

# Depression Severity Estimation from Multiple Modalities

Evgeny Stepanov  
University of Trento, Italy  
evgeny.stepanov@unitn.it

Stephane Lathuiliere  
INRIA Grenoble, France  
stephane.lathuiliere@inria.fr

Shammur Absar Chowdhury  
University of Trento, Italy  
shammur.chowdhury@unitn.it

Arindam Ghosh  
University of Trento, Italy  
arindam.ghosh@unitn.it

Radu-Laurențiu Vieriu  
University of Trento, Italy  
radulaurentiu.vieriu@unitn.it

Nicu Sebe  
University of Trento, Italy  
niculae.sebe@unitn.it

Giuseppe Riccardi  
University of Trento, Italy  
giuseppe.riccardi@unitn.it

## ABSTRACT

Depression is a major debilitating disorder which can affect people from all ages. With a continuous increase in the number of annual cases of depression, there is a need to develop automatic techniques for the detection of the presence and extent of depression. In this AVEC challenge we explore different modalities (speech, language and visual features extracted from face) to design and develop automatic methods for the detection of depression. In psychology literature, the PHQ-8 questionnaire is well established as a tool for measuring the severity of depression. In this paper we aim to automatically predict the PHQ-8 scores from features extracted from the different modalities. We show that visual features extracted from facial landmarks obtain the best performance in terms of estimating the PHQ-8 results with a mean absolute error (MAE) of 4.66 on the development set. Behavioral characteristics from speech provide an MAE of 4.73. Language features yield a slightly higher MAE of 5.17. When switching to the test set, our Turn Features derived from audio transcriptions achieve the best performance, scoring an MAE of 4.11 (corresponding to an RMSE of 4.94), which makes our system the winner of the AVEC 2017 depression sub-challenge.

## Keywords

Affective Computing; Depression Detection; Machine Learning; Speech; Natural Language Processing; Facial Expressions

## 1. INTRODUCTION

According to the World Health Organization (WHO), depression is a major mental disorder with about 300 million people of all ages affected worldwide. As per the Global Burden of Disease Study [17], depression is the second leading cause of disability worldwide and is on the rise. Depression affects every aspect of a person's life. People affected from depression often suffer from a certain extent of physical and social impairment. Side effects of depression include sleep disruptions or insomnia, drug or alcohol abuse, and overall loss of quality of life. If left untreated it can lead to complications such as reductions in the volume of the hippocampus [47]. Major clinical depression may even lead to

suicide and annually the burden of death due to depression is on the rise. There is growing evidence that depression can cause impairment of the immune function by affecting different immunological pathways such as the central nervous system (CNS), the endocrine system, and the cardiovascular system. This can lead to the development or aggravation of co-morbidities and worsen health conditions in other diseases. Nicholson et al [36], through a meta-analysis of 54 cohort studies which performed follow up analysis of coronary heart diseases (CHD) showed that patients with major depression had an increased risk of developing fatal CHD.

Diagnosis of depression still remains a challenge. Some symptoms of depression are not readily visible to others. Since depressed people often have decreased social contact, detection of the disease becomes difficult. Current diagnosis of depression is dependent on an evaluation by a psychiatrist supported by standard questionnaires to screen for depression. The Personal Health Questionnaire Depression Scale (PHQ-8) Scoring and the Hamilton Depression Rating Scale are two well established tools for the diagnosis of depression. However, these questionnaires need to be administered and interpreted by a therapist. The stigma around the disease and lack of understanding often prevents patients from seeking early psychiatric help.

The growing burden of this disease suggests that there is a need to develop technologies which can aid in automatic detection and effective care of patients suffering from depression. Affective computing focuses on the sensing, detection, and interpretation of affective states of people from interactions with computers or machines. Research on affective computing uses modalities ranging from overt signals such as speech, language and video to covert signals such as heart rate, skin temperature, galvanic skin response to understand the mental and affective states of humans. While the initial goal of affective computing research was to build better computers which could understand and empathize with humans, the same techniques have been applied to turn computers into tools for automatically identifying psychological states and mental health.

