathing), F (fillers), LN (lip noise), LA (laughter), SI (silence) and S (subject's speech). We also annotated the host-patient dialogue segment and removed this part from the training of the classifier.

For automatic segmentation modeling, we experimented with eGeMAPS [15] and Deep Spectrum (VGG-16 [50]) features. eGeMAPS LLDs were summarized over non-overlapping windows of {100, 200, 400, 500, 1000} milliseconds (ms) with mean and std functionals, while the window size is optimized using a portion of the annotated data as development set. The suprasegmental features are mapped to the majority class in the corresponding window. The organizer provided Deep Spectrum VGG-16 (DS-VGG) features are extracted from 4 second windows with 1 second shifts. We use each DS-VGG feature to represent a one second slice of the signal and map it to the majority segment class during training.

The preliminary experiments for seven-class classification for speech segmentation using the eGeMAPS LLDs showed that the classes of interest (especially breathing, laughter and fillers) are recognized either poorly or not at all. The reason of this is due to: i) relatively low number of samples ii) high acoustic similarity between the non-linguistic vocalizations and linguistic speech (see Figure 1). We therefore combined B, F, LA and LN classes into a new class, dubbed NLV (non-linguistic vocalizations) and proceeded the experiments with a four-class segmentation problem.

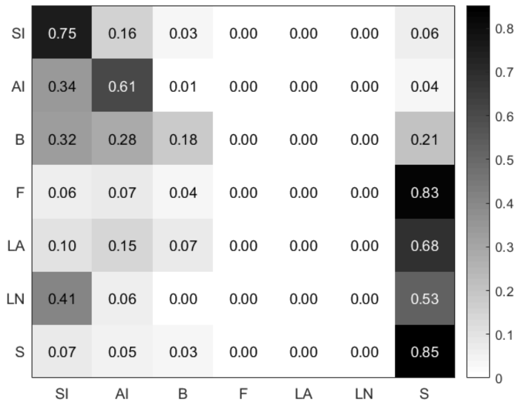

Figure 1: Confusion matrix using seven classes for speech segmentation.

Once the models using eGeMAPS and VGG-16 features are optimized, we applied the best models to the training, development and test set audio signals and extracted i) duration features for SI, NLV and S classes, ii) computed mean of the acoustic functional features corresponding to detected segments of SI, NLV and S, separately. The duration features extracted for each segment class (e.g. SI) consist of the number of turns, total duration, average, minimum, maximum and range of duration, as well as the ratio of the total duration for the class of interest to the signal length. The segment level duration and acoustic features are used to predict depression level (measured with PHQ) and post traumatic stress disorder (PTSD) level.

4.2 Model Learning

The clip level suprasegmental features are modelled using Kernel Extreme Learning Machines (KELM). ELM family is introduced as a non-backpropagation based, fast and robust learning strategy, initially for single hidden layer feedforward neural networks [24], and later extended to deep neural networks and kernel machines [23]. The main theme of the approach is to randomly generate the first layer weights and learn the second layer weights analytically, using least squares regression. A special pseudo-inverse, namely *Moore-Penrose Pseudo-inverse* [39] that minimizes both the norm of the projection weights and the error simultaneously, is used in the original *basic ELM* version [24]. Subsequently, a regularization hyper-parameter C is introduced for relaxing optimization and increasing generalization [23], relating it to ridge regression. This was further extended to include the kernel trick for the first later. Let K denote the $\mathbb{R}^{N \times N}$ training kernel and T denote $\mathbb{R}^{N \times L}$ target matrix (encoded using one-hot encoding for classification tasks), where N and L are the number of training set instances and the number of classes, respectively; the second layer projection matrix β is computed via $\mathbf{H}$ [23]:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}, \tag{1}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. We employ popularly used linear and Gaussian (radial basis function (RBF) kernels in our experiments, optimizing the RBF kernel scale and complexity parameter C on the challenge development set.

5 PROPOSED METHOD FOR CES

5.1 Feature extraction and pre-processing

According to baseline results for CES, presented in [17], more complex deep-learning based feature sets (Deep Spectrum, ResNet, VGG) do not provide better results compared to expert-knowledge ones, such as eGeMAPS and FAUs. Therefore, we have chosen the latter for our further experiments. BoAW-e have often provided better results compared to eGeMAPS based functionals, but they strongly rely on the data and require re-evaluation once something is changed. Moreover, they are harder to transfer for cross-corpus training.

