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

Alexandros Potamianos*
School of Electrical and
Computer Engineering,
National Technical University
of Athens, Greece

Shrikanth S. Narayanan
Signal Analysis and
Interpretation Laboratory,
Univ. of Southern California,
Los Angeles, CA, USA

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 understanding and 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

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 [54] and behavioral signal processing (BSP) [37]. 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 analysis and decision making. BSP methodologies have been applied to various medical/clinical domains, including marital conflict and couples therapy [7,8,26,28], addiction counseling [11,22,60], autism spectrum disorders [9,12], and metabolic health monitoring in obesity [4,15]. The Audio/Visual Emotion Challenge (AVEC) 2014 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.

[*] The author is also affiliated with Behavioral Informatix, LLC.

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

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 [17]. 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 [5, 38]. 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 [23].

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 [18, 36, 42]. 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 [58]. 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 [20, 57]).

Significant related work has been done on the automatic analysis and prediction of depression using speech and prosody [14, 27, 29, 35, 40, 59] gestures, head pose, and facial expressions [2, 24, 41, 45, 49, 56], and a multimodal combination of these cues [25, 33, 34, 61, 63]. As described in the literature, there are a variety of challenges in automatically determining whether a person is depressed or not, 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. In addition, there is also the difficulty in choosing the most appropriate way to label emotions for time-series data, i.e., *what* emotional labels to use and at what temporal *granularity*. Two of the most popular annotation schemes are the use of discrete categorical labels (e.g., angry, happy, sad, neutral) and continuous attribute-based ("dimensional") labels (arousal, valence, and dominance); the relationship between these categorical and dimensional descriptors is well documented (e.g., see [10]). Since emotions tend to evolve in a relatively slow manner, as compared to phonetic sequences in speech or hand/body movements, the majority of emotion recognition work has assumed that an emotion is uniform across a pre-determined interaction-specific temporal period (e.g., at the utterance-, turn-, or discourse-level). However, continuously tracking emotion is also of significant importance, particularly when an event occurs that immediately changes the emotional state of a person; it also has the advantage of eliminating the decision on when a judgment should be made (since judgments are made at *all* available points in time). This is especially relevant with multimodal data, where pertinent 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. Our system provides a subset of features most predictive of depression level under different experimental settings, but overfitting is an issue given the large feature dimensionality and relatively few training instances. Though our best system achieves a root mean squared error (RMSE) value of 7.44 (baseline: 9.26) on the development set, the best RMSE on the testing set is 10.33 (baseline: 10.86), suggesting overfitting. 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 [21], to process the outputs obtained from various modalities. Our proposed system achieves correlation coefficients of 0.427, 0.617, and 0.419 on the target valence, arousal, and dominance dimensions, respectively, on the testing set.

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 report our results and provide a discussion in Section 5, and we offer our conclusions in Section 6.

2. DEPRESSION DATASET

We use the Audio-Visual Emotion Challenge (AVEC) 2014 dataset for our evaluation [51]. 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 age. 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 an equitable distribution 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 [51].

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

3.1.1 Baseline features

The audio baseline features are adopted from the AVEC 2013 challenge, extracted using the openSMILE toolkit [16]. 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 [51].

3.1.2 Additional features

We construct an additional acoustic feature representation by combining the speech streams as proposed in [53]. 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 [52] complemented the baseline features proposed in the INTERSPEECH 2014 challenge [47].

3.2 Video

3.2.1 Baseline features

The AVEC 2014 challenge also provides local binary patterns (LBP) features, well known for describing facial expressions [39]. A LBP descriptor, centered at a pixel, is a binary vector computed by comparing the pixel's intensity with those of its neighbors. After the patterns are computed for each pixel, the LBP descriptor is 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). These "LGBPTOP" features are computed dynamically by considering a short video sequence around each frame to capture temporal changes in facial appearance.

3.2.2 Additional features

Additional LBP features: We compute an additional set of LBP features, but unlike the baseline features, we calculate LBP features in the pixel domain. Videos are first subjected to face detection and tracking to extract human faces at every frame using the standard Viola-Jones face detector in OpenCV. 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) of the video. The static features are local binary patterns [39] 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 LBPTOP feature [62] 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) of a video. This feature is intended 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; please see Fig. 1. The keypoints are detected using a corner detector algorithm [48]. Total motion per frame is computed by adding the amount of motion each keypoint has undergone, thus generating a scalar value per frame.

