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  
jw@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
In individuals with major depressive disorder, neuropsychological 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 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].

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 frequencies 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).
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>
<thead>
<tr>
<th>#</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Inner Brow Raise</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>Outer Brow Raise</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>Brow Lower</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>Eye Widen</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>Nose Wrinkle</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>Lip Raise</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>Lip Corner Pull</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>Dimpler</td>
<td>18</td>
</tr>
<tr>
<td>9</td>
<td>Lip Corner Depressor</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>Chin Raise</td>
<td>20</td>
</tr>
</tbody>
</table>
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...
Training set recordings with valid FAU features. The facial-based eigenvalue differences are similar to those found in Formant–CPP and CPP–HNR\textit{xcorr} features.

Figure 7. FAU\textit{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, $R$ as its correlation with BDI, and $w_i$ as its weight. Then, the fused duration measure is

$$d = \sum w_id_i,$$  \hfill (1)

where

$$w_i = \frac{\text{sign}(R)\sqrt{1 - R^2}}.$$

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.

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>North Wind</th>
<th>Free Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘l’</td>
<td>0.50</td>
<td>‘ng’ 0.38</td>
</tr>
<tr>
<td>‘ah’</td>
<td>0.45</td>
<td>‘t’ 0.34</td>
</tr>
<tr>
<td>‘n’</td>
<td>0.41</td>
<td>‘hh’ 0.33</td>
</tr>
<tr>
<td>‘ih’</td>
<td>0.34</td>
<td>‘ey’ 0.32</td>
</tr>
<tr>
<td>‘b’</td>
<td>0.34</td>
<td>‘ow’ 0.28</td>
</tr>
<tr>
<td>‘ow’</td>
<td>0.34</td>
<td>‘er’ 0.27</td>
</tr>
<tr>
<td>Fused</td>
<td>0.54</td>
<td>Fused 0.57</td>
</tr>
</tbody>
</table>

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 $R = 0.63$ (NW) and $R = 0.51$ (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\textit{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\textit{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.

<table>
<thead>
<tr>
<th>FAU #</th>
<th>North Wind</th>
<th>Free Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.24</td>
<td>2 –0.16</td>
</tr>
<tr>
<td>4</td>
<td>0.26</td>
<td>3 0.22</td>
</tr>
<tr>
<td>5</td>
<td>0.21</td>
<td>4 −0.27</td>
</tr>
<tr>
<td>7</td>
<td>−0.30</td>
<td>5 0.19</td>
</tr>
<tr>
<td>8</td>
<td>−0.28</td>
<td>8 −0.18</td>
</tr>
<tr>
<td>10</td>
<td>−0.23</td>
<td>9 0.17</td>
</tr>
<tr>
<td>11</td>
<td>−0.24</td>
<td>11 −0.15</td>
</tr>
<tr>
<td>14</td>
<td>0.27</td>
<td>12 0.14</td>
</tr>
<tr>
<td>15</td>
<td>−0.30</td>
<td>13 0.16</td>
</tr>
<tr>
<td>18</td>
<td>0.37</td>
<td>15 −0.22</td>
</tr>
<tr>
<td>Fused</td>
<td>0.58</td>
<td>Fused 0.46</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Data</th>
<th>Feature Type</th>
<th># Dim.</th>
<th>PCA #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NW</td>
<td>Formant−CPP xc ...</td>
<td>248</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>NW</td>
<td>CPP−HNR xc ...</td>
<td>88</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>NW</td>
<td>Delta MFCC xc ...</td>
<td>968</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>NW</td>
<td>Phonomv duration</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>NW</td>
<td>Pitch slope</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>NW</td>
<td>FAU rate</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>FS</td>
<td>FAU xc ...</td>
<td>1208</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>FS</td>
<td>Phonomv rate</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>FS</td>
<td>Pitch slope</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>FS</td>
<td>FAU rate</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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 $\ell^2$-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 initial 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 $\bar{y}_j$ and $\bar{y}'_j$ 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:

$$
\tilde{z} = \tilde{z} + 0.9 \sum_{j=1}^{N} \left( \bar{y}_j - \bar{y}'_j \right) / (0.1 + N)
$$

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:

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

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.

<table>
<thead>
<tr>
<th>Feature Sets</th>
<th>Classifier</th>
<th># Train</th>
<th># Test</th>
<th>Regression Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: 1-2</td>
<td>GMM</td>
<td>100</td>
<td>50</td>
<td>4</td>
</tr>
<tr>
<td>2: 3-8</td>
<td>GMM</td>
<td>80</td>
<td>45</td>
<td>4</td>
</tr>
<tr>
<td>3: 1-2</td>
<td>ELM</td>
<td>100</td>
<td>50</td>
<td>3</td>
</tr>
</tbody>
</table>

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
autoregressive moving average modeling and inference for


