Hybrid Depression Classification and Estimation from Audio Video and Text Information

Le Yang
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University(NPU)
Shaanxi Key Lab on Speech and Image Information Processing
127 Youyi Xilu, Xi'an 710072 China
yangle.cst@gmail.com

Hichem Sahli
NPU-VUB Joint AVSP Research lab
Deptartment of Electronics & Informatics(ETRO)
Vrije Universiteit Brussel(VUB)
Pleinlaan 2, 1050 Brussels, Belgium
Interuniversity Microelectronics Centre(IMEC)
Kepeldreef 75, 3001 Heverlee, Belgium
hsahli@etrovub.be

Xiaohan Xia
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University(NPU)
Shaanxi Key Lab on Speech and Image Information Processing
127 Youyi Xilu, Xi'an 710072 China
xiaohanxia@mail.nwpu.edu.cn

Ercheng Pei
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University(NPU)
Shaanxi Key Lab on Speech and Image Information Processing
127 Youyi Xilu, Xi'an 710072 China
peiercheng@mail.nwpu.edu.cn

Meshia Cédric Oveneke
NPU-VUB Joint AVSP Research lab
Dept. Electronics & Informatics (ETRO)
Vrije Universiteit Brussel(VUB)
Pleinlaan 2, 1050 Brussels, Belgium
mcovenek@etro.vub.ac.be

Dongmei Jiang
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University(NPU)
Shaanxi Key Lab on Speech and Image Information Processing
127 Youyi Xilu, Xi'an 710072 China
jiangdm@nwpu.edu.cn

ABSTRACT

In this paper, we design a hybrid depression classification and depression estimation framework from audio, video and text descriptors. It contains three main components: 1) Deep Convolutional Neural Network (DCNN) and Deep Neural Network (DNN) based audio visual multi-modal depression recognition frameworks, trained with depressed and not-depressed participants, respectively; 2) Paragraph Vector (PV), Support Vector Machine (SVM) and Random Forest based depression classification framework from the interview transcripts; 3) A multivariate regression model fusing the audio visual PHQ-8 estimations from the depressed and not-depressed DCNN-DNN models, and the depression classification result from the text information. In the DCNN-DNN based depression estimation framework, audio/video feature descriptors are first input into a DCNN to learn high-level features, which are then fed to a DNN to predict the PHQ-8 score. Initial predictions from the two modalities are fused via a DNN model. In the PV-SVM and Random Forest based depression classification framework, we explore semantic-related text features using PV, as well as global text-features. Experiments have been carried out on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset for the Depression Sub-challenge at the 2017 Audio-Visual Emotion Challenge (AVEC), results show that the proposed depression recognition framework obtains very promising results, with root mean square error (RMSE) as 3.088, mean absolute error (MAE) as 2.477 on the development set, and RMSE as 5.400, MAE as 4.359 on the test set, which are all lower than the baseline results.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
AVEC'17, October 23, 2017, Mountain View, CA, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5502-5/17/10…$15.00
https://doi.org/10.1145/3133944.3133950

CCS CONCEPTS

• **Pattern Recognition → Applications|signal processing, computer vision, speech processing**;

KEYWORDS

Depression recognition, Depression classification, DCNN-DNN, PV-SVM, Multi-modal

ACM Reference format:
Le Yang, Hichem Sahli, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Dongmei Jiang. 2017. Hybrid Depression Classification and Estimation from Audio Video and Text Information. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, October 23, 2017 (AVEC'17),* 7 pages.
https://doi.org/10.1145/3133944.3133950

1 INTRODUCTION

Depression is a state of low mood and aversion to activity that can affect a person's thoughts, behaviours, feelings, and sense of well-being. At present, depression and anxiety disorders are highly

prevalent worldwide causing burden and disability for individuals, families and society [5].