Features derived from facial landmarks: In addition, we employ the CSIRO Face analysis SDK [13] 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 [44] as shown in Fig. 2. 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 up-

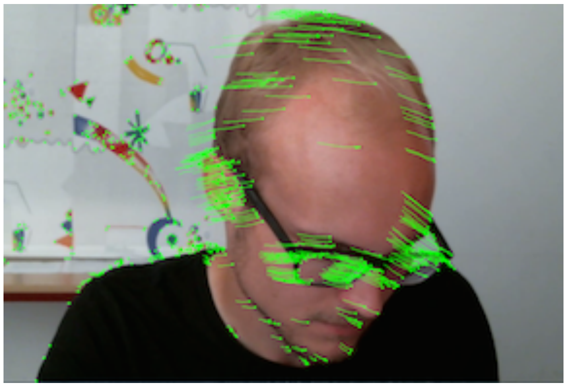

Figure 1: Example of overall motion tracking.

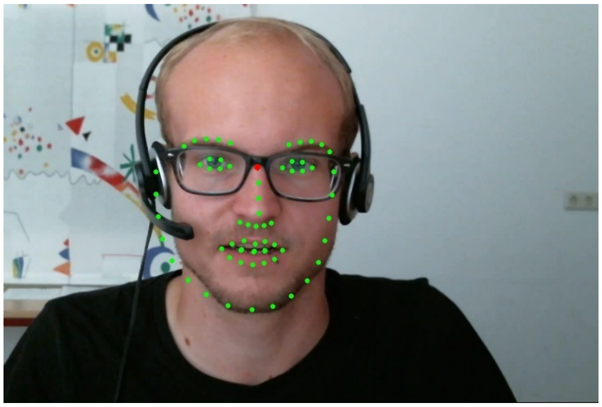

Figure 2: Example of facial landmark fitting.

per 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 centroid of the mouth, normalized by mouth width; (4) the animation of all facial landmarks with respect to the middle red point between the eyes, 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 frames/second, as compared to the audio feature extraction rate of 100 frames/second.

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.

3.3.1 Generating the lexicon

We use the affective lexicon generated by an automated algorithm of lexicon expansion as described in [30, 31, 50]. 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) [55] 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).

3.3.2 Extracting session-level features

Every session in the FreeForm data was part-of-speech tagged using the German version of Treetagger [46]. We collected all the 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

We designed separate methods for the depression and affect recognition sub-challenges, described in detail next.

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

Features	Data	Dimension	Selected
Audio	Northwind	865	41
Audio	Freeform	865	55
Audio deltas	Northwind	427	40
Audio deltas	**Freeform**	427	36
Add. LBP	**Northwind**	768	11
Add. LBP	Freeform	768	13
MFCC	Northwind	672	42
MFCC	Freeform	672	76
MFCC deltas	Northwind	304	9
MFCC deltas	**Freeform**	304	8
Motion	**Northwind**	100	8
Motion	**Freeform**	100	6
Text	**Freeform**	1836	42
Video baseline	**Northwind**	16992	67
Video baseline	Freeform	16992	12

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

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.

We propose a multi-stage feature selection approach: 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. 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 four different systems (two on the video features and two 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 features preparation, followed by the description of our affect prediction system.

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

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

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 and need to be synchronized with the ratings and the video frames. We derive local means of audio features computed over multiple frames and downsample them to be synchronized with the video features. We list the pseudo code for downsampling and synchronizing in Algorithm 1. M denotes the number of frames in the window used to compute the local mean. As there are 100 audio frames per 30 video frames (3.33 times), we set M to 4 for minimal audio frame overlap in computing the local means while retaining information from all the frames.

