Decision Tree Based Depression Classification from Audio Video and Language Information

Le Yang
NPU-VUB Joint AVSP Lab
School of Computer Science
Northwestern Polytechnical University(NPU)
127 Youyi Xilu, Xi'an 710072, China
yangle.cst@gmail.com

Dongmei Jiang
NPU-VUB Joint AVSP Lab
School of Computer Science
Northwestern Polytechnical University(NPU)
127 Youyi Xilu, Xi'an 710072, China
jiangdm@nwpu.edu.cn

Lang He
NPU-VUB Joint AVSP Lab
School of Computer Science
Northwestern Polytechnical University (NPU)
127 Youyi Xilu, Xi'an 710072, China
langhe@mail.nwpu.edu.cn

Ercheng Pei
NPU-VUB Joint AVSP Lab
School of Computer Science
Northwestern Polytechnical University (NPU)
127 Youyi Xilu, Xi'an 710072, China
peierch@mail.nwpu.edu.cn

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

Hichem Sahli
NPU-VUB Joint AVSP Lab
Dept. ETRO, VUB
Pleinlaan 2, 1050 Brussels
Interuniversity Microelectronics Centre
Kepeldreef 75, 3001 Heverlee, Belgium
hsahli@vub.ac.be

ABSTRACT

In order to improve the recognition accuracy of the Depression Classification Sub-Challenge (DCC) of the AVEC 2016, in this paper we propose a decision tree for depression classification. The decision tree is constructed according to the distribution of the multimodal prediction of PHQ-8 scores and participants' characteristics (PTSD/Depression Diagnostic, sleep-status, feeling and personality) obtained via the analysis of the transcript files of the participants. The proposed gender specific decision tree provides a way of fusing the upper level language information with the results obtained using low level audio and visual features. Experiments are carried out on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) database, results show that the proposed depression classification schemes obtain very promising results on the development set, with F1 score reaching 0.857 for class depressed and 0.964 for class not depressed. Despite of the over-fitting problem in training the models of predicting the PHQ-8 scores, the classification schemes still obtain satisfying performance on the test set. The F1 score reaches 0.571 for class depressed and 0.877 for class not depressed, with the average 0.724 which is higher than the baseline result 0.700.

Keywords

Depression classification, decision tree, multi-modal

1. INTRODUCTION

Nowadays, depression and anxiety disorders are highly prevalent worldwide causing burden and disability for individuals, families and society. According to the World Health Organization (WHO), depression will be the fourth mental disorder in 2020. The affective computing community has shown a growing interest in developing various systems, using audio and video features, to assist psychologists in the prevention and treatment of clinical depression.

Considering the audio features, researchers have found that depressed subjects are prone to possess a low dynamic range of the fundamental frequency, a slow speaking rate, a slightly shorter speaking duration, and a relatively monotone delivery [13], [5], [23], [16], [1]. Moreover, compared with health controls, the Harmonic-to-Noise Ratio (HNR) values of depressed subjects are higher [14]. Consequently, researchers formulated subtle changes in speech characteristics (e.g., differences in pitch, loudness, speaking rate, articulation, etc.) as indicators of depression. Low-level descriptors (LLDs) (such as energy, spectrum, and Mel frequency cepstrum coefficients-MFCC) based features were used as baseline audio features in the Audio Visual Emotion Challenge and Workshops (AVEC2013 and AVEC2014) [22], [21]. In [18] and [15], an i-vector based representation was computed to convert the frame-level features to a global representation. Experimental results revealed that i-vector level fusion of low-level features can result in more accurate systems for depression recognition.

