Detecting Depression using Vocal, Facial and Semantic Communication Cues

James R. Williamson
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420, USA

Elizabeth Godoy
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420, USA

Miriam Cha
Computer Science Dept.
Harvard University
Cambridge MA 02138, USA

Adrianne Schwarzentruber
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420, USA

Pooya Khorrami
Beckman Institute
University of Illinois
Champaign IL 61820, USA

Youngjune Gwon
MIT Lincoln Laboratory
244 Wood Street
Lexington MA, USA

H.T. Kung
Computer Science Dept.
Harvard University
Cambridge MA 02138, USA

Charlie Dagli
MIT Lincoln Laboratory
244 Wood Street
Lexington MA, USA

Thomas F. Quatieri
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420, USA

ABSTRACT[1]

Major depressive disorder (MDD) is known to result in neurophysiological and neurocognitive changes that affect control of motor, linguistic, and cognitive functions. MDD's impact on these processes is reflected in an individual's communication via coupled mechanisms: vocal articulation, facial gesturing and choice of content to convey in a dialogue. In particular, MDD-induced neurophysiological changes are associated with a decline in dynamics and coordination of speech and facial motor control, while neurocognitive changes influence dialogue semantics. In this paper, biomarkers are derived from all of these modalities, drawing first from previously developed neurophysiologically-motivated speech and facial coordination and timing features. In addition, a novel indicator of lower vocal tract constriction in articulation is incorporated that relates to vocal projection. Semantic features are analyzed for subject/avatar dialogue content using a sparse coded lexical embedding space, and for contextual clues related to the subject's present or past depression status. The features and depression classification system were developed for the 6th International Audio/Video Emotion Challenge (AVEC), which provides data consisting of audio, video-based facial action units, and transcribed text of individuals communicating with the human-controlled avatar. A clinical Patient Health Questionnaire (PHQ) score and binary depression decision are provided for each participant. PHQ predictions were obtained by fusing outputs from a Gaussian staircase regressor for each feature set, with results on the development set of mean F1=0.81, RMSE=5.31, and MAE=3.34. These compare favorably to the challenge baseline development results of mean F1=0.73, RMSE=6.62, and MAE=5.52. On test set evaluation, our system obtained a mean F1=0.70, which is similar to the challenge baseline test result. Future work calls for consideration of joint feature analyses across modalities in an effort to detect neurological disorders based on the interplay of motor, linguistic, affective, and cognitive components of communication.

Keywords

Affective Computing; Major Depressive Disorder; Text Mining; Speech; Facial Expression; Semantic Analysis; Challenge

1. INTRODUCTION

Major depressive disorder (MDD) is the most prevalent mood disorder, with a lifetime risk of 10–20% for women and 5–12% for men [7]. As the number of people suffering from MDD steadily increases, so too does the burden of accurate diagnosis. The growing global burden of MDD suggests that a convenient and automated method to evaluate depression severity would both simplify and standardize the task of diagnosing and monitoring depression, allowing for greater availability and uniformity in assessment. One advance in this direction presented by AVEC 2016 is use of a virtual avatar to interact with patients, albeit with human control due to current limitations in natural dialogue technologies. An automated approach for depression evaluation may reduce multiple in-office clinical visits, facilitate accurate measurement and identification, and quicken the evaluation of treatment. Toward these objectives, potential depression biomarkers of growing interest are vocal- and facial expression-based features, two categories of easily-acquired measures that have been shown to change with a patient's mental condition and emotional state [2,3,5,8,9,26,27,31,33,34,36]. Other higher-level aspects of speech such as language and cognitive expression are also known to be affected by depressed state [4,20]. In this paper, we bring together biomarkers derived from these modalities through voice, face and semantic analysis.

[1] This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government

AVEC'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2988257.2988263

Neurophysiological changes in depression impact the motor control functions that drive vocal and facial gesturing, while neurocognitive changes influence the quality and manner of response in dialogue. In neurophysiological change, many individuals suffering from MDD suffer from psychomotor retardation, which affects mechanisms controlling speech production and facial expression. Previously developed neurophysiologically motivated vocal and facial coordination and timing features have been shown to be powerful indicators of depression [31,33,34]. In particular, for the audio modality, formant correlations and phone-dependent duration features proved very effective in classifying depression. Here, we have extended this basis to also include features based on novel indicators of lower vocal tract constriction and measures of variability vocal quality. For the video modality, we exploited the facial action coding system (FACS), which quantifies localized changes in facial expression representing facial action units (FAUs) that correspond to distinct muscle movements of the face [18]. Combined, the consistently high depression classification performance of vocal and facial coordination and timing features indicate that these biomarkers are effectively capturing the influence of psychomotor retardation attributed to MDD [26].