Therefore, the motivation of the study, is to explore different sources of information, such as audio, video, language and behavioral cues, to predict the severity of depression. While doing so, we also investigate different feature repre-

sentation and modeling techniques corresponding to each modality for improving the performance of automatic prediction.

The paper is organized as follows. In Section 2, we present the literature review and the state of the art experiments performed for the detection of depression and affective disorders from speech, language, and facial expressions. This is followed by a brief description of the multi-modal data used for the study in Section 3. An overview of the features and experimental methodology used in this study are given in Section 4 and then we provide a conclusion in Section 5.

## 2. STATE OF THE ART - SPEECH, LANGUAGE AND FACIAL EXPRESSIONS

Speech, language and facial expressions are three of the major overt signals which have been widely used for interpreting human psychological states. Automatic analysis of speech has been used for emotion recognition [50, 37], stress detection [24, 55], and mood state characterisation [52, 8]. Natural language and speech processing from diaries and recordings have been used to detect the onset of dementia, alzheimers, and aphasia [49, 19]. Analysis of facial expressions have shown to be highly effective in tracking the progressive degeneration of cognitive health in patients suffering from schizophrenia and bipolar disorder [6].

### 2.1 Speech and Language

Several psychological conditions clearly manifest themselves through changes in speech patterns and language usage. Computational and automatic screening methods have the power to detect micro-changes in speech and language patterns which would otherwise have gone unnoticed. Properties such as speech rate, pause duration and usage of fillers can be indicative of cognitive decline in individuals. Changes in prosody, and fluency can also be useful in detecting mental health changes of depressive patients.

Research on the diagnosis of mental health from speech and language was pioneered by the German psychiatrist Zwirner [57] in early 1930. He designed a device capable of tracking fundamental frequency for the detection of mental health of patients suffering from depression. Newman and Mather [35] in 1938 carried out similar experiments to systematically record patient’s speech as they read pre-defined text and interacted with a psychiatrist. This data was analysed to show that there were distinct speech features such as speech tempo, prosodic pauses, absence of glottal rasping associated with patients suffering from affective disorders.

France et. al [18] performed multivariate feature and discriminant analyses on the speech data from 67 male and 48 female subjects to show that formant and power spectral density (PSD) based features demonstrated the highest discriminative powers for classification in both genders. Pope et al [40] investigated the relationship between anxiety and depression and speech patterns to show that anxiety was positively correlated with speech disturbances and resistivity in speech. He also found that silent pauses were positively correlated with depression [40].

Kaya et al [28] demonstrated that feature selection techniques based on canonical correlation analysis (CCA) can be effective in detecting depression from speech signals.

Wang et al [53] applied data mining techniques to build models which achieved a precision of 80% for detecting de-

pression based on sentiment analysis of users on a Chinese micro-blogging platform. Rumshisky et al [42] demonstrated through a study on 4687 patients that NLP techniques such as topic modeling can be used to improve prediction of psychiatric readmission.

### 2.2 Face Analysis

Facial expressions can be an extremely powerful medium used to convey human overt emotional feedback. In recent times, there has been significant progress in developing methods for facial feature tracking for the analysis of facial expressions and the detection of emotions. Studies have shown that it is possible to effectively detect the presence of pain shown on faces.

Machine learning techniques have been shown to be effective for the automatic detection of pain and mental state from facial expressions. Littleworth et al. [31] used a two-stage system to train machine learning algorithms to detect expressions of real and fake pain. Their classifier obtained an accuracy of 88% compared to an accuracy of 49% demonstrated by naive human subjects used in their study. Ambadar et al [4] demonstrated that analysis of facial expression can be used to classify smiles into three distinct categories - amused, polite and nervous.

One of the most popular technique used for capturing the subtlety and fine-grained variations in facial expression is the Facial Action Coding System (FACS) developed by Ekman and Freisen. The FACS is based on the consensus of the judgment human experts who observe pre-recorded facial expressions and perform manual annotation of FACS codes for each frame. These annotations, which are called action units (AUs) can belong to one of 44 different classes. FACS has been widely used in the field of psychology for measuring emotions, affect, and behavior [12, 5, 43]. More recently [20], FACS have been shown to be correlated with depression severity. Specifically, [20] found that severely depressed subjects are more likely to show fewer affiliative facial action units (AU12 and AU15) and more non-affiliative ones (AU14).