Video features, such as body movements and gestures, subtle expressions and periodical muscular movements, have been also widely explored for depression analysis. To describe the dynamics of facial appearance, AVEC2013 [22]

AVEC'16, October 16 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-4516-3/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2988257.2988269

adopted the Local Phase Quantisation (LPQ) features, while AVEC2014 adopted the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) feature as baseline visual features. Girard et al. [8] investigated the relationship between nonverbal behavior and severity of depression using Facial Action Coding System (FACS) action units (AUs) and head pose. They found that when symptom severity was high, participants made fewer affiliative facial expressions (AU12 and AU15), more non-affiliative facial expressions (AU14), and diminished head motion. Head pose analysis was also made in [12] and [2]. In [17], Scherer et al. proposed the vertical (head and eye) gaze directionality, smile intensity and average duration, as well as self-adaptors and leg fidgeting, as nonverbal behavior descriptors. Space-temporal interesting point (STIP) features, describing the spatio-temporal changes by taking into account of the movements from facial area, hands, shoulder, and head etc., have been also employed in [12, 6] for depression classification. Typical symptoms of depression can be well described by global variation information, therefore most approaches in depression analysis extract global feature vectors from the complete video by aggregating a large set of local descriptors. In [10], Motion History Histogram (MHH), bag of words (BOW) and Vector of Local Aggregated Descriptors (VLAD) have been performed on the LGBP features or STIP features to obtain such global features.

Apart from the audio and visual cues, some researchers analyzed depression from the text/language information. In [7], the authors explored the potential to use the social media to detect and diagnose major depressive disorder in individuals. To characterize the topical language of individuals detected positively with depression, the authors built a lexicon of terms that are likely to appear in postings from individuals discussing depression or its symptoms in online settings. Using the frequency of depression terms, an Support Vector Machine (SVM) classifier was built to provide estimates of the risk of depression before the reported onset. Application of topic and sentiment modeling was presented in [11] for online therapy for depression and anxiety. It was found that besides the discussion topic and sentiment, style and/or dialogue structure is also important for measuring the patient progress. Asgari et al. [3] explored the information from "what is said" (content) and "how it is said" (prosody). To extract features from text, they used a published table to tag each word in an utterance with an arousal and a valence rating. Finally the speech prosody features and text features are fused to detect depression by a SVM classifier.

In [19], Stratou et al. showed that gender plays an important role in the automatic assessment of psychological conditions such as depression and PTSD, and a gender dependent approach significantly improves the performance over a gender agnostic one.

In this paper, we target the Depression Classification Sub-Challenge (DCC) task of AVEC2016 [20], and inspired by [19], we build classification models for females and males, respectively. First, a gender-specific multimodal framework, combining audio features, visual features as well as AU evidences and emotion evidences, is proposed for the prediction of PHQ-8 scores. Then, a depression classification decision tree is constructed according to the distribution of the multimodal predicted PHQ-8 scores and participants' characteristics obtained via the analysis of their transcript files. Four criteria, namely, PTSD/Depression Diagnostic, sleep-status, feeling and personality have been defined via content analysis of the participant's transcripts. The proposed decision tree provides a way of fusing the upper level language information with the predicted PHQ-8 scores obtained using low level audio and visual features. Experiments are carried out on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) database [9].

The remainder of this paper is organized as follows. The audio and video features, as well as the multimodal prediction framework of the PHQ-8 scores, are addressed in section 2. Section 3 describes the approach we used for content analysis of the transcripts to characterize the participants. Using the training set and based on basic summary statistics of the considered participant's characteristics and the PHQ-8 scores, we introduce in Section 4 the proposed decision tree based depression classification scheme for females and males, respectively. Section 5 analyzes the experimental results, and finally conclusions are drawn in Section 6.

2. PREDICTION OF THE PHQ-8 SCORES

2.1 Audio and Video Features

For the audio and video features, we make use of the baseline features provided by AVEC 2016. The baseline audio features, consist of 5 formant features and 74 prosodic and voice quality features, denoted here after as "covarep" features. From the video, based on the OpenFace [4] framework, AVEC 2016 provides histogram of oriented gradients (HOG) features, eye gaze features, and head pose features. In our implementation the eye gaze and head pose features have been concatenated into a "Gaze-pose" feature vector. AVEC2016, also provides (i) emotion evidence measures for the set {Anger, Contempt, Disgust, Joy, Fear, Neutral, Sadness, Surprise, Confusion, Frustration}. The evidence for an expression channel is a number (typically between -5 and +5) that represents the odds, in logarithmic (base 10) scale, of a target expression being present. And (ii) AUs: {AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, AU28, AU43} for AUs.