In addition to drawing from and extending findings motivated by neurophysiological changes in MDD, our work further seeks to link analysis of motor function to neurocognitive indicators in communication. Specifically, text-based *semantic content* features are used to exploit word representations using GloVe model embedding in combination with a high-level feature learning method [25]. In highlighting semantic content, this methodology shows that embedding of text allows for indicators of neurocognitive state, with the interviewer (avatar) content representing the most powerful indicator, presumably because her words avoid sources of intersubject variability that are not related to depression. This observation highlights a limitation of previous research on text-based sentiment analysis that focuses solely on a subject's text (e.g. sentiment classification using Twitter messages [1,16] and movie rating classification based on review texts [15]). In this paper, we introduce text-based methods that infer semantic content from the interviewer's questions as well as from the interviewee's answers, leading to a better estimation of neurocognitive state (i.e., depression). In addition, text-based *semantic context* features are used that focus on specific points in the interview dialogue, accumulating indicators of potential depression reoccurrence along with commitment to therapy and the individual's self-described assessment of his or her current state.

Data for feature and system development is provided by the 6th International Audio/Video Emotion Challenge (AVEC) consisting of audio and corresponding transcribed text elicited from participant interviews with the USC Creative Technology Center avatar, "Ellie." For each participant in the training and development sets, a Patient Health Questionnaire (PHQ) score and binary depression decision are provided. The challenge evaluation criterion is mean F1 score, which is a combined measure of precision and recall in detecting both depressed and non-depressed subjects [31]. Using a Gaussian staircase regressor with fusion of the multi-modal features designed in-house, our system achieves a mean F1=0.81, RMSE=5.31, and MAE=3.34 on the development set, compared to the challenge baseline mean F1=0.73, RMSE=6.62, and MAE=5.52. Additionally, our best PHQ correlation results were obtained by combining predictions from audio, video and text modalities. This motivates continued study into joint multi-modal feature analysis to capture interplay of motor, linguistic, and cognitive components of communication.

In Section 2 we discuss our set of audio speech features and Section 3 describes the video facial features for estimating depressive state and Section 4 outlines the text-based semantic features. Section 5 describes the classification approaches and provides results for audio, video, semantic and ensemble systems. Discussion and concluding remarks are in Section 6.

2. AUDIO FEATURES

2.1 Audio Data Preprocessing

2.1.1 Audio Data Considerations

Careful examination of the AVEC 2016 Depression dataset revealed some issues related to the audio acquisition and segmentation. First, the segmentation of participant speech, taken from time stamps on provided transcripts, contained substantial errors for some subjects, resulting in audio-transcript miss-alignments. Subsequently, forced alignment of the participant speech to acquire phone and word boundaries was sensitive to these errors and could not be used reliably. Some segmentation inconsistencies also result in Ellie the avatar's speech being included in audio intended to be from the Participant. This could distort audio feature statistics according to the percentage of Ellie's speech included in the Participant segments. There also appeared to be a change in Ellie's behavior after approximately one third of the participants audio had been collected. This was observed in the transcript text of Ellie's statements and questions as well as in turn duration statistics based on the segmentation. As in many applications, separation of Ellie and the Participant's speech might benefit from tailored diarization for this dataset.

Additionally, audio quality varied across subjects, with what seemed like initial (low numbered) collects suffering from higher noise levels and consequently lower SNR. This might be due to a common and understandable real-world problem of adjusting microphone, subject and room set-ups accordingly as a collect gets underway. Audio levels across subjects were also sensitive to microphone taps, coughs and other impulse-like sounds. Finally, for a few of the subjects, the avatar Ellie's speech had higher energy levels and thus could not be considered low-level crosstalk. Ultimately, as in most real-world contexts, treatment of the audio channel requires certain algorithmic adaptations to the data. The following audio pre-processing and feature design sought to limit the impact of any data acquisition and segmentation issues.

2.1.2 Audio Preprocessing

For each file, the Participant's speech was extracted using the segmentation provided in the transcript, despite the presence of some errors described in section 2.1.1. The final waveform was amplitude normalized to adjust for level differences in the recordings. Specifically, to mitigate observed disturbances in the audio due to coughs, sniffles, and what seem to be microphone taps, the SNR output level for each file was first limited. With informal listening, a moderate cutoff value of -15dB was selected. Finally, the Participant's audio file was normalized to have a maximum absolute value of 1.

2.2 Spectral Features

2.2.1 Correlation structure of formant tracks

Properties of vocal tract resonances over time contain information about speech dynamics related to articulatory properties of the depressed voice. A formant tracking algorithm based on Kalman filtering was used to obtain smooth estimates of the first three resonant frequencies over time [21]. Formant frequencies were extracted at every 10 ms from the audio signal, which was unprocessed other than being segmented based on the transcripts to include subject speech. Embedded in the formant tracking algorithm is a voice-activity detector that allows a Kalman smoother to smoothly coast through non-speech regions. Estimates of the third formant that went above a threshold of 4.5k Hz were truncated.

