Multimodal Prediction of Affective Dimensions and Depression in Human-Computer Interactions

Rahul Gupta
Signal Analysis and Interpretation Laboratory, Univ. of Southern California

Nikolaos Malandrakis
Signal Analysis and Interpretation Laboratory, Univ. of Southern California

Bo Xiao
Signal Analysis and Interpretation Laboratory, Univ. of Southern California

Tanaya Guha
Signal Analysis and Interpretation Laboratory, Univ. of Southern California

Maarten Van Segbroeck
Signal Analysis and Interpretation Laboratory, Univ. of Southern California

Matthew Black*
Information Sciences Institute, Univ. of Southern California, Marina del Rey, CA, USA

Alexandros Potamianos*
School of Electrical and Computer Engineering, NTUA, Athens, Greece

Shrikanth Narayanan
Signal Analysis and Interpretation Laboratory, Univ. of Southern California

ABSTRACT

Depression is one of the most common mood disorders. Technology has the potential to assist in screening and treating people with depression by robustly modeling and tracking the complex behavioral cues associated with the disorder (e.g., speech, language, facial expressions, head movement, body language). Similarly, robust affect recognition is another challenge which stands to benefit from modeling such cues. The Audio/Visual Emotion Challenge (AVEC) aims toward understanding the two phenomena and modeling their correlation with observable cues across several modalities. In this paper, we use multimodal signal processing methodologies to address the two problems using data from human-computer interactions. We develop separate systems for predicting depression levels and affective dimensions, experimenting with several methods for combining the multimodal information. The proposed depression prediction system uses a feature selection approach based on audio, visual, and linguistic cues to predict depression scores for each session. Similarly, we use multiple systems trained on audio and visual cues to predict the affective dimensions in continuous-time. Our affect recognition system accounts for context during the frame-wise inference and performs a linear fusion of outcomes from the audio-visual systems. For both problems, our proposed systems outperform the video-feature based baseline systems. As part of this work, we analyze the role played by each modality in predicting the target variable and provide analytical insights.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*signal processing, computer vision, text processing*; J.3 [**Computer Applications**]: Life and Medical Sciences—*health*

Keywords

Behavioral Signal Processing (BSP); Multimodal signal processing; Depression; Arousal; Valence; Dominance; Fusion

*The author is also affiliated with Behavioral Informatix.

AVEC'14, November 7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2661806.2661810 .

1. INTRODUCTION

Interdisciplinary research efforts in computational paralinguistics have increased dramatically in the past decade, leading to the emergence of fields such as social signal processing [41] and behavioral signal processing (BSP) [30]. One of the central foci of BSP is on addressing societally-significant health-related problems by applying engineering techniques and enriching them through collaborations with domain experts (e.g., psychologists, doctors, clinical providers). State-of-the-art multimodal signal processing techniques are first used to extract relevant cues ("features") from human behavioral signals (e.g., speech, language, gestures, physiology). Machine learning techniques are then used to map these features to relevant higher-level descriptions, which can be used by human experts for informing their decision making. BSP methodologies have been applied to various medical/clinical domains, including marital conflict and couples therapy [7, 22], addiction counseling [47, 17], autism spectrum disorders [8, 9], and metabolic health monitoring in obesity [4]. The 2014 Audio/Visual Emotion Challenge (AVEC) provides a platform to explore the relevant application domains of depression and affect recognition. These studies can inform mental health researchers how depression and affect are associated with several mood disorders and mental well-being. We use BSP techniques to investigate these problems, contributing both toward understanding the two phenomena as well as advancing the methodologies.

Major depressive disorder, also known as clinical depression, is a mood disorder that is characterized by persistent feelings of sadness, low self-esteem, and loss of interest [3]. It is prevalent in both men and women and across all ages and ethnicities worldwide [12]. In the United States alone, it is estimated that 7 percent of the population suffers from depression in a given year and 17 percent at one point in their lives. In addition to taking a toll on the individual and family, depression also causes an enormous economic burden,

costing tens of billions of dollars each year in the United States [31, 5]. As with many mental health disorders, early diagnosis and appropriate treatment (e.g., medication, psychotherapy) can help alleviate symptoms and allow people with depression to live healthier and happier lives [18].

There is a tremendous opportunity for human-centered engineering to aid in the diagnosis, intervention, and treatment of depression. It is well established that there are differences in speech production in people with depression, validated through measurable changes in pitch, loudness, speaking rate, and articulation after treatment [29, 13]. Language use has also been shown to differ for those diagnosed with depression, with one study showing that people with depression said "I" more frequently than those unaffected [45]. There are also several nonverbal cues that have been shown to be indicative of depression severity (e.g., an increase in withdrawing gestures and fewer smiles [44, 15]).

Significant related work has been done on the automatic analysis and prediction of depression using speech and prosody [28, 21, 23, 46], gestures, head pose, and facial expressions [2, 33, 43, 19], and a multimodal combination of these cues [49, 20, 27, 26]. As described in the literature, there are a variety of challenges in automatically determining whether a person is depressed, including: 1) each behavioral signal provides only partial information, which must be combined to form a more realistic model for recognizing behaviors indicative of depression; 2) some information that is relevant may not be available or may be inherently hidden; 3) defining baseline behavior can be difficult with limited behavioral data.