In order to isolate speaker's speech from noise and speech of the interlocutor, we cleaned the audio data, according to turn metadata provided by organizers. After that, we re-extracted suprasegmental features, trying different window lengths (2, 4, 6, 8 seconds). However, we have not noticed any significant performance gains, hence we decided to proceed with the original 4 seconds window. Further pre-processing includes feature transformation with PCA and Canonical Correlation Analysis (CCA) for unsupervised cross-corpus feature adaptation, as suggested by Sagha et al. [44]. For PCA, we tried reducing the feature set, keeping 95%, 98%, 99%, 99.99% of the total variance. Transformation usually led to improvements in terms of development set performance with 99.99% variance being best for audio features (eGeMAPS) and 98% for video (FAUs). The number of CCA components were analysed and selected for each experiment independently via grid search, further reducing the PCA features. The general trends for CCA showed better performance on the development set with a higher number of components (close to the number of components in PCA), which may be a result of high level of data representation provided by expert-knowledge sets and the small amount of redundancy.

Feature sets are normalized with the Z-transform. For cross-corpus experiments, each corpus is normalized separately, in order to level the corpora differences, especially for audio, in terms of recording devices, environmental noises, etc.

5.2 Model Learning

We use a methodology similar to the one provided in the baseline paper, with changes in network architecture. Our network consists of three gated recurrent unit (GRU) [9] layers with 200, 100 and 200 nodes (linear activation function), respectively. The network has less nodes in the middle layer in order to create a bottleneck effect [61]. The model is optimized with Adam using CCC-based loss function for 20 epochs with a learning rate of 0.001.

In addition to direct modeling used in the baseline, we use two additional approaches: cross-corpus training and interlocutor dependent modeling, respectively.

For cross-corpus modelling, we train the model in two stages: (i) after separate pre-processing of each corpus, we combine data from additional corpora (i.e. RECOLA and SEMAINE) and train the model on them; (ii) then we fine-tune the model with train partition of SEWA corpus, keeping track of the development performance. For fine-tuning, we fix the weights of the first two layers and adjust only the parameters of the last GRU layer and fully connected regression output layer. Assuming that our model will extract representational maps relevant to emotional dimensions from additional corpora before the bottleneck layer (second GRU layer), we then allow it to adapt its time-continuous behaviour to the target corpus.

For interlocutor-dependent modeling, we assume that the emotional flow of the conversation is comprised of mutually created emotions of both interlocutors. Therefore, the analysis of both sources of data jointly may allow us to describe the flow more precisely. In this paper, we use feature-level fusion of audio-visual data from both interlocutors prior to training the model. Each audio recording in the SEWA database corresponds to visual data from two speakers, recorded with two different cameras. We automatically detect the interacting pairs from audio content/length similarity and use both video streams for each audio file. We first combine features within each modality, and then use PCA for dimensionality reduction. The test set is used to learn the PCA space, which corresponds to target domain adaptation. This can only be done when the testing is done wholesale, and in the offline mode. Both the training and test sets are processed with PCA, and the model described above is applied.

6 EXPERIMENTAL RESULTS

6.1 DDS Experiments

From the baseline feature sets, we select the best performing audio modality feature (DS-VGG) and best performing video modality feature (ResNet), based on their CCC performance on the development set. These features are summarized over the whole clip using mean, std and curvature functionals, which are described in Table 1. Different subsets of functionals with each feature, feature- and decision-level fusion strategies are experimented with. For all experiments, regression on depression severity level is first probed, and the best systems (with same or very similar hyper-parameters) are used for PTSD level prediction.

DS-VGG features that are reported to have the highest CCC performance (0.305) on the development set, performed very poorly with KELM (0.07 CCC). On the other hand, the ResNet Features summarized using the aforementioned three functionals (abbrv. ResNetX3) reach a development set CCC performance of 0.468. Note that when only mean and std functionals are used with ResNet, the KELM performance is 0.364, which is higher than the corresponding performance reported in the challenge paper (0.269) but lower compared to the use of mean, std and curvature.