Next, speaker turn segments above one second in duration were used for further feature processing. For each of these segments, formant correlation structure features were computed as follows. A channel-delay correlation matrix was computed from the formant tracks using time-delay embedding. The correlation matrix, with dimensionality (45 x 45), based on three formant channels and 15 time delays per channel, with 3-frame (30 ms) delay spacing. From this matrix the 45-dimensional rank-ordered eigenspectrum was computed, characterizing the within-channel and cross-channel distributional properties of the multivariate formant time series. These articulatory coordination measures have previously been used for estimating depression severity, Parkinson's disease, age-related cognitive decline, mild TBI, and cognitive load [12,27,33,34,36].

Dimensionality reduction of the 45-dimensional correlation structure feature vectors was done as follows. Each feature element was z-scored across the training set so that features at each eigenvalue index are afforded equal weight, and then principal component analysis (PCA) was used to produce a four-dimensional feature vector. The z-scoring and PCA transformations obtained from the training set were then applied to the test set feature vectors.

2.2.2 Correlation structure of delta MFCCs

To introduce vocal tract spectral magnitude information, a standard set of 16 MFCCs was generated by Opensmile from segmented but otherwise unprocessed audio files [6]. Delta MFCCs (dMFCCs) were then computed, which reflect dynamic velocities of the MFCCs over time. Delta coefficients were computed using a delay parameter of 2 (regression over two frames before and after a given frame).

From each speaker turn over one second in duration, a channel-delay correlation matrix was computed from the dMFCCs using time-delay embedding, with dimensionality (240 x 240), based on 16 dMFCC channels and 15 delays per channel with 1-frame (10 ms) delay spacing. From this matrix the 240-dimensional rank-ordered eigenspectrum was computed, which characterizes the within-channel and cross-channel distributional properties of the multivariate dMFCC time series. These spectral coordination measures have previously been used to estimate the severity of depression, Parkinson's disease, and cognitive load from speech [27,33,34,36]. Dimensionality reduction of the 240-dimensional correlation structure feature vectors to four principal components was done using the same procedure used for the formant correlation structure features.

2.3 Lower Vocal Tract Physiology Features

Though often overlooked in speech processing applications, the lower Vocal Tract (VT) plays a key role in speech production [13,29,30]. Situated between the glottis and pharynx, the lower VT cavities regulate flow throughout the speech production system, thus representing a region of high source-filter coupling [30]. Most notably in the voice community is the lower VT's critical role in the production of a singer's/speaker's formant or a form of production sometimes referred to as "resonant" voice. In this type of production mode, the individual makes a concerted effort to project his or her voice by narrowing the epilarynx cavity and strengthening constrictions at the epilarynx and piriform openings. Spectrally, this results in an enhancement of the distinct lower VT resonance pattern [13], specifically amplifying the epilarynx resonance (typically around 3kHz) and deepening the piriform null (typically around 5kHz). On the contrary, for patients with vocal issues (e.g. hoarseness), the opposite trend has been observed [23]. In the context of MDD, the hypothesis follows similarly that the opposite of a loud and deliberate production mode trend will be observed in depressed subjects, corresponding to less vocal effort manifested as a relaxation of control of the lower VT cavities by muscles at and adjacent to the larynx.

In order to quantify the relative degree of speakers formant phenomena, the following metric was used. First, the mean (in dB) spectral envelope (True Envelope [28] order 40 over 3 pitch periods [10]) was calculated for voiced frames in two frequency regions, 3-4kz and 4-5kHz, respectively approximating epilarynx and piriform bands. Then, to quantify contrast in the relative degree of enhancement between the epilarynx resonance and piriform null, the ratio (difference in dB) of energy between the lower 3-4kHz and upper 4-5kHz band was calculated. As expected, for depressed participants, this localized spectral energy difference was less than for the non-depressed subjects. This represents a novel finding, both in objectively quantifying the degree of enhancement of the lower VT resonance patterns and in application for depression classification.

2.4 Loudness Variation Features

As a gross indicator of loudness linked to waveform shapes, a peak-to-rms measure was calculated on a segmental level, reflecting a local loudness metric related to waveform shape across a few pitch periods (with a standard analysis window of 30ms). In order to capture indications that might be linked to prosodic variation across the session, the global standard deviation of local (mean, std, range) peak-to-rms statistics were the features used for classification. The local mean, standard deviation and range (difference between top and bottom 5% values) statistics were calculated for voiced frames in 2 second time intervals sliding with 50% overlap. For depressed speech, the variations in peak2rms levels across the session were higher, potentially indicating mixed regions of both modal and non-modal phonation. While the peak-to-rms statistic variations were statistically significant when correlated with the PHQ score, a complementary loudness feature was not statistically significant even though it provided indications of a trend towards overall softer speaking levels for depressed subjects.