In this paper, we attempt to overcome these challenges by applying BSP methodologies to robustly predict self-assessed depression severity, fusing multiple sources of information across modalities. The second objective of this paper is to recognize affect in continuous-time from the same multimodal data using similar modeling techniques. Since depression is a mood disorder, these two objectives are related; moment-to-moment changes in people's emotions may shed light onto their underlying state of depression, and affective state may be an important cue to track when screening for mental health disorders such as depression.

Arguably the most investigated computational paralinguistics topic, emotion recognition consists of predicting the affective nature of a person (or group of people) using only the raw behavioral signals (e.g., audio, video, text, physiological signals) and a representative labeled training corpus. Technical challenges include accounting for: 1) the inherent subjectivity in the perception of human emotion; 2) intra-person and inter-person variability and individual idiosyncrasies; 3) the context in which the human behavior or interaction took place. Furthermore, continuously tracking emotion (as opposed to classifying emotion of a pre-determined temporal period) has additional challenges but is particularly relevant with multimodal data, where important events may occur asynchronously across different modalities.

Inspired by previous work, we implement BSP methods to predict depression and recognize the three most common dimensional attributes of emotion (valence, arousal, dominance) in continuous-time. For the depression challenge, we derive visual, audio, and text based cues and experiment with several feature selection strategies to obtain the optimal set of features predictive of depression. For the emotion challenge, our methods are based on fusing frame-wise predictions over multiple modalities. We hypothesize that different cues carry complementary information with respect to the target affective dimension and perform a weighted combination to obtain our final predictions. Moreover affect evolves steadily over time and sudden changes are unlikely over a time window containing only a few frames. We leverage this contextual information during prediction using similar techniques as proposed in [16], to process the outputs obtained from various modalities.

Section 2 describes the depression dataset (used for all experiments), and Section 3 discusses the audio, video, and text features we extracted. Section 4 describes our proposed methodologies for the emotion and depression challenges. We discuss results in Section 5 and conclude in Section 6.

2. DEPRESSION DATASET

We use the 2014 Audio-Visual Emotion Challenge (AVEC) dataset for our evaluation [38]. This dataset is a subset of the audio-visual depressive language corpus (AViD-Corpus) and is composed of 300 webcam video recordings of human-computer interaction tasks. The total number of subjects is 84, ranging from 18 to 63 years. In each recording, the participant completes either the task of reading aloud a paragraph (Northwind) or responding to a number of questions (Freeform), both in German. The duration of the recordings range from 6 to 248 seconds. The challenge organizers equally partitioned the 150 Northwind-Freeform pairs into training, development, and testing sets, maintaining balance across the subjects' age, gender, and depression levels.

Each recording was labeled in terms of affective dimensions (valence, arousal, and dominance) and level of depression. The three affective dimensions were annotated continuously by a team of five naive raters, to obtain a value per video frame (30 frames/second). The depression level was labeled as a single value per recording as derived from self-report analysis using the Beck Depression Inventory (BDI)-II [6]. The AVEC 2014 challenge consists of predicting these affective dimensions (Affect recognition sub-challenge) and depression levels (Depression recognition sub-challenge) as two separate sub-challenges. For more details on the challenge data set and labels, please refer to [38].

3. FEATURES

We use an assembly of audio, video, and text based features for the two sub-challenges, described in detail next.

3.1 Audio

Baseline features: The audio baseline features are adopted from the AVEC 2013 challenge, extracted using the openSMILE toolkit [11]. Feature vectors consist of various acoustic low-level descriptors (LLDs), such as relative spectral (RASTA) MFCCs, spectral energies, and voicing/unvoiced related features. For the affect recognition sub-challenge (ASC), we use the frame-wise LLDs, augmented by a few window-wise functionals (total count: 79), computed at 100 frames/second. For the depression recognition sub-challenge (DSC), the LLDs are augmented by utterance-level static functionals to obtain a total of 2268 baseline features per recording [38].

Additional features We construct an additional acoustic feature representation by combining the speech streams as proposed in [40]. Each of these streams models a different cue in the auditory spectrum that is related to human speech production: (i) spectral shape, (ii) spectro-temporal modulations, (iii) periodicity structure due to the presence of pitch harmonics, and (iv) the long-term spectral variability profile. For each audio frame, these streams are stacked in

one feature vector and subsequently decorrelated and dimensionality reduced by applying Principal Component Analysis (PCA). The principal components are computed on all training data, and only the components that correspond to the 88 largest singular values are retained, accounting for 90% of the variance. All deployed feature representations are subsequently mean variance normalized on a per utterance basis. Application of these features in [39] complemented the baseline features proposed in the INTERSPEECH 2014 challenge [36].

3.2 Video

Baseline features: The AVEC2014 challenge provides a set of local binary patterns (LBP) features well known for describing facial expressions [32]. An LBP descriptor, centered at a pixel, is a binary vector computed by comparing the pixel's intensity with those of its neighbors. After the descriptors are computed for each pixel, a feature is computed as a histogram with each bin corresponding to a different binary pattern. Given a video $V(x, y, t)$, the baseline LBP video features are computed in the Gabor domain along three orthogonal planes: xy, yt, and xt (LGBPTOP). These features are computed dynamically by considering a short video sequence around each frame to capture temporal changes in facial appearance.

