Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing

James R. Williamson
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
(781) 981-5374
jrw@ll.mit.edu

Thomas F. Quatieri
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
(781) 981-7487
quatieri@ll.mit.edu

Brian S. Helfer
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
(781) 981-7962
brian.helfer@ll.mit.edu

Gregory Ciccarelli
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
(781) 981-3474
gregory.ciccarelli@ll.mit.edu

Daryush D. Mehta
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
(781) 981-5818
daryush.mehta@ll.mit.edu

ABSTRACT[1]

In individuals with major depressive disorder, neurophysiological changes often alter motor control and thus affect the mechanisms controlling speech production and facial expression. These changes are typically associated with psychomotor retardation, a condition marked by slowed neuromotor output that is behaviorally manifested as altered coordination and timing across multiple motor-based properties. Changes in motor outputs can be inferred from vocal acoustics and facial movements as individuals speak. We derive novel multi-scale correlation structure and timing feature sets from audio-based vocal features and video-based facial action units from recordings provided by the 4th International Audio/Video Emotion Challenge (AVEC). The feature sets enable detection of changes in coordination, movement, and timing of vocal and facial gestures that are potentially symptomatic of depression. Combining complementary features in Gaussian mixture model and extreme learning machine classifiers, our multivariate regression scheme predicts Beck depression inventory ratings on the AVEC test set with a root-mean-square error of 8.12 and mean absolute error of 6.31. Future work calls for continued study into detection of neurological disorders based on altered coordination and timing across audio and video modalities.

Categories and Subject Descriptors

G.3 [Mathematics of Computing]: Probability and Statistics−*correlation and regression analysis, time series analysis*
I.5.4 [Computer Methodologies]: Pattern Recognition−*signal processing*

Keywords

major depressive disorder, motor control, vocal biomarker, facial biomarker, incoordination and timing, correlation structure, Gaussian mixture model, extreme learning machine

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 [6]. As the number of people suffering from MDD steadily increases, so too does the burden of accurate diagnosis. Currently, the diagnosis of MDD requires a comprehensive assessment by a professional with significant clinical experience. However, the inter-clinician variability of these assessments makes the tracking of medication efficacy during clinical trials difficult. 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. An automated approach 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 [3, 7, 8, 21, 25–28, 34].

Figure 1 illustrates the categorization of vocal characteristics into three components: speech excitation (source), vocal tract (system), and pattern of stress and intonation (prosody). Depression-related changes in speech reflect the perception of qualities such as monotony, slur, slowness, hoarseness, and breathiness in the speech of depressed individuals. Hoarseness and breathiness may be associated with speech source characteristics (at the level of the vocal folds). Monotony may be

[1] This work is sponsored by the Assistant Secretary of Defense for Research & Engineering 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'14, November 7, 2014, Orlando, FL, USA.

http://dx.doi.org/10.1145/2661806.2661809

associated with prosody (e.g., modulation of speech rate, pitch, and energy) and slur with speech system characteristics (e.g., vocal tract articulators).

Characterizing the effects of depression on facial movements is an active research area. Early work found measurable differences between facial expressions of people suffering from MDD and facial expressions of non-depressed individuals [8]. EMG monitors can register facial expressions that are imperceptible during clinical assessment [9], and have found acute reductions in involuntary facial expressions in depressed persons [31]. The facial action coding system (FACS) quantifies localized changes in facial expression representing facial action units (FAUs) that correspond to distinct muscle movements of the face [5].

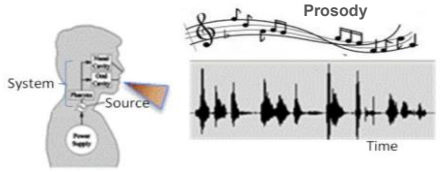

Figure 1. Illustration of speech source (at the vocal folds), system (vocal tract), and prosody (melody).

Although there has been significant effort in studying vocal and facial biomarkers for emotion classification, there has been little or no study into changes in coordination, movement, and timing using speech and facial modalities for depression classification or severity prediction. In individuals suffering from MDD, neurophysiological changes often alter motor control and thus affect mechanisms controlling speech production and facial expression. Clinically, these changes are typically associated with psychomotor retardation, a condition of slowed neuromotor output manifested in altered coordination and timing across multiple observables of acoustics and facial movements during speech.

Figure 2 displays a block diagram of the developed system for predicting depression severity from the Beck depression inventory (BDI) rating scale. Incorporating audio and video features reflecting manifestations of altered coordination and timing into a novel machine learning scheme is the focus of this study.

2. AUDIO/VIDEO DATABASE

The 2014 Audio/Video Emotion Challenge (AVEC) uses a depression corpus that includes audio and video recordings of subjects performing a human-computer interaction task [35]. Data were collected from 84 German subjects, with a subset of subjects recorded during multiple sessions: 31 subjects were recorded twice and 18 subjects were recorded three times. The subjects' age varied between 18 and 63 years, with a mean of 31.5 years and a standard deviation of 12.3 years.