Algorithm 1 Obtaining audio features synchronized with the video features.

```
Define: $\{v_1, .., v_k, .., v_K\}$ : Time stamps for video frames.
$\{a_1, .., a_j, .., a_J\}$ : Time stamps for audio frames.
$\{A_1, .., A_j, .., A_J\}$ : Original audio features.
$\{A'_1, .., A'_k, .., A'_K\}$ : Local means audio features synchronized with video frames.
$M$: Local mean window length
for $k = 1$ to $K$ do
    Compute $a_j$ closest to $v_k$.
    $A'_k = \text{mean}(A_{(j-\frac{M}{2})}, ..., A_j, ..., A_{(j+\frac{M}{2})})$
end for
```

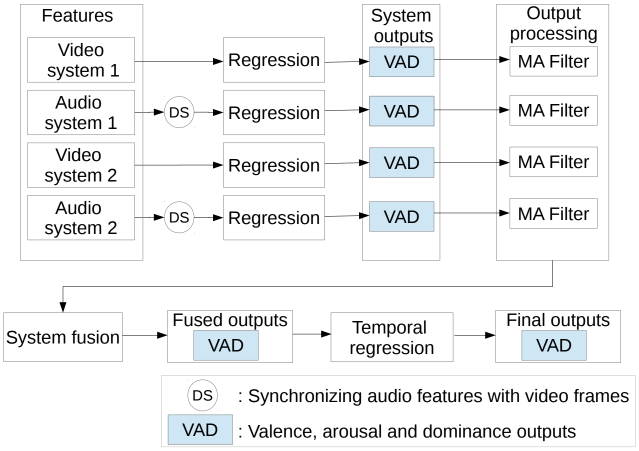

Figure 3: Block diagram of the valence, arousal, and dominance (VAD) prediction system.

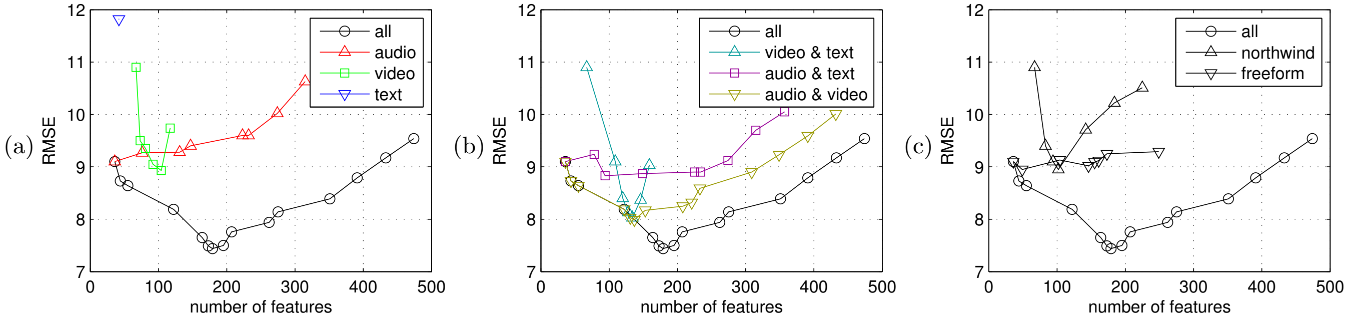

Figure 4: Depression prediction RMSE performance on the development set vs. the number of features used for: (a) single modalities, (b) pairs of modalities, and (c) all modalities only for Northwind or Freeform data.

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. 3. 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 × 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

We list the outcome of feature selection per feature group in Table 1. We then select a subset of feature groups based on brute-force strategy. Furthermore, we use a best-first forward search strategy on the combination of features obtained from selected feature groups, this time based on the performance on the development set. 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. 4 shows the prediction performance with addition of each new feature as obtained by best-first forward search strategy on the development set. Performance achieved using features from a single modality is shown in Fig. 4(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. 4(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. 4(c) and is a nice verification of the importance of both datasets during inference.