Additional LBP features: Videos are first subjected to face detection and tracking to extract human faces at every frame using standard implementations in OpenCV. An additional set of LBP features are computed in the pixel domain. For each video, we compute two types of LBP features: 1) static features per frame, and 2) LBP features computed along three orthogonal planes (LBPTOP). The static features are the local binary patterns [32] extracted from the facial region in each frame, without using any temporal information. These features (a 256 dimensional vector per frame) are computed to complement the dynamic baseline features, which capture short-term temporal variation but may be affected by misaligned or missing faces. The LBP-TOP feature [48] consists of a single 768-dimension feature vector for each video, calculated every other frame. This is obtained by computing and concatenating LBP features along the spatial (xy) and two temporal planes (xt and yt) to capture an overall signature of a subject's facial expressions.

Motion features: We hypothesize that overall facial and head motion of the subjects carry important information about their affective state and depression level. To capture motion information, optical-flow-based motion vectors are computed between a pair of consecutive frames at keypoints detected using a corner detection algorithm [37]; please see Fig. 1. Total motion per frame is computed by adding the amount of motion each keypoint undergoes, thus generating a scalar value per frame.

Features derived from facial landmarks: In addition, we employ the CSIRO Face analysis SDK [10] to extract facial landmarks from the video data. We fit 66 landmark points to the face for each frame, using mean-shift based deformable model fitting [34], as shown in Fig. 1. Based on the results, we compute the following frame-wise features: 1) general head motion of the landmark point (marked red between the eyes), normalized by face size; 2) averaged two eyes open-close state, computed as the mean distance of upper and lower landmarks of the eyelids, normalized by eye width; 3) mouth animation state, calculated as the mean-squared distance of all mouth landmarks to the mouth centroid, normalized by mouth width; 4) animation of all facial

Figure 1: Examples of overall motion tracking (left) and facial landmark fitting (right).

landmarks with respect to the red landmark points, normalized by face size. The average frame-wise success rate of facial landmark fitting per session is 93%. We pad zeros for frames with missing feature values. Note that all video features are computed at a lower frame rate of 30 fps, as compared to the audio feature extraction rate of 100 fps.

3.3 Text

Language use is an important window into the mind and therefore potentially an indicator of mental and affective state. To make use of language features, we need transcriptions, which are not provided in the challenge data. To acquire transcriptions, we posted the FreeForm data on Amazon Mechanical Turk (Language analysis is trivial for the Northwind data). Our goal was three transcriptions per sample, in order to disambiguate in the case of disagreements, but we were only able to obtain two annotations for most of the 150 samples. The mean word "error" rate between transcriptions is 30.6%, although this figure is artificially inflated by the inconsistent use of umlauts and special characters as in the double "s" case. Lacking enough information to disambiguate in the case of disagreement, we randomly picked one transcription for each sample. The language features we extract are inspired by sentiment analysis research, where the word sequence is converted to a signal (sequence of numbers) by looking up emotional ratings in an affective lexicon and then functionals are computed across the signal. We next describe the generation of affective lexicon, followed by the extraction of session-level features.

Generating the lexicon: We use the affective lexicon generated by an automated algorithm of lexicon expansion as described in [24]. We assume that the continuous valence and arousal ratings ($\in [-1, 1]$) of any term t_j can be represented as a linear combination of its semantic similarities d_{ij} to a set of seed terms w_i, as shown in (1). $a_i(w_i)$ represents the weight corresponding to the seed term w_i.

$$\hat{v}(t_j) = a_0 + \sum_{w_i \in \text{seed terms}} a_i(w_i)\, d_{ij} \qquad (1)$$

For the purposes of this work, d_{ij} is the cosine similarity between context vectors computed over a corpus of 170 million sentences, created by collecting web snippets (up to 500 for each word in the German Aspell [1] spellchecker) using the Yahoo! search engine.

The starting point for the lexicon creation was the manually annotated lexicon *Berlin Affective Word List Reloaded* (BAWL-R) [42] that contains continuous valence and arousal ratings for 2902 German words. These words were used to form the dimensions of a Distributional Semantics Model (DSM), a space where each dimension corresponds to the semantic similarity to a word or concept. We then applied

PCA to create a new DSM of concepts based on the original space, using the first 600 principal components. The d_{ij} terms in (1) are calculated on this component space. Using the entirety of BAWL-R as training samples and the 600 concepts as seeds, we created a system of linear equations that, when solved using Least Squares Estimation, gives us the weights a_i. Thus we obtain a model that can be used to generate valence and arousal ratings for any ngram.

For each new term, we created arousal and valence ratings by calculating their semantic similarities with the 2902 words in BAWL-R, transforming them to the component DSM space, and then applying equation (1).

Extracting session-level features: Every session in the FreeForm data was part-of-speech tagged using the German version of Treetagger [35]. We collected all token unigrams and bigrams and created ratings using the lexicon expansion algorithm. Before replacing the terms with their ratings, we applied multiple selection criteria to select a fraction of the terms: ngram level (unigram or bigram), and in the case of unigrams, further filtering based on the part-of-speech tag. The outcome is a multiple filtered version of each transcript, which are then converted to signals by replacing the terms with their ratings.

