Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 55 (2019) 40–56

www.elsevier.com/locate/csl

Tracking depression severity from audio and video based on speech articulatory coordination☆,☆☆

James R. Williamson[a,*], Diana Young[b], Andrew A. Nierenberg[c,d], James Niemi[b], Brian S. Helfer[a], Thomas F. Quatieri[a]

[a] *MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02420*

[b] *Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA 02138*

[c] *Dauten Family Center for Bipolar Treatment Innovation, Massachusetts General Hospital, 50 Staniford Street, Suite 580, Boston, MA 02114*

[d] *Harvard Medical School, 25 Shattuck St, Boston, MA 02115*

Received 14 December 2017; received in revised form 7 August 2018; accepted 9 August 2018

Available online 17 August 2018

Abstract

The ability to track depression severity over time using passive sensing of speech would enable frequent and inexpensive monitoring, allowing rapid assessment of treatment efficacy as well as improved long term care of individuals at high risk for depression. In this paper an algorithm is proposed that estimates the articulatory coordination of speech from audio and video signals, and uses these coordination features to learn a prediction model to track depression severity with treatment. In addition, the algorithm is able to adapt its prediction model to an individual's baseline data in order to improve tracking accuracy. The algorithm is evaluated on two data sets. The first is the Wyss Institute Biomarkers for Depression (WIBD) multi-modal data set, which includes audio and video speech recordings. The second data set was collected by Mundt et al (2007) and contains audio speech recordings only. The data sets are comprised of patients undergoing treatment for depression as well as control subjects. In its within-subject tracking of clinical Hamilton depression (HAM-D) ratings, the algorithm achieves root mean squared error (RMSE) of 5.49 with Spearman correlation of $r = 0.63$ on the WIBD data set, and achieves RMSE = 5.99 with $r = 0.48$ on the Mundt data set.

Keywords: Depression; Speech; Articulation; Coordination; Audio; Video

☆ This paper has been recommended for acceptance by Roger Moore.

☆☆ DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited. This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

The authors wish to thank Dr. James Mundt for the use of his data collection.

* Corresponding author.

E-mail address: jrw@ll.mit.edu (J.R. Williamson).

https://doi.org/10.1016/j.csl.2018.08.004

0885-2308/

1. Introduction

In recent years there has been an upsurge of interest in the ability to automatically estimate depression from speech (Cummins et al., 2015a). Most research in this area has focused on depression screening: the ability to classify or estimate depression severity in a test subject based on statistical models trained on a separate population. Another practical application for automated speech analysis in depression is within-subject longitudinal tracking. This would enable short-term monitoring of the efficacy of pharmaceutical and therapeutic interventions as well as long term monitoring of individuals in high risk populations.

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 patientâs mental condition and emotional state (Darby et al., 1984; Dejonckere and Lebacq, 1996; Ekman et al., 1980; Fava and Kendler, 2000; France et al., 2000; Gaebel and Wölwer, 1992; Low et al., 2010; Moore et al., 2003; Mundt et al., 2007; Ozdas et al., 2004; Quatieri and Malyska, 2012; Trevino et al., 2011). The group of vocal features observed to change with a patient's mental condition and emotional state is motivated by perception of monotony, hoarseness, breathiness, glottalization, and slur in the voice of a depressed subject. Vocal characteristics studied include prosody (e.g., fundamental frequency and speaking rate), spectral features, glottal (vocal fold) excitation flow patterns, timing jitter, amplitude shimmer, and "spirantization", a measure that reflects aspirated leakage at the vocal folds. Interest in vocal biomarkers is exemplified by the growing number of depression challenges such as the AVEC challenges (Valstar et al., 2014). Although not always consistent across studies, many statistical relationships of vocal features with depression have been found, and in some cases applied towards developing automatic classifiers. Discrepancies across studies are due to differences in patient population (no two studies have used the same population), small data sets, different forms of depression ("agitated" and "slowed"), and differences in signal processing methods for vocal feature extraction (Darby et al., 1984; Dejonckere and Lebacq, 1996; Ekman et al., 1980; Fava and Kendler, 2000; France et al., 2000). An exhaustive study of the state-of-the-art in vocal biomarkers for depression is given in Cummins et al. (2015a).

Characterizing the effects of depression on facial movements is also an active research area. Early work found measurable differences between facial expressions of people suffering from major depressive disorder (MDD) and facial expressions of non-depressed individuals (Gaebel and Wölwer, 1992). EMG monitors can register facial expressions that are imperceptible during clinical assessment (Greden and Carroll, 1981) and have found acute reductions in involuntary facial expressions in depressed persons (Schwartz et al., 1976). 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 (Ekman et al., 1980).

There have been numerous efforts in tracking depression over time with therapy and treatment. For example, physiological and behavioral (including speech) measures have been used in monitoring depression state through electro-convulsive therapy, with subjects wearing continuous ambulatory technology (Sung et al., 2005). Fairly accurate predictions of state were found during treatment over a three-week period. Other examples of tracking depression during treatment, as quantified by HAM-D assessments, have used changes in fundamental frequency in the early work of Nilsonne et al. (1988), and syllabic rate in the more recent work of Sahu and Espy-Wilson (2015).

In this paper, we also address longitudinal tracking of depression severity with treatment. Our hypothesis in this voice and face study is that neurophysiological change in depression generally affects motor coordination, including articulatory control and dynamics (Williamson et al., 2013; 2014). The tracking algorithm uses measures of speech articulatory coordination from both audio and video modalities to estimate depression levels. The speech articulatory coordination features are designed based on the hypothesis that depression often leads to psychomotor retardation, which reduces coordination in complex motor systems such as speech (Quatieri and Malyska, 2012; Quatieri et al., 2017; Mundt et al., 2007; Association et al., 2013; France et al., 2000; Low et al., 2010; Moore et al., 2003; Ozdas et al., 2004; Caligiuri and Ellwanger, 2000). The prediction model can be individualized using one or more baseline sessions from a test subject in which both speech recordings and depression ratings are available. The algorithm is then able to estimate depression severity and changes in severity ratings based on subsequent speech recordings alone.

Table 1
Speech recording types and durations from the two evaluated data sets.

Data set	Speech	Modality	Mean (s)	Min (s)	Max (s)
WIBD	Read	Audio & Video	49.4	32.0	70.8
WIBD	Free	Video	707.9	42.2	1,772.4
Mundt	Read	Audio	49.4	32.9	60.6
Mundt	Free	Audio	153.7	43.6	392.1

In this paper the algorithm is evaluated on two datasets that consist primarily of patients undergoing pharmacological and/or psychotherapeutic treatment for depression. In both datasets, clinical depression scores are obtained for subjects in multiple sessions spaced two or more weeks apart. The algorithm uses a statistical prediction model trained from other subjects in the same dataset, as well as a single baseline session from the test subject, to predict the depression scores in subsequent sessions. The depression data sets are described in Section 2, the depression tracking algorithm is described in Section 3, and results are presented in Section 4.

2. Data collection

Table 1 shows the types of recorded speech, signal modalities, and recording duration statistics from the WIBD and Mundt data sets. These data sets are described in detail below. The Hamilton Depression Rating Scale (HAM-D) was the clinical depression scale used in both studies (Hedlund and Vieweg, 1979). The total HAM-D score is the sum of the subscores from 17 questions. Eight of these items are scored on a 5-point scale (0–4) and nine are scored on a 3-point scale (0–2). Table 2 illustrates clinical interpretations of different ranges of HAM-D total scores.