Head pose and eye gaze have also been shown to encode information about depression. For instance [20] observes that an increase in the severity of depression comes with a diminished head motion. Other works [3, 27, 45] have also investigated the link between head pose, eye gaze and depression, all evidence that such a link exists and it is all worth considering.

### 2.3 Combination

Combination of facial expressions, speech and multimodal information can be used to enhance the recognition of human mental state. Busso et al. [7] demonstrated that both feature fusion (early fusion) and decision fusion (late fusion) from the different modalities outperformed individual features-based classification.

Dibeklioglu et al [13] combined speech, facial movement and head movement to achieve an accuracy of 88.9% for the detection of depression from clinical interviews. The accuracy of the combined signal streams exceeded the accuracy of single modalities to show that multimodal measures can be powerful for detection of depression. Alghowinem et al [2] also demonstrated similar findings in their research to show that a combination of head pose, eye gaze and paralinguistic features yielded better performance than unimodal schemes.

**Table 1: Distribution of the AVEC data set into training and development sets for depressed (D) and non-depressed (ND) classes, and overall (ALL).**

	ND		D		ALL
<i>Training</i>	77	(72%)	30	(28%)	107
<i>Development</i>	23	(66%)	13	(34%)	35

### 3. AVEC AUDIO VIDEO DATABASE

The 2017 Audio/Video Emotion Challenge and Workshop (AVEC 2017) “Real-life depression” provides a corpus comprising of audio and video recordings and transcribed speech from the Distress Analysis Interview Corpus (DAIC) [21].

The dataset comprises of recordings from 189 sessions of human agent interaction where each subject was interviewed by a virtual psychologist (see Table 1 for the distribution of labels in the training and development sets). The audio files, transcripts and continuous facial features of the human subject is provided as part of the challenge. The Personal Health Questionnaire Depression Scale (PHQ-8) score of the subjects is also provided in the dataset. The PHQ-8 [30] is a set of 8 short multiple choice questions which has been established as a diagnostic tool for the measurement of the severity of depressive disorders. Automatic estimation of the PHQ-8 score from different modalities such as speech and video can aid in the early detection of depression and monitoring of depressive states. In the AVEC challenge, the goal is to look at different streams of data recorded during a session with the subject to predict the PHQ-8 scores, and to classify the subject as depressed or not.

## 4. EXPERIMENTS

In this section we describe the experiments conducted for the feature extraction and regression experiments conducted on the speech, behavioral, language and

### 4.1 Speech and Behavioral Characteristic Features

#### 4.1.1 Acoustic Features

To understand the predictive characteristics of low-level acoustic feature groups to assess the depression severity of the participant, we extracted low-level descriptors (LLDs) from the participant’s turns in each conversation. For this, we have extracted different groups of low-level features using openSMILE [16], motivated by their successful utilization in several paralinguistic tasks [46, 1, 10, 9]. These sets of acoustic features were extracted with approximately 100 overlapping frames per second and with 25 milliseconds of window. The low-level features are extracted as three groups including:

- Spectral features (**S**) such as energy in spectral bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz), roll-off points (25%, 50%, 70%, 90%), centroid, flux, max-position and min-position.
- Prosodic features (**P**) such as pitch (Fundamental frequency  $f_0$ ,  $f_0$ -envelope), loudness, voice-probability.
- Voice Quality features (**VQ**) such as jitter, shimmer, logarithmic harmonics-to-noise ratio (logHNR).

These low-level features are then projected on 24 statistical functionals, which include range, absolute position of max and min, linear and quadratic regression coefficients and their corresponding approximation errors, zero crossing rate, peaks, mean peak distance, mean peak, geometric mean of non-zero values, number of non-zeros and moments-centroid, variance, standard deviation, skewness, and kurtosis.

#### 4.1.2 Behavioral Characteristics Features

Apart from extracting low-level features from raw speech signals, we also explored the transcription.