Accurate depression classification and estimation are of great importance and has broad application prospects. Within the past several years, various methods have been investigated for depression classification and estimation. Cohn *et al.* [1] compared clinical diagnosis of major depression with facial actions and vocal prosody, and adopted support vector machine (SVM) and logistic regression to classify depression or non-depression. In [2], Nicholas *et al.* investigated several speech features for depression detection, indicating that the combination of Mel-frequency cepstral coefficient (MFCC) and formant based features obtained 80% classification accuracy. The Audio Visual Emotion Challenge (AVEC) 2016 [17] also focused on the depression classification task. Ma *et al.* [11] proposed a deep model, named as DepAudioNet, to encode the depression related characteristics and achieve the final classification result. In [13], Pampouchidou *et al.* implemented depression classification by fusing the high level and low level features from audio, video and text modalities. In AVEC2014 [18] where the main task was estimation of the BDI-II scores, Jain *et al.* [8] focused on the visual descriptors and utilized a Fisher Vector to estimate the depression levels. In [7], the authors adopted a Support Vector Regress to model the relationship between the audio, visual, linguistic information and depression scores. It is worth noting that the methods mentioned above only studied the depression from a single point of view, either depression classification, or depression estimation. However, experimental results showed that when depression classification and depression estimation are considered at the same time, better performance could be obtained [20] [15] [19] [21].

In this paper, we target the Depression Sub-Challenge (DSC) task of AVEC2017 [4], and design a hybrid depression classification and depression estimation framework from audio, video and text descriptors, as shown in Figure 2. From the work of [21] and [19] on AVEC2016, we observed that the text information from the dialogues between the participants and Ellie plays an important role in depression classification, because the answers to Ellie's questions relate the symptoms associated with psychoanalytic aspects of depression, such as whether the participant has been diagnosed as depressed or post-traumatic stress disorder (PTSD), sleeping disorder, feelings, etc. Therefore, in this work, we firstly make depression/non-depression classification from the text information. As text descriptors of the selected sentences (answers to Ellie's questions), we make use of the Paragraph Vector (PV), which has been introduced in [10] to directly learn the distributed representations of sentences and documents. Different from most of the conventional text feature extraction methods, such as Bag-of-words, PV takes consideration of context semantics using low-dimensional representations. Apart from these semantic-related text features, we also consider the global text-features. Both text features are used to improve the classification accuracy of depression/non-depression. As far as we know, our approach is the first which applies Paragraph Vectors to textual transcripts for depression analysis.

To improve the depression estimation accuracy, we build an audio-video fusion deep learning model, composed of deep convolutional neural networks (DCNN) and deep neural networks (DNN) to learn low dimensional global feature vectors with compact dynamic information, and improve the estimation accuracy of the eight-item Patient Health Questionnaire (PHQ-8) scores. It contains two input streams: the audio network processing audio descriptors with one DCNN-DNN model, and the visual network processing visual descriptors with another DCNN-DNN model. The initial estimations of the PHQ-8 scores from these two modalities are fused via a DNN model. To deal with the behavioral variability between depressed and not-depressed subjects, we train two separate audio-visual deep learning models, one for each group. The predictions of PHQ-8 from the depressed and not-depressed DCNN-DNN models, and the depression classification result from the text information, are fused via a multivariate regression model for the final PHQ-8 estimation.

Experiments are carried out on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) database [6] of AVEC 2017, results indicate that our hybrid depression recognition framework presents promising performance.

The outline of this paper is as follows. The proposed multi-modal depression analysis framework is introduced in Section 2, together with the used audio, video and text features. Section 3 illustrates and analyzes the experimental results. Finally, conclusions and future works are given in Section 4.

2 MULTI-MODAL DEPRESSION ANALYSIS

2.1 Audio/Video PHQ-8 Prediction

We firstly propose an audio/visual uni-modal depression recognition framework, as shown in Figure 1. The audio and video networks are trained individually using the ground truth labels of the PHQ-8 scores. For each audio (video) segment, we pass the corresponding feature descriptors through a DCNN, which has n convolutional layers, followed by one ReLU, Pooling and Dropout layers, while the last convolutional layer is followed by two fully connected layers. In the training process, we add a fully-connected layer to produce the prediction score. The loss function associated to the output of the model is the Euclidean loss. After training the DCNN, we freeze the weights, discard the last layer and connect the second fully connected layer to the visible layer of a DNN with m layers. Here also, the loss function associated to the output of the model is the Euclidean loss.