2.1. WIBD data set

The Wyss Institute Biomarkers for Depression (WIBD) multi-modal data set was collected at Massachusetts General Hospital to support the development of technology for combining multiple sensing modalities for assessing depression severity. Written informed consent was obtained in accordance with required federal and institutional guidelines, including Institutional Review Board oversight. In an initial screening session, medication (10–20 mg/day of escitalopram) was prescribed, which was to be started immediately following the first baseline speech recording session. In addition to the baseline speech recording session, there were up to four additional followup sessions, which were spaced every two weeks, and in which read speech (Grandfather passage) and free speech (clinical interview to determine HAM-D rating) were recorded. Control subjects had their speech recorded twice, in the first session and again after six weeks. In all, the data set contains data from 12 depressed subjects with 46 usable sessions of speech recordings, and six control subjects with 10 usable sessions of speech recordings. Each session consists of the recordings of read speech followed by free speech. The subject pool is evenly balanced between men and women.

Audio recordings were obtained using a lapel microphone (DPA 4061). Video recordings were obtained using a Canon XH A1S HD video camera at a conventional standard definition (SD) NTSC 720 × 480, with a frame rate of

Table 2
HAM-D total scores and interpretations.

HAM-D total score	Interpretation
0–7	Normal
8–13	Mild depression
14–18	Moderate depression
19–22	Severe depression
≥ 23	Very severe depression

30 frames-per-second. The camera was placed directly in front of the subject at a distance of about 6-feet. A blue screen was positioned behind the subject in order to facilitate subsequent image analysis.

The data used for analysis were the Grandfather passage (audio and video) and the free speech interview (video only). Audio was not used from the free speech interview due to the fact that it contains audio signals from the clinician, and the focus of the current analysis is on the subject's speech only.

2.2. Mundt data set

This data set was originally collected by Mundt et al. (2007) for a depression-severity study involving interactive voice recording technology. The study consists of 35 physician-referred subjects (20 women and 15 men, mean age 41.8 years). The subjects were predominately Caucasian (88.6%), with four subjects being of other descent. The subjects had all recently started on pharmacotherapy and/or psychotherapy for depression and continued treatment over the 6-week-assessment period. This study includes clinical HAM-D ratings and speech recordings (sampled at 8 kHz) in four sessions at weeks 0, 2, 4, and 6. The recordings consist of read speech (the Grandfather passage) and free response speech. More details of the collection process are given in Mundt et al. (2007).

3. Depression tracking algorithm

The depression tracking algorithm proposed in this paper is based on the depression estimation algorithms used in the MIT Lincoln Laboratory systems in the AVEC 2013 and AVEC 2014 Depression sub-challenges (Williamson et al., 2013; 2014). The MIT Lincoln Laboratory systems took first place in both challenges, obtaining the highest accuracy among all competitors in predicting Beck depression inventory (BDI) ratings. The competitions involved 100 training sessions of read and free speech, recorded in audio and video modalities, used to learn a statistical model for predicting BDI scores on a held out test set of 50 sessions. Using audio only, the 2013 system achieved RMSE = 8.50 with Spearman correlations $r = 0.65$ on the held-out test data set. Using both audio and video, but with shorter speech recordings available, the 2014 system achieved RMSE = 8.11 with $r = 0.71$ on the held-out test data. For comparison, baseline RMSE values of 11.49 and 11.60, respectively, are obtained by predicting mean BDI scores on the held-out test data sets.

The algorithm used in this paper incorporates the subset of the audio- and video-based features from the MIT Lincoln Laboratory systems that are designed to characterize articulatory coordination. The articulatory coordination features are designed to capture levels of dynamical interrelations among the multiple articulators involved in speech production. This feature approach is premised on the hypothesis that a prominent symptom of depression is psychomotor retardation, which reduces the magnitude, complexity, and coordination of articulator movements during speech (Quatieri and Malyska, 2012; Quatieri et al., 2017; Mundt et al., 2007; Association et al., 2013; France et al., 2000; Low et al., 2010; Moore et al., 2003; Ozdas et al., 2004; Caligiuri and Ellwanger, 2000).

More specifically, the articulatory coordination features are derived from the eigenspectra of correlation and covariance matrices, which in turn are constructed from the time series of audio formant frequencies and video facial action units (FAUs), with expanded dimensionality based on time delay embedding. The feature extraction 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 approach was first introduced for analysis of spatiotemporal coordination among EEG signals for predicting epileptic seizures (Williamson et al., 2011; 2012).

3.1. Low-level features

Audio formant frequencies. Vocal tract resonances contain information about speech dynamics over time that are related to depressed speech (Cummins et al., 2015b; Scherer et al., 2015; 2016). A formant tracking algorithm based on Kalman filtering was used to obtain smooth estimates of the first three resonant frequencies over time (Mehta et al., 2012). Formant frequencies are extracted every 10 ms from the audio signal. Embedded in the formant tracking algorithm is a voice-activity detector that allows a Kalman smoother to smoothly coast through non-speech

regions. Post processing was also applied, in which estimates of the third formant frequency that are above a threshold of 4.5 kHz are truncated.

On the WIBD data set, the audio signal was successfully obtained from the read Grandfather passage in 55 out of the total number of 56 sessions. On the Mundt data set, the audio signal was successfully obtained in the read passage and free speech passage in all 132 sessions.

Video facial action units. The facial action coding system (FACS) provides a formalized method for identifying changes in facial expression. However, its implementation for the analysis of large quantities of data is impractical due to the need for trained annotators to mark individual frames of a recorded video session. For this reason, the University of California San Diego developed a computer expression recognition toolbox (CERT) for the automatic identification of facial action units (FAUs) from individual video frames (Littlewort et al., 2011). CERT extracts FAUs independently from each video frame at every 33 ms. The FAUs output by CERT represent the following 20 localized expressions: inner brow raise, outer brow raise, brow lower, eye widen, nose wrinkle, lip raise, lip corner pull, dimpler, lip corner depressor, chin raise, lip stretch, cheek raise, lids tight, lip pucker, lip tightener, lip pressor, lips part, jaw drop, lips suck, and blink/eye closure.

Post-processing was also applied, in which FAU feature vectors containing *NaN*s were marked as invalid and removed. If the duration of the remaining FAU time series was less than either 30 s or 40% of the original recording length, then the entire set of FAUs for that recording was not used. Using this procedure on the WIBD data set, FAUs from the read passage were available in 52 out of 56 sessions, and FAUs from the free speech passage were available in 54 out of 56 sessions. Each of the 20 FAU features was further converted from a support vector machine (SVM) hyperplane distance to a posterior probability using a logistic model, which was trained on a separate database of video recordings (Lucey et al., 2010). Henceforth, the term FAU refers to these frame-by-frame estimates of FAU posterior probabilities. Additionally, an outlier removal procedure applied a 2nd-order Kalman filter to each of the 20 FAU time series to compute a smoothed, local average representation. Any frame in which one or more vector element deviated by more than 0.3 from the smoothed signal was then automatically removed.