We crafted features that can capture information regarding the participant’s non-vocal behavior (NB) along with their turn-taking behaviors (TB) and participants’ Previous Diagnosed Information (PDI) features. The non-vocal behavior ( $|NB| = 3$ ) includes:

- frequency of laughter in participant’s turns.
- percentage of disfluencies in the participant’s turns, which might indicate hesitations.
- counts of cues that might suggest inconvenience like whistling, mumbling, whispering or taking deep breaths among others.

The features that are used to describe the turn-taking behaviors, ( $|TB| = 6$ ) are the first and third quartiles and the median duration of respond time (in seconds) of the participants. Similarly we also extracted statistics for the with-in speaker silence (pause). The respond time represents how long the participants took to respond to the previous turn of the agent.

The PDI feature set ( $|PDI| = 3$ ) contained numerical representations of the response of the participants to queries such as having any Post-traumatic Stress Disorder (PTSD), `ptsd`, depression `dep`, even having any military backgrounds `mb`. Each individual feature is encoded into three values (-1,0,1) where -1 represents the query is not present in the session, 0 presents a disconfirmation (e.g `ptsd=0` means the participant responded as “no” to the previous turn query) and 1 presents confirmation of the query.

#### 4.1.3 Methodology and Results

For the regression task, we studied the performance of acoustic and behavioral characteristics features. For modeling individual acoustic feature groups and their linear combination we used support vector machine for regression, implemented in weka [23] using Radial Basis Function (RBF) kernel with  $\gamma = 0.01$  and  $C = 1.0$ .

As for the linear combination of different acoustic feature groups, we first merged all the feature vectors linearly to obtain vector  $M$ , as shown in Equation 1

$$\mathbf{M} = P \cup S \cup VQ = \{p_1, \dots, p_m, s_1, \dots, s_n, v_1, \dots, v_l\} \quad (1)$$

where feature vectors  $P$ ,  $S$  and  $VQ$  stands for prosody, spectral and voice quality as presented in Equations 2-4.

$$P = \{p_1, p_2, \dots, p_m\} \quad (2)$$

$$S = \{s_1, s_2, \dots, s_n\} \quad (3)$$

$$VQ = \{v_1, v_2, \dots, v_l\} \quad (4)$$

From the merged feature vector we selected relevant feature subset  $F_s - M$  using training set only. For the automatic feature selection, we used Relief feature selection

**Table 2: Results of individual acoustic feature groups with linearly merged feature groups and with Relief feature selection for depression severity estimation on the development set. \* represents results tuned using 3-fold cross validation on the training set.  $|F|$  represent feature set dimension.**

Feature set, F	$ F $	RMSE	MAE
<i>Spectral</i>	864	<b>6.32</b>	<b>4.96</b>
<i>Voice Quality</i>	288	7.05	5.70
<i>Prosody</i>	288	7.10	5.75
<i>Merged</i>	1440	6.43	5.40
<i>Merged+Feat.Selection*</i>	20	6.70	5.20

**Table 3: Result for depression severity estimation using behavioral characteristic features on development set.  $|F|$  represent feature set dimension.**

Feature set, F	$ F $	RMSE	MAE
<i>Behavioral characteristic</i>	12	<b>5.54</b>	<b>4.73</b>

technique [29, 41], successfully used in paralinguistic tasks [1, 11]. The technique calculates the weight of the features based on the nearest  $k$  instances ( $k = 20$ , used for this study) of the same and different classes to rank each features. Then by using a threshold,  $th = 0.02$ , we selected top 20 features to use for the regression task. These parameters ( $th=0.02$ , 0, -0.02 and  $k=5,10,15,20$ ) are tuned using 3-fold cross validation of the training set.

As for predictor using behavioral characteristic feature group, we used Reduced Error Pruning Tree (“REPT”) implemented in weka [23], which is a fast regression tree learner that uses information of variance reduction and prunes it using reduced error pruning.