Subjects performed two speech tasks in the German language: (1) reading a phonetically-balanced passage and (2) replying to a free-response question. The read passage (NW) was an excerpt of the fable *Die Sonne und der Wind* (*The North Wind and the Sun*). The free speech section (FS) asked the subjects to respond to one of a number of questions (prompted in written German), such as "What is your favorite dish?" "What was your best gift, and why?" and "Discuss a sad childhood memory." The NW and FS passages ranged in duration from 00:31 to 01:29 (mm:ss) and 00:06 to 03:50 (mm:ss), respectively.

Video of the subjects' face was captured using a webcam at 30 frames per second and a spatial resolution of 640 x 480 pixels. Audio was captured with a headset microphone connected to a laptop soundcard at sampling rates of 32 kHz or 48 kHz using the AAC codec. For each session, the self-reported BDI score was available. The recorded sessions were split into three partitions (training, development, and test) with 50 recordings in each set. We combined the training and development sets into a single 100-session data set, which is henceforth termed the Training set.

3. LOW-LEVEL FEATURE EXTRACTION

We exploit dynamic variation and inter-relationships across speech production systems by computing features that reflect complementary aspects of the speech source, system, and prosody. In the video domain, FAUs yield measures reflecting facial movements during speech that can contribute to depression characterization.

3.1 Voice Source Properties

Harmonics-to-noise ratio (HNR): A spectral measure of harmonics-to-noise ratio was performed using a periodic/noise decomposition method that employs a comb filter to extract the harmonic component of a signal [17–19]. This "pitch-scaled harmonic filter" approach uses an analysis window duration equal to an integer number of local periods (four in the current work) and relies on the property that harmonics of the fundamental frequency exist at specific frequency bins of the short-time discrete Fourier transform (DFT). In each window, after obtaining an estimate of the harmonic component, subtraction from the original spectrum yields the noise component, where interpolation fills in gaps in the residual noise spectrum. The time-domain signals of the harmonic and noise components in each frame are obtained by performing inverse DFTs of the respective spectra. Overlap-add synthesis is then used to merge together all the short-time segments. The short-time harmonics-to-noise ratio is the ratio, in dB, of the power of the decomposed harmonic signal and the power of the decomposed speech noise signal.

Cepstral peak prominence (CPP): Recent research has focused on developing improved acoustic measures that do not rely on an accurate estimate of fundamental frequency, as required for jitter and shimmer measures. Several studies have reported strong correlations between cepstral peak prominence (CPP) and overall dysphonia perception [4, 13, 23], breathiness [14, 15], and vocal fold kinematics. CPP is defined as the difference, in dB, between the magnitude of the highest peak and the noise floor in the power cepstrum for quefrencies greater than 2 ms (corresponding to a range minimally affected by vocal tract–related information) and was computed every 10 ms. .

3.2 Speech System Properties

Formant frequencies: We associate vocal tract resonance information with speech dynamics as a means to represent articulatory changes in the depressed voice. We have selected an algorithm based on the principle that formants are correlated with one another in both frequency and time domains [24, 30]. Formant frequencies are computed every 10 ms. Embedded in the algorithm is a voice-activity detector that allows a Kalman smoother to smoothly coast through non-speech regions. Because we are using only the frequencies of formants, these features are approximately immune to slowly-varying linear channel effects.

Mel frequency cepstral coefficients (MFCCs): To introduce vocal tract spectral magnitude information, we use standard MFCCs provided by AVEC. We also derived 16 corresponding delta MFCCs to 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).

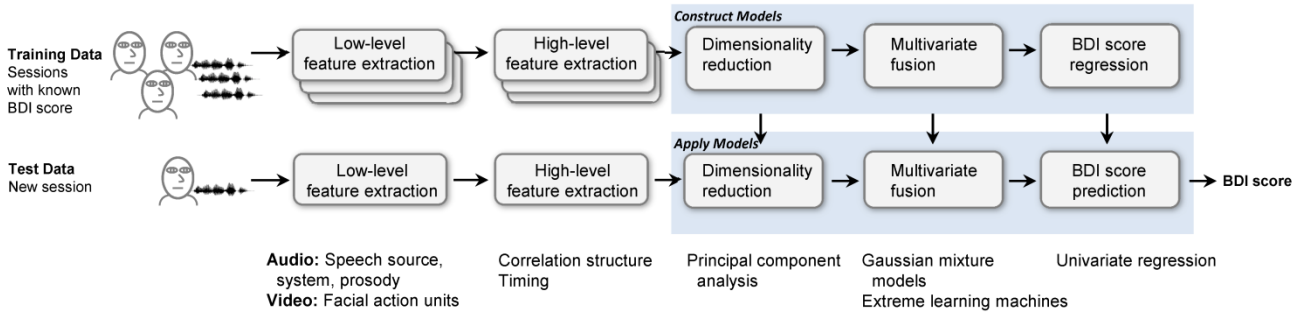

Figure 2. Block diagram of the developed system for predicting the Beck depression inventory (BDI) score.

3.3 Speech Prosody Properties

Phoneme durations: We have found that computing phoneme-specific characteristics, rather than average measures of speaking rate, reveals stronger relationships between speech rate and depression severity [34]. Using an automatic phoneme recognition algorithm [32], we detect phonetic boundaries and phoneme-specific durations that are associated with each instance of the 40 classes of defined phonetic speech units.

