INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Predicting Affective Dimensions based on Self Assessed Depression Severity

Rahul Gupta, Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, USA

Abstract

Depression is a state of severe despondency and affects a person's thoughts and behavior. Depression leads to several psychiatric symptoms such as fatigue, restlessness, insomnia as well as other mood disorders (e.g. anxiety and irritation). These symptoms have a resultant impact on the subject's emotional expression. In this work, we address the problem of predicting the emotional dimensions of valence, arousal and dominance in subjects suffering from variable levels of depression, as quantified by the Beck Depression Inventory-II (BDI-II) index. We investigate the relationship between depression severity and affect, and propose a novel method for incorporating the BDI-II index in affect prediction. We validate our models on two datasets recorded as a part of the AViD (Audio-Visual Depressive language) corpus: Freeform and Northwind. Using the depression severity and a set of audio-visual cues, we obtain an average correlation coefficient of .33/.52 for affective dimension prediction in the Freeform/Northwind datasets, against baseline performances of .24/.48 based on using the audio-visual cues only. Our experiments suggest that the knowledge of depression severity significantly improves the emotion dimension prediction, however the BDI-II score incorporation scheme varies between the two datasets of interest.

Index Terms: Depression, Affect prediction, Valence, Arousal, Dominance, Multi-modal fusion

1. Introduction

Depression is a clinical condition characterized by a state of low mood, despondency and dejection [1]. Depression can impact a patient's mood leading to symptoms such as sadness, anxiety and restlessness [2]. The National Institute of Mental Health identifies various forms of depression (such as major depressive disorder, dysthemic disorder and psychotic depression) and links their impact to the patient's life including his personal relationships, professional life as well as daily habits such as eating and sleeping [3]. Depression also has a direct impact on the patients affective expression and their association has been widely studied in relation to depression therapy [4], genetic analysis of depression [5] and perception of emotions [6]. In this work, we address the problem of relating affective expression to the severity of a patient's depressive disorder. Tracking affective dimensions (valence, arousal and dominance) is a classic problem in the study of emotions [7] and we incorporate the severity of depression quantized by a self-assessed metric, the Beck Depression Inventory-II (BDI-II) index [8] in prediction of the affective dimensions. Through our proposed affect prediction system, we aim to exploit the impact that depression has on a patient's emotional states with an overarching goal of assisting the analysis and treatment of depression disorders.

Several previous works have analyzed the relationship between depression and emotions including various cross-cultural studies [9], in neurology [10] and psychology [11]. Greenberg et al. [4] describe an emotion focused therapy for depression in their book and Izard [12] and Blumberg et al. [13] analyzed patterns of emotions with respect to depression. Considering the application of machine learning to the analysis of depression, researchers have investigated the relation between depression and various audio-visual cues using Canonical Correlation Analysis (CCA) [14], i-vectors [15] and acoustic volume analysis [16]. Tracking affect is another problem that is widely studied in emotion research. For instance, Metallinou et al. [17] incorporated body language and prosodic cues in tracking continuous emotion and Nicolaou et al. [7] proposed an output-associative relevance vector machine regression for continuous emotion prediction. Ringeval et al. [18] utilized physiological data in predicting emotions and Gunes et al. [19] presented an analysis of trends and future directions in affect analysis. The Audio Visual Emotion Challenges (AVEC) [20, 21] led to particular interest in the study of depression disorder and emotions. Several interesting approaches were presented as a part of the challenge in predicting depression and tracking affective dimensions. A few proposed methods for tracking emotions include using ensemble CCA [22], regression based multi-modal fusion [23] and use of application dependent meta knowledge [24]. Methods proposed for rating depression include the use of vocal and facial biomarkers [25], facial expression dynamics [26] and Fisher vector encoding [27].