The results are presented in Table 2 for individual feature set and their combinations. The result indicated that spectral features are a good predictor of PHQ score compared to all other settings presented in the table. It is observed that even feature selection on the merged vector also performed better than other sets except spectral and is above the baseline, i.e.,  $MAE = 5.36$  and  $RMSE = 6.74$  on the same development set. The selected features include features from spectral group (75%), prosodic group (20%) and voice quality (5%) group.

It is also observed that using behavioral characteristic features, we obtained a decrease of both MAE and RMSE by a magnitude of 0.63 and 1.20 respectively compared to all the results reported in the AVEC2017 baseline manuscript. Further analysis using feature ranking technique, Relief, indicated that the PDI features especially *dep* and *ptsd* are the top ranked features followed by the median of the response time, the quartiles of the within-speaker silence duration and laughter frequency.

## 4.2 Language

Additional to the speech-based features, we explore text-based representations to predict depression severity estimates. The widely used representation of a document in NLP is bag-of-words, where a document is represented by word occurrences ignoring the order in which they appear. We experiment both with binary (BOOL) and tf-idf (TFIDF) weighted representations. While the binary representation encodes

**Table 4: Root mean square error (RMSE) and mean absolute error (MAE) for depression severity regression using lexical features and Support Vector regression with linear kernel on the development set for the mean baseline (BL: mean), binary (BOOL), tf-idf weighted (TFIDF) bag-of-words representations, and averaged word embedding vectors (WE). We also provide the audio and audio-video feature-based baselines (BL: Audio and BL: Audio-Video) using Random Forests.**

	RMSE	MAE
<i>BL: mean</i>	6.57	5.50
<i>BL: Audio</i>	6.74	5.36
<i>BL: Audio-Video</i>	6.62	5.52
<i>BOOL</i>	<b>6.31</b>	<b>5.17</b>
<i>TFIDF</i>	6.78	5.40
<i>WE</i>	6.84	5.41

words that are present in the document regardless of their frequency, tf-idf weighted representation considers both the frequency of the term ( $tf$ ) in a document and the inverse document frequency ( $idf$ ) – which lowers the weight of the very frequent terms in a collection and increases the weight of the rare terms with respect to the equations 5-6.

$$tf - idf(t, d) = tf(t, d) * idf(t) \quad (5)$$

$$idf(t) = \log \frac{n_d}{df(d, t)} + 1 \quad (6)$$

Where  $tf(t, d)$  is the term frequency,  $n_d$  is the total number of documents, and  $df(d, t)$  is the frequency of documents containing the term.

Besides bag-of-words representation, we also experiment with the word embedding representation (WE) [34], where pre-trained per-word embedding vectors are averaged for a document. We make use of the SKIPGRAM embedding vectors pre-trained on GoogleNews with a embedding dimension 300 and window 10.

Since the provided speech transcripts are of human-machine conversations, we first extract human turns and convert them into bag-of-words representation. The transcripts contain annotations for the speech phenomena such as laughter, sigh, etc., which were treated as any other token. Thus, the representation implicitly encodes the presence of these phenomena in the conversation; and also its frequency in the case of tf-idf based representations. For the word embedding representation, however, this is not the case, as there are no pre-trained vectors for these.

The algorithm of our choice for text-based representations is Support Vector Regression (SVR) with linear kernel, implemented in scikit-learn [38]. The regression results for each of the document representations are given in Table 4 in terms of RMSE and MAE. We also provide a mean baseline (BL:mean) and the audio and audio-video feature-based baselines<sup>1</sup>. As it can be observed, the only representation that outperforms all the baselines is the binary bag-of-word representation that yields RMSE=6.31 and MAE=5.17.

## 4.3 Visual Features

<sup>1</sup>Cite the baseline paper

Inspired by [51] and the success reported in [54], we use the 68 3D facial keypoints and compute geometric features as follows: for every facial representation, we first remove the 3D bias (equal to a translation in the Euclidean space by subtracting the mean value in 3D), then we normalize the resulting representation so that the average distance to the center (origin) is equal to 1. Finally, we compute Euclidean distances between all possible pairs of 3D normalized points and add them to the normalized representation. This results in a feature vector of size 2482. Consequently, we reduce this dimension by applying PCA and keeping over 99.5% of variance, resulting in a feature vector of size 33.