We next experiment with our proposed ASR transcription-based features. 42 duration and word-count-based features (abbrv. ASR_WordDur) reach a development set CCC score of 0.382, which outperforms all development set scores reported in the challenge paper [17]. The ASR transcription-based simple BoW features are transformed with PCA, and the number of eigenvectors are optimized in the [10, 160] range with steps of 10. BoW features with the top 30 PCA dimensions (abbrv. ASR_BoWPCA30) provided the highest CCC performance on the development set (0.444). A summary of PHQ and PTSD prediction performances for the best performing features are shown in Table 2. Although the features and classifier hyper-parameter ranges are optimized for the PHQ task, PTSD prediction performances are always higher than those of PHQ.

Table 2: Top performing feature types and their development set CCC performances on depression severity (PHQ) and PTSD prediction tasks.

Model	Feature	PHQ	PTSD
M1	ResNetX3	0.468	0.526
M2	ASR_BoWPCA30	0.444	0.508
M3	ASR_WordDur	0.382	0.431

Our first two test set submissions for the DDS are simple and weighted fusions of single-feature type systems reported in Table 2. Our first submission takes the average of the predictions from ASR_BoWPCA30 and ASR_WordDur based KELM models. Our second submission uses a weighted fusion of the three models' predictions, where a pool of randomly generated fusion weights are applied and the one yielding the highest CCC score is selected. The development and test set CCC performances of these fusion systems are reported in Table 3. Here, we observe that despite the outstanding development set performance of the ResNet based model in single and combined settings, the fusion system that employs ResNet performs slightly poorer than the challenge baseline (0.120) on the test set. The best test set performance (0.344 CCC) is obtained using simple ASR transcription-based features (i.e. word count, duration and BoW).

Table 3: Development and Test Set CCC Scores for ASR- and ResNet-features based Fusion Systems. EF: Equal weighted fusion, WF: Weighted Fusion.

	PHQ		PTSD	
System	Devel.	Test	Devel.	Test
Baseline [17]	0.269	0.120	NA	NA
WF(M1, M2, M3)	0.606	0.118	0.625	0.141
EF(M2, M3)	0.481	**0.344**	0.554	0.376

Motivated by the results of the ASR feature-based system, we next experimented with automatic segmentation based features. To obtain an automatic speech segmentation we trained classifiers with varying window lengths, two acoustic feature types as explained in Section 4.1.2. In addition to KELM, we use a special variant of KELM dubbed *Weighted Kernel ELM* [64] that gives higher importance weights to minority class instances and hence inherently tries to maximize unweighted average recall (UAR). The best segmentation UAR performance (65.75%) on the validation set, which is composed of 4 out of 17 annotated audio files, is obtained with 500 ms non-overlapping windows using functionals of eGeMAPS LLDs and Weighted KELM as classifier. The corresponding confusion matrix is shown in Figure 2. We should note that the recall for NLV and SI are higher compared to the use of KELM, but still low for subsequent feature extraction from predicted segments.

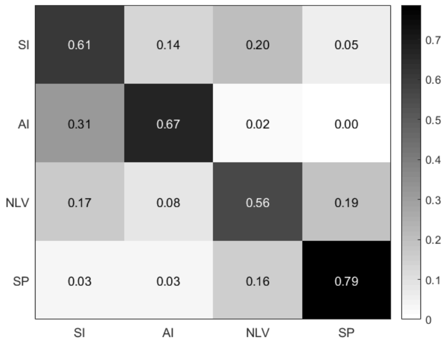

Figure 2: Confusion matrix of the best model for speech segmentation.

We have extracted the segment-dependent duration and acoustic features and carried out experiments for predicting depression severity level. The best performance of these features on the development set was 0.193 CCC; a score below what is reached with simple ASR transcript-based word duration features (ASR_WordDur). When these features were combined with the ASR_WordDur features, the performance of ASR_WordDur features dropped severely. The performance of ASR_WordDur features increased only when duration features corresponding to the predicted silence episodes were combined. When combined with only six silence segment duration-based features, CCC performance of ASR_WordDur features (0.382) increased to 0.420 and 0.431, using Linear (Model M4) and RBF kernels (Model M5), respectively.