Pitch slopes: The fundamental frequency (pitch) was estimated using a time-domain autocorrelation method over 40 ms Hanning windows every 1 ms [2]. Within each phone segment, a linear fit was made to these pitch values, yielding a pitch slope feature (ΔHz/s) associated with each instance of phonetic speech units.

3.4 Facial Action Units (FAUs)

Although the FACS provides a formalized method for identifying changes in facial expression, its implementation for the analysis of large quantities of data has been impeded by the need for trained annotators to mark individual frames of a recorded video session. For this reason, the University of California San Diego has developed a computer expression recognition toolbox (CERT) for the automatic identification of FAUs from individual video frames [20].

Table 1 lists the FAUs output by CERT used for the video-based facial expression analysis. All frames marked as invalid by the program and values considered outliers were removed. In addition, each frame of data is retained only if it is marked valid across all 20 FAUs. If the duration of the remaining FAU time series was less than 30 s or 40% of their original length, the entire set of FAUs for that recording was not used. With this procedure, FAUs from five NW and 17 FS passages in the Training set and from three NW and five FS passages in the Test set were omitted from processing.

Each FAU feature was converted from a support vector machine (SVM) hyperplane distance to a posterior probability using a logistic model trained on a separate database of video recordings [22]. Henceforth, the term FAU refers to these frame-by-frame estimates of FAU posterior probabilities.

4. HIGH-LEVEL FEATURE EXTRACTION

Our high-level features are designed to characterize properties of coordination and timing from the low-level features. The measures of coordination use assessments of the multi-scale structure of correlations among the low-level features. This approach is motivated by the observation that auto- and cross-correlations of measured signals can reveal hidden parameters in the stochastic-dynamical systems that generate the time series. This multivariate feature construction approach—first introduced for analysis of EEG signals for epileptic seizure prediction [36, 37]—has since been successfully applied to speech analysis for estimating depression [33], the estimation of cognitive performance associated with dementia [39], and the detection of changes in cognitive performance associated with mild traumatic brain injury [11].

Channel-delay correlation and covariance matrices are computed from multiple time series channels (of given vocal and facial parameters). Each matrix contains correlation or covariance coefficients between the channels at multiple relative time delays. Changes over time in the coupling strengths among the channel signals cause changes in the eigenvalue spectra of the channel-delay matrices. The matrices are computed at four separate time scales, in which successive time delays correspond to frame spacings of 1, 3, 7, and 15. A detailed description of the cross-correlation approach can be found in [37]. Overall covariance power (logarithm of the trace) and entropy (logarithm of the determinant) are also extracted from the channel-delay covariance matrices at each scale.

For vocal-based timing features we use cumulative phoneme-dependent durations and pitch slopes, obtained using estimated phoneme boundaries. For facial-based timing features, we use FAU rates obtained from their estimated posterior probabilities.

4.1 Speech Correlation Structure

Figure 3 illustrates the cross-correlation (*xcorr*) technique applied to a formant time series. For each above time scale (10, 30, 70, 150 ms for a 10-ms frame), correlation coefficients are computed among signals shifted in time relative to each other, with 15 time-delays used per scale.

After investigating multiple combinations of the low-level vocal features as input to the *xcorr* analysis, we found the best overall

Table 1. The 20 facial action units from CERT.

#	Description	#	Description
1	Inner Brow Raise	11	Lip Stretch
2	Outer Brow Raise	12	Cheek Raise
3	Brow Lower	13	Lids Tight
4	Eye Widen	14	Lip Pucker
5	Nose Wrinkle	15	Lip Tightener
6	Lip Raise	16	Lip Presser
7	Lip Corner Pull	17	Lips Part
8	Dimpler	18	Jaw Drop
9	Lip Corner Depressor	19	Lips Suck
10	Chin Raise	20	Blink/Eye Closure

performance using the following three combinations: 1) Formant–CPP, 2) CPP–HNR, and 3) delta MFCC. Figures 4–6 show example results at a single time scale for each of the feature combinations. Two speech recordings of the NW passage

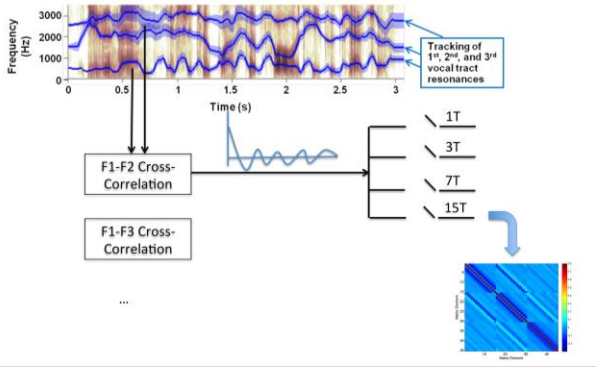

Figure 3. Diagram of cross-correlation analysis of articulatory coordination, as performed through formant-based features using channel-delay correlation matrices at multiple delay scales. A channel-delay matrix from one scale is shown.

illustrate the typical effect of depression on these *xcorr* channel-delay matrices and eigenspectra feature vectors. These recordings are of a non-depressed individual (BDI = 0) and a depressed individual (BDI = 35) from the Training set. The lower-left in each figure gives the eigenvalues for the normal and depressed subject cases, while the lower-right plot in each figure shows the mean normalized eigenvalues for all Training sessions grouped into four different BDI score ranges: 0–8 (blue), 9–19 (cyan), 19–28 (green), and 29–45 (red).