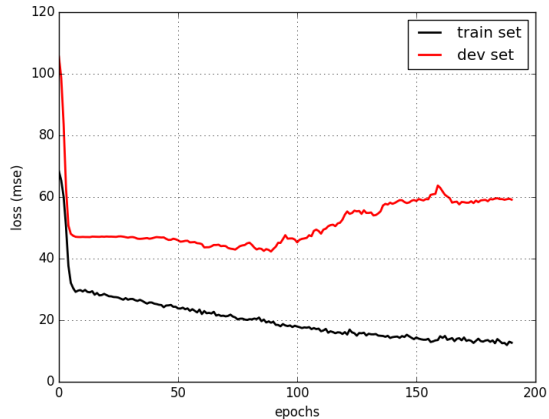
Since we are dealing with video sequences, we propose to regress depression using models naturally designed for temporal data. Specifically, we propose the use of LSTMs [25] for this task. LSTMs have emerged as an effective and scalable model for several learning problems related to sequential data, such as handwriting recognition [39, 14], generation of handwritten characters [22], language modeling and translation [56, 32], audio [33] and video [15] signal analysis, acoustic speech modeling [44] and others. They have proved effective at capturing long-term temporal dependencies without suffering from the optimization hurdles that plague simple recurrent neural networks (RNNs).

In order to build our training set, we apply a sliding window approach to the video sequences, using windows of size  $W$ , overlapped by  $O$  samples. We use the *success* flag provided by the dataset creators which models the tracking confidence for each frame. We adopt a 0-tolerance strategy and discard all windows for which at least one failed tracking is present. We do this to exclude the risk of introducing artifacts into the feature space, that the model might misleadingly exploit for solving the task. We set the values for  $W$  and  $O$  empirically to 60 and 30, respectively. We down-sample the data to 1 second, which makes our windows 1 minute long, with an overlap of 30 seconds. During testing, we apply the same window-ing scheme and average the window-level predictions over the length of the test sequence.

Next, we train a double layered LSTM model on regressing depression at window level on the training set. The model is composed of two stacked layers of size 16, followed by a *Dense* layer with a *linear* activation function. We use dropout [48] equal to 0.5 to control overfitting and batch normalization [26] to limit internal covariance shift. As loss function, we use the mean squared error. In order to validate our LSTM model, we perform a leave-one-sequence-out cross-validation scheme on the training set. After 100 epochs, our models achieve an MAE of 4.97 and an RMSE of 6.26, which we find encouraging. We further retrain the model on the full training set and monitor the performance on the development partition.

Figure 1 shows the learning plots of the loss function during training for both training (black) and validation (red) sets. We observe a monotonic decrease of the loss function on the training set, while on the validation, the behavior is a typical decrease, followed by an increase of the same loss. We use the validation set to early stop the training, thus resulting in a model (*lstm\_opt*) with the best performance on this set.

Following the baseline manuscript, we report in Table 5 as performance measures the RMSE and MAE of *lstm\_opt* on both train as well as test sets. In addition to the requested quantities, we also report the explained variation regression



**Figure 1: LSTM learning curves: trainset (black) and development set (red). We note the existence of a turning point in the validation loss, typically used as a good compromise between underfitting and overfitting**

**Table 5: Performance measures obtained using our LSTM model on the training set as well as on the development set. We also report the explained variance regression score (EVS), which measures the degree to which the model "explains" the variation of the ground truth labels using the predictions (see Equation 7 for a formal definition)**

	RMSE	MAE	EVS
train set	3.17	2.32	0.66
dev set	6.09	4.66	0.15

score (EVS), defined as:

$$evs(y, \hat{y}) = 1 - \frac{Var(y - \hat{y})}{Var(y)} \quad (7)$$

where  $Var$  represents the statistical variance. EVS measures the degree to which a model (in our case *lstm\_opt*) accounts for the variance of a given set of labels through the predictions it makes. The upper bound of EVS is 1 and corresponds to a perfect modeling.

As can be observed from Table 5, our LSTM model fits well the training set and manages to score a promising MAE on the development partition, better than all reported values in the AVEC2017 baseline manuscript as well as in the last year’s winning paper [54].