We then generated features by computing functionals across the signals: length (cardinality), min, max, max amplitude, sum, average, range, standard deviation, and variance. We also created normalized versions by dividing by the same statistics calculated over all tokens, e.g., the maximum of adjectives over the maximum of all unigrams. This results in features like, "maximum of valence over unigram proper nouns" and "range of arousal over all bigrams."

4. EXPERIMENTAL METHODS

4.1 Depression Recognition Sub-Challenge

Our proposed DSC method utilizes multiple modalities, with models trained on functionals of the frame-level LLDs. Our goal is to combine information across modalities and data types (Northwind and Freeform experiments) to predict subjects' depression ratings.

4.1.1 Modeling

One could envision a hierarchical model, where modalities and experiments are represented by stand-alone systems. However, the limited sample size of the data would make training such a model difficult. Instead, we used simple models and feature-level fusion to keep the complexity in check: all features are combined into a single vector and used for classification. Our DSC model is a Support Vector Regressor (SVR), with a second-degree normalized polynomial kernel, a setup that performed particularly well when confronted with the vastly different features produced for the different modalities.

4.1.2 Features and Selection

We use the session-level baseline audio features, session-level means of baseline video features, session-level statistics (deltas and functionals similar to audio baseline features) over additional video features, and the text-based features. All of these features, apart from the text features, are extracted independently from the Freeform and Northwind samples, creating a pool of 42092 candidate features. As the feature dimensionality is extremely high, given the sample count, we resort to feature selection methods for this sub-challenge to reduce feature dimensionality.

Table 1: DSC feature groups and the number of candidate features. Sets in bold are selected at the next stage using brute force strategy.

Features	Northwind	Freeform
Audio	865	865
Audio deltas	427	**427**
Add. LBP	**768**	768
MFCC	672	672
MFCC deltas	304	**304**
Motion	**100**	**100**
Text	0	**1836**
Video baseline	**16992**	16992

Table 2: List of features in the 4 ASC systems.

System name	Frame-wise features	Dimensionality
Video system 1	Baseline video features	16992
Video system 2	Additional video features	256+1+68
Audio system 1	Baseline audio features	79
Audio system 2	Additional audio features	88

We evaluate multiple approaches to feature selection. The first is a two-stage forward selection approach implemented by: 1) splitting the set of features into smaller groups based on modality, data (Northwind/Freeform), and the delta functionality; 2) applying supervised feature selection (best-first, multiple correlation criterion) to each feature group; 3) performing brute-force selection of a few groups and combining the member features selected in step 2; and 4) performing a second stage feature selection based on the development set performance to prevent overfitting. The second approach is a backward selection scheme, similar to the previous approach, but using single correlation as the selection criterion. Our goal is to select a reduced set of the most predictive features across modalities, thereby addressing data sparsity.

Table 1 lists the feature groups. Northwind and Freeform derived features are considered separately, along with features from different modalities. The audio baseline features are split into MFCC, MFCC deltas, audio (all other features), and audio deltas. The video baseline features are kept as one group. The rest of the groups correspond to additional LBPTOP, overall motion, and text features.

4.2 Affect Recognition Sub-Challenge

Unlike DSC, we train separate models for each modality in ASC, using the audio and video features to predict the continuous valence, arousal, and dominance ratings. Despite the availability of textual features closely related to affect, we were unable to use them due to the lack of time-aligned transcripts. We design 4 different systems (2 on the video features and 2 on the audio features) and perform a fusion of their outputs. Table 2 shows a list of the systems and the constituent features. Baseline and additional feature systems are configured in order to observe the additive advantage of using the new set of features across both modalities. In the next section, we describe the feature preparation, followed by the description of our affect prediction system.

4.2.1 Processing system features

By design, the features from video system are synchronized with the frame-level annotations. However, the audio features are computed at a higher frame rate (100 fps) and need to be synchronized with the ratings and the video frames (30 fps). We derive local means of audio features computed over multiple frames and downsample them to be synchronized with the video features.

4.2.2 *Predicting Valence, Arousal, and Dominance*

We chose a multi-layered system for valence, arousal, and dominance (VAD) prediction, comprised of four stages: 1) individual system prediction, 2) processing system outputs, 3) system fusion, and 4) temporal regression. A graphical representation of the prediction system is shown in Fig. 2. We describe each stage in greater detail next.

(i) **Individual system training**: We use all features from each system and train a linear regression model to predict the VAD outcomes. We train separate systems for each dataset (Freeform and Northwind), since they are collected under different protocols. Therefore, we obtain a total of six models (3 target outcomes $\times$ 2 datasets) for each system. Our regressor performs an independent frame-wise prediction that does not take context from neighboring frames into account. We address this issue in the next step.

(ii) **Processing system outputs**: Independent frame-wise prediction proposed in step (i) is counterintuitive, since affective states typically evolve smoothly over time. We exploit this correspondence by linearly combining the VAD outcomes over a temporal window. We chose a moving average (MA) filter and tune the filter length (W_P) on the development set. This low pass operation obtains prediction for frame k based on the unweighted combination of predictions over a window of length W_P centered at k. This operation has the benefit of removing any high frequency noise in the individual system prediction introduced during features extraction, downsampling, and/or prediction.