The third and fourth submissions are dedicated to these systems. In the third submission, a weighted fusion of M2 (ASR BoW based model), M4 and M5 is employed, whereas submission four uses an unweighted average of M4 and M5. The development and test set CCC scores of these systems are summarized in Table 4, where we observe reduced performance with respect to the first submission using only ASR transcription-based features. Furthermore, without ASR BoW features, ASR duration features combined with silence duration features completely fail to generalize (test set CCC score: 0.028). This performance reduction is largely attributed to the automatic segmentation problems, which will be a focus in our future work.

Table 4: CCC Performance of Systems Using ASR and Automatic Segmentation based Features

	PHQ		PTSD	
System	Devel.	Test	Devel.	Test
WF(M2, M4, M5)	0.515	0.252	0.561	0.365
EF(M4, M5)	0.450	0.028	0.458	0.169

The test set results indicating the importance and robustness of linguistic features in the depression severity level prediction are in line with recent works reporting results on DAIC-WOZ corpus [38, 51, 52]. In particular, Stepanov et al. [51] used BoW representation of the transcripts without feature reduction and modeled the high dimensional features with Support Vector Regressor (SVR). They also extracted behavioural features (such as response duration, count of non-verbal signals and laughter events) from transcripts. The BoW representation and transcript-based behavioural characteristics features were the top performing on the test set, where the authors did not combine the modalities. We should note that our BoW representation gave a low (around 0.17 CCC) development set performance without PCA transformation, which was not experimented in [51]. Moreover, the provided transcripts in AVEC 2019 DDS do not contain non-verbal signals such as laughter and the virtual agent's speech segments were not annotated. Our manual annotation of a portion of the training set for extracting behavioral patterns such as response time after virtual agent's questions, silences and non-linguistic vocalizations was not sufficient for segmentation. We will work in this direction using the DAIZ-WOZ and other corpora for extracting a compact set of higher-level, explainable and predictive features.

6.2 CES Experiments

We follow a different approach in cross-cultural emotion sub challenge, where silences and breathing are not supposed to play such a prominent role. According to pre-processing and modeling procedures, mentioned in Section 5, we first apply PCA to eGeMAPS based functionals with 4 seconds window length, extracted from turn-cleared audio data and fused at feature level with FAUs (S1). This approach provides an improvement in both arousal and valence for German and Chinese datasets (see Table 5). On the Hungarian dataset, such a system performs worse, as well as on liking dimension for Chinese.

As an extension of the approach described above, we implement a cross-corpus system by integrating two additional languages, English and French. PCA transformation parameters were trained on target corpus and then applied to these additional corpora after normalization (S2). As no "liking" dimension is provided for RECOLA or SEMAINE, we make predictions only for arousal and valence. This approach leads to further improvements for German language and a slight improvement for Hungarian, over the S1 model. However, the performance for Hungarian language is still lower than baseline and significant performance loss is noticeable for Chinese data.

We further extend the second approach and add CCA into the pre-processing pipeline on the same conditions (S3). This results in performance losses for both German and Hungarian languages, but the Chinese part performance remains at the same level (even slightly higher).

Our fourth approach is PCA applied to feature-level fusion of speaker's and the interlocutor's data (S4). Here, we have only the target corpus for training, therefore predictions are available for each dimension. In general, this approach performs worse than the baseline with one exception for German language on the arousal dimension.

For the last system, we use weighted fusion of predictions obtained with our best system on the Chinese data for each dimension and baseline predictions, provided by organizers (S5). For arousal, we fuse S1 with eGeMAPS-BoAW and VGG with weights optimized on development set: 0.66, 0.17, and 0.17, respectively. For valence we fuse S1 with FAUs and VGG, using the following weights: 0.54, 0.22, and 0.24, respectively. For liking we use S4 and only eGeMAPS-functionals with weights 0.64 and 0.36, respectively. Late fusion provided constant performance gain for each language on valence, having two best results out of three (highlighted in bold in Table 5). However, it did not provide better results for arousal and liking. We provide performance on the development set as well (German and Hungarian combined) for the reference.