In this work, instead of using a large number of descriptors as inputs to the DCNNs, we select two informative features. For audio,

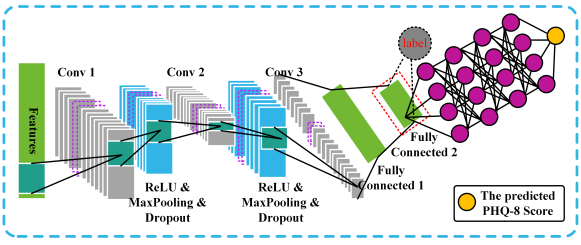

Figure 1: Unimodal DCNN-DNN model for depression recognition.

we utilize the openSMILE toolkit [14] to extract for each audio segment, 238 low level descriptors (LLDs), comprising 211 spectral and energy related features and 27 voicing related dynamic features. The LLDs, which are extracted with the frame length of 60ms and frame shift of 10ms, are shown in Table 1, where the numbers between brackets are the dimensions of the extracted features vectors, and $\triangle$, $\triangle\triangle$ denote the first and second order derivatives, respectively. 25 statistical functionals and 4 regression functionals, as shown in Table 2, have been performed on the extracted LLDs, the resulting 6902 dimensional feature vector for each speech segment are used as inputs to the DCNN.

Table 1: 238 Low level descriptors from openSMILE.

Energy and Spectral related (211)	
PLPCC(5)+$\triangle$+$\triangle\triangle$	MFCC-sma(15)+$\triangle$+$\triangle\triangle$
LOGenergy(1)+$\triangle$+$\triangle\triangle$	Chroma(12)
LspFreq(8)+$\triangle$	SpectralRollOff(4)+$\triangle$
LengthL1norm(2)+$\triangle$	LpcCoeff(11)
LogRelF0(2)+$\triangle$	Amplitude(3)+$\triangle$
SpectralEnergy(2)+$\triangle$	SpectralSlope(2)+$\triangle$
Zcr(1)+$\triangle$	Loudness(1)+$\triangle$
RASTA-filtered(26)+$\triangle$	RMSenergy(1)+$\triangle$
Spectral(Flux, Centroid, Entropy,	Hammarberg(1)+$\triangle$
Variance, Skewness, Kurtosis,	LpGain(1)
Harmonicity, Flatness)(8)+$\triangle$	AlphaRatio(1)+$\triangle$
Voicing related (27)	
F0(2)+$\triangle$ LogHNR(1)+$\triangle$	JitterLocal(1)+$\triangle$
FormantFreqLpc(6)	ShimmerLocal(1)+$\triangle$
FormantBandwidthLpc(6)	FormantFrameIntensity(1)
VoicingFinalUnclipped(1)+$\triangle$	JitterDDP(1)+$\triangle$

Table 2: 29 Functionals from openSMILE.

Statistical functionals (25)
max, min, arithmetic mean, norm, variance, stddev, skewness, kurtosis, numPeaks, meanPeakDist, peakMean, peakMeanMeanDist, quartiles(1-3), samplepos, numSegments, meanSegLen, maxSegLen, minSegLen, upleveltime25, upleveltime50, upleveltime75, risetime, falltime
Regression functionals (4)
linregc1, linregc2, linregerrA, linregerrQ