3. VIDEO FEATURES

To capture the joint dynamical properties across multiple facial action units (FAUs) during speech, the correlation structure of FAUs was computed. From each speaker turn over one second in

duration, a channel-delay correlation matrix was computed from the FAUs using time-delay embedding, with dimensionality (300 x 300), based on 20 FAU channels and 15 delays per channel with 1-frame (33 ms) delay spacing. From this matrix the 300-dimensional rank-ordered eigenspectrum was computed, which characterizes the within-channel and cross-channel distributional properties of the multivariate FAU time series. These facial coordination measures have previously been used to estimate the severity of depression and cognitive load from speech [27,34]. Dimensionality reduction of the 300-dimensional correlation structure feature vectors to four principal components was done using the same procedure used for the formant and delta MFCC correlation structure features.

4. SEMANTIC FEATURES

Two approaches for text-based semantic processing were done. The first approach is *semantic content* analysis, based on summarizing spoken content by projecting transcribed sentences into high dimensional word spaces and learning statistical regression models that relate the word embeddings to PHQ scores. The second approach is *semantic context* analysis, which obtains a coarse characterization of contextual evidence related to depression, based on factors such as previous depression diagnoses, indications of ongoing therapy, and indicators of negative emotional state and feelings.

4.1 Semantic Content Features

The semantic content analysis approach uses the provided transcripts to discover, using mappings to a word embedding space followed by sparse coding, correlations between the semantic content of both interviewer and interviewee with the PHQ scores.

4.1.1 Preprocessing

The following preprocessing of raw transcripts was first done:

Question/Answer pair extraction. Given a raw transcript, question-answer pairs were formed each time Ellie asked a new question. Examples are shown in Figure 1.

Filled pause extraction. A substantial portion of human communication is non-verbal. Non-verbal cues (e.g., postures, facial expression, eye gaze, gestures) provide useful information in determining a participant's psychological distress. For example, a participant's laughter can indicate a positive affective response in a conversation. Along with verbal features, commonly occurring filled pauses were extracted, inclusive of [laughter], [sniffle], [cough], [sigh], [deep breath].

Stopword elimination. The most frequently occurring words often do not carry much information. Common words such as 'a', 'on', and 'through', were automatically removed based on the hypothesis that these words contain little discriminative information.

4.1.2 GloVe embedding of questions and answers

Global vectors for word representation, or GloVe, is a distributed text representation method in which words are embedded in a high dimensional space, with distances between the words arising from their co-occurrence statistics within documents [25]. The embedding can be trained using word2vec or obtained from a preexisting model trained on a large text corpus [22]. A pre-existing GloVe model with a 50-dimensional embedding

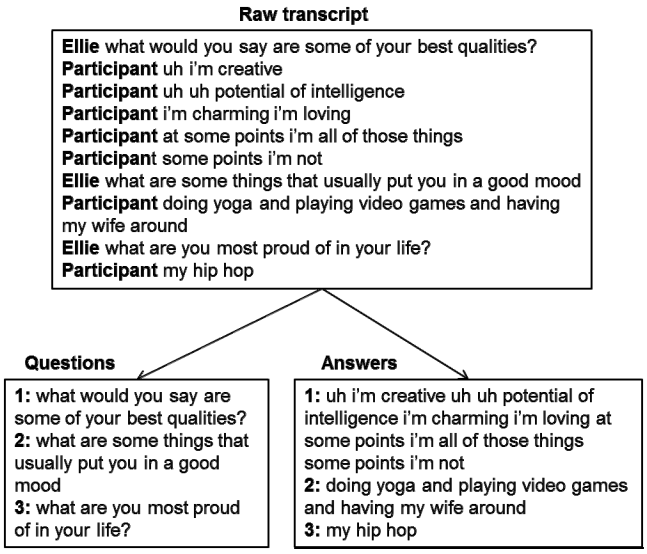

Figure 1. Example of automated extraction of questions and answers.

space, which was trained on the Wikipedia 2014+Gigaword 5 data set, was used as an embedding space. A question or answer sentence was then represented by the average of its embedded word vectors.

The embedded vector space from representing the set of questions and answers were next transformed into two alternative representations, using 1) principal components analysis (PCA) and 2) the whitening (ZCA) transform.

4.1.3 High-level feature learning

Next, sparse coded feature representations were formed from the PCA- and ZCA-transformed embedded spaces. With sparse coding, more complex topic-related features can be automatically learned [24], often leading to better generalization performance in classification and regression problems than is obtained from hand-engineered algorithms.