In the case of this sub-challenge, the best performance across languages was achieved using the simpler models (S1 and S2) alone or in combination with the baseline prediction on different feature sets. Although liking values moderately correlate with valence values, we did not manage to get an improvement on this dimension, using audio and video modality only. It is obvious, that for better results, one should consider using textual modality as a primary one for liking. However, in the case of the lacking transcription or in the real-life scenario, it is crucial to have a multi-language ASR with high level of confidence.

Table 5: CCC Performance of Systems used for CES

	Baseline	S1	S2	S3	S4	S5
			Arousal			
Devel	-	0.620	0.600	0.575	0.579	-
Test-DE	0.562	0.583	**0.641**	0.538	0.632	0.621
Test-HU	**0.527**	0.484	0.507	0.451	0.441	0.501
Test-CN	0.355	**0.466**	0.406	0.407	0.349	0.391
			Valence			
Devel	-	0.598	0.556	0.551	0.524	-
Test-DE	0.646	0.715	0.734	0.688	0.635	**0.750**
Test-HU	**0.548**	0.346	0.359	0.270	0.277	0.462
Test-CN	0.468	0.483	0.376	0.384	0.389	**0.499**
			Liking			
Devel	-	0.218	-	-	0.226	-
Test-DE	0.074	**0.176**	-	-	0.132	0.106
Test-HU	**0.089**	0.038	-	-	-0.008	0.016
Test-CN	**0.041**	-0.073	-	-	-0.051	-0.032

7 CONCLUSIONS

In this paper, we presented our proposed methods and results for the AVEC 2019 sub-challenges on Depression with AI (DDS) and Cross-cultural Emotion (CES) [17].

Our results on depression detection show that while the patterns in non-verbal parts of the signal are important, combining this with linguistic information produces the best results, without using state of the art acoustic or visual features. While it is possible to hear distinct patterns of breathing in the speech of subjects with depression and high PTSD levels [2], automatic approaches did not suffice to recognize these breathing episodes accurately. For future work, we plan to increase the amount of annotated data, and introduce other methods to deal with the class-imbalance problem to exploit the non-linguistics vocalizations and silence episodes.

In the challenging cross-cultural emotion recognition task, we cannot emphasize one particular approach that leads to the highest results across modalities and languages. Instead we proposed to utilize several methods of data pre-processing and modeling, through which we have achieved 31.3% of relative improvement for arousal and 6.6% for valence on Chinese data with cross-cultural setting. Better results were also obtained within the original culture for German language: 14.1% on arousal and 16.1% on valence. Moreover, the cross-cultural liking dimension did not benefit from audio and video features. Including interlocutor's data into modelling and treating interactions as mutually effecting did not show the best results in this challenge, although, provided minor improvements over baseline models in some cases for German language. It may be useful for future work to not consider this data simultaneously, but track the general emotional flow of interaction independently. PCA-CCA based analysis in cross-corpus scenario brought minor improvements over only PCA based approach, but did not exceed other methods in any separate case.

ACKNOWLEDGMENTS

This study was partially conducted within the framework of the Russian Science Foundation project No. 18-11-00145.

REFERENCES

[1] Mohammed Abdelwahab and Carlos Busso. 2015. Supervised domain adaptation for emotion recognition from speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5058–5062. https://doi.org/10.1109/ICASSP.2015.7178934

[2] Alkim Almila Akdag Salah, Meral Ocak, Heysem Kaya, Evrim Kavcar, and Albert Ali Salah. 2019. Hidden in a Breath: Tracing the Breathing Patterns of Survivors of Traumatic Events. In *Digital Humanities*.

[3] 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. https://doi.org/10.1162/tacl_a_00111

[4] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[5] 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.

[6] James T. Becker, François Boiler, Oscar L. Lopez, Judith Saxton, and Karen L. McGonigle. 1994. The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis. *JAMA Neurology* 51, 6 (06 1994), 585–594. https://doi.org/10.1001/archneur.1994.00540180063015

[7] Linlin Chao, Jianhua Tao, Minghao Yang, and Ya Li. 2015. Multi task sequence learning for depression scale prediction from video. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 526–531. https://doi.org/10.1109/ACII.2015.7344620

[8] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC '17)*. ACM, New York, NY, USA, 19–26. https://doi.org/10.1145/3133944.3133949

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[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 *Proceedings of the 3rd biannual Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, Amsterdam, Netherlands.