As visual features, we make use of the Action Units provided by AVEC2017. Facial Action Units (FACS), is the standard reference in facial action annotation widely used in psychology to measure emotion, pain, and behavioural measures of psychopathology [3]. They have been suggested as domain knowledge to guide depression classification [1] [16]. To aggregate over the whole video the AU descriptors from all the frames, inspired by the idea of Motion History Histogram (MHH) used by Meng *et al.* [12], we propose a method to extract the dynamics from the AUs. This involves estimating the changes on each AU component of the feature vector sequence, from which a histogram of B equally spaced bins, $\{R_b, b = 1, \cdots, B\}$, spanning the range $[-20, 20]$ is created. For temporal modelling, we estimate the AU's changes (displacements) at different time intervals $M_k, k = 1, \cdots, K$ and concatenate the obtained histograms. Formally, for a given time interval M_k, let $D(i, j) = AU^j_{i+M_k} - AU^j_i$ be the change (displacement) of AU^j between frame i and frame $i + M_k$, each bin of the histogram contains the number of occurrences of displacements $D(i, j), i = 1, \cdots, N, j = 1, \cdots, N_{AU}$ in the corresponding range, with N the number of frames, and $N_{AU} = 20$ the number of AUs. In our experiments we consider 5 time intervals (M_k), as 10, 20, 30, 40, and 50 frames, respectively. With respect to the histogram bins, we consider 4 equally spaced bins (R_i), spanning the range $[-20, 20]$. In total we obtain 200 visual features.

2.2 Text-based Depression Classification

In addition to non-verbal behaviour, transcript files provided by the DAIC-WOZ contains attitudinal, and more specifically, answers to the questions related the symptoms associated with psychoanalytic aspects of depression such as sleeping disorder, feelings, etc. In this work we propose a text-based depression classification framework combining both text-features and semantic (content) features, as illustrated in the lower part of Figure 2. The final text-based depressed / not-depressed classification, denoted as D_c, is obtained as a logical *AND* of the two results of text streams.

The global-text features include the length of the conversation (in seconds), number of sentences, number of segments, number of words, number of laughs, number of sighs, number of deep breaths, and number of filler words (i.e "mhm", "nuh", "nmm", "nhmm"). In total 8 global text features are fed to a random forest for depression / non-depression classification.

Table 3: Example of the question-answer pairs in transcript files.

start_time	stop_time	speaker	value
...	...	...	...
148.94	150.8	Ellie	do you consider yourself an introvert?
153.04	155.01	Participant	um i was an extrovert
...	...	...	...
326.63	328.83	Ellie	how easy is it for you to get a good night's sleep?
329.53	331.41	Participant	mm it isn't
...	...	...	...
384.442	386.302	Ellie	have you been diagnosed with depression?
386.952	387.372	Participant	yes
...	...	...	...
438.69	440.071	Ellie	how have you been feeling lately?
442.61	443.52	Participant	i guess sorta depressed
...	...	...	...

For the semantic features, we conduct content analysis of the transcripts (as shown in Table 3) to select the patient's answers to questions related to 5 symptoms associated with psychoanalytic aspects of depression:

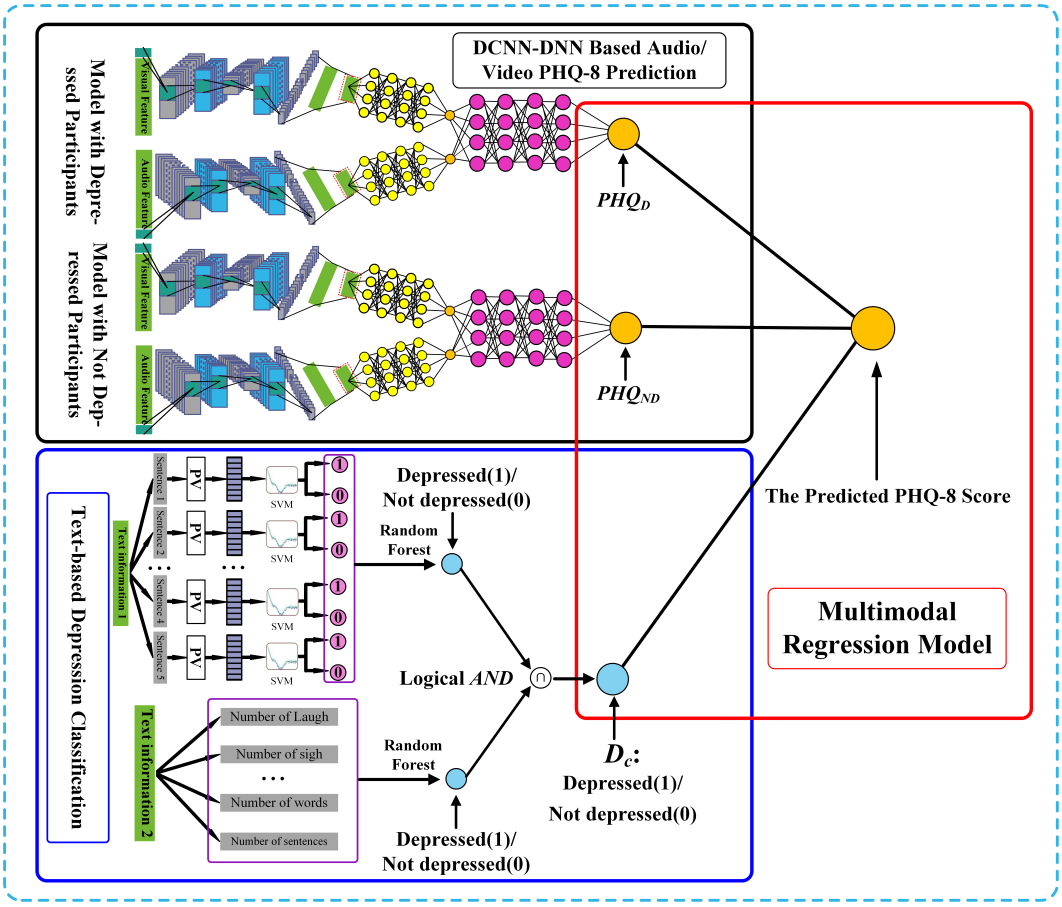

Figure 2: The structure of the proposed audio-visual-text hybrid depression classification and recognition

.

(1) Prior depression diagnosis (Yes/No)
(2) Prior post-traumatic stress disorder (PTSD) diagnosis (Yes/No)
(3) Sleep disorder (Yes/No)
(4) Feelings (Bad/Good)
(5) Personality (Extrovert/Introvert).

The goal is to automatically detect the presence or absence of these considered psychoanalytic factors from the selected sentences, for subsequent usage in depression / non-depression classification. As illustrated in the lower part of Figure 2, from the selected 5 sentences, we extract Paragraph Vector (PV) descriptors [10], which are fed to SVMs for presence/absence classification. E.g. for the sentences related to Sleep disorder, we automatically classify the participant's answer as *Yes* (presence) or *No* (absence). For training the SVMs we conduct content analysis of the training transcripts and label the sentences according to the presence/absence of the considerer psychoanalytic factor. E.g. for the sentences related to sleep disorder, we label a sentence *Yes* if the participant's answer contains the following expressions "not had a good sleep", "really hard", "kind a difficult", "never easy"; the label is *No* if the answer contains "no problem", "pretty good", "get a good night's sleep", etc. The binary outputs of the 5 SVMs are concatenated into a vector of binary values, which is used as input to a Random Forest for depression / non-depression classification. The output of the latter is fused with the output of the global-text features model using the logical *AND* operator to get the final classification result. Our motivation of using Paragraph Vectors [10], is that they can generate the representation of any sentence without considering the text's length, they are learned from unlabelled data, they inherit an important property of the Word2Vec model, being the semantics of the words, and importantly consider the word order in the sentence.

2.3 Hybrid Depression Recognition and Classification Framework

The proposed hybrid multi-modal depression recognition and classification framework is as shown in Figure 2. It contains three parts: the prediction of PHQ-8 score from audio and video features; the classification of depression/non-depression from text information; and the multimodal regression for final depression prediction.

To deal with the behavioral variability between *depressed* and *non-depressed* subjects, we train two separate audio-visual deep learning models for depression prediction, one for each group. We denote by PHQ_D the output of the model trained with *depressed*

participants, and PHQ_{ND} the output of the model trained with *non-depressed* participants. The proposed audio-video depression prediction model contains two input streams: the audio network processing audio descriptors with one DCNN-DNN model, and the visual network processing visual descriptors with another DCNN-DNN model. The initial predictions of the PHQ-8 scores from these two modalities are fused via a DNN model.