For Formant–CPP *xcorr* features, vectors consist of 248 elements (4 channels, 4 time scales, 15 delays per scale, and 2 covariance features per scale). For CPP–HNR *xcorr* features, vectors consist of 88 elements (2 channels, 4 scales, 15 delays per scale, top 20 eigenvalues per scale, and 2 covariance features per scale). For delta MFCC *xcorr* features, the vectors consist of 968 elements (16 channels, 4 scales, 15 delays per scale, and 2 covariance features per scale).

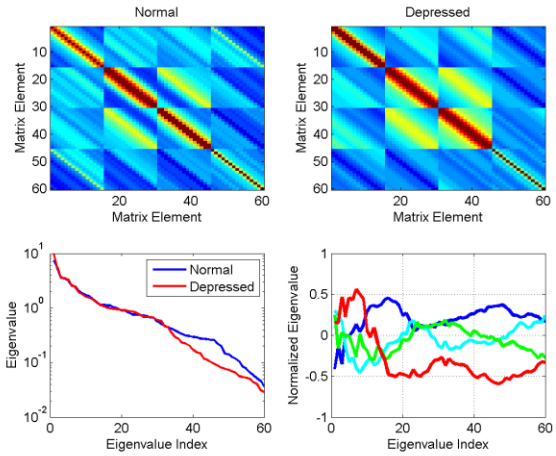

Figure 4. Formant–CPP *xcorr* features. Top: Channel-delay correlation matrices from NW passage for a normal and a depressed subject. Red denotes high and blue low (auto-) cross-correlation values. Bottom: Eigenvalues for these subjects (left) and average normalized eigenvalues for four BDI ranges in Training set (right).

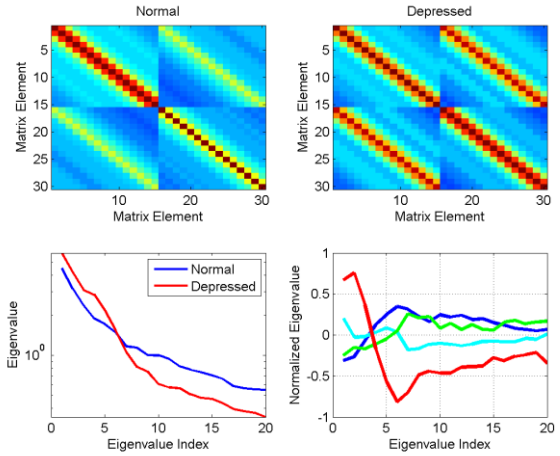

Figure 5. CPP–HNR *xcorr* features. Top: Channel-delay correlation matrices from NW passage for a normal and a depressed subject. Bottom: Eigenvalues for these subjects (left) and average normalized eigenvalues for four BDI ranges in Training set (right).

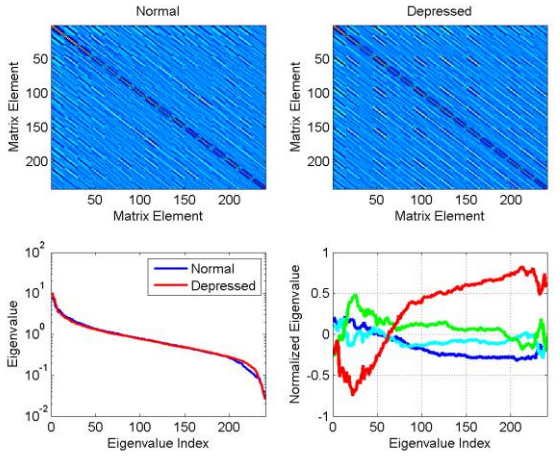

Figure 6. Delta MFCC *xcorr* features. Top: Channel-delay correlation matrices from NW passage for a normal and a depressed subject. Bottom: Eigenvalues for these subjects (left) and average normalized eigenvalues for four BDI ranges in Training set (right).

4.2 Facial Correlation Structure

Facial coordination features are obtained by applying the *xcorr* technique to the FAU time series using the same parameters that were used to analyze the vocal-based features. Because of the 30 Hz FAU frame rate, spacing for the four time scales correspond to time sampling in increments of approximately 33 ms, 100 ms, 234 ms, and 500 ms.

Figure 7 (top) shows example FAU channel-delay matrices at a single time scale from the same normal and depressed subjects that were used for illustration in Figures 4–6. These matrices are derived from the FS passage.

As with Formant–CPP and CPP–HNR *xcorr* features, Figure 7 (bottom-left) shows that the eigenspectra of the depressed subject contain less power in the small eigenvalues. This effect is observed across a spectrum of BDI scores in all 83 free-response

Training set recordings with valid FAU features. The facial-based eigenvalue differences are similar to those found in Formant–CPP and CPP–HNR *xcorr* features.