Despite the progress in the study of relating depression and emotions, existing models do not take depression severity into account while tracking affect. The challenge lies in incorporating a single patient specific depression assessment value in the models for tracking affect. We address this challenge in this work by performing feature transformation based on the self-assessed depression severity. We perform experiments on two datasets obtained from the Audio-Visual Depressive language (AViD) corpus: Freeform and Northwind datasets [21]. Both of these datasets contain sessions involving human-computer interaction, where either the patients discuss freely on a given question (Freeform) or read aloud an excerpt (Northwind). In order to establish the relationship between depression severity and affect in the datasets, we initially perform preliminary correlation analysis between the patients' BDI-II index and statistical functionals computed over their affective dimensions. This is followed by the design of the affect prediction system, where we first develop a baseline system based on audio-visual features only. We then extend the model to incorporate feature transformation based on the depression severity (as quantified by BDI-II index) for the specific individual patient in the session. The motivation behind adding the feature transformation is to train a joint model, incorporating the scalar depression severity value within the audio visual prediction system. We test several feature transformation schemes and our best models obtain a mean correlation coefficient values of .33 and .52 (baseline system performance: .24 and .48) computed over the affective dimensions (valence, arousal and dominance) in the Freeform and Northwind datasets, respectively. Finally, we discuss the feature transformations applied, interpret the results for the two datasets and propose a few future directions.

2. Database

We use a subset of the AViD (Audio-Visual Depressive language) corpus in this work, also used in the Audio-Visual Emo-

http://dx.doi.org/10.21437/Interspeech.2016-187

tion Challenges (AVEC), 2013-14 [20,21]. The corpus includes microphone and webcam recordings of subjects performing a human-computer interaction task. A single recording session contains only one subject. The subset of the corpus we use is divided into two parts: Freeform and Northwind datasets. Both these datasets contain the same set of subjects, with 100 session recordings. In the Freeform dataset, the participants respond freely to a question, while the Northwind dataset is more structured in the sense that the participants read aloud a given excerpt. 3-5 naive annotators rate every session with three affective dimensions: valence, arousal and dominance, at a rate of 30 Frames Per Second (FPS). The final affect annotations are obtained as the mean over all the annotator ratings, computed per frame. The subjects participating in the sessions also complete the standardized self-assessment based Beck Depression Inventory-II (BDI-II) questionnaire [8]. The BDI-II index is a single score between 0-63 determined based on a set of 21 questions, with a higher score implying more severe depression. For more details regarding the corpus, please refer to [20].

3. Experiments

We divide our experiments into two parts: (i) Investigating the relationship between affective dimensions and the depression severity using correlation analysis and, (ii) Affect prediction incorporating self-assessed depression severity. We describe our experiments in detail below.

3.1. Investigating relationship between affective dimensions and depression severity

As discussed in section 1, existing literature offers an in depth exploration of the relationship between affect and depression and suggests several links [5,6]. In this experiment, we perform an analysis to validate the relationship between affective dimensions and depression severity on the Freeform and Northwind datasets. We compute session-level statistics (mean, variance, range and median) over the time series of affective dimensions (valence, arousal and dominance) and look for any significant correlation with the BDI-II index. Table 1 lists the values of correlation coefficient between each of these statistics and the BDI-II score for the two datasets. Significance of the correlation coefficient is computed using the Student's t-distribution test at 5% level against a null hypothesis of no correlation. Since we are performing multiple hypothesis tests, we apply the Bonferroni correction [28]. We limit ourselves to a few statistical functionals as the Bonferroni correction is likely to give more false negatives with increasing number of significance tests.

From the table 1, we observe that the severity of the depression correlates significantly with several statistics of the affective dimensions for both the datasets. In particular, the mean and median statistics correlate well with the BDI-II score for both the datasets. The variance and range statistics correlate with the BDI-II score only for the Northwind dataset sessions. As the set of subjects is same across the two datasets, this difference in correlation suggests that affective expression may be affected by the nature of the task (dataset collection). Spontaneous versus read elicitation exercise different aspects of the neurocognitive system and hence the resulting affective vocal/visual behavior can be differently affected by depression. Nevertheless, significance of several correlation coefficients validate the relationship between emotions and depression and motivates our next experiment in predicting affective dimensions based on depression severity assessment.

3.2. Predicting Affective dimensions

In this section, we propose a model to predict the frame-wise affective dimension ratings conditioned on the depression severity. Initially, we develop a multi-modal system for predicting affective dimensions and use it as a baseline. We then extend the baseline model to incorporate the BDI-II index as a parameter in affect prediction. We describe these models below.

Table 1: Correlation coefficient ρ between a subject's BDI-II score and statistical functionals computed over the affective dimensions for his session. Significance of $\rho \neq 0$ is shown in bold.