[11] 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 (2015), 10 – 49. https://doi.org/10.1016/j.specom.2015.03.004

[12] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. 2018. Semisupervised Autoencoders for Speech Emotion Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 1 (Jan 2018), 31–43. https://doi.org/10.1109/TASLP.2017.2759338

[13] Jun Deng, Zixing Zhang, Florian Eyben, and Björn Schuller. 2014. Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition. *IEEE Signal Processing Letters* 21, 9 (Sep. 2014), 1068–1072. https://doi.org/10.1109/LSP.2014.2324759

[14] 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.

[15] 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.

[16] 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.

[17] Fabien Ringeval and Björn Schuller and Michel Valstar and Nicholas Cummins and Roddy Cowie and Leili Tavabi and Maximilian Schmitt and Sina Alisamir and Shahin Amiriparian and Eva-Maria Messner and Siyang Song and Shuo Lui and Ziping Zhao and Adria Mallol-Ragolta and Zhao Ren, and Mohammad Soleymani, and Maja Pantic. 2019. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In *Proceedings of the 9th International Workshop on Audio/Visual Emotion Challenge, AVEC'19, co-located with the 27th ACM International Conference on Multimedia, MM 2019*, Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, and Maja Pantic (Eds.). ACM, Nice, France.

[18] Silvia Monica Feraru, Dagmar Schuller, and Björn Schuller. 2015. Cross-language acoustic emotion recognition: An overview and some tendencies. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 125–131. https://doi.org/10.1109/ACII.2015.7344561

[19] 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 '17)*. ACM, New York, NY, USA, 69–76. https://doi.org/10.1145/3133944.3133945

[20] 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 Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Languages Resources Association (ELRA), Reykjavik, Iceland, 3123–3128. http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

[21] Aron Halfin. 2007. Depression: the benefits of early and appropriate treatment. *The American journal of managed care* 13, 4 Suppl (2007), S92–7.

[22] Ali Hassan, Robert Damper, and Mahesan Niranjan. 2013. On Acoustic Emotion Recognition: Compensating for Covariate Shift. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 7 (July 2013), 1458–1468. https://doi.org/10.1109/TASL.2013.2255278

[23] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. 2012. Extreme Learning Machine for Regression and Multiclass Classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 42, 2 (2012), 513–529.

[24] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. 2006. Extreme learning machine: theory and applications. *Neurocomputing* 70, 1 (2006), 489–501.

[25] 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 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC'18)*. ACM, New York, NY, USA, 57–64. https://doi.org/10.1145/3266302.3266304

[26] Asim Jan, Hongying Meng, Yona Falinie Binti A Gaus, and Fan Zhang. 2018. Artificial Intelligent System for Automatic Depression Level Analysis Through Visual and Vocal Expressions. *IEEE Transactions on Cognitive and Developmental Systems* 10, 3 (Sep. 2018), 668–680. https://doi.org/10.1109/TCDS.2017.2721552

[27] 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 (01 Nov 2013), 217–228. https://doi.org/10.1007/s12193-013-0123-2

[28] Heysem Kaya, Dmitrii Fedotov, Ali Yeşilkanat, Oxana Verkholyak, Yang Zhang, and Alexey Karpov. 2018. LSTM Based Cross-corpus and Cross-task Acoustic Emotion Recognition. In *Proc. Interspeech 2018*. 521–525. https://doi.org/10.21437/Interspeech.2018-2298

[29] Heysem Kaya and Alexey A. Karpov. 2018. Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing* 275 (2018), 1028 – 1034. https://doi.org/10.1016/j.neucom.2017.09.049

[30] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Björn W. 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* abs/1901.02839 (2019).

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

[32] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 1 (2012), 5–17.

[33] A. Metallinou, A. Katsamanis, M. Wöllmer, F. Eyben, B. Schuller, and S. Narayanan. 2015. Context-sensitive learning for enhanced audiovisual emotion classification (Extended abstract). In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 463–469. https://doi.org/10.1109/ACII.2015.7344611

[34] David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 Shared Task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA, 118–127. https://doi.org/10.18653/v1/W16-0312

[35] 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 Psychology — From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Vancouver, BC, 1–12. https://doi.org/10.18653/v1/W17-3101

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

[37] Sameer Pradhan, Noemie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. 54–62. https://doi.org/10.3115/v1/S14-2007

[38] Syed Arbaaz Qureshi, Mohammed Hasanuzzaman, Sriparna Saha, and Gaël Dias. 2019. The Verbal and Non Verbal Signals of Depression–Combining Acoustics, Text and Visuals for Estimating Depression Level. *arXiv preprint arXiv:1904.07656* (2019).