Overall the minimum RMSE achieved on the development set was 7.44, which is a notable improvement over the 9.26 baseline. However, we end up with 179 features (> Number of training samples) to achieve this performance, so overfitting may still be an issue. Along with this model, we also tested two additional experimental systems on the testing

System	RMSE	
	Development	Testing
Baseline system	**9.26**	**10.86**
Proposed system 1	11.42	10.35
Proposed system 2	**7.44**	10.56
Proposed system 3	8.51	**10.33**

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

Affective Dimension	*Video system 1* Raw	*Video system 1* Proc.(W_P)	*Video system 2* Raw	*Video system 2* Proc.(W_P)	*Audio system 1* Raw	*Audio system 1* Proc.(W_P)	*Audio system 2* Raw	*Audio system 2* Proc.(W_P)	Fusion	Temporal Regression(W_T)
Freeform										
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										
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 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.

data: i) *System 1* – Correlation based feature selection is applied over all features using all data. No grouping of features is done. ii) *System 2* – We use the groups indicated in Table 1 and apply the correlation based feature selection per group. We select a few groups and do not perform any further feature selection. iii) *System 3* – This system is as proposed above. Feature selection is applied per feature group, and then feature selection is performed based on the development set.

Table 3 lists the results of the three systems on the development and testing sets. Note that for results on the development set, the training data are used for selection and training, while for the results on the test set, the concatenation of training and development data is used. Despite large differences in system performance on the development set, all three proposed systems achieve virtually the same performance on the testing set.

5.2 Affect recognition sub-challenge

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

System fusion obtains better results combining outputs of individual systems. To analyse this, we list the statistics over the individual system outputs scaled with corresponding fusion regression coefficients in Table 7. The final outcome is the sum of all such scaled vectors from the four systems. Thus, each represents the contribution of the corresponding system toward the fused output. From Table 7, we observe that individual systems with higher correlations

System	Affective Dimension	ρ Development	ρ Testing
Baseline	Valence	.355	.188
	Arousal	.412	.206
	Dominance	.319	.196
	Average	**.362**	**.196**
Proposed	Valence	.443	.427
	Arousal	.651	.617
	Dominance	.377	.419
	Average	**.490**	**.488**

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

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

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.

		Freeform		
Affective Dimension	*Video system 1*	*Video system 2*	*Audio system 1*	*Audio system 2*
Valence	1.31 (4.22)	0.27 (1.94)	0.16 (1.85)	0.02 (0.93)
Arousal	2.63 (3.42)	-1.25 (1.59)	4.73 (3.91)	0.09 (1.89)
Dominance	0.27 (2.10)	0.23 (4.54)	1.21 (2.24)	0.01 (0.74)
		Northwind		
Affective Dimension	*Video system 1*	*Video system 2*	*Audio system 1*	*Audio system 2*
Valence	0.10 (0.94)	0.30 (0.91)	0.34 (0.57)	0.01 (0.43)
Arousal	0.09 (2.63)	0.21 (3.46)	1.29 (4.54)	0.00 (0.07)
Dominance	0.42 (0.42)	0.57 (3.48)	1.46 (2.09)	0.00 (0.30)

Table 7: Mean (standard deviation) over system-wise contributions (regression coefficient × system output) from the four ASC systems.

with the target variable have a higher mean contribution toward the final outcome. Audio system 2 makes little contribution for most of the dimensions, with mean of the scaled outputs very close to 0. Also, a low standard deviation implies that the values do not deviate too far from 0. All contributions are positive except for arousal for video system 2 in the Freeform dataset; this may be due to a poorly trained system model in arousal prediction, with the obtained predictions only used to offset the predictions from the other modalities. 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 [19,32]. 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.

8. 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] M. Annavaram, N. Medvidovic, U. Mitra, S. S. Narayanan, G. Sukhatme, Z. Meng, S. Qiu, R. Kumar, G. Thatte, and D. Spruijt-Metz. Multimodal sensing for pediatric obesity applications. In *Proceedings of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems (UrbanSense)*, pages 21–25, Raleigh, NC, Nov. 2008.

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

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

[7] M. P. Black, P. G. Georgiou, A. Katsamanis, B. R. Baucom, and S. S. Narayanan. 'You made me do it': Classification of blame in married couples' interactions by fusing automatically derived speech and language

information. In *Proc. Interspeech*, Florence, Italy, 2011.

[8] M. P. Black, A. Katsamanis, B. Baucom, C.-C. Lee, A. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*, 55(1):1–21, 2013.

[9] D. Bone, M. Black, C.-C. Lee, M. Williams, P. Levitt, S. Lee, and S. S. Narayanan. The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research*, 2013.