Using the provided 68 2D facial landmarks, we follow the approach of [20] to extract geometric features. We first calculate the mean shape of 51 stable points, for female and male respectively, using the samples of the training, development and test sets. Then, the feature points are aligned with the mean shape, and difference between the coordinates of the aligned landmarks and those from the mean shape, and also between the aligned landmark locations in the previous and the current frame, are computed, resulting in 204 features. The Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame is also calculated, resulting in 51 features. Finally, the facial landmarks are splitted into three groups of different regions: the left eye and left eyebrow, the right eye and right eyebrow, and the mouth. For each of these groups, the Euclidean distances and the angles between the points are computed, providing 75 features. In total, we obtain 330 geometric features.

As the sizes of the HOG feature vector and geometric feature vector are high, we apply PCA (with 99.9% of the variance) and obtain the following reduced features sets: 43 HOG-PCA features and 43 GEO-PCA features for female, 62 HOG-PCA features and 62 GEO-PCA features for male, respectively.

Finally, for each type of feature, we take its average over the entire screening interview as the global feature of the considered video.

2.2 Multimodal Prediction

According to [19], female and male has different symptoms on depression, therefore in our work, the PHQ-8 scores prediction is done separately for male and female. For each input feature stream, i.e. formant, covarep, HOG-PCA and GEO-PCA we train a separate Support Vector Regression (SVR) model with Radial basis function (RBF) kernel to predict the PHQ-8 score, the parameters cost and gama of the SVM were optimised in the range [2^{-8} – 2^{8}]. For the emotion and AU measures we train several SVRs considering as input feature all the combinations of 1 to 4 evidences, and select the combination producing the lowest root mean square error (RMSE) and mean absolute error (MAE) between the predicted and reported PHQ-8 scores, averaged over all sequences. In our experiment, the (disgust, fear, sadness) emotion evidence combination and the (AU5, AU17, AU25) AU evidence combination obtained the lowest RMSE and MAE among all the combinations for female. While for male, (joy, baseline, confusion) and (AU5, AU20, AU25) obtain the best prediction performance on the development set. The output of the 7 unimodal (GEO-PCA, Gaze-pose, HOG-PCA, covarep, formant, best emotion evidence combination, best AU evidence combination) SVRs, are input to a second level SVR model, or a local linear regression (LLR) model, for the final (multimodal) PHQ-8 score predication as follows. As the AU evidence stream and the emotion evidence stream provide promising prediction results on the training set and development set, we use these two streams as inputs to the second level SVR model and select among the other 5 streams (GEO-PCA, Gaze-pose, HOG-PCA, covarep, and formant) the ones providing the lowest RMSE and MAE (see Table 10).

3. PARTICIPANT CHARACTERISTICS

Apart from the above described features, we conducted content analysis of the transcripts to characterize the participants following four criteria: PTSD/Depression Diagnostic (Yes/No), sleep-status (Normal/Abnormal), Feeling (Bad/Good) and Personality (Shy/Extrovert). The analysis has been made as follows:

- **Sleep Status.** if the participant answers the “easy sleep” question with positive words such as “no problem”, “pretty good”, “get a good night’s sleep”, “pretty easy”, “easy”, “I’m ok”, “fairly easy”, etc., or he/she does not answer this question, the sleep status is marked as “normal”. In case the answer contains the words such as “not had a good sleep”, “really hard”, “kinda difficult”, “never easy”, etc., the sleep status is marked as “abnormal”. Moreover, according to the reason of not having a good sleep, with such as “disturbing thought”, “mind will be racing a lot”, “thoughts running through my mind”, “hard to keep my thoughts”, etc., the sleep status is further marked as “sleep abnormal/mind reason”. If there is no information about the reason, the sleep status is considered as “sleep abnormal/other reason”.