(iii) **System fusion**: We linearly fuse the processed outputs from the four systems produced in step (ii). The weights for system fusion are again determined by linear regression, thus minimizing the mean squared error between the fused and target outcome values. We perform linear regression on the development set to prevent overfitting to the training set.

(iv) **Temporal regression**: As the final step, we process the fused outputs to obtain the final prediction for a frame using prediction values over a window. For prediction on the frame k, we perform a linear combination on the fused outputs, obtained in step (iii), over a window of length W_T centered at k. The weights for combination are determined using linear regression. In order to tune W_T and obtain the regression coefficients, we evenly split the development set, tuning W_T and training the regressor on the first half and obtaining predictions on the second. This system accounts for any contextual dependencies introduced after fusion.

5. RESULTS AND DISCUSSION

5.1 Depression Recognition Sub-challenge

Feature selection is performed per feature group listed in Table 1. We then select a subset of feature groups based on brute-force strategy. The overall theme of the feature selection experiments was that the better features from each modality came from different experiments, with audio and text from Freeform and video features from Northwind combining into the best set (motion features extracted from Freeform being the exception). The limited utility of audio features extracted from the Northwind experiment is somewhat perplexing, given that it shares many characteristics with typical speech experiments.

Fig. 3 shows the prediction performance with addition of each new feature as obtained by best-first forward search

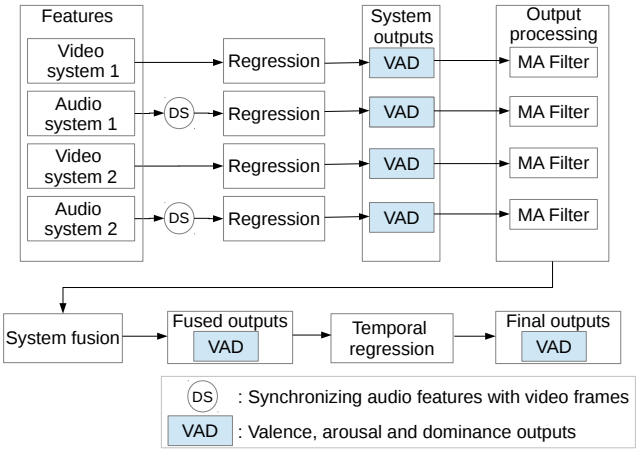

Figure 2: Block diagram of the valence, arousal, and dominance (VAD) prediction system.

Table 3: DSC RMSE performance of the 3 submitted systems on the development and testing sets.

System	RMSE	
	Development	Testing
Baseline system	**9.26**	**10.86**
Forward - all	11.42	10.35
Forward - groups	**7.44**	10.56
Forward - groups - partial	8.51	10.33
Backward - all	9.68	**8.99**
Backward - groups	**7.44**	9.85

strategy on the development set. Performance achieved using features from a single modality is shown in Fig. 3(a), with audio and video features both reaching about a 9 RMSE and text features performing far worse. It should be noted that the SVR model used was not the best tested model for each individual modality, only for the combination thereof, so the differences shown may be exaggerated. Performance achieved using two modalities (i.e., excluding one modality) is shown in Fig. 3(b). Excluding text or audio features results in an RMSE of around 8, while removing video has the worst effect. Still, these plots shows that all modalities contribute. Finally, performance achieved when using only one of the two available experiments (Northwind or Freeform) is shown in Fig. 3(c) and is a nice verification of the importance of both datasets during inference.

Overall five systems using different selection methods were created and submitted for the final challenge, as listed in Table 3. The first system uses a simple one stage forward selection on all features and results in overfitting. The second and third system use the two stage forward selection scheme, but system three skips selecting again after the feature groups have been merged. System four uses a minimal backward selection, where only useless (zero variance) features are rejected. Finally system five uses backward selection on the selected groups. The two systems utilizing only the selected groups were best throughout our development experiments, though that is clearly not the case on the test set. The minimum RMSE achieved on both sets is notably better than the baseline.

5.2 Affect recognition sub-challenge

We list the correlation coefficients (ρ) with the target valence, arousal, and dominance labels on the development

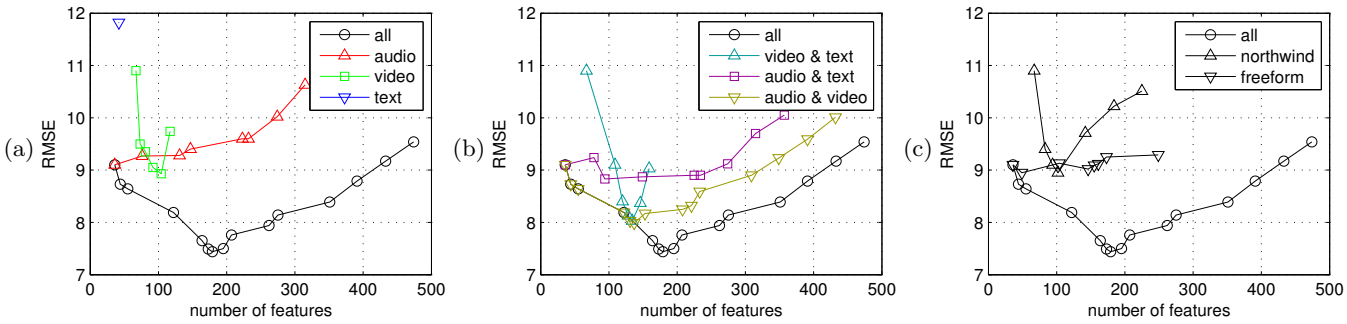

Figure 3: Depression prediction RMSE performance on the development set vs. the number of features used for: (a) single modalities, (b) pairs of modalities, and (c) all modalities only for Northwind or Freeform data.