[39] Calyampudi Radhakrishna Rao and Sujit Kumar Mitra. 1971. *Generalized inverse of matrices and its applications.* Vol. 7. Wiley New York.

[40] 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.

[41] Fabien Ringeval, Björn W. Schuller, Michel F. 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 *AVEC@MM*.

[42] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[43] 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 *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5800–5804. https://doi.org/10.1109/ICASSP.2016.7472789

[44] 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.

[45] Albert Ali Salah, Heysem Kaya, and Furkan GürpÄśnar. 2019. Chapter 17 - Video-based emotion recognition in the wild. In *Multimodal Behavior Analysis in the Wild*, Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe (Eds.). Academic Press, 369 – 386. https://doi.org/10.1016/B978-0-12-814601-9.00031-6

[46] 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.

[47] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. 2010. Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing* 1, 2 (July 2010), 119–131. https://doi.org/10.1109/T-AFFC.2010.8

[48] Björn Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll. 2011. Using multiple databases for training in emotion recognition: To unite or to vote?. In *Twelfth Annual Conference of the International Speech Communication Association*.

[49] Björn W. Schuller, Zixing Zhang, Felix Weninger, and Gerhard Rigoll. 2011. Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization.

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

[51] Evgeny A Stepanov, Stephane Lathuiliere, Shammur Absar Chowdhury, Arindam Ghosh, Radu-Laurenţiu Vieriu, Nicu Sebe, and Giuseppe Riccardi. 2018. Depression severity estimation from multiple modalities. In *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 1–6.

[52] Bo Sun, Yinghui Zhang, Jun He, Lejun Yu, Qihua Xu, Dongliang Li, and Zhaoying Wang. 2017. A random forest regression method with selected-text feature for depression assessment. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 61–68.

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[54] Daniel T. Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. 2017. Universals and Cultural Variations in 22 Emotional Expressions Across Five Cultures. *Emotion* 18 (06 2017). https://doi.org/10.1037/emo0000302

[55] 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.

[56] 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.

[57] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*. ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/2661806.2661807

[58] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge (AVEC '13)*. ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/2512530.2512533

[59] Oxana Verkholyak, Dmitrii Fedotov, Heysem Kaya, Yang Zhang, and Alexey Karpov. 2019. Hierarchical Two-level Modelling of Emotional States in Spoken Dialog Systems. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6700–6704. https://doi.org/10.1109/ICASSP.2019.8683240

[60] 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 2018 on Audio/Visual Emotion Challenge and Workshop (AVEC'18)*. ACM, New York, NY, USA, 47–55. https://doi.org/10.1145/3266302.3266314

[61] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King. 2015. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4460–4464.

[62] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. 2018. Visually Interpretable Representation Learning for Depression Recognition from Facial Images. *IEEE Transactions on Affective Computing* (2018), 1–1. https://doi.org/10.1109/TAFFC.2018.2828819

[63] Y. Zhu, Y. Shang, Z. Shao, and G. Guo. 2018. Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics. *IEEE Transactions on Affective Computing* 9, 4 (Oct 2018), 578–584. https://doi.org/10.1109/TAFFC.2017.2650899

[64] Weiwei Zong, Guang-Bin Huang, and Yiqiang Chen. 2013. Weighted extreme learning machine for imbalance learning. *Neurocomputing* 101 (2013), 229 – 242. https://doi.org/10.1016/j.neucom.2012.08.010

[65] Yuan Zong, Wenming Zheng, Tong Zhang, and Xiaohua Huang. 2016. Cross-Corpus Speech Emotion Recognition Based on Domain-Adaptive Least-Squares Regression. *IEEE Signal Processing Letters* 23, 5 (May 2016), 585–589. https://doi.org/10.1109/LSP.2016.2537926