	Freeform			Northwind		
	Val.	Aro.	Dom.	Val.	Aro.	Dom.
Mean	**-.40**	**-.46**	**-.35**	**-.34**	**-.50**	-.20
Median	**-.39**	**-.46**	**-.35**	**-.33**	**-.51**	-.20
Variance	-.09	-.08	.08	-.23	**-.28**	-.09
Range	-.03	-.08	.07	**-.31**	**-.39**	-.13

3.2.1. Baseline: Multi-modal affective dimension prediction

We initially develop a system for affect prediction based on audio-visual cues. We perform a frame-wise extraction of several audio-visual features and develop a multi-layered system for affect prediction. Below, we describe the audio visual cues used in prediction followed by the model description.

Multi-modal cues: We use a similar set of audio visual cues as was used in the AVEC challenge 2014 [21]. A brief description of the audio-visual features is given below.

a) Audio features: We adopt the set of audio features proposed in the AVEC 2014 challenge baseline paper [21]. The set of features include low level descriptors such as Mel Frequency Cepstral Coefficients, loudness, jitter and shimmer. For a complete list of features used in the audio model, please refer to the Table 1 in [21]. The list consists of 32 energy and spectral features and 6 voicing related features. We further append delta features and window-wise statistical functionals to these 38 features as described in the AVEC 2014 baseline paper [21]. Note that the audio features are obtained at a sample rate five times the annotation frame rate (30 FPS). Thus, we downsample the audio features by sampling every fifth frame, as is suggested in [21]. We represent the audio features for the t^{th} frame in the session s as the row vector $\boldsymbol{x}_a^s(t)$.

b) Video features: The set of video features used in the baseline system is also borrowed from the AVEC challenge 2014 [21]. The proposed set of frame-wise Local Binary Pattern (LBP) features is well known for describing facial expressions. The LBP descriptors computed for a pixel compare the pixel's intensity to it's neighbors. After computing the descriptors per pixel, a histogram feature is computed with each bin as different binary pattern. For a complete description of the LBP features, please refer to section 4.2 in [21]. We represent the video features for the t^{th} frame in the session s as the row vector $\boldsymbol{x}_v^s(t)$.

Affect prediction system: Our baseline affect prediction system uses the aforementioned audio-visual cues for frame-wise prediction of the affective dimensions. A schematic of the baseline system is shown in Figure 1. We describe various components of the system below.

a) Input audio/video features: The bottom-most layer of the system in Figure 1 serves as the input for the audio/video feature values. Note that we have separate inputs for the audio and video features. In the session s, to predict the affective dimensions for the t^{th} frame, the system uses a window of audio/video features centered at the t^{th} frame. That is, for the t^{th} frame, the audio (video) features used is the concatenated set of vectors $[\boldsymbol{x}_a^s(t-n),..,\boldsymbol{x}_a^s(t),..,\boldsymbol{x}_a^s(t+n)]$ $([\boldsymbol{x}_v^s(t-n),..,\boldsymbol{x}_v^s(t),..,\boldsymbol{x}_v^s(t+n)])$ where the window length is given by $2n+1$.

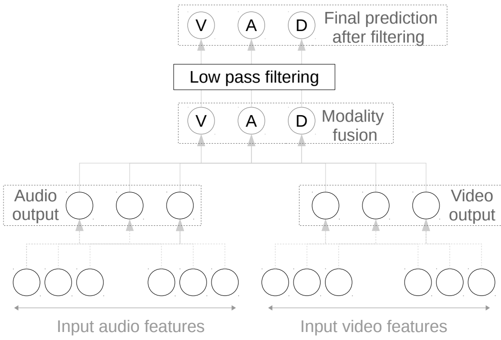

Figure 1: Baseline system with audio-video features as inputs.

b) Audio/video outputs: The audio/video outputs are the output values obtained from the respective modalities. The motivation for such an untied system is for an independent evaluation of each modality. We represent the audio (video) output for the t^{th} frame in the session s as $\boldsymbol{y}_a^s(n)$ ($\boldsymbol{y}_v^s(n)$). The dimensionality of $\boldsymbol{y}_a^s(n)$ and $\boldsymbol{y}_v^s(n)$ is the same as the number of affective dimensions, as represented by the 3 nodes in the audio/video output layer in Figure 1. We chose a linear system to obtain $\boldsymbol{y}_a^s(n)$ and $\boldsymbol{y}_v^s(n)$ from the window of audio and video features as shown in equations (1) and (2). $\boldsymbol{w}_a$ and $\boldsymbol{w}_v$ represent the weight vectors multiplied with the audio and video features, respectively, and $\boldsymbol{b}_a$ and $\boldsymbol{b}_v$ are the bias terms. The strategy for training the system and obtaining the parameters $\boldsymbol{w}_a, \boldsymbol{w}_v, \boldsymbol{b}_a$ and $\boldsymbol{b}_v$ is described in the next section.