3.2. High-level features

High-level articulatory coordination features are extracted from the low-level formant and FAU time series data. In this feature approach, channel-delay correlation and covariance matrices are computed from low-level multichannel signals using time-delay embedding at multiple different delay scales. Then, the features consist of the eigenspectra from the correlation matrices and summary statistics from the covariance matrices. Specifically, a channel-delay correlation matrix at delay scale j is computed as

$$\mathbf{R_j} = \begin{bmatrix} \begin{bmatrix} r_{1,1}(j) & \cdots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \cdots & r_{N,N}(j) \end{bmatrix}_{1,1} & \cdots & \begin{bmatrix} r_{1,1}(j) & \cdots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \cdots & r_{N,N}(j) \end{bmatrix}_{1,M} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} r_{1,1}(j) & \cdots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \cdots & r_{N,N}(j) \end{bmatrix}_{M,1} & \cdots & \begin{bmatrix} r_{1,1}(j) & \cdots & r_{1,N}(j) \\ \vdots & \ddots & \vdots \\ r_{N,1}(j) & \cdots & r_{N,N}(j) \end{bmatrix}_{M,M} \end{bmatrix} \tag{1}$$

where M is the number of channels, N is the number of delays per channel, and the matrix consists of correlation coefficients between channels m_1 and m_2 at time delays n_1 and n_2 and at delay scale j. The correlation coefficients are defined as

$$[r_{n_1,n_2}(j)]_{m_1,m_2} = \frac{\sum_{i=1}^{n} [\mathbf{x}_{\mathbf{m_1}}(i-n_1\delta_j) - \overline{\mathbf{x}}_{\mathbf{m_1}}][\mathbf{x}_{\mathbf{m_2}}(i-n_2\delta_j) - \overline{\mathbf{x}}_{\mathbf{m_2}}]}{\sqrt{[\mathbf{x}_{\mathbf{m_1}}(i-n_1\delta_j) - \overline{\mathbf{x}}_{\mathbf{m_1}}]^2 [\mathbf{x}_{\mathbf{m_2}}(i-n_2\delta_j) - \overline{\mathbf{x}}_{\mathbf{m_2}}]^2}}, \tag{2}$$

where δ_j is the delay spacing. For audio, $M = 3$, which is the number of formant features. For video, $M = 20$, which is the number of FAU features. The articulatory coordination features are the rank-ordered eigenvalues of the correlation matrix, $\mathbf{R_j}$,

$$\lambda_{\mathbf{j}} = \text{eig}(\mathbf{R_j}). \tag{3}$$

In addition, a covariance matrix $\mathbf{C_j}$ is constructed in the same way as R_j, except the matrix elements are covariance coefficients rather than correlation coefficients,

$$[c_{m_1,m_2}(j)]_{n_1,n_2} = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{x_{m_1}}(i-n_1\delta_j)-\overline{\mathbf{x}}_{\mathbf{m_1}}][\mathbf{x_{m_2}}(i-n_2\delta_j)-\overline{\mathbf{x}}_{\mathbf{m_2}}]. \tag{4}$$

From the $\mathbf{C_j}$ matrix, two summary-statistic features are computed, which are related to the total power and the entropy, respectively, of the set of time-embedded signals,

$$\rho_j = \log(\text{tr}(\mathbf{C_j})), \tag{5}$$

$$h_j = \log(|\mathbf{C_j}|). \tag{6}$$

Eqs. (1)–(6) are implemented at four different delay scales, $j = 1, 2, 3, 4$, using delay spacings of $\{\delta_1, \delta_2, \delta_3, \delta_4\} = \{1, 3, 7, 15\}$. The low-level formant features are sampled every 10 ms, so the four scales involve delays in the formant signals of 10 ms, 30 ms, 70 ms, and 150 ms increments, respectively. The low-level FAU features are sampled every 33.33 ms, so the four scales involve delays in the FAU signals of 33 ms, 100 ms, 233 ms, and 500 ms increments.

3.3. Feature examples

Next, formant-based examples from two read speech sessions (Grandfather passage) are used to illustrate the properties of the high-level articulatory coordination features. The formant tracks, shown in Fig. 1, are obtained from two successive sessions of a subject in the WIBD data set. These sessions were chosen due to a remarkably large improvement in HAM-D total score between the two sessions, from HAM-D = 23 to HAM-D = 2. Notice in Fig. 1 that there are no obvious differences between the formant tracks in the two sessions, except that F_3 has a higher average frequency, with larger and slower fluctuations, in the HAM-D = 23 session.

Fig. 2 (top) shows the channel-delay correlation matrices, described in Eq. (1), that are obtained from these two examples at the 3*rd* delay scale. The matrices contain nine 45×45 element submatrices, each of which contains correlation coefficients between two pairs of formants channels. The F_1 to F_1 correlations are in the upper left submatrix. In the 3*rd* delay scale, each successive submatrix element (moving horizontally or vertically) is a correlation coefficient obtained using a relative time delay that is shifted seven frames (70 ms) with respect to the time shift used to compute the correlation coefficient in the previous matrix element. A prominent difference between the two matrices in Fig. 2 is that there is more variability in the low HAM-D matrix, across time delays, in the submatrices containing correlations between F_1 and F_2 and between F_2 and F_3.

The articulatory coordination features are the matrix eigenvalues, shown in Fig. 2 (bottom). The eigenvalues represent the rank-ordered sizes of the principal axes of the 45-dimensional time-embedded scatter distribution represented by each correlation matrix. Notice that the low-rank eigenvalues from the low HAM-D session are larger than those from the high HAM-D session. This means that lower depression is associated with a time-embedded formant scatter distribution that is more isotropic, which can be thought of as more "complex".

Therefore, the formant articulatory coordination features have some similarities to feature approaches that capture a tightening of the formant vowel space associated with depression (Cummins et al., 2015b; Scherer et al., 2015; 2016). The nature of formant space tightening captured by the articulatory coordination features is a reduction in lower rank eigenvalues, which corresponds to a flattening of the multivariate time-embedded formant scatter distribution. This representation is invariant to the orientation of the scatter distribution with respect to the coordinate axes in the time delay embedding space. Another difference between this approach and typical formant vowel space approaches is the inclusion of F_3. In previous unpublished analyses on the AVEC data set, we have found that F_3 provides additional discriminative information, although it is not as

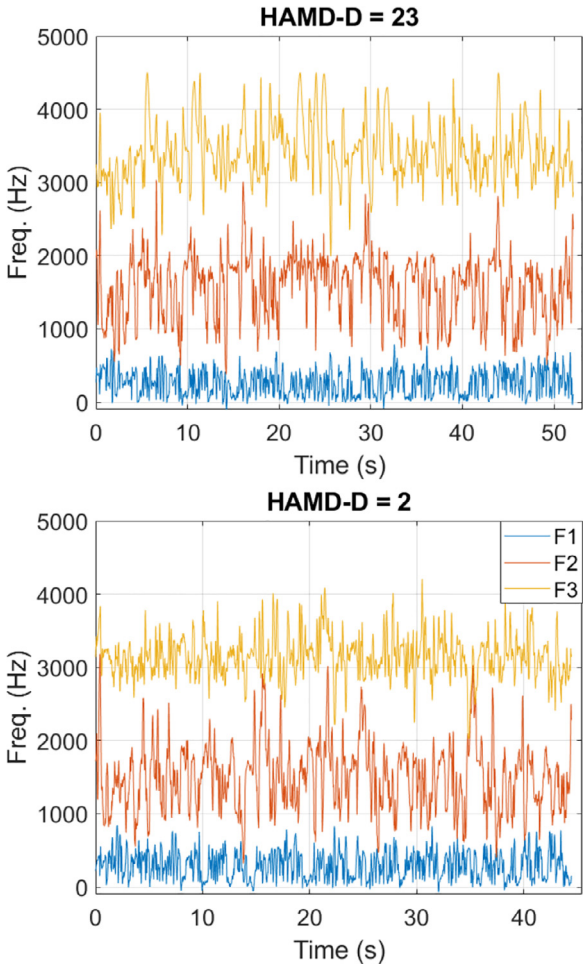

Fig. 1. Formant tracks for read speech (Grandfather passage) in two successive sessions from the same subject in the WIBD data set.

important as F_1 or F_2 for discriminating depression. It is possible that an extension of this approach to a larger number of resonant frequencies could be useful.