For fusing the depression recognition results from the audio visual modalities, and the depression classification result from the text modality, we propose a simple multivariate regression model of the form:

$$y_i = b_0 + b_1 * PHQ_D + b_2 * PHQ_{ND} + b_3 * x_1 + b_4 * x_2 + e \quad (1)$$

with y_i the final PHQ-8 prediction, PHQ_D and PHQ_{ND} the outputs of the audio-visual hybrid deep learning models trained with depressed and not-depressed participants, respectively. $x_1 = D_c * PHQ_D$, $x_2 = (1 - D_c) * PHQ_{ND}$, where D_c is the output of the text-based depression / non-depression classification. This simple model considers the behavioral variability between depressed and non-depressed subjects, and reinforces the final decision according to the text analysis: when $D_c = 1$, $x_1 = PHQ_D$ and $x_2 = 0$.

The multivariate regression model is trained using the training results of the audio/video (PHQ_D and PHQ_{ND}), and text (D_c) models, respectively, and the ground-truth labels y_i.

The structural hyperparameters of the used models are discussed in Section 3.3. All of our DCNN and DNN models were trained using Caffe [9].

3 EXPERIMENTS AND ANALYSIS

In this section, we carry out experiments on the AVEC 2017 depression dataset. To improve the performance, we train the gender specific models, and do depression recognition and classification experiments for female and male, respectively. Finally the results of females and males are put together for the evaluation on all participants.

3.1 The AVEC 2017 Depression Dataset

The AVEC2017 depression dataset [6] consists of 189 segments of clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. A segment corresponds to an audio/video recording of the participant's answering a question from Ellie, the animated virtual interviewer. The recorded segments were split into a training set of 107 segments, a development set of 35 segments and a test set of 47 segments. For each segment of the training and development sets, AVEC 2017 provides the label of PHQ-8 score ranging from 0 to 24, and a binary value of depressed/not-depressed. Participant with higher PHQ-8 score has higher probability of being depressed. In AVEC 2017 depression dataset, segments with PHQ-8 score exceeding 10 are labelled as depressed.

3.2 Data Expansion

Table 4: Number of the depressed/not depressed samples in the training set(inside the brackets are the numbers after sampling).

Gender	Depressed Samples	Not Depressed Samples
Female	17(170)	27(270)
Male	13(130)	50(500)

The training set of the AVEC2017 is imbalanced with very few samples, as shown in Table 4. Such data distribution may decrease the recognition performance and cause over-fitting. In order to increase the number of samples and improve the robustness of the models, in this work we re-sample the dataset to obtain a larger scale data. The sampling is made as follows: we first remove the *non-speaking* segments when the participant listens to Ellie. Then for the *speaking* segments from each session, the longest ten segments in each session are taken as ten new samples.

For the development and test sets, we adopt the same expansion method, therefore for each session we get ten prediction results. Eventually, the average of the ten predictions will be the final prediction result of the session.

3.3 Model Parameters

For designing the audio and video DCNN-DNN networks, we have tested several network architectures and selected the ones which provided the best PHQ-8 prediction, in terms of root mean square error (RMSE) and mean absolute error (MAE), on the development set. For the DCNNs, we evaluated architectures with 3, 4 and 5 convolutional layers (CV). The convolution kernel size was set to 1×5 and $stride = 1$, and the number of feature maps $N_{maps} \in \{24, 48, 64, 128\}$. Each CV layer is followed by a ReLU, MaxPooling and Dropout layer, except for the last CV. The pooling kernel size was set to 2×2, $stride = 2$, and for the Dropout layer, we evaluated dropout ratios in the range $[0.3, 0.5]$.

For the DNN models we evaluated architectures with number of layers $N_{dlayer} \in \{3, 4, 5\}$. In each layer, the number of hidden nodes was set $N_{nodes} \in \{15, 20, 25, 30, 35, 50, 100, 150\}$, followed by ReLU as the activation function, and Euclidean loss as loss function.

For the text modality, PV-SVM and Random Forest have been mainly used. The Paragraph Vector model can be trained from a large text corpus. In this work, we train the Paragraph Vector model with 402, 325 dialogues collected from *Friends*, *The Big Bang Theory* and *Game of Thrones*. Most of these dialogues are simple and short.