#### 4.4 Results on the test set

We submitted four trials for evaluation on the held out test set. Results are depicted in Tab. 6. The behavioral characteristic features extracted from audio transcriptions achieve the lowest errors on the test partition, which is unsurprising considering the promising cross-validation results obtained on the the development set (*i.e.* RMSE of 5.54 and MAE of 4.73). What is slightly surprising though is the performance of the visual features. Despite achieving an encouraging MAE on the development set, our LSTM model failed to generalize well enough to unseen data.

**Table 6: Results on the test set**

	RMSE	MAE
Spectral features (speech)	6.63	5.08
Turn features (speech)	4.94	4.11
Text features	5.83	4.88
Video features	6.72	5.36

## 5. CONCLUSIONS

In this paper we address the depression sub-challenge problem formulated in AVEC2017, *i.e.* regressing PHQ-8 depression scores from multi-modal data. We process different modalities (audio, language, visual) accompanying the corpus and developed regression systems separately. In the audio domain, we find the spectral features to be most suited for this task, achieving an MAE score of 4.96 on the development set (RMSE = 6.32) while lexical features score no lower than 5.17 (MAE) and 6.31 (RMSE). Despite being the worst performing modality in the baseline manuscript, visual features achieve the smallest errors on the development set in our experiments. Using a sliding window approach and temporal modeling, we obtain an MAE of 4.66 (RMSE = 6.09). We also observed that behavioral cues extracted from transcripts achieve smaller errors (MAE = 4.73, RMSE = 5.54) compared to audio and language features and are good predictor of the depression severity scores. When studied further, we found that previous diagnosed information cues, participants' response time to the agent among others are one of the most informed feature to predict the depression PHQ-8 scores. This is indeed confirmed by the results obtained on the test set, where behavioral cues scored the smallest MAE values among all other feature sets.

In this paper, we have studied each modality individually to understand its strength in estimating the depression severity. In future work, we plan investigating how we can combine individual modalities to improve the overall performance.

## 6. REFERENCES

- [1] F. Alam and G. Riccardi. Comparative study of speaker personality traits recognition in conversational and broadcast news speech. In *INTERSPEECH*, pages 2851–2855, 2013.
- [2] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 2016.
- [3] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear. Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 283–288. IEEE, 2013.
- [4] Z. Ambadar, J. F. Cohn, and L. I. Reed. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of nonverbal behavior*, 33(1):17–34, 2009.
- [5] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 223–230. IEEE, 2006.
- [6] G. Bersani, E. Polli, G. Valeriani, D. Zullo, C. Melcore, E. Capra, A. Quartini, P. Marino, A. Minichino, L. Bernabei, et al. Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: a partially shared cognitive and social deficit of the two disorders. *Neuropsychiatric disease and treatment*, 9:1137, 2013.
- [7] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM, 2004.
- [8] K.-h. Chang, D. Fisher, J. Canny, and B. Hartmann. How's my mood and stress?: an efficient speech analysis library for unobtrusive monitoring on mobile phones. In *Proceedings of the 6th International Conference on Body Area Networks*, pages 71–77. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011.
- [9] S. A. Chowdhury, M. Danieli, and G. Riccardi. Annotating and categorizing competition in overlap speech. In *Proc. of ICASSP*. IEEE, 2015.
- [10] S. A. Chowdhury and G. Riccardi. A deep learning approach to modeling competitiveness in spoken conversation. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.
- [11] S. A. Chowdhury, G. Riccardi, and F. Alam. Unsupervised recognition and clustering of speech overlaps in spoken conversations. In *Proc. of Workshop on Speech, Language and Audio in Multimedia*, 2014.
- [12] S. D. Craig, S. D'Mello, A. Witherspoon, and A. Graesser. Emote aloud during learning with autotutor: Applying the facial action coding system to cognitive-affective states during learning. *Cognition and Emotion*, 22(5):777–788, 2008.
- [13] H. Dibeklioglu, Z. Hammal, Y. Yang, and J. F. Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 307–310. ACM, 2015.
- [14] P. Doetsch, M. Kozielski, and H. Ney. Fast and robust training of recurrent neural networks for offline handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 279–284. IEEE, 2014.
- [15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.
- [16] F. Eyben, F. Wenginger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.
- [17] A. J. Ferrari, F. J. Charlson, R. E. Norman, S. B.