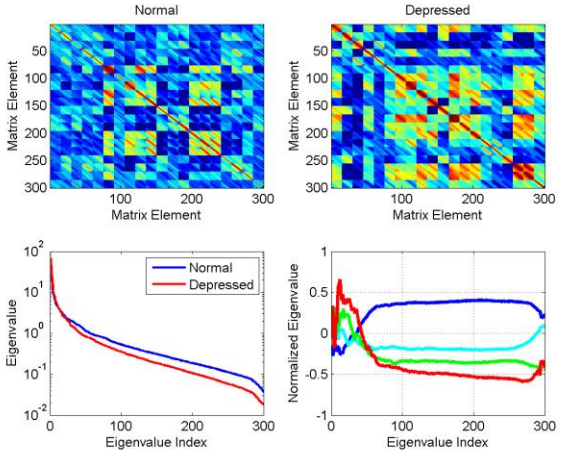

Figure 7. FAU *xcorr* features. Top: Channel-delay correlation matrices from FS passage for a normal and a depressed subject. Bottom: Eigenvalues for these subjects (left) and average normalized eigenvalues for four BDI ranges in Training set (right).

4.3 Phoneme Duration

Building on previous work [34], the summed durations of certain phonemes are linearly combined to yield fused phoneme duration measures. A subset of phonemes whose summed durations are highly correlated with BDI scores on the Training set are selected to create these fused measures, with weights based on the strength of their individual correlations. Table 2 lists the selected phonemes for the North Wind passage (left) and the first six of the ten selected phonemes for the Free Speech passage (right), along with their individual BDI correlations. The correlations of the fused measures for each passage are shown at the bottom.

The linear combination used to obtain the fused measures is as follows. For each phoneme category i, we denote d_i as the cumulative duration over a recording, R_i as its correlation with BDI, and w_i as its weight. Then, the fused duration measure is

$$\hat{d} = \sum_i w_i d_i, \qquad (1)$$

where

$$w_i = \text{sign}(R_i)/(1 - R_i^2). \qquad (2)$$

Equation (2) is a modification of the combination rule used in [34] that causes highly correlating phoneme durations to be weighted more strongly than weakly correlating phoneme durations. For the NW passage, d_i is the total duration of phoneme i. For the FS passage, d_i is the total phoneme duration divided by the total passage duration (speech only) and thus provides a rate measure (percent time present).

4.4 Pitch Slope

A fused phoneme-dependent pitch slope measure is obtained using essentially the same procedure as described above. For each phoneme, we compute the sum of *valid* pitch slopes across all instances of that phoneme. Invalid slopes are those with absolute value greater than eight, resulting in the exclusion of most slopes that are computed from discontinuous pitch contours.

Table 2. Correlation coefficients (R, $p < 0.01$) between fused phoneme durations and BDI scores in the Training set. Fusion is done using linear combinations of phoneme durations. Only 6 of the 10 Free Speech phonemes are shown.

North Wind		Free Speech	
Phoneme	***R***	**Phoneme**	***R***
'l'	0.50	'ng'	0.38
'ah'	0.45	't'	0.34
'n'	0.41	'hh'	0.33
'ih'	0.34	'ey'	0.32
'b'	0.34	'ow'	0.28
'ow'	0.34	'er'	0.27
Fused	**0.54**	**Fused**	**0.57**

For each passage, the set of phonemes with the highest correlating summed pitch slopes are then selected. The summed pitch slopes are combined using equations (1) and (2) to obtain fused measures for the NW and FS passages. Using 20 phonemes for NW and 15 phonemes for FS, these fused measures have BDI correlations of $\boldsymbol{R = 0.63}$ (NW) and $\boldsymbol{R = 0.51}$ (FS).

4.5 Facial Activation Rate

The facial activation rate feature is obtained by computing mean FAU values (an estimate of percent time present via posteriori probabilities) over each passage and combining several of these into a fused FAU rate measure. Weights are based on FAU correlations with BDI scores using the combination rule in Equations 1 and 2. Table 3 shows the FAUs used in this fusion process, their individual correlations, and the correlations of the fused measures.

4.6 Dimensionality Reduction

The *xcorr* feature vectors typically contain highly correlated elements. To obtain lower-dimensional uncorrelated feature vectors for machine learning techniques, we apply principal component analysis. Table 4 lists the number of principal components we chose for each *xcorr* feature type, along with phonetic and FAU rate features. The number of principal components in each case was determined empirically by cross-validation performance.

Table 3. Correlation coefficient (R) between mean FAU posterior probabilities and BDI in the Training set ($p < 0.05$ for all $|R| \geq 0.21$). Fusion is done using linear combinations of the mean FAU posterior probabilities. See Table 1 for FAU descriptions.

North Wind		Free Speech	
FAU #	***R***	**FAU #**	***R***
3	0.24	2	−0.16
4	−0.26	3	0.22
5	0.21	4	−0.27
7	−0.30	5	0.19
8	−0.28	8	−0.18
10	−0.23	9	0.17
11	−0.24	11	−0.15
14	0.27	12	0.14
15	−0.30	13	0.16
18	0.37	15	−0.22
Fused	**0.58**	**Fused**	**0.46**

5. MULTIVARIATE FUSION AND PREDICTION

Our next step involves mapping the features described in Section 4 into univariate scores that can be easily mapped into BDI predictions. To do this, we use both generative Gaussian mixture models (GMMs), which have been widely used for automatic speaker recognition [29] and have recently been extended to vocal-based depression classification [12, 33, 38], and discriminative extreme learning machines (ELMs), a single layer feedforward neural network architecture with randomly assigned hidden nodes [10, 16].