Given an input feature patch $\mathbf{x} \in R^N$, sparse coding solves for a representation $\mathbf{y} \in R^K$ in the following optimization problem:

$$\min_{\{\mathbf{D},\mathbf{y}\}} (\|\mathbf{x} - \mathbf{D}\mathbf{y}\|_2)^2 + \rho\|\mathbf{y}\|_1 \quad \text{s.t } \|\mathbf{d}_i\| \le 1, \forall I, \tag{4}$$

where $\mathbf{d}_i$ is the i^{th} dictionary atom in an overcomplete dictionary $\mathbf{D} \in R^{N \times K}$ and where $\rho > 0$ is a regularization parameter. In this paper least angle regression (LARS) was used in the dictionary learning algorithm from the SPAMS toolbox [14,19]. Parameters used were K=200 and ρ=0.2.

4.1.4 Multivariate regression

A linear support vector regressor (SVR) was used to learn a mapping from the text-based feature vectors to subjects' PHQ scores. Separate evaluations were done for question- and answer-based feature vectors, and based on using raw GloVe features, GloVe+PCA, and GloVe+ZCA features prior to sparse coding. The SVR was learned from the training set and validated on the development set. Because the SVR produces a separate PHQ prediction for each question or answer, a method is needed for fusing these predictions across a subject's session. The adopted method is to use the average of the top *n* predicted PHQ scores for each subject, based on the hypothesis that the questions or answers that provide the strongest predictions of depression level

are also the most indicative of depression. Validation testing on the development set resulted in the use of *n=1* for the PCA-based feature vectors and *n=6* for the ZCA-based feature vectors.

To evaluate the relative usefulness of the different feature approaches, the development set PHQ predictions were converted to depression predictions using thresholds, resulting in mean F1 scores as shown in Table 1. Notice that the highest mean F1 scores were obtained for question-based features, with PCA and ZCA producing better results than using the raw GloVe features (as input to sparse coding). Based on these results, the question-based GloVe+PCA and GloVe+ZCA PHQ predictions were selected as the text content analysis features used by the depression classification system described in Section 5. Dimensionality reduction of this two-dimensional feature vector to a single principal component was done using the same procedure used for the correlation structure feature vectors in Sections 2 and 3.

Table 1. Semantic content performance on Development set.

Text Type	Embedded Space	Mean F1
Question	GloVe	0.24
	Glove+PCA	0.62
	Glove+ZCA	0.75
Answer	GloVe	0.46
	Glove+PCA	0.62
	Glove+ZCA	0.54

4.2 Semantic Context Features

The semantic context indicators use the provided transcripts to infer a subject's status with respect to four conceptual classes. The first conceptual class seeks priors on depression based on prior diagnosis. The second class seeks assessment of current or prior involvement in therapy. For both of the above context classes, the indicator relies on Ellie's questions and her responses to participant answers to provide consistent measures across all participants. The third conceptual class seeks indicators of the participant's present state and feelings, focusing on the PHQ 2 and 8 questionnaire categories. Ellie's questions and responses are considered as well as keywords in the participant's answer. The fourth class seeks indications of past or present suicidal ideation. Figure 2 summarizes the four semantic context indicators. The semantic context feature is the sum of points accrued from all four indicators. If none of the indicators' conditions are satisfied, then the participant receives a feature value of zero.

5. DEPRESSION CLASSIFICATION SYSTEM

5.1 Dataset and Evaluation

The experiments in this section use the Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) dataset [32] for train and test provided for the AVEC 2016 depression challenge. The depressed state of subjects is based on the PHQ-8 metric [17]. Our approach to depression classification is to learn a statistical regression model for predicting the PHQ scores from the training set, and to classify as depressed subjects that have suprathreshold PHQ predictions. The training, development, and test sets contain 107, 35, and 47 subjects, respectively, containing 21, 7, and 9 depressed subjects in each set. Development set

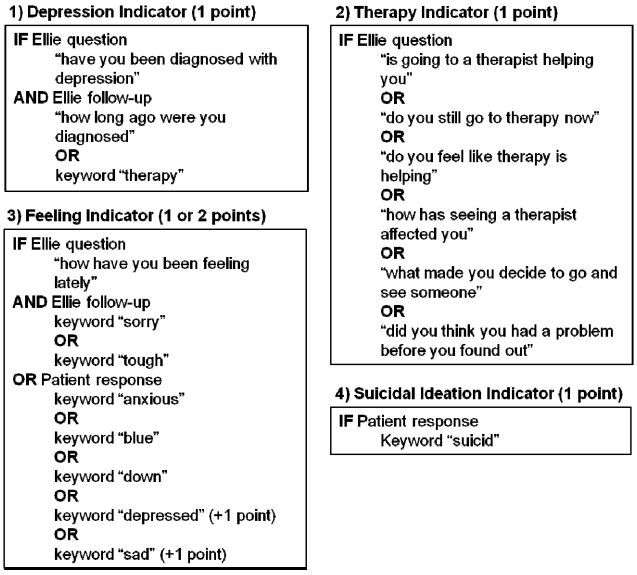

Figure 2. Semantic context indicators.

predictions were obtained by training the model on the training set. Test set predictions were obtained by training the model on the combined training/development set.