For designing the PV-SVM and Random Forest networks, we also tested several network architectures and selected the ones which provided the best classification results on the development set. The parameters used in Paragraph Vector model are the dimension L of the PV for each sentence, and the window size S. In our experiments we evaluated $L \in \{50, 100, 150\}$ and $S \in \{5, 10\}$. For selecting the parameters c and g for SVM, we adopted a grid search approach with 2 as the base and y as the index, $y \in [-12, 12]$. For the Random Forest, the only hyper-parameter in our experiment was the number of trees which was set as $N_{trees} \in \{1, 2, 3, 4, 5, 10, 15, 20\}$.

Table 5: The prediction of PHQ-8 score based on AU and Audio features on the development set

Gender	Model type	Feature	DCNN		DCNN-DNN		DNN fusion	
			RMSE	MAE	RMSE	MAE	RMSE	MAE
Female	Depressed	AU	4.789	4.226	4.614	4.134	4.463	3.876
		Audio	4.590	3.589	4.516	3.633		
	Not Depressed	AU	2.486	2.033	2.457	1.986	2.499	2.061
		Audio	2.864	2.393	2.767	2.350		
Male	Depressed	AU	3.559	2.988	3.330	2.739	1.566	1.266
		Audio	1.802	1.690	1.467	1.226		
	Not Depressed	AU	2.972	2.320	2.903	2.238	2.809	2.167
		Audio	2.827	2.575	2.694	2.092		

3.4 Depression Recognition Results

In this section, we will introduce our depression recognition results. As shown in Figure 2, the proposed hybrid depression classification and recognition framework can be divided into three parts: the prediction of PHQ-8 score from audio and video features; the classification of depression/non-depression from text information; and the multivariate regression for final depression prediction.

3.4.1 PHQ-8 score prediction from audio and video features. As mentioned in Section 2, we build two separate audio-visual hybrid deep learning models to predict the PHQ-8 scores, one is trained with depressed participants and the other with not-depressed participants. The depression recognition results are shown in Table 5. One can notice that: 1) For single modality, the hybrid DCNN-DNN framework outperforms a DCNN framework trained for PHQ-8 prediction, indicating that our proposed DCNN-DNN framework effectively combines the advantages of DCNN and DNN. 2) The multi-modal DNN fusion scheme obtains comparable results to the single modality models.

3.4.2 Text-based depression/non-depression classification. The presence/absence classification accuracies from the paragraph vectors and SVM models of the 5 considered psychoanalytic factors are listed in Table 6. We can see that the overall classification accuracies are satisfying. In addition, the classification accuracy for the questions related to prior diagnoses of depression and PTSD are very high compared to the other topics (sleep, feelings, and personality), with 84.21% and 100% for female, 81.25% and 93.75% for male. This is mainly due to the fact that in the case of the prior diagnoses of depression and PTSD, the participant answers with discriminative words, such as "Yes" or "No", while for Sleep, Feeling lately and Personality issues, the answers are often more complex and ambiguous.

As mentioned in Section 2, apart from the text which can reflect the symptoms associated with psychoanalytic aspects of depression, we also exploit global text features, such as the number of laughs in the whole session. Table 7 lists the classification accuracies by fusing the results from PV features and global text features, in terms of $F1$ score, *precision*, and *recall*, for the "depressed" class (between brackets are results for the "non-depressed" class). One can notice from the precisions and recalls that all the samples classified as "depressed" are correct, no "not-depressed" participant is classified as "depressed". Moreover, the classification results for females are better than those of males.