$$\boldsymbol{y}_a(t) = \boldsymbol{w}_a[\boldsymbol{x}_a(t-n), .., \boldsymbol{x}_a(t), .., \boldsymbol{x}_a(t+n)]^T + \boldsymbol{b}_a \quad (1)$$

$$\boldsymbol{y}_v(t) = \boldsymbol{w}_v[\boldsymbol{x}_v(t-n), .., \boldsymbol{x}_v(t), .., \boldsymbol{x}_v(t+n)]^T + \boldsymbol{b}_v \quad (2)$$

c) Modality fusion: Modality fusion performs a weighted combination of the outputs $\boldsymbol{y}_a^s(t)$ and $\boldsymbol{y}_v^s(t)$ to provide the fused output $\boldsymbol{y}_f^s(t)$ for the t^{th} frame in session s. The fusion is again chosen to be linear and the output $\boldsymbol{y}_f^s(t)$ is obtained as shown in equation (3). $\boldsymbol{w}_f$ and $\boldsymbol{b}_f$ represent the weight and bias vectors used for fusion, respectively.

$$\boldsymbol{y}_f^s(t) = \boldsymbol{w}_f[\boldsymbol{y}_a^s(t), \boldsymbol{y}_v^s(t)]^T + \boldsymbol{b}_f \quad (3)$$

One could chose one of the several strategies for training the model shown in Figure 1. For instance, the bottom three layers in Figure 1 represent a neural network and can be trained using the standard back-propagation algorithm [29]. However, we chose to train each layer independently using data bootstrapping [30]. That is, the audio and video system parameters ($\boldsymbol{w}_a, \boldsymbol{b}_a$ and $\boldsymbol{w}_v, \boldsymbol{b}_v$) are optimized independently on randomly sampled portions of the training set to predict the affective dimensions; using the Minimum Mean Squared Error (MMSE) criteria [31]. The fusion parameters ($\boldsymbol{w}_f, \boldsymbol{b}_f$) are then obtained to predict affective dimensions on another independently sampled subset of the training data by fusing audio and video outputs ($\boldsymbol{y}_a$ and $\boldsymbol{y}_v$), again using the MMSE criteria. We randomly sample 80% of the training data for each optimization. We chose this training strategy because of the following reasons: (i) This strategy allows for independent evaluation of audio and video systems, as well as their fusion, (ii) our preliminary experiments suggested that while data bootstrapping and back-propagation results are comparable, the former is faster to perform. Next, we describe our final low pass filtering step to obtain the final predictions.

d) Final prediction after filtering: In the final step of the baseline system, we low pass filter the time-series of predicted affective dimensions. Note that this is a post processing step after predictions for each analysis frame has been obtained. This step is motivated from the fact that affective dimensions evolve smoothly over time without abrupt changes, as is also observed in several other works [23, 32]. In our experiments, we use a moving average filter (length: k) as the low pass filter. In the next section, we extend the current system to use the depression severity in predicting the affective dimensions.

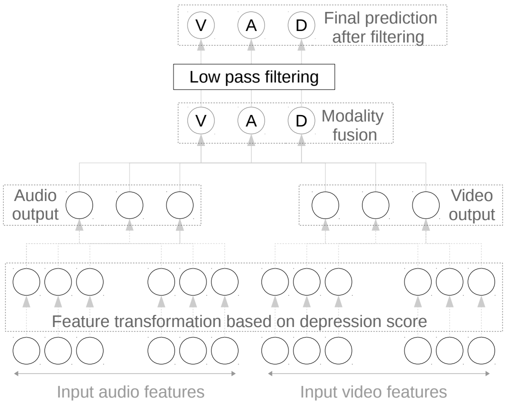

Figure 2: Proposed system with a feature transformation layer appended to the baseline system.

3.2.2. *Affective dimension prediction incorporating depression severity*

In this section, we propose an extension to the baseline model by incorporating the BDI-II depression index within the affect prediction system. The motivation of the system design is to perform a joint learning on the self-assessed depression severity and audio-visual cues to predict affective dimensions. Since the BDI-II index is a single value associated with every subject, the challenge lies in using the index in the frame-wise affective dimension prediction. We propose the inclusion of the subject-specific BDI-II score as a model parameter in predicting affective dimension for that subject. Using the same set of audio-visual cues in the baseline system, we incorporate the BDI-II score in the model as described below.