5.2 Gaussian Staircase Model

The Gaussian staircase is a statistical modeling approach that generalizes the use of Gaussian distributions for binary classification into the domain of multivariate regression [33,34,35]. This is done by partitioning the outcome variable into multiple nested ranges with binary class labels for "lower" and "higher" being associated with complementary ranges at each nested partition. A multivariate normal distribution is used to model the class-conditioned features in each partition, and the class-conditioned likelihoods are computed by summing the likelihoods across all the partitions.

In the current work PHQ is the outcome variable. PHQ scores are partitioned into five staircase levels for each class C_k. The PHQ ranges for these levels are [0-5, 0-7, 0-10, 0-12, 0-15] for C_1 (non-depressed) and [6-23, 8-23, 11-23, 13-23, 16-23] for C_2 (depressed). The likelihood model for each class at each partition is defined using a multivariate normal distribution obtained from the training data.

The final class-conditioned likelihoods, $p(x_i \mid C_k)$, are obtained by summing over staircase likelihoods. The resulting two-class log-likelihood ratio,

$$y_i = \log(p(x_i \mid C_2)) - \log(p(x_i \mid C_1)), \quad (3)$$

is used as a basis for PHQ prediction. For the correlation structure features there is a separate feature vector per speaker turn within each session, so further processing is needed to obtain a single log-likelihood ratio from the entire session. Better discrimination performance is obtained by using the median log-likelihood ratio from only the highest *n* log-likelihood ratios per speaker. This is done based on the hypothesis that the speaker turns with the highest log-likelihood ratios will more reliably capture depression-related speech differences. The median of the same number of log-likelihood ratios, *n=25*, is computed for all three correlation structure feature sets. Next, a linear regression model is constructed from the training set log-likelihood ratios and applied to the test set log-likelihood ratios, and depression

predictions made when the regression outputs, z_i, exceed the following empirically derived threshold,

$$z_{thresh} = mean_i(z_i) + 0.9\ std_i(z_i). \quad (4)$$

5.3 Feature Set Results

Table 1 summarizes the results on the development data of the seven individual feature sets used in our depression classification system. The mean F1 scores are shown based on assigning depression predictions using equation (4). An alternative measure of detection accuracy, the area under the receiver operating characteristic (ROC) curve, or AUC, is also shown. Finally, the Pearson correlations, *r*, of the PHQ predictions to the actual PHQ scores are shown.

Examining Table 1, we see that the Correlation Structure (CS) features for audio and video perform well, as is consistent with previous results for depression classification. However, the CS features' performance is lower than might be expected here, likely due to the fact that the analyzed speech segments are, on average, of very short duration. Also, with regard to the organizer-provided transcript-based segmentations, the lower VT features appear to be affected by changes in turn segmentation that occurred starting with subject 363 in the data set. Though analysis on voiced frames might have limited some of these sensitivities, hindsight would suggest isolating a subset of segments as in the approaches for the CS-based features. Finally, semantic features prove to be highly informative for depression classification, indicating a powerful modality to be exploited in future work on depression classification. Overall, the audio, video and semantic feature sets in Table 1 provide complementary information relating to motor function timing and dynamics as well as dialogue semantics.

Table 1. Performance of individual feature sets on development data.

Modality	Feature Set	Mean F1	AUC	*r*
Audio	1. CS-Formant	0.55	0.71	0.41
	2. CS-dMFCC	0.45	0.56	0.33
	3. Lower VT	0.51	0.58	0.07
	4. Loudness Var.	0.76	0.69	0.32
Video	5. CS-FAU	0.53	0.63	0.44
Semantic	6. Content	0.81	0.93	0.81
	7. Context	0.76	0.91	0.78

5.4 Fused Results

As described in Section 5.3, separate PHQ predictions are made from each feature set. Fusion of these predictions is done using a weighted average of PHQ predictions. Using the method from [34], fusion weights are determined based on Pearson correlations on the development set: $w=1/(1-r^2)$. This weighting system allows substantial weighting of all feature sets (with a minimum weight of one), but also provides stronger weighting of the more useful feature sets. Notice that, while it would be easy to design weights that give better fused performance on the development set, such an approach is avoided due to concerns of overfitting.