3.4. Machine learning

The same parameter values and machine learning methods were used in extracting and processing features in both modalities (audio and video), on both data sets, and on both types of speech (read and free speech) within each data set.

Dimensionality reduction. The same dimensionality reduction procedure is used for all feature vectors. Each feature vector is the set of articulatory coordination features obtained from a speech recording in the audio or video modality and concatenated across the four delay scales. Specifically, a feature vector for the ith observation, v_i, is obtained by concatenating the features from Eqs. (3), (5), and (6) across the four delay scales,

$$\mathbf{v_i} = [\lambda_{\mathbf{1}}, \lambda_{\mathbf{2}}, \lambda_{\mathbf{3}}, \lambda_{\mathbf{4}}, \rho_1, \rho_2, \rho_3, \rho_4, h_1, h_2, h_3, h_4]_i. \tag{7}$$

Prior to dimensionality reduction, the elements of the $\mathbf{v_i}$ feature vector are z-scored across all subjects and sessions in the training set so that all elements are weighted equally. Dimensionality reduction using principal components analysis (PCA) is then performed to produce the same number of components ($K = 8$) from all feature vectors. This

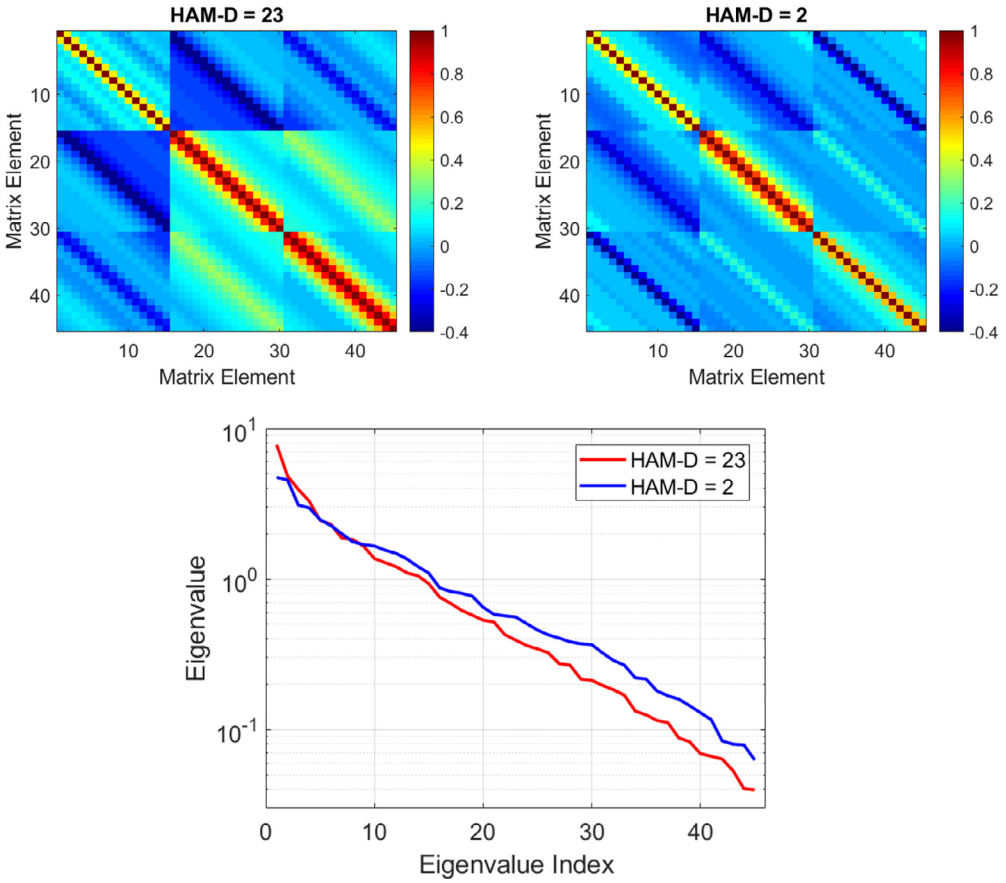

Fig. 2. Top: Time-delay correlation matrices at the 3*rd* delay scale, obtained from the formant tracks in two successive sessions (see Fig. 1). Bottom: Eigenvalues, obtained from the two correlation matrices, are ordered from largest to smallest. The low rank eigenvalues are larger for low depression (blue).

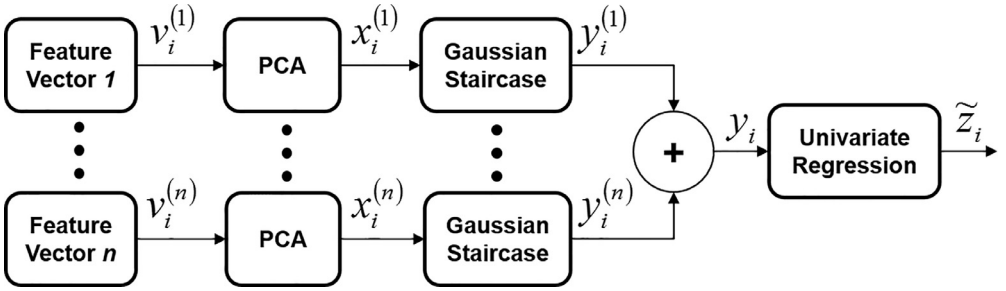

Fig. 3. Diagram showing processing steps involved in transforming multiple feature vectors from a single test observation *i* into a HAM-D prediction.

number of principal components explains 83–88% of the variance on the formant-based articulatory features in the read speech and free speech passages of the WIBD and Mundt data sets, and 96–97% of the variance of the FAU-based articulatory features in the read speech and free speech passages of the WIBD data set. The principal components are then z-scored before being input into the regression system. All of these transforms are computed from the training set and applied to the test set.

Table 3
Partitions used by the staircase regression algorithm to create multiple statistical models for each output class on the WIBD and Mundt data sets.