Affect prediction system: The proposed system transforms the audio-visual features based on the subject's BDI-II score values. Figure 2 shows a schematic of the proposed model and below we describe each component of the model.

a) Input audio/video features: The feature input scheme is same as in the baseline system. The system takes in a window of audio/video features and transforms them based on the BDI-II score as discussed next.

b) Feature transformation based on depression score: In this layer, we transform the features for a session based on the corresponding subject's BDI-II score. Although there are several feature transformations that one can apply, we test three transformations in this work which modify the features means and/or variances in a session based on the corresponding subject's depression severity. We discuss these transformations below.

1. Feature shifting: In this transformation, we shift the features values for a session by adding the BDI-II score for the corresponding subject. This transformation alters the feature means for a session based on the subject's depression severity. The shifting transformation $\mathcal{T}_1$ for audio features for the session s with the corresponding subject's BDI-II index, d_s can be represented as shown in equation (4) (the same transformation holds for video features). $\mathbf{1}$ represents a vector of ones of the same dimensionality as the input feature.

$$\begin{aligned}\mathcal{T}_1\big([\boldsymbol{x}_a^s(t-n), .., \boldsymbol{x}_a^s(t), .., \boldsymbol{x}_a^s(t+n)]\big) = \\ [\boldsymbol{x}_a^s(t-n), .., \boldsymbol{x}_a^s(t), .., \boldsymbol{x}_a^s(t+n)] + \big(d_s \times \mathbf{1}\big)\end{aligned} \quad (4)$$

2. Feature scaling: In this transformation, we scale the frame-wise feature values for each session by the correspond-

Table 2: Mean of the correlation coefficients, ρ (and ρ per affective dimension: valence: val., arousal: aro, dominance: dom.) between the ground truth and system prediction. Best performing system for each data is shown in bold. Best systems are significantly better than the baseline at 5% level using the Student's t-statistics test (number of samples = number of frames).

Dataset	System	Audio features mean (val./ aro./ dom.)	Video features mean (val./ aro./ dom.)	Fused output mean (val./ aro./ dom.)	Filtered output mean (val./ aro./ dom.)
Freeform	Baseline	.12 (.10/.25/.01)	.21 (.21/.19/.22)	.19 (.18/.22/.18)	.24 (.22/.28/.23)
	Proposed: $\mathcal{T}_1$	**.25 (.28/.35/.11)**	**.24 (.26/.23/.25)**	**.28 (.27/.33/.25)**	**.33 (.31/.37/.31)**
	Proposed: $\mathcal{T}_2$	.05 (.04/.20/-.09)	.13 (.16/.20/.05)	.12 (.11/.24/.02)	.16 (.15/.31/.03)
	Proposed: $\mathcal{T}_3$	.19 (.23/.30/.04)	.21 (.21/.27/.15)	.21 (.22/.30/.11)	.27 (.25/.35/.24)
Northwind	Baseline	.19 (.12/.26/.18)	.36 (.37/.37/.33)	.38 (.38/.41/.36)	.48 (.47/.50/.47)
	Proposed: $\mathcal{T}_1$	.36 (.32/.43/.33)	.43 (.41/.46/.43)	.45 (.42/.49/.45)	.50 (.46/.53/.52)
	Proposed: $\mathcal{T}_2$	.19 (.08/.32/.17)	.38 (.35/.51/.29)	.39 (.36/.52/.30)	.45 (.41/.59/.36)
	Proposed: $\mathcal{T}_3$	**.37 (.35/.45/.30)**	**.46 (.38/.55/.45)**	**.48 (.41/.57/.46)**	**.52 (.45/.61/.51)**

ing subject's BDI-II score. This transformation alters the feature variances for a session based on the subject's depression severity. The scaling transformation $\mathcal{T}_2$ is represented in equation (5). $*$ represents element-wise multiplication and d_s is the BDI-II index for the subject in session s.

$$\mathcal{T}_2\big([\boldsymbol{x}_a^s(t-n),..,\boldsymbol{x}_a^s(t),..,\boldsymbol{x}_a^s(t+n)]\big) = [\boldsymbol{x}_a^s(t-n),..,\boldsymbol{x}_a^s(t),..,\boldsymbol{x}_a^s(t+n)] * \big(d_s \times \mathbf{1}\big) \tag{5}$$