Table 4: System-wise correlation coefficient ρ in predicting valence, arousal, and dominance outcomes on the Freeform and the Northwind development datasets; please see Fig. 3 for system block diagram.

Freeform										
Affective	*Video system 1*		*Video system 2*		*Audio system 1*		*Audio system 2*		Fusion	Temporal
Dimension	Raw	Proc.(W_P)	Raw	Proc.(W_P)	Raw	Proc.(W_P)	Raw	Proc.(W_P)		Regression(W_T)
Valence	.377	.420(250)	.260	.327(1220)	.108	.199(170)	.024	.127(290)	.455	**.460**(60)
Arousal	.366	.416(430)	.064	.091(290)	.264	.551(600)	.088	.320(290)	**.612**	**.612**(0)
Dominance	.225	.275(1200)	.254	.329(310)	.060	.148(690)	.006	.071(260)	.368	**.369**(10)
Northwind										
Affective	*Video system 1*		*Video system 2*		*Audio system 1*		*Audio system 2*		Fusion	Temporal
Dimension	Raw	Proc.(W_P)	Raw	Proc.(W_P)	Raw	Proc.(W_P)	Raw	Proc.(W_P)		Regression(W_T)
Valence	.285	.316(730)	.246	.293(110)	.163	.286(240)	.035	.104(90)	.408	**.428**(10)
Arousal	.413	.450(140)	.428	.508(320)	.273	.534(260)	.035	.146(130)	**.695**	**.695**(0)
Dominance	.332	.375(4200)	.287	.390(2040)	.180	.321(260)	.019	.062(80)	.515	**.518**(10)

Table 5: Combined correlation coefficient (ρ) on the development and testing sets in ASC.

System	Affective Dimension	ρ	
		Development	Testing
Baseline	Valence	.355	.188
	Arousal	.412	.206
	Dominance	.319	.196
	Average	**.362**	**.196**
Proposed	Valence	.453	.493
	Arousal	.670	.620
	Dominance	.397	.453
	Average	**.506**	**.522**

Table 6: Histogram feature counts with $|\rho|$ in the range 0.1-0.2/0.2-0.3/0.3-0.4 for the 4 ASC systems.

Dim.	Video 1	Video 2	Audio 1	Audio 2
N	16992	325	79	88
Freeform				
Val.	3259/148/0	43/7/6	3/0/0	0/0/0
Aro.	1812/22/0	76/16/0	8/0/0	0/0/0
Dom.	8380/2547/49	62/20/1	12/0/0	0/0/0
Northwind				
Val.	3906/417/14	81/11/0	4/0/0	0/0/0
Aro.	3569/323/1	105/13/1	4/0/0	0/0/0
Dom.	6164/1648/103	82/18/0	7/0/0	0/0/0

set of Freeform and Northwind datasets separately in Table 4. ρ obtained on the development and testing sets after combining prediction from both the datasets are shown in Table 5. We do not have stand alone system performance on the testing partition due to unavailability of test labels.

From the results on the development set (Table 4), we observe that the performance varies across the systems. Processing the raw outputs from individual systems provides substantial gains, particularly in the case of audio features. This increase in ρ suggests that even though the constituent features may not be highly correlated with target affective dimensions, using context in prediction does add information, particularly for the audio features. Also, the low pass filter operation removes the noisy variations in the audio system outcomes introduced during feature downsampling. The values of W_P and W_T vary widely even for same target variable, across the two datasets. We speculate this to be an artifact of the data and more robust models may be designed by defining certain constraints on W_P and W_T (e.g. a prior). Combined results (Table 5) over valence and arousal settle close to the mean of performances on individual datasets. However, the dominance value does not follow the same pattern. This suggests that the dynamic range of dominance over the two datasets is different and thus a data specific evaluation is recommended.

We analyze the performance of each system based on the correlation of member features to the target affective dimensions. Features with high values of absolute correlation are desirable as we train linear systems. We list the histogram count of features falling under various absolute correlation coefficient ($|\rho|$) ranges in Table 6. We observe that the ρ values obtained for video systems is fairly good due to the presence of several features which are highly correlated with the target variables. There are no features with $|\rho|$ higher than 0.2 in audio system 1 and 0.1 in audio system 2. Therefore, the raw performance of these systems is poor. However, we observe poor performance for a few systems even in the presence of highly correlated features (e.g., Freeform arousal video system 2, Freeform dominance video system 1). We speculate that this situation arises due to: 1) mismatch between the training and development set and 2) multicollinearity in features leading to poor regression coefficient estimation.

Finally, we observe minor gains using temporal regression, implying context in fused outcome provides little information. Overall, our proposed system generalizes well, as exemplified from the final combined results on the testing set in Table 5.