	WIBD Partitions		Mundt Partitions	
Index	Class 1	Class 2	Class 1	Class 2
1	0-2	3–50	0–8	9–50
2	0–6	7–50	0–11	12–50
3	0–10	11–50	0–14	15–50
4	0–13	14–50	0–17	18–50
5	0–17	18–50	0–20	21–50
6	0–21	22–50	0–23	24–50

Staircase regression. Fig. 3 depicts the processing flow that converts articulatory coordination feature vectors from multiple feature vectors into a HAM-D total score prediction. Machine learning is used to form statistical models in the Gaussian staircase modules and in the univariate regression module as described below.

A staircase regression algorithm is used to map the PCA features into HAM-D total score predictions (Williamson et al., 2013; 2014). This algorithm uses multiple partitions of the outcome variable (HAM-D scores) to define an ensemble of *lower outcome* statistical models on one side of each partition (Class 1), and *higher outcome* models on the other side (Class 2). By partitioning the outcome variable into multiple staircase levels, then constructing statistical models at each level, and then combining them into a composite model, additional flexibility is obtained in adjusting to statistical changes in feature-to-outcome mappings at different ranges of the outcome variable (Cummins et al., 2017).

The staircase approach is compatible with the use of any reasonable statistical model at each level. In this paper, multivariate Gaussian models are used to reduce the risk of overfitting the small data sets. Evenly spaced partitions on the outocme variables (HAM-D scores) are used, such that there is a roughly equal number of observations (i.e., subject sessions) between each partition delimiter. These partitions are listed in Table 3 for the WIBD and Mundt data sets. A training observation's output value (HAM-D score), z_i, determines which set of partitions it maps into. Each observation typically maps into multiple different partitions from both Class 1 and Class 2.

The statistical models are formed from training data as follows. Let $P_{j,\,k}$ represent the jth partition of the kth class ($k = 1$ or 2). Let each data observation i contain a feature vector $\mathbf{x_i}$ and an associated output value z_i (i.e., the HAM-D score). Then, the sample mean and covariance for partition j of class k are

$$\mu_{\mathbf{j},\mathbf{k}} = \frac{1}{n_{j,k}} \sum_{i:z_i \in P_{j,k}} \mathbf{x_i}, \tag{8}$$

$$\mathbf{C}_{\mathbf{j},\mathbf{k}} = \frac{1}{n_{j,k}-1} \sum_{i:z_i \in P_{j,k}} (\mathbf{x_i}-\mu_{\mathbf{j},\mathbf{k}})(\mathbf{x_i}-\mu_{\mathbf{j},\mathbf{k}})', \tag{9}$$

where

$$n_{j,k} = \sum_{i:z_i \in P_{j,k}} 1. \tag{10}$$

There is also a subject-dependent adaptation step, in which the sample mean is moved toward the mean of those training vectors that are from the same subject and that also belonged to the same staircase partition. This adaptation step is modulated by the relevance parameter α,

$$\widehat{\mu}_{\mathbf{i},\mathbf{j},\mathbf{k}} = \frac{\sum_{i':z_{i'} \in P_{j,k}} \delta(s_i - s_{i'})\mathbf{x_{i'}} + \alpha\mu_{\mathbf{j},\mathbf{k}}}{\sum_{i':z_{i'} \in P_{j,k}} \delta(s_i - s_{i'}) + \alpha}, \tag{11}$$

Table 4
Parameters used by the depression tracking algorithm on the WIBD and Mundt data sets. For the last two parameters, α and σ, candidate values considered in a 2-D grid search (using nested cross-validation) are listed.

Parameter	Value	Eq.	Description
M	3 or 20	(1)	# channels
N	15	(1)	# delays per channel
$\{\delta_1, \delta_2, \delta_3, \delta_4\}$	{1, 3, 7, 15}	(2), (4)	delay spacings per scale
K	8	—	# PCA features per set
L	6	(14)	# staircase levels
α	1,2,4,8, or 16	(11)	relevance
σ	1,2,4,8, or 16	(12)	regularization

where s_i is the subject index for observation i. Smaller values of α result in a higher relative weighting of the feature mean, obtained from training data of the same subject, and concomitantly a lower relative weighting of the sample mean from training data of other subjects. Notice that, due to the experimental design used in this paper, there is never more than one training observation from a test subject in the training set.

There is also a regularization parameter, σ, which spherically inflates the sample covariance distribution and which is useful for buffering against statistical fluctuations in small datasets,

$$\widehat{\mathbf{C}}_{\mathbf{j,k}} = \sigma^2\mathbf{I} + \mathbf{C}_{\mathbf{j,k}}. \tag{12}$$

Finally, the likelihood for observation i conditioned on partition j of class k is obtained using

$$p_{j,k}(x_i) = \frac{1}{(2\pi)^{D/2}|\widehat{\mathbf{C}}_{\mathbf{j,k}}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x_i}-\widehat{\mu}_{\mathbf{i,j,k}})\widehat{\mathbf{C}}_{\mathbf{j,k}}^{-1}(\mathbf{x_i}-\widehat{\mu}_{\mathbf{i,j,k}})'\right\}, \tag{13}$$

and the total likelihood for observation i conditioned on class k is obtained by summing across the staircase partitions,

$$p_k(x_i) = \sum_{j=1}^{L} p_{j,k}(x_i), \tag{14}$$

for $k = 1, 2$, and where $L = 6$ is the number of partitions (see Table 3 for details on the partitions).

The two-class log-likelihood ratio,

$$y_i = \log\Big(p_2(x_i)\Big) - \log\Big(p_1(x_i)\Big), \tag{15}$$

is mapped into an outcome prediction, $\widetilde{z}_i$, using a 2nd-order least squares regression model. The regression model is constructed from the log-likelihood ratios obtained on the training set. Note that the subject-dependent mean adaptation step in Eq. (11) is used to adjust the likelihoods of test data, but not of training data.

Feature vectors are fused by summing their log-likelihood scores prior to univariate regression. For each session, there are three feature vectors in the WIBD data set, one audio-based and one video-based feature vector from read speech, and one video-based feature vector from free speech. On each session of the Mundt data set there is one audio-based feature vector from read speech and one audio-based feature vector from free speech.

Cross-validation methodology. The same set of parameters was used on both data sets, as summarized in Table 4. Values for two of the parameters were selected independently within each cross-validation fold using a nested leave-one-subject-out cross-validation selection procedure, applied solely to the training data. This was done to adjust, in an unbiased way, to the unique statistical properties of each training set. The two parameters are the relevance parameter α from Eq. (11) and the regularization parameter σ from Eq. (12). The values for these parameters were chosen jointly in a two dimensional grid search from the values [1, 2, 4, 8, 16] for each parameter. The criterion for selecting the pair of parameter values was minimization of the average longitudinal tracking prediction RMSE on