- Patten, G. Freedman, C. J. Murray, T. Vos, and H. A. Whiteford. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS medicine*, 10(11), 2013.
- [18] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837, 2000.
- [19] K. Fraser, F. Rudzicz, N. Graham, and E. Rochon. Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pages 47–54, 2013.
- [20] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014.
- [21] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128, 2014.
- [22] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on*, pages 6645–6649. IEEE, 2013.
- [23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
- [24] J. H. Hansen. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech communication*, 20(1-2):151–173, 1996.
- [25] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [26] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.
- [27] J. Joshi, R. Goecke, G. Parker, and M. Breakspear. Can body expressions contribute to automatic depression analysis? In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.
- [28] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller. Cca based feature selection with application to continuous depression recognition from acoustic speech features. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3729–3733. IEEE, 2014.
- [29] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*, pages 171–182. Springer, 1994.
- [30] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1):163–173, 2009.
- [31] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12):1797–1803, 2009.
- [32] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*, 2014.
- [33] E. Marchi, G. Ferroni, F. Eyben, L. Gabrielli, S. Squartini, and B. Schuller. Multi-resolution linear prediction based features for audio onset detection with bidirectional lstm neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2164–2168. IEEE, 2014.
- [34] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [35] S. Newman and V. G. Mather. Analysis of spoken language of patients with affective disorders. *American journal of psychiatry*, 94(4):913–942, 1938.
- [36] A. Nicholson, H. Kuper, and H. Hemingway. Depression as an aetiologic and prognostic factor in coronary heart disease: a meta-analysis of 6362 events among 146 538 participants in 54 observational studies. *European heart journal*, 27(23):2763–2774, 2006.
- [37] P. Patel, A. Chaudhari, R. Kale, and M. Pund. Emotion recognition from speech with gaussian mixture models & via boosted gmm. *International Journal of Research In Science & Engineering*, 3, 2017.
- [38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.
- [39] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour. Dropout improves recurrent neural networks for handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 285–290. IEEE, 2014.
- [40] B. Pope, T. Blass, A. W. Siegman, and J. Rahe. Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology*, 35(1p1):128, 1970.
- [41] M. Robnik-Sikonja and I. Kononenko. An adaptation of relief for attribute estimation in regression. In D. H. Fisher, editor, *Fourteenth International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1997.
- [42] A. Rumshisky, M. Ghassemi, T. Naumann, P. Szolovits, V. Castro, T. McCoy, and R. Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, 6(10):e921, 2016.
- [43] J. A. Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102, 1994.
- [44] H. Sak, A. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual*

*Conference of the International Speech Communication Association*, 2014.

- [45] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, L.-P. Morency, et al. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658, 2014.
- [46] B. Schuller, A. Batliner, S. Steidl, and D. Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087, 2011.
- [47] Y. I. Sheline, M. H. Gado, and H. C. Kraemer. Untreated depression and hippocampal volume loss. *American Journal of Psychiatry*, 160(8):1516–1518, 2003.
- [48] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- [49] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *Mechatronics and Automation, 2005 IEEE International Conference*, volume 3, pages 1569–1574. IEEE, 2005.
- [50] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5200–5204. IEEE, 2016.
- [51] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2016.
- [52] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanata, and E. P. Scilingo. Speech analysis for mood state characterization in bipolar patients. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 2104–2107. IEEE, 2012.
- [53] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 201–213. Springer, 2013.
- [54] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli. Decision tree based depression classification from audio video and language information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 89–96. ACM, 2016.
- [55] C. Yogesh, M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, C. Berkai, and K. Polat. A new hybrid pso assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Systems with Applications*, 69:149–158, 2017.
- [56] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.
- [57] E. Zwirner. Contribution to the speech of depressives. phonometry iii. special applications i, 1930.