- **PTSD/Depression Diagnosed.** The value of this criteria is “yes” or “no” according to the characteristic “depression_diagnosed” and “ptsd_diagnosed” transcriptions.

- **Feeling.** This attribute takes the values “Bad” or “Good” following the transcript “feel_lately”. The value “Bad” is given when the transcript contain negative words such as “feeling depressed”, “little depressed”, “tired”, “sad”, “depressed blue”, “not okay”, “frustrated”, “angry”, “down”, we mark the participant as “feel bad”. If it contains the positive words like “fine”,“good”, “pretty good”, “great”, “okay”, or the participant does not answer this question, the feeling status will be considered as “Good”.

- **Personality.** This criteria takes the value “Shy” if the transcript contains the words “shy”, “introvert”, “more shy” and “probably shy”. If the words like“outgoing”, “extrovert”, “mostly outgoing”, are used, we mark the participant as “outgoing”. If the answer contains “the middle”, “a little bit of both”, “depends on the situation”, the personality is considered “extrovert”.

4. DECISION TREE BASED DEPRESSION CLASSIFICATION

The research results of [19] have shown that contributions of different behavioral indicators to depression and PTSD are different for males and females. This finding implies that a decision tree-based classification method maybe improves the recognition accuracy of depression. Most of the methods that generate decision trees for a specific problem use examples of data instances in the decision tree generation process. To this aim we examined the statistics of the above defined participants characteristics, which are summarized in the following sections.

4.1 Females

Based on the training set we computed basic summary statistics for each of the defined characteristics:

- **Sleep Status.** From Table 1, one can notice that most (67.74%) of the not depressed females are marked as “sleep normal”. While 84.62% of the depressed females are marked as “sleep abnormal”, among which 61.54% are because of the “mind reason”, showing that depressed females think a lot when they sleep.

Table 1: Sleep Status - Females

classes	sleep normal (%)	sleep abnormal(%)	
		mind reason	other reason
not depressed	21(67.74)	1(3.23)	9(29.03)
depressed	2(15.38)	8(61.54)	3(23.08)

- **PTSD/Depression Diagnosed.** Statistics on whether the females have been diagnosed with depression or PTSD are listed in Table 2, which indicates that almost all (92.31%) of the depressed females have been diagnosed with either depression or PTSD before, or even both, while only 25.81% of the not depressed females have been diagnosed with depression or PTSD.

Table 2: PTSD/Depression - Females

classes	no ptsd/ depression(%)	ptsd/ depression(%)
not depressed	23(74.19)	8(25.81)
depressed	1(7.69) (no answer)	12(92.31)

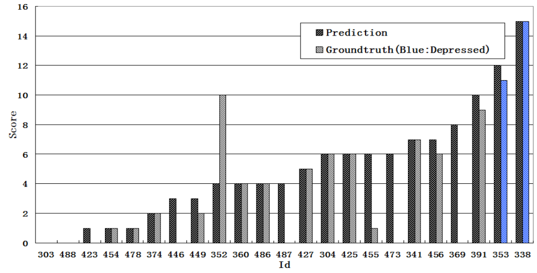

Figure 1: PHQ-8 scores of the "sleep normal" females

- **Feeling.** Table 3 showing that 83.87% of the not depressed females feel good, while 69.23% of the depressed females feel bad.

Table 3: Feeling - Females

classes	feel bad (%)	feel good (%)
not depressed	5(16.13)	26(83.87)
depressed	9(69.23)	4(30.77)

- **PHQ-8 Scores.** Figure 1 depicts the predicted and ground truth PHQ-8 scores of the participants marked as "sleep normal". As it can be seen, all the 21 not depressed females have a PHQ-8 score lower than 11, while the PHQ-8 scores of the two depressed females are higher or equal to 11.

From the above statistics, the decision tree based depression classification for females is given in Figure 2.