Table 4. Total number of dimensions (# Dim.) and number of dimensions selected after principal component analysis (PCA #) for each of the eight features sets.

Feature Set	Data	Feature Type	# Dim.	PCA #
1	NW	Formant–CPP *xcorr*	248	4
	NW	CPP–HNR *xcorr*	88	2
	NW	Delta MFCC *xcorr*	968	5
2	NW	Phoneme duration	1	1
3	NW	Pitch slope	1	1
4	NW	FAU rate	1	1
5	FS	FAU *xcorr*	1208	6
6	FS	Phoneme rate	1	1
7	FS	Pitch slope	1	1
8	FS	FAU rate	1	1

5.1 Gaussian Mixture Model

Gaussian staircase: To train the generative GMMs, we utilize the *Gaussian staircase* approach in which each GMM is comprised of an ensemble of Gaussian classifiers [38]. The ensemble is derived from six partitions of the training data into different ranges of depression severity for low (Class 1) and high (Class 2) depression. Given a BDI range of 0 to 45, the Class 1 ranges for the six Gaussian classifiers are: 0−4, 0−10, 0−17, 0−23, 0−30, and 0−36, with the Class 2 ranges being the complement of these. The Gaussian classifiers comprise a single, highly regularized GMM classifier, with feature densities that smoothly increase in the direction of decreasing (Class 1) or of increasing (Class 2) levels of depression. Additional regularization of the densities is obtained by adding 0.1 to the diagonal elements of the normalized covariance matrices.

Subject-based adaptation: Individual variability in the relationships between features and BDI are partially accounted for within the GMMs using Gaussian-mean subject-based adaptation. If one or more sessions in the Training set have the same subject ID as the Test subject and are in the same BDI-based partition, the mean of the Gaussian for that partition is assigned to the mean of the data from that subject only, rather than the mean of the data from all subjects within the partition [38].

5.2 Extreme Learning Machine

ELMs are used to provide a complementary discriminative approach for predicting depression level. The ELM is a feedforward neural network with a single hidden layer, in which the weights and biases of the nodes are randomly assigned. The number of hidden nodes and the activation function were empirically selected, and ridge regression output was used to map the transformed feature space into depression scores. This was done by solving an L^2-norm regularized least squares problem. Feature Set 1 for the ELM uses a hidden layer with 395 nodes and a hyperbolic tangent activation function, while Feature Set 2 uses a network with 62 nodes and an inverse triangular basis activation function. These values were selected empirically.

The ELM was chosen in place of more traditional multilayer perceptron or deep neural network architectures for the advantages it provides with the AVEC audio-visual data. Due to the limited number of Training sessions, and the noisiness of multidimensional maps between features and BDI scores, gradient descent learning algorithms typically converge to highly suboptimal solutions. The use of a least squares constraint in the ELM avoids such a stepwise iterative procedure, and instead uses a matrix inverse operator to directly solve for the output mapping. An additional benefit with ELM is that it tends to learn output weights with a small norm, thereby providing better generalization performance according to Bartlett's theory [1].

5.3 Predictors

Table 4 lists the eight Feature Sets used in the prediction system. A separate GMM classifier is used for each Feature Set, outputting a log-likelihood ratio score for Class 1 (Normal) and Class 2 (Depressed). Separate ELM classifiers are used for Feature Sets 1 and 2.

Initial BDI predictions are obtained from three Predictors, which use different combinations of the eight Feature Sets and two types of classifiers (Table 5). Within each Predictor, the classifier outputs from Feature Sets are summed together. Following this, a univariate regression model is created from the Training set and applied to the classifier output from the Test data. The resulting univariate regression output is the intitial BDI score prediction from each Predictor.

For Predictors 1 and 2, subject-based adaptation is then applied to adjust this initial prediction by correcting for consistent biases seen in the BDI Training set predictions of the same subject. If there are any Training sessions from a given Test subject, then the average Training set error from that subject is used to adjust the prediction, as follows. Let N denote the number of repeat sessions in the Training set, and let y_j and $\hat{y}_j$ indicate the true and predicted BDI score, respectively, for the j^{th} repeat Training session. Then, the prediction is adjusted using the following equation, which contains empirically derived parameters:

$$\tilde{z} = \hat{z} + 0.9\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)/\left(0.1+N\right). \quad (3)$$

5.4 Fusing Predictors

The outputs of the three Predictors are fused to create a final BDI prediction, using weights based on each Predictor's accuracy, quantified by BDI correlations, in predicting BDI scores on the Training set:

$$w_i = R_i^2/\left(1-R_i^2\right). \quad (4)$$

For Predictor 2 (Feature Sets 3−8), only those Training and Test sessions that contain valid data from all of its Feature Sets are used, resulting in 80 valid Training sessions and 45 valid Test sessions. In developing our prediction system, we obtain separate estimates of performance for novel Test subjects and for repeat Test subjects, who have sessions in the Training set. This is done using *non-repeat subjects* and *repeat subjects* cross-validation evaluation. With *non-repeat subjects* evaluation, the system is trained only on subjects other than the one being tested. With *repeat subjects* evaluation, the system is tested only on subjects who have multiple sessions in the Training set, and sessions from those subjects are included in training.