Table 6: Classification accuracy of the symptoms for female and male on the development set

	Female (%)	Male (%)
Depression diagnoses	84.21	81.25
PTSD diagnoses	100	93.75
Sleep	47.37	68.75
Feelings	73.68	62.50
Personality	68.42	68.75

Table 7: Text-based depression classification results for female and male on the development set

Gender	F1 Score	Precision	Recall
Female	0.727(0.889)	1.000(0.800)	0.571(1.000)
Male	0.571(0.880)	1.000(0.786)	0.400(1.000)
All	0.667(0.885)	1.000(0.793)	0.500(1.000)

Table 8: Depression Recognition Results

Gender	Data set	RMSE	MAE
Female	training	0.826	0.626
	Dev.	3.416	2.696
Male	training	2.390	1.928
	Dev.	2.646	2.217
All	Dev.(baseline)	6.620	5.520
	Dev.(proposed)	**3.088**	**2.477**
	test.(baseline)	6.970	6.120
	test.(proposed)	**5.400**	**4.359**

3.4.3 Multimodal regression results. The final audio visual and text fusion estimations of PHQ-8 score for gender specific and all participants are given in Table 8. On both development set and test set, our proposed hybrid classification and recognition framework obtains better performance than the baseline results, with RMSE as 5.400 and MAE as 4.359 on the test set.

4 CONCLUSIONS

In this paper, we target the AVEC2017 Depression Challenge, and propose a hybrid depression classification and recognition framework. Initial PHQ-8 score predictions are obtained from audio and visual descriptors based on DCNN-DNN fusion models trained on depressed and non-depressed participants, respectively. At the same time, depression classification is performed using the Paragraph Vectors of 5 text sentences associated with psychoanalytic aspects of depression, as well as the global-text features. The final PHQ-8 prediction is obtained by a multivariate regression model from the initial predictions of depressed and non-depressed DCNN-DNN models, and the depression classification results. As far as we know, our approach is the first which applies Paragraph Vectors to textual transcripts for depression analysis.

In our future work, we will include an improved text analysis model to account for all the interview conversations, and also investigate end-to-end learning strategies to further boost performance.

ACKNOWLEDGMENTS

This work is supported by the Shaanxi Provincial International Science and Technology Collaboration Project (grant 2017KW-ZD-14), the National Natural Science Foundation of China (grant 61273265), and the VUB Interdisciplinary Research Program through the EMO-App project.

REFERENCES

[1] 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 *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 1–7.

[2] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*.

[3] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[4] Ringeval Fabien, Bjorn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Mozgai Sharon, Cummins Nicholas, Schmitt Maximilian, , and Maja Pantic. 2017. AVEC 2017 - Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge*.

[5] Lynne Friedli, World Health Organization, et al. 2009. Mental health, resilience and inequalities. (2009).

[6] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The Distress Analysis Interview Corpus of human and computer interviews.. In *LREC*. 3123–3128.

[7] Rahul Gupta, Nikos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth S. Narayanan. 2014. Multimodal Prediction of Affective Dimensions and Depression in Human-Computer Interactions. In *AVEC@MM*.

[8] Varun Jain, James L Crowley, Anind K Dey, and Augustin Lux. 2014. Depression estimation using audiovisual features and fisher vector encoding. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 87–91.

[9] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

[10] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[11] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 35–42.

[12] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed AI-Shuraifi, and Yunhong Wang. 2013. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 21–30.

[13] Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, Panagiotis Simos, Kostas Marias, et al. 2016. Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 27–34.

[14] Björn W Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load.. In *INTERSPEECH*. 427–431.

[15] Mohammed Senoussaoui, Milton Orlando Sarria Paja, João Felipe Santos, and Tiago H. Falk. 2014. Model Fusion for Multimodal Depression Classification and Level Detection. In *AVEC@MM*.

[16] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. 2015. Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender. *Journal on Multimodal User Interfaces* 9, 1 (2015), 17–29.

[17] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis 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*. ACM, 3–10.

[18] 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*. ACM, 3–10.

[19] James R. Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie K. Dagli, and Thomas F. Quatieri. 2016. Detecting Depression using Vocal, Facial and Semantic Communication Cues. In *AVEC@ACM Multimedia*.

[20] James R Williamson, Thomas F Quatieri, Brian S Helfer, Gregory Ciccarelli, and Daryush D Mehta. 2014. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 65–72.

[21] Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2016. Decision Tree Based Depression Classification from Audio Video and Language Information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 89–96.