As shown in Table 1, 84.62% of the depressed females can not sleep well, therefore the sleep status is firstly checked. If it is "sleep normal", then we check the predicted PHQ-8 score, if the score is lower than a threshold (10 in our experiment to release the influence of inaccuracy in predicting the PHQ-8 score), the participant is considered not depressed (class "0"). If the score is higher than the threshold, we further check if the participant has been already diagnosed (variable "ptsd/depression diagnosed").

On the other hand, if the sleep status is "sleep abnormal", we further check the reasons. From Table 1, we can see that 8 depressed females can not sleep well because of mind reason, therefore, we further test the "sleep reason", followed by the "ptsd/depression diagnosed" status and finally the "feeling".

4.2 Males

The statistical analysis of the characteristics variables for males are summarized here after.

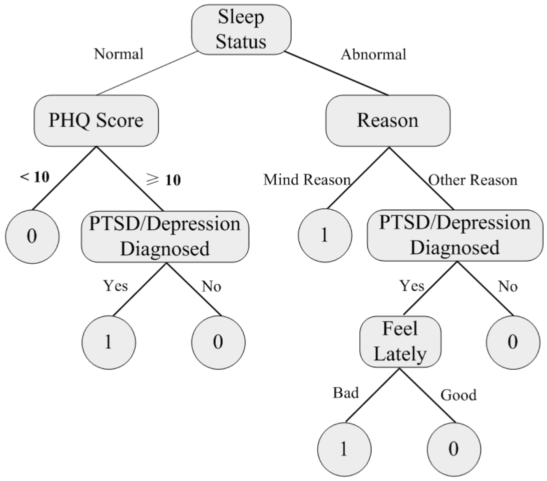

Figure 2: Decision Tree for Females

- **Sleep Status.** Among the 8 depressed males of the training samples, 1 does not answer the "easy sleep" question, 3 have normal sleep, and 4 have abnormal sleep. Therefore, we did not use this characteristic for the final classification.
- **PTSD/Depression Diagnosed.** Among the depressed males, 5 (62.5%) have been diagnosed with depression or PTSD before, while 3 (37.5%) have not been diagnosed before. Therefore we did not use this characteristic for the final classification.
- **Feeling.** From Table 4. One can notice that 87.5% of the depressed males feel bad, while only 20% of the not depressed males feel bad. Therefore the "feeling" characteristics is discriminative to classify depressed/not depressed males.

Table 4: Feeling - Males

classes	feel bad (%)	feel good (%)
not depressed	11(20)	44(80)
depressed	7(87.5)	1(12.5)

- **Personality.** Statistics on the males personalities are listed in Table 5. We should notice here that among the 26 "shy", 17 of them explicitly used the words related to "shy", and the other 9 did not answer the question. Therefore, we consider that for the not depressed males, their personality of being "shy", "Extrovert" and "both" is evenly distributed. For the depressed males, 4(50%) think themselves as being "shy" (2 participants do not answer this question), and no one answered about "Extrovert".

Table 5: Personality - Males

classes	shy(%)	outgoing(%)	both(%)
not depressed	26(47.27)	16(29.09)	13(23.64)
depressed	6(75)	0(0)	2(25)

- **PHQ-8 Scores.** The predicted and ground-truth PHQ-8 scores of the "shy" (or "both") males who feel bad recently are shown in Figure 3. We can see that when