Intermediate prediction: The BDI correlations in Equation (4) are computed separately from *non-repeat subjects* and *repeat subjects* evaluations, resulting in *non-repeat subject* weights for Predictors 1 and 2 of $w_1 = 0.36, w_2 = 0.59,$ and *repeat subject* weights of $w_1 = 1.68, w_2 = 1.33$. For each Test session, the appropriate weights are applied to the outputs of Predictors 1 and 2, and the weighted sum is then normalized by the sum of the weights. Nineteen of 50 Test sessions contain repeat subjects.

Final prediction: The procedure described above for obtaining a fused output from Predictors 1 and 2 is now repeated, to fuse with the output from Predictor 3. The *non-repeat subject* weights for this fusion are $w_{1,2} = 0.68, w_3 = 0.49$, and the *repeat-subject* weights are $w_{1,2} = 1.64, w_3 = 0.72$.

6. RESULTS

The prediction system described above was used for our best submission in the AVEC 2014 Challenge, with test RMSE = 8.12 and MAE = 6.31. These results are an improvement on the winning submission in the AVEC 2013 competition, which was test RMSE = 8.50 and MAE = 6.52. Last year's result was obtained using a read passage (*Homo Faber*) that was much longer than the NW passage made available this year. Our introduction this year of new vocal and facial features, as well as improved machine learning techniques, has resulted in improved performance despite the relative lack of data in this year's AVEC challenge. In our five submissions for AVEC 2014 we investigated various Feature Set and Predictor combinations, resulting in test RMSE values of (in order): 8.71, 9.08, 8.12, 8.36, and 8.27.

Table 5. Three Predictors consisting of different combinations of Feature Sets and Classifiers.

Feature Sets	Classifier	# Train	# Test	Regression Order
1: 1-2	GMM	100	50	4
2: 3-8	GMM	80	45	4
3: 1-2	ELM	100	50	3

7. DISCUSSION

In addition to the introduction of several novel feature combinations, this work is technically significant because it demonstrates benefits of shifting from basic low-level features to high-level features that characterize and emphasize interactions and timing. These benefits may be due to the information inherent in a holistic analysis in which neurocognitive changes are manifested not just in subsystems of expression but in the degraded coordination among the subsystems. The benefits may also be due to increased robustness to channel effects because of little direct dependence on speech spectral magnitude. Clinically, the ability to achieve a low test RMSE has important implications for automatic and frequent at home monitoring to assess patient state and quickly adapt treatment.

Our future work will emphasize three analysis branches. The first branch is continued analysis of the vocal and facial modalities. For vocal analysis, we are considering more sophisticated versions of prosodic characterization that jointly consider pitch and intensity. For facial analysis, we are considering the emotional state classification outputs provided by CERT. Both feature sets may provide additional insight into the arousal and valence of subjects. Second, we will continue to improve feature fusion methodologies to handle individually noisy or missing data. Third, we seek a unified neurocognitive model which links the observed features to mechanistic changes in the brain that correspond to neural circuit changes associated with depression. This 'holy-grail' of neurocognitive research has the potential to relate the feature sets we have identified to neurological substrates.

8. CONCLUSION

In summary, we have presented a multimodal analysis pipeline that exploits complementary information in audio and video signals for estimating depression severity. We investigated how speech source, system, and prosody features, along with facial action unit features, correlate with MDD using the AVEC 2014 database, consisting of a read passage and free-response speech segment from subjects with varying depression levels according to their self-reported Beck depression inventory assessment.

Specifically, we selected speech features that reflect movement and coordination in articulation from formant frequencies and delta mel-cepstra, aspects of the voice source including degree of source irregularity (CPP and HNR), and changes in phonetic durations and pitch slopes as properties of prosody. We explored how certain facial expression–based features correlated with depression severity by showing the importance of coordination across FAUs. These coordination measures were obtained from auto- and cross-correlations of the multichannel speech and video signals. We also obtained fused phoneme duration features, and applied similar fusion techniques to obtain novel pitch slope and FAU rate features. Finally, combining GMM classifiers created from a Gaussian staircase training procedure with ELM classifiers, we achieved a test RMSE of 8.12.

9. ACKNOWLEDGMENTS

The authors thank Dr. Nicolas Malyska, Dr. Charles Dagli, and Bea Yu of MIT Lincoln Laboratory for their contributions to software development required in phoneme-based rate and facial action unit features.

10. REFERENCES

[1] Bartlett, P.L. 1998. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. Information Theory, IEEE Transactions on. 44, 2 (1998), 525–536.

[2] Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences. 17, (1993), 97–110.

[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] Dejonckere, P. and Lebacq, J. 1996. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. ORL. 58, 6 (1996), 326–332.

[5] Ekman, P., Freisen, W.V. and Ancoli, S. 1980. Facial signs of emotional experience. Journal of personality and social psychology. 39, 6 (1980), 1125.

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

[7] 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.

[8] 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.