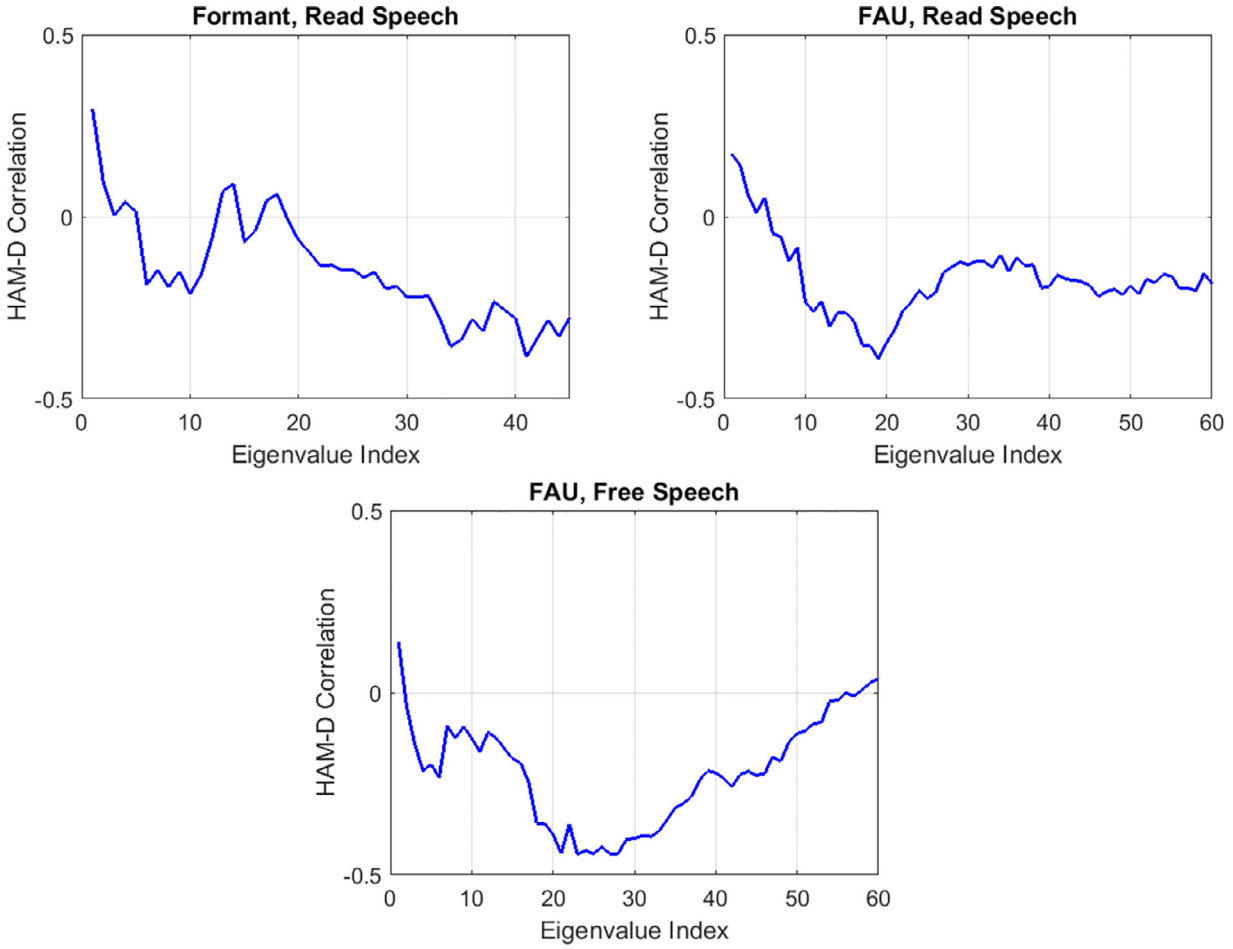

Fig. 4. The Spearman correlations of formant and FAU eigenvalue features with HAM-D scores on the WIBD data set is shown as a function of eigenvalue rank (largest to smallest).

held-out validation sets within each training set. These parameters were selected using fusion of all available feature vectors.

The most commonly chosen values of σ and α within the nested cross-validations were as follows. For the WIBD data set, $\sigma = 16$ was selected 100% of the time and $\alpha = 16$ was selected 82% of the time, with $\alpha = 8$ and $\alpha = 4$ each selected 9% of the time. For the Mundt data set, $\sigma = 1$ was selected 48% of the time and $\sigma = 2$ was selected 52% of the time, with $\alpha = 1$ selected 100% of the time. Table 4 summarizes the common set of parameters used in the high-level feature extraction and machine learning components of the algorithm.

4. Results

4.1. Feature distributions

In previous work, a characteristic signature of depression has been found in audio- and video-based articulatory coordination features (rank-ordered eigenvalues), in which depression severity is negatively correlated with lower rank eigenvalues (Williamson et al., 2013; 2014). See Fig. 2 for an example of this signature. A similar signature has also been found associated with other neurological disorders such as Parkinsonâs disease, mTBI, and age-related cognitive decline (Williamson et al., 2015; Smith et al., 2017; Helfer et al., 2014; Yu et al., 2014). The magnitude of low rank eigenvalues is an indicator of the dimensionality of the time-embedded feature space, with larger values in the low rank eigenvalues indicating greater complexity of articulatory coordination.

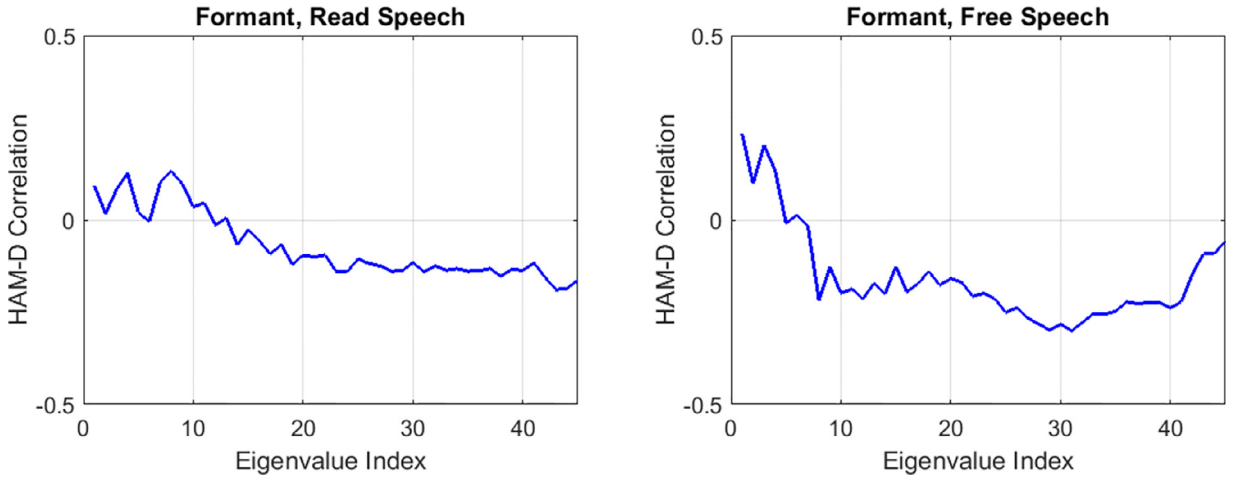

Fig. 5. The Spearman correlations of formant and FAU eigenvalue features with HAM-D scores on the Mundt data set is shown as a function of eigenvalue rank (largest to smallest).

Figs. 4 and 5 indicate the degree to which these signatures were found in the WIBD and Mundt datasets by plotting the Spearman correlations between the eigenvalue features and the HAM-D scores. For the formant features, the correlations are plotted for the third delay scale. For the FAU features, the correlations of the first 60 eigenvalues are plotted for the fourth delay scale.