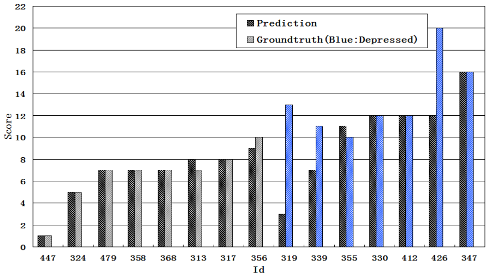

Figure 3: PHQ-8 Scores of the males

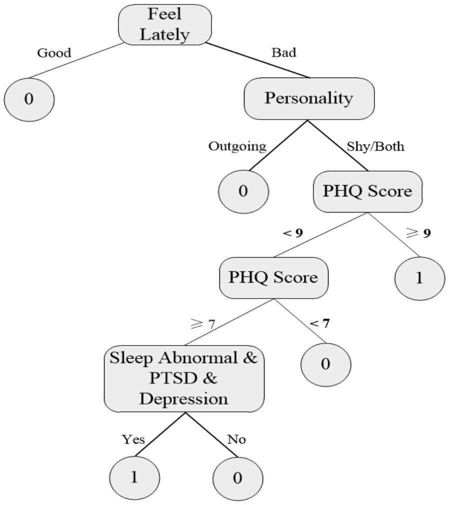

Figure 4: Decision Tree for Males

the predicted scores are higher than 9, the participants are depressed, while when the scores are lower than 7, the participants are not depressed. For the participants with ID319 and ID339, the PHQ-8 scores are far from being accurately predicted.

The decision tree of depression classification for males is as shown in Figure 4.

5. EXPERIMENTS AND ANALYSIS

5.1 Prediction of PHQ-8 Scores

The root mean square error (RMSE) and mean absolute error (MAE) averaged over all sequences are being used to assess the proposed approach.

5.1.1 Prediction from Audio and Visual Features

We report in Table 6 and Table 7 the PHQ-8 scores prediction accuracy, using single stream input features. The dimension of the feature vectors are given between brackets. One can see that the audio and visual features obtain close performances. For the females, the RMSEs and MAEs on the training set and development set are very close. For the males, the differences of RMSEs and MAEs between the training set and development set are high, showing that the SVR models are somehow over-fitting in the training process.

Table 6: Audio/Visual Prediction - Female

Features	Dataset	RMSE	MAE
GEO-PCA(43)	Train	5.778	4.705
	Dev.	6.387	5.105
Gaze-pose(9)	Train	5.800	4.727
	Dev.	6.362	5.105
HOG-PCA(43)	Train	5.891	4.886
	Dev.	6.391	5.158
covarep(74)	Train	5.560	4.545
	Dev.	6.224	4.842
formant(5)	Train	5.778	4.705
	Dev.	6.320	5.000

Table 7: Audio/Visual Prediction - Male

Features	Dataset	RMSE	MAE
GEO-PCA(62)	Train	4.832	3.825
	Dev.	6.982	5.750
Gaze-pose(9)	Train	4.020	2.762
	Dev.	6.955	5.750
HOG-PCA(62)	Train	4.832	3.825
	Dev.	6.982	5.750
covarep(74)	Train	4.339	3.111
	Dev.	6.910	5.750
formant(5)	Train	4.761	3.714
	Dev.	6.833	5.563

5.1.2 Prediction from Evidence Features

We combine different AU evidences or emotion evidences as input features to predict the PHQ-8 scores. In our experiment, we train several SVRs considering as input feature all the combinations of 1 to 4 evidences, and select the combination producing the lowest RMSE and MAE between the predicted and reported PHQ-8 scores, averaged over all sequences. Table 8 and Table 9 lists some of the obtained results, for female and male, respectively. One can notice that for both female and male, and for both emotion evidence and AU evidence, the combination with 3 evidences obtains the lowest RMSEs and MAEs. Moreover, in the case of 4 AU evidences for male, the RMSE (MAE) on the training set is 0.756 (0.222), while being 4.191 (3.563) on the development set, showing that the SVR model is over-fitting. Therefore in our experiments, we use the combinations of 3 evidences, indicated in bold in Table 8 and Table 9, to predict the PHQ-8 scores from emotion evidences and AU evidences, respectively.

5.1.3 Multimodal Prediction of the PHQ-8 Scores

The best multimodal prediction results of the PHQ-8 scores on the training and development sets are listed in Table 10. The LLR model has been used for the multimodal prediction of PHQ-8 scores of females, and SVR for males. The RMSE and MAE on the test set are also reported. One can notice that they are quite high compared to those on the training set and development set.