[10] C. Busso, B. Murtaza, and S. S. Narayanan. Toward effective automatic recognition systems of emotion in speech. In J. Gratch and S. Marsella, editors, *The Role of Prosody in Affective Speech*. Oxford University Press, 2013.

[11] D. Can, P. G. Georgiou, D. Atkins, and S. S. Narayanan. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In *Proceedings of InterSpeech*, Sept. 2012.

[12] T. Chaspari, D. Bone, J. Gibson, C.-C. Lee, and S. S. Narayanan. Using physiology and language cues for modeling verbal response latencies of children with asd. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013.

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

[14] N. Cummins, J. Epps, M. Breakspear, and R. Goecke. An investigation of depressed speech detection: Features and normalization. In *Proc. of Interspeech*, 2011.

[15] A. Emken, M. Li, G. Thatte, S. Lee, M. Annavaram, U. Mitra, S. S. Narayanan, and D. Spruijt-Metz. Recognition of physical activities in overweight hispanic youth using knowme networks. *Journal of Physical Activity & Health*, 9(3):432–441, Mar. 2012.

[16] 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.

[17] 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.

[18] 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 Transactions on Biomedical Engineering*, 47(7):829–837, 2000.

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

[20] 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.

[21] 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.

[22] R. Gupta, P. G. Georgiou, D. Atkins, and S. S. Narayanan. Predicting client's inclination towards target behavior change in motivational interviewing and investigating the role of laughter. In *Proceedings of InterSpeech*, Sept. 2014.

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

[24] 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. *Journal of Neuroscience Methods*, 200(2):237–256, 2011.

[25] 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. of the Internation Conference on Pattern Recognition Applications and Methods*, 2014.

[26] A. Katsamanis, J. Gibson, M. P. Black, and S. S. Narayanan. Multiple instance learning for classification of human behavior observations. In *Affective Computing and Intelligent Interaction*, Memphis, TN, USA, 2011.

[27] 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.

[28] C.-C. Lee, A. Katsamanis, M. P. Black, B. Baucom, A. Christensen, P. G. Georgiou, and S. 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, Mar. 2014.

[29] 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. of ICASSP*, 2010.

[30] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. Kernel models for affective lexicon creation. In *Proc. Interspeech*, pages 2977–2980, 2011.

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

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

[33] A. Metallinou, A. Katsamanis, and S. 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, Feb. 2013.

[34] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion

classification. *IEEE Transactions on Affective Computing*, 3(2):184–198, Apr. 2012.

[35] 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 Transactions on Biomedical Engineering*, 55(1):96–107, 2008.

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

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

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

[39] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[40] K. E. B. Ooi, L.-S. A. Low, M. Lech, and N. Allen. Early prediction of major depression in adolescents using glottal wave characteristics and teager energy parameters. In *Proc. of ICASSP*, 2012.

[41] 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.

[42] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, 51(9):1530–1540, 2004.

[43] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[44] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *IEEE Conference on Automatic Face and Gesture Recognition*, 2013.

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

[46] 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.

[47] J. Shi and C. Tomasi. Good features to track. In *Proc. Computer Vision and Pattern Recognition 1994*, pages 593–600. IEEE, 1994.

[48] G. Stratou, S. Scherer, J. Gratch, and L. Morency. Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences. In *International Conference on Affective Computing and Intelligent Interaction*, 2013.

[49] P. Turney and M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. technical report ERC-1094 (NRC 44929). National Research Council of Canada, 2002.

[50] 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 *Proc. 4th ACM International Workshop on Audio/Visual Emotion Challenge*, 2014.

[51] M. Van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan. Classification of cognitive load from speech using an i-vector framework. In *Proc. Interspeech*, 2014. accepted.

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

[53] 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.

[54] 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.

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

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

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

[58] 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. of the ACM International Workshop on AVEC*, 2013.

[59] 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.

[60] Z. Yang, A. Metallinou, and S. S. Narayanan. Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues. *IEEE Transactions on Multimedia*, 2014.

[61] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence,*, 29(6):915–928, 2007.

[62] 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. of the Workshop on the Semantics and Pragmatics of Dialogue*, 2013.