5.3 General Discussion

In our experiments for the two sub-challenges, we observe that even though we surpass the video features-based baseline for both sub-challenges by using multiple modalities, the performance gain varies. In ASC, the addition of more features and modalities provides significant gains. However, for DSC, we suffer from data sparsity, and learning becomes more challenging with the addition of more features. In this work, we focus on developing separate systems for affect and depression prediction. This may not be optimal, as depression state prediction is correlated with affect recognition [25, 14]. Moreover, we use similar modalities and feature derivatives to predict the outcomes for both the target variables. This provides further encouragement to investigate similarities and inter-relationships between the LLDS, depression level, and affective state. Another challenge lies in mapping the continuous affective dimensions to a single global label of depression over the entire interaction. One suggested approach may involve the use of a few intermediate variables in coupling the two outcomes. The inherent difference between the recording conditions of Northwind and Freeform of dataset is a further point of investigation, as the latter is performed under a higher cognitive load. Such factors may impact the outcomes of the designed systems and need to be further investigated in the future.

6. CONCLUSIONS

We address the AVEC 2014 – 3D dimensional affect and depression recognition challenges, proposing methods toward robust prediction of both variables. For ASC, we present a four-stage affect recognition model trained over multiple systems involving audio-visual cues. Our model accounts for context in prediction to achieve better results. We analyze the contributions from each modality toward the final outcome and observe that introduction of audio features helps us surpass the baseline model trained solely on video features. Our experiments suggest that the modalities complement each other in the prediction, albeit to different extents. In addition, we present a depression recognition system, combining audio, visual, and linguistic cues into a single discriminative model for DSC. While the system proved capable of high performance, overfitting was a problem. Our experiments show an interesting complementarity between the Freeform and Northwind experiments at the modality level, with linguistic information from Freeform and visual information from Northwind combining to form our best performing system.

There is a wide scope for improvements to our current approaches, both in improving the individual sub-challenge systems and designing a combined system toward joint prediction. A major barrier for DSC is the insufficient number of samples to evaluate a large number of features from multiple modalities. We proposed a multi-stage feature selection scheme which may benefit from more informed results on a larger dataset. We train linear models for the affective dimension recognition system, but we believe that these are highly simplified mappings between the low-level descriptors and affective state. A more sophisticated system may be used for capturing the non-linear mappings between the LLDs and affective state. Moreover, we need to further investigate models to better capture context. Finally, one may develop a model to leverage the correlation between depression severity and affective state. This may provide better clinical predictions alongside explaining the relationship between the two complex phenomena.

7. ACKNOWLEDGMENTS

This research is supported by NSF, NIH, and DARPA.

References

[1] Gnu aspell. `http://www.aspell.net`.

[2] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear. Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction*, 2013.

[3] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders*. 5th edition, VA: American Psychiatric Publishing, 2013.

[4] Murali Annavaram, Nenad Medvidovic, Urbashi Mitra, Shrikanth S. Narayanan, Gaurav Sukhatme, Zhaoshi Meng, Shi Qiu, Rohit Kumar, Gautam Thatte, and Donna Spruijt-Metz. Multimodal sensing for pediatric obesity applications. In *Proc. Int. Workshop UrbanSense*, pages 21–25, Raleigh, NC, November 2008.

[5] Anxiety and Depression Association of America. Depression, January 2014. `http://www.adaa.org/understanding-anxiety/depression`.

[6] Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *J. of personality assessment*, 67(3):588–597, 1996.

[7] Matthew P. Black, Athanasios Katsamanis, Brian Baucom, Chi-Chun Lee, Adam Lammert, Andrew Christensen, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*, 55 (1):1–21, 2013.

[8] Daniel Bone, Matthew Black, Chi-Chun Lee, Marian Williams, Pat Levitt, Sungbok Lee, and Shrikanth S. Narayanan. The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *J. of Speech, Language, and Hearing Research*, 2013.

[9] Theodora Chaspari, Daniel Bone, James Gibson, Chi-Chun Lee, and Shrikanth S. Narayanan. Using physiology and language cues for modeling verbal response latencies of children with asd. In *Proc. ICASSP*, May 2013.

[10] M Cox, J Nuevo-Chiquero, JM Saragih, and S Lucey. Csiro face analysis sdk. *Brisbane, Australia*, 2013.

[11] F. Eyben, M. Wöllmer, and B. Schuller. OpenSMILE - The Munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462, Firenze, Italy, 2010.

[12] A. J. Ferrari, F. J. Charlson, R. E. Norman, S. B. Patten, G. Freedman, C. J. L. Murray, and H. A Whiteford. Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study 2010. *Public Library of Science Medicine*, 10(11), 2013.

[13] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomedical Engineering*, 47 (7):829–837, 2000.

[14] Tülin Gençöz. Discriminant validity of low positive affect: is it specific to depression? *Personality and Individual Differences*, 32(6):991–999, 2002.

[15] J. Girard, J. Cohn, M. H. Mahoor, S. M. Mavadati., Z. Hammal, and D. P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analysis. In *Image and Vision Computing*, 2013.

[16] R. Gupta, K. Audhkhasi, S. Lee, and S. S. Narayanan. Speech paralinguistic event detection using probabilistic time-series smoothing and masking. In *Proc. Interspeech*, 2013.