Table 2 summarizes the fused results within each sensor modality and across all modalities. In addition to the statistics shown in Table 1, the RMSE and MAE values for the PHQ predictions are also shown, which facilitates comparison with the baseline benchmark results. In Figure 3 the predicted PHQ scores for our ensemble system are plotted as a function of true PHQ score. Non-depressed subjects are plotted in blue and depressed in red. Green circles are plotted around the eight subjects predicted to have depression. Notice that the false positive classifications are subjects with reasonably high PHQ scores of 9, 10 and 15.

Table 2. Performance on development set of fused feature sets both within and across sensor modalities.

Modality	Mean F1	AUC	r	RMSE	MAE
Audio	0.57	0.72	0.44	6.38	5.32
Video	0.53	0.63	0.44	6.45	5.33
Semantic	0.84	0.94	0.83	4.46	3.34
Ensemble	**0.81**	**0.92**	**0.84**	**5.31**	**4.18**

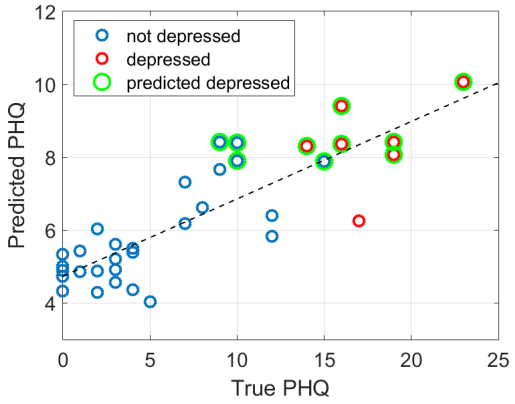

Figure 3. Results on development set for our ensemble PHQ prediction system, with predicted PHQ plotted as a function of true PHQ. Depressed subjects are plotted in red and predicted depressed subjects plotted with surrounding green circles.

Table 3 lists the organizer baseline results on the development set within the audio and video modalities, as well as baseline ensemble results. From audio, our system obtains comparable performance to the baseline system, albeit with a small advantage in mean F1 and RMSE. From video, our system performs worse in terms of mean F1 but better in terms of RMSE and MAE. Finally, our ensemble system far outperforms the baseline system due to its inclusion of a semantic analysis component.

Table 3. AVEC 2016 Depression baseline performance on development set [32].

Modality	Mean F1	RMSE	MAE
Audio	0.50	6.74	5.36
Video	0.72	7.13	5.88
Ensemble	0.72	6.62	5.52

5.5 Test Set Results

Our ensemble system was also used to predict depressed subjects on the held out test set. Based on the threshold rule in equation (4), our system predicted 13 subjects as being depressed, of which six were actually labeled depressed, resulting in a mean F1 score

of 0.70. Thus, we see a moderate performance degradation from the development mean F1 score of 0.81. Access to the true PHQ scores would allow a better understanding of how much of this performance degradation is due to inability to generalize to the test set versus variability in the distribution of depression labels relative to PHQ scores. The mean F1 score that we obtained on the test set is the same as the baseline results from the video and ensemble systems. However, any direct comparisons between these results must be tentative, as the recall values reported in the baseline paper [32] indicate that the baseline system was evaluated on a data set with a different number of subjects labeled as depressed.

6. DISCUSSION

In this work, fusion of audio, video and semantic features which were motivated by neurophysiological and neurocognitive effects of MDD allowed for a high performing depression detection system. Previously developed audio and video motor control features based on correlation structure performed moderately well on this AVEC 2016 dataset and were complemented by novel physiologically-based features linked to vocal projection through control of the lower vocal tract. Semantic analyses of dialogue transcripts provided the highest performing features, suggesting that future work on depression classification should exploit semantic features. Also, interestingly enough, the most informative indicators of the dialogue content and context were obtained by analyzing the avatar's text, which avoids sources of inter-subject variability unrelated to depression. This observation could play an important role in the design of automatic screenings for depression diagnostics.

Future work calls for joint analyses across speech, facial gestures and semantic content and context in a dialogue. For example, degrees of motor coordination might vary as a function of communicative intent in a conversation. Specifically, the question arise of how a person's articulation and facial gesturing are impacted by affectively stressing or neurocognitively challenging dialogue turns (e.g. recalling a difficult experience in life). That is, audio and visual modalities conditioned on semantics might reveal differences based on the individual's affective and/or neurological state. Ultimately, with grounding in neuro-physiological and cognitive analyses, we seek to exploit the interplay between what and how a person communicates.

7. REFERENCES

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media. pp. 30–38 (2011)

[2] Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., & Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and cognition*, *56*(1), 30-35.