Fig. 4 shows that, on the WIBD dataset, the depression signature from the audio formants is mostly consistent with the expected pattern, with low rank eigenvalues negatively correlated with HAM-D, except for a small unexpected reversal in eigenvalues in ranks 12–18. The video FAU eigenvalues show strong agreement with the expected pattern, with particularly large negative correlations found for the free speech FAUs. Fig. 5 shows the depression signatures obtained from the Mundt data set. The read speech and free speech signatures are both consistent with the expected pattern, with larger negative correlations in the lower rank eigenvalues found for free speech. Overall, the results indicate a strong agreement with, and replication of, the previously discovered depression signatures in articulatory coordination features.

4.2. Tracking of depression severity

The WIBD data set comprises 56 sessions with recorded speech data from 12 depressed and six control subjects. The read speech audio is available in 55 sessions, the read speech video in 52 sessions, and the interview free speech

Table 5
The accuracy of longitudinal tracking of HAM-D scores on the WIBD data set is shown for three feature vectors individually, and in combination. Results are shown both with and without individualization from a baseline session. Accuracy is quantified based on RMSE and Spearman correlation. The best performance of RMSE = 5.49 and $r = 0.63$ is obtained with individualization by combining all three feature vectors. In comparison, the benchmark approach of predicting static HAM-D scores as baseline results in RMSE = 9.96.

		Individualization		No individualization	
Features	Speech Type	RMSE	r (p)	RMSE	r (p)
Formant	Read	7.21	0.21 (0.20)	7.37	0.13 (0.43)
FAU	Read	6.54	0.26 (0.14)	6.88	0.22 (0.21)
FAU	Interview	6.18	0.60 (0.00)	5.92	0.60 (0.00)
Combined	Combined	5.49	0.63 (0.00)	5.52	0.62 (0.00)

Table 6
The accuracy of longitudinal tracking of HAM-D scores on the Mundt data set is shown for formant features on read speech and on free speech. Results are shown both with and without individualization from a baseline session. The best performance of RMSE = 5.99 and $r = 0.48$ is obtained with individualization by combining the two feature vectors. In comparison, the benchmark approach of predicting static HAM-D scores as baseline results in RMSE = 6.89.

		Individualization		No Individualization	
Features	Speech Type	RMSE	r (p)	RMSE	r (p)
Formant	Read	6.26	0.31 (0.00)	7.16	−0.05 (0.62)
Formant	Free	6.30	0.34 (0.00)	6.90	0.03 (0.75)
Combined	Combined	5.99	0.48 (0.00)	7.05	0.01 (0.95)

video in 54 sessions. All 56 sessions contain at least two out of the three feature vectors, i.e., formant features from read speech, FAU features from read speech video, and FAU features from free speech video. The baseline sessions for the 12 depressed and six control subjects are used to create individualized prediction models, via Eq. (11). The algorithm's HAM-D predictions on 38 followup sessions (34 for depressed subjects and four for control subjects) are then evaluated by comparing them to actual clinical ratings for those sessions. The root mean square difference of HAM-D scores on followup sessions compared to the baseline is 10.50 for depressed subjects and 2.23 for control subjects.

The accuracy of WIBD HAM-D predictions are quantified in Table 5 with RMSE and Spearman correlation values. Results are shown based on each feature vector individually as well as based on the three feature vectors combined by summing log-likelihoods from Eq. (15). For each of these conditions, results are also shown with and without the availability of a baseline session for training and individualization. The video feature vectors are more effective in terms of prediction accuracy than the audio feature vector, with the most effective individual feature vector being the video from the free speech interview. Individualization via Eq. (11) provides a relatively small benefit. The best overall result of RMSE = 5.49 with $r = 0.63$ is obtained from the combined system with individualization. A benchmark comparison is to use each subject's baseline HAM-D score as the HAM-D prediction in followup sessions. This benchmark produces a much lower accuracy of RMSE = 9.96. This shows that the use of speech to track depression is far more effective than relying on prior knowledge.

The accuracy of HAM-D predictions on the Mundt data set is shown in Table 6. The Mundt data set contains 97 followup sessions, in which the root mean square difference of HAM-D scores compared to baseline is 6.89. Results are based on each feature vector individually as well as on the two feature vectors (fused via log-likelihood summation), both with and without individualization using the test subject's baseline session. On this data set, individualization is essential for producing good performance, with the best overall result of RMSE = 5.99 and $r = 0.48$ obtained from the combined system with individualization. Once again, this is an improvement over the benchmark constant-depression approach, which yields RMSE=6.89.

Fig. 6 summarizes the best tracking results on the two data sets by showing scatter plots of predicted HAM-D scores as a function of true HAM-D scores on the follow up sessions. These are shown for the WIBD data set (left) and for the Mundt data set (right), with linear regression fits shown in red. These plotted results are based on the combined feature vectors (bottom rows of Tables 5 and 6).

In addition to results plotted at the group level, it is also instructive to observe tracking of HAM-D scores per individual over time. To illustrate this, prediction results are plotted in Fig. 7 for the nine depressed subjects from the WIBD data set who have four or more sessions each. For each of these subjects, the baseline and follow up true HAM-D scores are plotted in blue, and the follow up predicted HAM-D scores are plotted in red. Subjects 1 and 2 are medication non-responders whose HAM-D scores remain high throughout. The algorithm's predictions show a lack of improvement that is similar to the true HAM-D scores. Subjects 3 through 6 are successful medication responders whose HAM-D scores monotonically reduce over time. The algorithm tracks these changes reasonably well. Subject 7 is an interesting case of a successful responder who suffers a severe but temporary relapse in session 4. Sessions 4 and 5 for this subject are the two cases that are illustrated in Figs. 1 and 2 for the formant-based features. The algorithm does remarkably well in tracking these nonmonotonic changes in HAM-D scores. Finally,

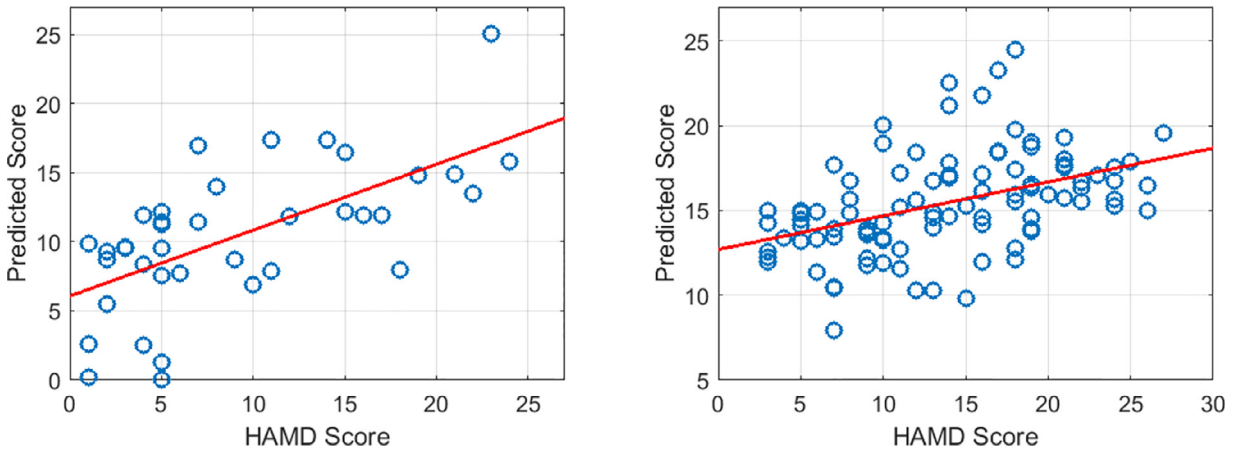

Fig. 6. Predicted HAM-D scores are plotted as a function of true HAM-D scores on the WIBD data set (left; $r = 0.63$) and the Mundt data set (right; $r = 0.48$).