[17] Rahul Gupta, Panayiotis G. Georgiou, David Atkins, and Shrikanth S. Narayanan. Predicting client's inclination towards target behavior change in motivational interviewing and investigating the role of laughter. In *Proc. InterSpeech*, September 2014.

[18] A. Halfin. Depression: The benefits of early and appropriate treatment. *American J. of Managed Care*, 13(4):S92–S97, 2007.

[19] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *J. of Neuroscience Methods*, 200(2): 237–256, 2011.

[20] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In *Proc. Int. Conf. on Pattern Recognition Applications and Methods*, 2014.

[21] M. Lech, L.-S. Low, and K. E. Ooi. Detection and prediction of clinical depression. *Mental Health Informatics, Studies in Computational Intelligence*, 491:185–199, 2014.

[22] Chi-Chun Lee, Athanasios Katsamanis, Matthew P. Black, Brian Baucom, Andrew Christensen, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions. *Computer, Speech, and Language*, 28(2):518–539, March 2014. doi: 10.1016/j.csl.2012.06.006. URL `www.sciencedirect.com/science/article/pii/S0885230812000472?v=s5`.

[23] 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 *Proc. ICASSP*, 2010.

[24] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. Distributional semantic models for affective text analysis. *IEEE Trans. Audio, Speech, and Language Processing*, 21(11):2379–2392, 2013.

[25] M. Mandal and B. Bhattacharya. Recognition of facial affect in depression. *Perceptual and motor skills*, 61 (1):13–14, 1985.

[26] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affective Computing*, 3(2):184–198, April 2012.

[27] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth S. Narayanan. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2): 137–152, February 2013. doi: dx.doi.org/10.1016/j.imavis.2012.08.018. URL `www.sciencedirect.com/science/article/pii/S0262885612001710`.

[28] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Trans. Biomedical Engineering*, 55(1): 96–107, 2008.

[29] M. C. Mundt, A. P. Vogel, D.E. Feltner, and W. R. Lenderking. Vocal acoustic biomarkers of depression severity and treatment response. *J. of Biological Psychiatry*, 72(7):580–587, 2012.

[30] Shrikanth S. Narayanan and Panayiotis G. Georgiou. Behavioral Signal Processing: Deriving human behavioral informatics from speech and language. *Proc. of IEEE*, 101(5):1203–1233, 2013.

[31] National Institute of Mental Health. Depression, January 2014. `http://www.nimh.nih.gov/health/topics/depression/index.shtml`.

[32] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI*, 24(7):971–987, 2002.

[33] K. E. B. Ooi, L. S. A. Low, M. Lech, and N. B. Allen. Prediction of clinical depression in adolescents using facial image analysis. In *Image Analysis for Multimedia Interactive Services*, 2011.

[34] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011.

[35] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proc. Int. Conf. on New Methods in Language Processing*, volume 12, pages 44–49, 1994.

[36] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang. The INTERSPECH 2014 computational paralinguistics challenge: Cognitive & physical load. In *Proc. Interspeech*, Singapore, Singapore, 2014.

[37] Jianbo Shi and Carlo Tomasi. Good features to track. In *Proc. CVPR 1994*, pages 593–600. IEEE, 1994.

[38] Michel Valstar, Bjoern Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. AVEC 2014 – 3D Dimensional Affect and Depression Recognition Challenge. In *Proc. ACM AVEC*, 2014.

[39] M. Van Segbroeck, R. Travadi, Colin Vaz, Jangwon Kim, Matthew P. Black, Alexandros Potamianos, and S. S. Narayanan. Classification of cognitive load from speech using an i-vector framework. In *Proc. Interspeech*, 2014.

[40] Maarten Van Segbroeck, Andreas Tsiartas, and Shrikanth S. Narayanan. A robust frontend for VAD: Exploiting contextual, discriminative and spectral cues of human voice. In *Proc. Interspeech*, 2013.

[41] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'ericco, and M. Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Trans. Affective Computing*, 3(1):69–87, 2012.

[42] M. VÃţ, M. Conrad, L. Kuchinke, K. Urton, M. Hofmann, and A. Jacobs. The berlin affective word list reloaded (bawl-r). *Behavior Research Methods*, 41: 534–538, 2009.

[43] P. Wang, F. Barrett, Martin E., M. Milonova, R. E. Gur, C. Gur, and C. Kohler. Automated video-based facial expression analysis of neuropsychiatric disorders. *J. of Neuroscience Methods*, 168(1):224–238, 2008.

[44] P. Waxer. Nonverbal cues for depression. *J. of Abnormal Psychology*, 83(3):319, 1974.

[45] W. Weintraub. *Verbal Behavior: Adaptation and Psychopathology*. New York: Springer, 1981.

[46] J. R. Williamson, T. F. Quatieri, R. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proc. ACM AVEC*, 2013.

[47] B. Xiao, P. G. Georgiou, Z. E. Imel, D. Atkins, and S. S. Narayanan. Modeling therapist empathy and vocal entrainment in drug addition counseling. In *Proc. Interspeech*, 2013.

[48] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. PAMI*, 29(6):915–928, 2007.

[49] Y. Zhou, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell. Multimodal prediction of psychological disorder: Learning nonverbal commonality in adjacency pairs. In *Proc. Workshop on Semantics and Pragmatics of Dialogue*, 2013.