[3] Darby, J.K., Simmons, N. and Berger, P.A. 1984. Speech and voice parameters of depression: A pilot study. Journal of Communication Disorders. 17, 2 (1984), 75–85.

[4] De Choudhury, M., Gamon, M., Counts, S., Horvitz, E., 2013. Predicting depression via social media, Association for the Advancement of Artificial Intelligence.

[5] Ellgring, H., & Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, *20*(2), 83-110.

[6] Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459-1462). ACM.

[7] Fava, M. and Kendler, K.S. 2000. Major depressive disorder. Neuron. 28, 2 (2000), 335–341.

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

[9] Gaebel, W. and Wölwer, W. 1992. Facial expression and emotional face recognition in schizophrenia and depression. European archives of psychiatry and clinical neuroscience. 242, 1 (1992), 46–52.

[10] Godoy, E., Malyska, N., & Quatieri, T. F. (2015). Estimating lower vocal tract features with closed-open phase spectral analyses. Interspeech, 771–775.

[11] Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, *7*(515-546).

[12] Helfer, B. S., Quatieri, T. F., Williamson, J. R., Keyes, L., Evans, B., Greene, W. N., Vian, T., Lacrignola, J., Shenk, T., Talavage, T., Palmer, J., & Heaton, K. (2014). Articulatory dynamics and coordination in classifying cognitive change with preclinical mTBI. In *INTERSPEECH* (pp. 485-489).

[13] Honda, K., Kitamura, T., Takemoto, H., Adachi, S., Mokhtari, P., Takano, S., Nota, Y., Hirata, H., Shimada, Y., Fujimoto, I., Masaki, S., Fujita, S., & Dang, J. (2010). Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling. Comp. Methods in Biomech. & Biomed. Engineering, 13(4), 443–453.

[14] INRIA. Sparse modeling software. http://spams-devel.gforge.inria.fr/.

[15] Joshi, M., Das, D., Gimpel, K., & Smith, N. A. (2010, June). Movie reviews and revenues: An experiment in text regression. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 293-296). Association for Computational Linguistics.

[16] Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the ICWSM (2011)

[17] Kroenke, K. et al, *The PHQ-8 as a Measure of Current Depression in the General Population*, Journal of Affective Disorders, 114(1-3):163-173, April 2009.

[18] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011, March). The computer expression recognition toolbox (CERT). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (pp. 298-305). IEEE.

[19] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online Dictionary Learning for Sparse Coding. In ICML, 2009.

[20] Mayberg HS. Limbic-cortical dysregulation: a proposed model of depression. J, Neuropsychiatry Clin Neurosci 1997; 9: 471–81.

[21] Mehta, D. D., Rudoy, D., & Wolfe, P. J. (2012). Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. *The Journal of the Acoustical Society of America*, *132*(3), 1732-1746.

[22] Mikolov, T., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

[23] Nawka, T., Anders. C, Cebulla, M. and Zurakowski, D. (1997) The speaker's formant in male voices. Journal of Voice, 11, 4, 422-428.

[24] Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1?. *Vision research*, *37*(23), 3311-3325.

[25] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global Vectors for Word Representation. In *EMNLP* (Vol. 14, pp. 1532-43).

[26] Quatieri, T.F. and Malyska, N. 2012. Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity. Interspeech.

[27] Quatieri, T. F., Williamson, J. R., Smalt, C. J., Patel, T., Perricone, J., Mehta, D. D., & Palmer, J. (2015). Vocal biomarkers to discriminate cognitive load in a working memory task. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[28] Röbel, A., & Rodet, X. (2005). Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In *International Conference on Digital Audio Effects* (pp. 30-35).

[29] Sundberg, J. (1974) Articulatory interpretation of the singing formant. Journal of Acoustical Society of America. 55, 4, 838-844.

[30] Titze, I. and B. Story (1996). Acoustic interactions of the voice source with the lower vocal tract. Journal of Acoustical Society of America. 101, 4, 2234-2243.

[31] Trevino, A. C., Quatieri, T. F., & Malyska, N. (2011). Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, *2011*(1), 1-18.

[32] Valstar, M. et al, *AVEC 2016 --- Depression, Mood, and Emotion Recognition Workshop and Challenge*, arXiv:1605.01600, May 2016.

[33] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., & Mehta, D. D. (2013, October). Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge* (pp. 41-48). ACM.

[34] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., & Mehta, D. D. (2014, November). Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (pp. 65-72). ACM.

[35] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Perricone, J., Ghosh, S. S., Ciccarelli, G., & Mehta, D. D. (2015, September). Segment-dependent dynamics in predicting Parkinson's disease. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[36] Yu, B., Quatieri, T. F., Williamson, J. R., & Mundt, J. C. (2014). Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers. In *INTERSPEECH* (pp. 1038-1042).