[9] Greden, J.F. and Carroll, B.J. 1981. Psychomotor function in affective disorders: An overview of new monitoring techniques. The American journal of psychiatry. (1981).
[10] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding and Rui Zhang 2012. Extreme Learning Machine for Regression and Multiclass Classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 42, 2 (Apr. 2012), 513–529.
[11] Helfer, B.S., Quatieri, T.F., Williamson, J.R., Keyes, L., Evans, B., Greene, W.N., Vian, T., Lacirignola, J., Shenk, T., Talavage, T., Palmer, J. and Heaton, K. 2014. Articulatory Dynamics and Coordination in Classifying Cognitive Change with Preclinical mTBI. Interspeech (2014).
[12] Helfer, B.S., Quatieri, T.F., Williamson, J.R., Mehta, D.D., Horwitz, R. and Yu, B. 2013. Classification of depression state based on articulatory precision. (2013).
[13] Heman-Ackah, Y.D., Heuer, R.J., Michael, D.D., Ostrowski, R., Horman, M., Baroody, M.M., Hillenbrand, J. and Sataloff, R.T. 2003. Cepstral peak prominence: a more reliable measure of dysphonia. Annals of Otology Rhinology and Laryngology. 112, 4 (2003), 324–333.
[14] Heman-Ackah, Y.D., Michael, D.D. and Goding Jr, G.S. 2002. The relationship between cepstral peak prominence and selected parameters of dysphonia. Journal of Voice. 16, 1 (2002), 20–27.
[15] Hillenbrand, J. and Houde, R.A. 1996. Acoustic Correlates of Breathy Vocal QualityDysphonic Voices and Continuous Speech. Journal of Speech, Language, and Hearing Research. 39, 2 (1996), 311–321.
[16] Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K. 2006. Extreme learning machine: Theory and applications. Neurocomputing. 70, 1-3 (Dec. 2006), 489–501.
[17] Jackson, P.J. and Shadle, C.H. 2000. Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. The Journal of the Acoustical Society of America. 108, 4 (2000), 1421–1434.
[18] Jackson, P.J. and Shadle, C.H. 2000. Performance of the pitch-scaled harmonic filter and applications in speech analysis. Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on (2000), 1311–1314.
[19] Jackson, P.J. and Shadle, C.H. 2001. Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. Speech and Audio Processing, IEEE Transactions on. 9, 7 (2001), 713–726.
[20] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. and Bartlett, M. 2011. The computer expression recognition toolbox (CERT). Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on (2011), 298–305.
[21] Low, L.-S., Maddage, M., Lech, M., Sheeber, L. and Allen, N. 2010. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on (2010), 5154–5157.
[22] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (2010), 94–101.
[23] Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N. and De Bodt, M. 2010. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. Journal of voice. 24, 5 (2010), 540–555.
[24] Mehta, D.D., Rudoy, D. and 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 (Sep. 2012), 1732–46.
[25] Moore, E., Clements, M., Peifer, J. and Weisser, L. 2003. Analysis of prosodic variation in speech for clinical depression. Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE (2003), 2925–2928.
[26] Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K. and Geralts, D.S. 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of neurolinguistics. 20, 1 (Jan. 2007), 50–64.
[27] Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K. and Wilkes, D.M. 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. Biomedical Engineering, IEEE Transactions on. 51, 9 (2004), 1530–1540.
[28] Quatieri, T.F. and Malyska, N. 2012. Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity. Interspeech (2012).
[29] Reynolds, D.A., Quatieri, T.F. and Dunn, R.B. 2000. Speaker verification using adapted Gaussian mixture models. Digital signal processing. 10, 1 (2000), 19–41.
[30] Rudoy, D., Spendley, D.N. and Wolfe, P.J. 2007. Conditionally linear Gaussian models for estimating vocal tract resonances. INTERSPEECH (2007), 526–529.
[31] Schwartz, G.E., Fair, P.L., Salt, P., Mandel, M.R. and Klerman, G.L. 1976. Facial expression and imagery in depression: an electromyographic study. Psychosomatic medicine. 38, 5 (1976), 337–47.
[32] Shen, W., White, C.M. and Hazen, T.J. 2009. A comparison of query-by-example methods for spoken term detection. DTIC Document.
[33] Sturim, D.E., Torres-Carrasquillo, P.A., Quatieri, T.F., Malyska, N. and McCree, A. 2011. Automatic Detection of Depression in Speech Using Gaussian Mixture Modeling with Factor Analysis. Interspeech (2011), 2981–2984.
[34] Trevino, A.C., Quatieri, T.F. and Malyska, N. 2011. Phonologically-based biomarkers for major depressive disorder. EURASIP Journal on Advances in Signal Processing. 2011, 1 (2011), 1–18.
[35] Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R. and Pantic, M. 2013. AVEC 2014–3D Dimensional Affect and Depression Recognition Challenge. (2013).
[36] Williamson, J.R., Bliss, D.W. and Browne, D.W. 2011. Epileptic seizure prediction using the spatiotemporal correlation structure of intracranial EEG. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (2011), 665–668.
[37] Williamson, J.R., Bliss, D.W., Browne, D.W. and Narayanan, J.T. 2012. Seizure prediction using EEG spatiotemporal correlation structure. Epilepsy & Behavior. 25, 2 (2012), 230–238.
[38] Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B. and Mehta, D.D. 2013. Vocal biomarkers of depression based on motor incoordination. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (2013), 41–48.
[39] Yu, B., Quatieri, T.F., Williamson, J.R. and Mundt, J.C. 2014. Prediction of cognitive performance in an animal fluency task based on rate and articulatory markers. Interspeech (2014).