5.2 Classification Results

Based on the decision trees of Figure 2 and Figure 4, depression classification experiments are carried out on the development set and the test set of the DAIC-WOZ database, respectively. The confusion matrix on the development set

Table 8: Evidence Based Prediction - Female

Evidence		Dataset	RMSE	MAE
Emotion	disgust	Train	6.094	5.0
		Dev.	5.699	4.579
	sadness, frustration	Train	5.811	4.182
		Dev.	5.161	3.895
	disgust, fear, sadness	Train	**3.519**	**1.932**
		Dev.	**4.377**	**3.368**
	anger, joy, fear, frustration	Train	4.026	2.432
		Dev.	4.894	3.737
	All	Train	3.908	2.273
		Dev.	5.943	4.579
AU	AU10	Train	4.975	3.341
		Dev.	5.201	4.211
	AU17, AU25	Train	5.379	3.477
		Dev.	4.322	3.526
	AU5, AU17, AU25	Train	**3.879**	**2.046**
		Dev.	**3.974**	**3.263**
	AU9, AU17, AU25, AU28	Train	4.647	2.955
		Dev.	4.383	3.421
	All	Train	5.796	4.818
		Dev.	6.279	4.895

Table 9: Evidence Based Prediction - Male

Evidence		Dataset	RMSE	MAE
Emotion	confusion	Train	4.595	3.206
		Dev.	6.093	5.25
	contempt, joy	Train	4.271	2.429
		Dev.	5.534	4.250
	joy, baseline, confusion	Train	**3.462**	**1.952**
		Dev.	**5.466**	**4.500**
	contempt, joy, sadness, confusion	Train	4.106	2.127
		Dev.	5.673	4.563
	All	Train	4.483	3.365
		Dev.	6.942	5.688
AU	AU23	Train	4.595	2.921
		Dev.	5.511	4.125
	AU4, AU14	Train	3.581	2.095
		Dev.	4.323	3.188
	AU5, AU20 AU25	Train	**3.625**	**1.492**
		Dev.	**4.294**	**3.188**
	AU1, AU10, AU17, AU18	Train	0.756	0.222
		Dev.	4.191	3.563
	All	Train	4.832	3.825
		Dev.	6.982	5.750

is shown in Table 11, as it can be seen, most females, as well as males, have been correctly classified. Among the 35 participants of the development set, only 2 participants have not been correctly classified.

The F1 score, precision, and recall for the "depressed" class, and between brackets for the "non depressed" class, are reported in Table 12. We can see that on the development set, the decision trees obtain very promising results for both males and females, with the overall F1 score reaching 0.857 for the "depressed" class, and 0.964 for the "non depressed" class, which are much higher than the baseline results. On the test set, the F1 score reaches 0.571 for the "depressed" class and 0.877 for the "non depressed" class, with the average 0.724 which is also higher than the baseline results. However, the obtained results using the test set are not so promising, this could be due to the over-fitting of the SVR models, which influences the classification in the decision trees.

6. CONCLUSIONS

In this paper, with the purpose improving the recognition accuracy of the Depression Classification Sub-Challenge (D-CC) of the AVEC 2016, we proposed a decision tree for depression classification. Two decision trees have been proposed, one for males and one for females. The decision trees have been constructed according to the distribution of the multimodal prediction of PHQ-8 scores and participants' characteristics (PTSD/Depression Diagnostic, sleep-status, feeling and personality) obtained via the analysis of the transcript files of the participants. The proposed gender specific decision tree provides a way of fusing the upper level language information with the results obtained using low level audio and visual features.

In our current implementation we considered a manual decision tree generation process, in future work we planned investigating automatic approaches, also other regression approaches will be investigated for the PHQ-8 scores.