3. Feature scaling and shifting This transformation both scales and shifts the feature values, thereby affecting both means and variances for the features. This transformation T_3 is shown below.

$$\mathcal{T}_3\big([\boldsymbol{x}_a^s(t-n),..,\boldsymbol{x}_a^s(t),..,\boldsymbol{x}_a^s(t+n)]\big) = [\boldsymbol{x}_a^s(t-n),..,\boldsymbol{x}_a^s(t),..,\boldsymbol{x}_a^s(t+n)] * \big(d_s \times \mathbf{1}\big) + \big(d_s \times \mathbf{1}\big) \tag{6}$$

c) *Audio/video outputs:* Following the feature transformation, we obtain the audio/video outputs using similar linear models as in the baseline model (equations (1), (2)). However instead of the explicit audio/video features, we use one of the feature transformations.

d) *Modality fusion:* The modality fusion strategy is again same as the baseline system. We perform training using data bootstrapping as discussed in the section 3.2.1(c).

e) *Final prediction after filtering:* The low pass filtering step is also same as the baseline system to avoid abrupt changes in tracking the affective dimensions. In the next section, we discuss the evaluation scheme and present our results.

3.2.3. Evaluation

We perform independent evaluations on the Freeform and Northwind datasets. For the 100 sessions in each dataset, we use a 10 fold cross-validation, with 8 partitions as the training dataset, and 1 each as development testing sets. For all our experiments, the features in the training set are normalized to be of zero mean and unit variance. During testing, we normalize the testing set features using feature means and variances computed on the training set. The BDI-II scores are also normalized to a range of 0-1 and during feature transformation they scale and shift the feature values accordingly. We use mean correlation coefficient ρ over the three affective dimensions computed over all the sessions as the evaluation metric, as was also used in the AVEC challenge 2014 [21]. We tune the feature window length n and length of the moving average filter length k on the development set. Table 2 shows the results for each of the dataset using the baseline and various feature transformations in the proposed system.

3.2.4. Discussion

In our first experiment in section 3.1, we observed that depression severity is correlated with affect and therefore provides a complementary source of information in affect prediction. The results in Table 2 vary in the two datasets with better prediction in the Northwind dataset. This may be due to the difference in structural formats of the two datasets, with each dataset calling for a different cognitive planning mechanism [33, 34]. The Freeform dataset incorporates the exercise of lexical planning to form an answer where as Northwind dataset involves sensory input and sentence reproduction. We also notice that the best feature transformation scheme varies for the two datasets. Scaling the features alone (transformation $\mathcal{T}_2$) does not perform well, implying changing feature variance based on the depression severity does not help, particularly in the case of Freeform dataset. This can be attributed to the lack of correlation between depression severity and variances of affective dimensions, as seen in Table 1. Since changing feature variances has a direct impact on output affective dimension prediction due to affine projections during prediction (Figure 2), scaling feature values based on depression serves as a noisy operation. However, we observe that changing the feature means via the shifting transformation ($\mathcal{T}_1$) helps in both the cases. For the Northwind dataset, the best results are obtained after applying the shifting and scaling transformation ($\mathcal{T}_3$). Apart from these observations, we also notice that modality fusion performs better suggesting that audio and video modalities carry complementary information. Also, the low pass filtering improves performance by removing high frequency components from the affective dimension prediction.

4. Conclusion

Researchers have investigated the impact of depressive disorders on emotion and discovered several patterns [12, 13]. In this work, we develop an affect prediction model with the subject's depression severity incorporated as a model parameter. We use two datasets for this purpose and initially test the relationship between depression severity and emotions using correlation analysis. We then develop an audio-visual feature based baseline model to predict affect. We modify the model to use the BDI-II depression index to perform session-wise feature transformations which shift or/and scale the feature values for a session based on the subject's depression severity. We test our model on two datasets and observe that the best performing transformation variation varies between them.

In the future, we will investigate other models for incorporating depression severity in predicting affective dimensions. As of now, the depression severity only affects the first layer of the prediction system. This BDI-II index could also be incorporated as a parameter in other layers of the model and the optimization problem could be framed accordingly. We also aim to apply the model to other problem domains involving time series prediction with an accompanied static label, e.g., tracking engagement based on autism severity [35]. Finally, the model could also be extended to datasets with ratings available at multiple temporal granularities.