Table 10: Multimodal Prediction of PHQ-8 Scores

Gender	Features	Data	RMSE	MAE
Female (LLR)	disgust, fear sadness, AU5, AU17, AU25, Geo-PCA	Train Dev.	3.286 3.770	2.023 2.632
Male (SVR)	joy, baseline, confusion AU5, AU20, AU25, formant, covarep	Train Dev.	2.705 3.666	1.000 2.938
All		test	9.106	6.702

Table 11: Confusion Matrix on the Development Set

Gender	Class	Depressed	Not Depressed
Female	Depressed	3	0
	Not Depressed	1	15
Male	Depressed	3	1
	Not Depressed	0	12
All	Depressed	6	1
	Not Depressed	1	27

7. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (grant 61273265), the Research and Development Program 863 of China (No. 2015AA016402), and the VUB Interdisciplinary Research Program through the EMO-App project.

8. REFERENCES

[1] S. Alghowinem. From joyous to clinically depressed: mood detection using multimodal analysis of a person's appearance and speech. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 648–654. IEEE, 2013.

[2] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear. Head pose and movement analysis as an indicator of depression. In *Affective Computing*

Table 12: Depression Classification Results

Gender	Data	F1 Score	Precision	Recall
Female	Dev.	0.857(0.968)	0.750(1.000)	1.000(0.938)
Male	Dev.	0.857(0.960)	1.000(0.923)	0.750(1.000)
All	Dev.(proposed)	0.857(0.964)	0.857(0.964)	0.857(0.964)
	Dev.(baseline)	0.58(0.86)	0.47(0.94)	0.78(0.79)
	test(proposed)	0.571(0.877)	0.500(0.914)	0.667(0.842)
	test(baseline)	0.50(0.90)	0.60(0.87)	0.43(0.93)

and Intelligent Interaction (ACII), 2013 Humaine Association Conference, pages 283–288. IEEE, 2013.

[3] M. Asgari, I. Shafran, and L. B. Sheeber. Inferring clinical depression from speech and spoken utterances. In *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, pages 1–5, 2014.

[4] T. Baltru, P. Robinson, L.-P. Morency, et al. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[5] N. Cummins, J. Epps, M. Breakspear, and R. Goecke. An investigation of depressed speech detection: features and normalization. In *Interspeech*, pages 2997–3000, 2011.

[6] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps. Diagnosis of depression by behavioural signals: a multimodal approach. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 11–20. ACM, 2013.

[7] M. Gamon, M. D. Choudhury, S. Counts, and E. Horvitz. Predicting depression via social media. In *AAAI*, 2013.

[8] J. M. Girard, J. F. Cohn, and M. H. Mahoor. Nonverbal social withdrawal in depression: evidence from manual and automatic analyses. *Image and Vision Computing*, 32(10):641–647, 2014.

[9] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, B. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency. The Distress Analysis Interview Corpus of human and computer interviews. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 3123–3128, 2014.

[10] L. He, D. Jiang, and H. Sahli. Multimodal depression recognition with dynamic visual and audio cues. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 260–266. AAAC, 2015.

[11] C. Howes, M. Purver, R. Mccabe, and R. Mccabe. Linguistic indicators of severity and progress in online text-based therapy for depression. In *ACL Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal To Clinical Reality*, pages 7–16, 2014.

[12] 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.

[13] L.-S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5154–5157. IEEE, 2010.

[14] L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen. Detection of clinical depression in adolescentsąŕ speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–86, 2011.

[15] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, and M. Graciarena. The SRI AVEC-2014 evaluation system. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 93–101. ACM, 2014.

[16] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics*, 20(1):50–64, 2007.

[17] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, R. Albert, and L.-P. Morency. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658, 2013.

[18] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk. Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 57–63. ACM, 2014.

[19] G. Stratou, S. Scherer, J. Gratch, and L. P. Morency. Automatic nonverbal behavior indicators of depression and PTSD: the effect of gender. *Journal on Multimodal User Interfaces*, 9(1):1–13, 2014.

[20] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. 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*, 2016.

[21] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014: 3D dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014.

[22] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013: the continuous audio/visual emotion and

depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[23] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 41–48. ACM, 2013.