5. References

[1] S. Sandra, "Depression: Questions you have-answers you need," *Peoples Medical Society*, 1997.

[2] American Psychiatric Association, "Diagnostic and statistical manual of mental disorders," 1980.

[3] National Institute of Mental Health, "Depression," http://www.nimh.nih.gov/health/publications/depression-what-you-need-to-know-12-2015/depression-what-you-need-to-know-pdf_151827.pdf.

[4] L. S. Greenberg and J. C. Watson, *Emotion-focused therapy for depression.* American Psychological Association, 2006.

[5] I. Myin-Germeys, N. Jacobs, F. Peeters, G. Kenis, C. Derom, R. Vlietinck, P. Delespaul, J. Van Os *et al.*, "Evidence that moment-to-moment variation in positive emotions buffer genetic risk for depression: a momentary assessment twin study," *Acta Psychiatrica Scandinavica*, vol. 115, no. 6, pp. 451–457, 2007.

[6] A. L. Bouhuys, E. Geerts, P. P. A. Mersch, and J. A. Jenner, "Nonverbal interpersonal sensitivity and persistence of depression: perception of emotions in schematic faces," *Psychiatry research*, vol. 64, no. 3, pp. 193–203, 1996.

[7] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012.

[8] A. T. Beck, R. A. Steer, G. K. Brown *et al.*, "Manual for the beck depression inventory-ii," 1996.

[9] M. Brandt and J. D. Boucher, "Concepts of depression in emotion lexicons of eight cultures," *International Journal of Intercultural Relations*, vol. 10, no. 3, 1986.

[10] W. Heller, N. S. Koven, and G. A. Miller, "Regional brain activity in anxiety and depression, cognition/emotion interaction, and emotion regulation." 2003.

[11] J. J. Gross and R. F. Muñoz, "Emotion regulation and mental health," *Clinical psychology: Science and practice*, vol. 2, no. 2, pp. 151–164, 1995.

[12] C. E. Izard, *Patterns of emotions: A new analysis of anxiety and depression.* Academic Press, 2013.

[13] S. H. Blumberg and C. E. Izard, "Discriminating patterns of emotions in 10-and 11-yr-old children's anxiety and depression." *Journal of personality and social psychology*, vol. 51, no. 4, p. 852, 1986.

[14] 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.* IEEE, 2014.

[15] N. Cummins, J. Epps, V. Sethu, and J. Krajewski, "Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on.* IEEE.

[16] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression," in *Annual Conference of the International Speech Communication Association*, 2014.

[17] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.* IEEE, 2011.

[18] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, 2014.

[19] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[20] 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.* ACM, 2013, pp. 3–10.

[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.* ACM, 2014, pp. 3–10.

[22] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.* ACM, 2014, pp. 19–26.

[23] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.* ACM, 2014, pp. 33–40.

[24] M. Kächele, M. Schels, and F. Schwenker, "Inferring depression and affect from application dependent meta knowledge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.* ACM, 2014.

[25] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "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, 2014, pp. 65–72.

[26] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.* ACM, 2014, pp. 73–80.

[27] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge.* ACM, 2014.

[28] R. J. Cabin and R. J. Mitchell, "To bonferroni or not to bonferroni: when and how are the questions," *Bulletin of the Ecological Society of America*, pp. 246–248, 2000.

[29] R. Rojas, *Neural networks: a systematic introduction.* Springer Science & Business Media, 2013.

[30] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application.* Cambridge university press, 1997.

[31] L. L. Scharf, *Statistical signal processing.* Addison-Wesley Reading, MA, 1991, vol. 98.

[32] R. Gupta, N. Kumar, and S. Narayanan, "Affect prediction in music using boosted ensemble of filters," in *The 2015 European Signal Processing Conference, Nice*, 2015.

[33] J. O. Greene and J. N. Cappella, "Cognition and talk: The relationship of semantic units to temporal patterns of fluency in spontaneous speech," *Language and Speech*, vol. 29, no. 2, pp. 141–157, 1986.

[34] G. Beattie and A. Ellis, *The psychology of language and communication.* Psychology Press, 2014.

[35] R. Gupta, D. Bone, S. Lee, and S. Narayanan, "Analysis of engagement behavior in children during dyadic interactions using prosodic cues," *Computer Speech & Language*, vol. 37, pp. 47–66, 2016.