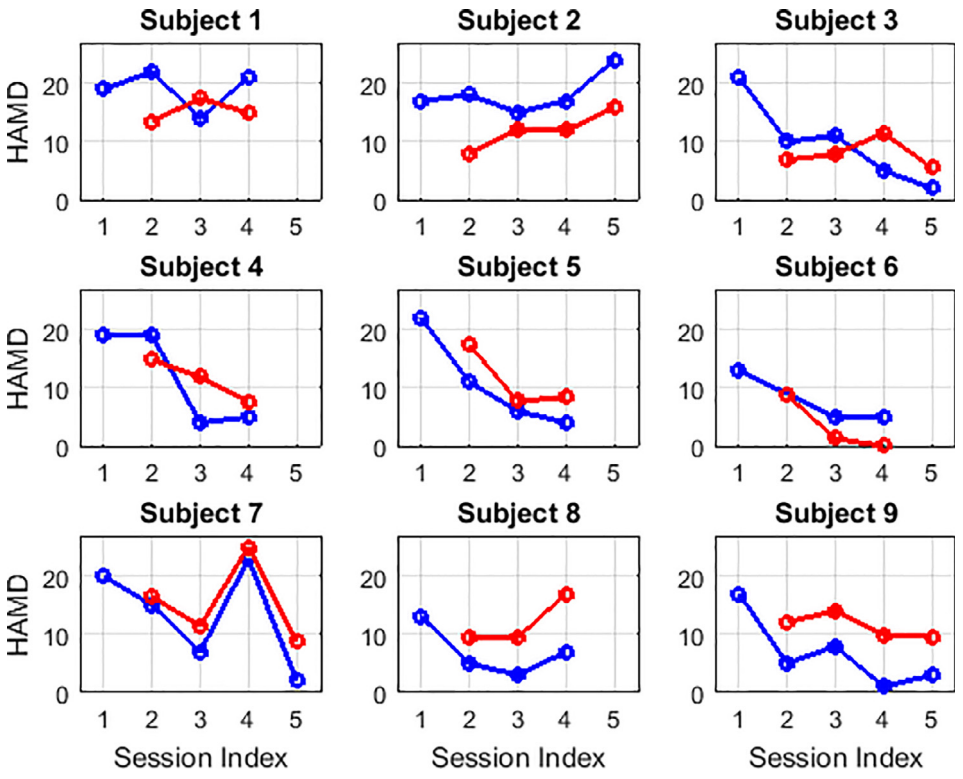

Fig. 7. For the nine subjects in the WIBD data set who have four or more sessions, longitudinal tracking of HAM-D scores is shown. True HAM-D scores are plotted in blue, and predicted scores in red.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

subjects 8 and 9 are responders whom the algorithm tracks well in a relative sense, but somewhat poorly in an absolute sense, by consistently overestimating depression severity.

5. Discussion

The advent of ubiquitous sensing promises to revolutionize both the detection and the ongoing monitoring of chronic diseases and medical disorders. Previous work has demonstrated the ability to detect major depressive disorder from speech, and to estimate it's severity, by characterizing the level of articulatory coordination (Cummins et al., 2015a; Williamson et al., 2013; 2014; 2016). This approach is motivated by the hypothesis that a central symptom of MDD, psychomotor retardation, causes a degradation in articulatory coordination during speech.

In this paper, depression prediction is done from both audio and video signals. From audio, time varying estimates are obtained of the first three formant frequencies during speech. The eigenspectra of correlation and covariance matrices, constructed from time-delay embedding of these signals, are used to predict depression severity. In previous work, depression severity has been associated with a lowered dimensionality, or flattening, of scatter distributions in this delay embedding space. These findings are replicated on the WIBD and Mundt data sets in this paper.

In video, time varying estimates of facial action units (FAUs) during speech provide the basis for characterizing articulatory coordination. Using the same analysis approach that is used on formants, it has also previously been found that MDD is associated with a lowered dimensionality in FAU scatter distributions in a time-delay embedding space (Williamson et al., 2014). This finding is also replicated on the WIBD data set in this paper.

The current work extends previous findings on depression classification and estimation based on these feature approaches by exploring the ability to longitudinally track depression severity levels over time. In the data sets explored in this paper there are large within-subject changes in depression ratings in patients undergoing treatment over the course of several weeks.

Two data sets are analyzed. The WIBD data set contains audio and video signals from read speech and video from free conversational speech. The Mundt data set contains audio only from read speech and free speech. Using the same set of feature extraction and machine learning parameters, the prediction algorithm is able to track depression severity levels on both data sets, generating predictions that correlate with true HAM-D scores. On the WIBD data set, the best prediction result is $r = 0.63$ and RMSE = 5.49 (given true scores with range 1−24). On the Mundt set the best result is $r = 0.48$ and RMSE = 5.52 (given true scores with range 3−27). Video-based features are found to be the most effective, especially from free conversational speech. Individualizing the prediction model using a baseline session from the test subject improves prediction performance on both data sets.

This work therefore demonstrates the viability of using properties of articulatory coordination, which are extracted from audio and video during read or free speech, as a basis for longitudinal tracking of depression severity. The feature approach in this work uses holistic characterizations of the delay-embedding dimensionality across all formant frequencies and across all FAUs. An avenue for future research is to conduct a more specific analysis of how changes among particular formants and among particular FAUs correlate with specific symptoms of depression, such as various subcomponents of the HAM-D assessment. Likewise, it will be useful to investigate the effects on articulatory coordination of depression subtypes such as agitated depression, wherein psychomotor retardation may be a less prominent symptom.

Acknowledgments

The authors thank Dr. James Mundt for providing the Mundt dataset and Dr. Daryush Mehta for providing the KARMA formant tracking software, as well as for discussions on their use.

References

Association, A.P., et al., 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.

Caligiuri, M.P., Ellwanger, J., 2000. Motor and cognitive aspects of motor retardation in depression. J. Affect. Disord. 57 (1), 83–93.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F., 2015a. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 71, 10–49.

Cummins, N., Sethu, V., Epps, J., Schnieder, S., Krajewski, J., 2015b. Analysis of acoustic space variability in speech affected by depression. Speech Commun. 75, 27–49.

Cummins, N., Sethu, V., Epps, J., Williamson, J.R., Quatieri, T.F., Krajewski, J., 2018. Generalized two-stage rank regression framework for depression score prediction from speech. IEEE Trans. Affect. Comput. Early Access.

Darby, J.K., Simmons, N., Berger, P.A., 1984. Speech and voice parameters of depression: a pilot study. J. Commun. Disord. 17 (2), 75–85.

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

Ekman, P., Freisen, W.V., Ancoli, S., 1980. Facial signs of emotional experience. J. Person. Soc. Psychol. 39 (6), 1125.

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

France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, M., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. 47 (7), 829–837.

Gaebel, W., Wölwer, W., 1992. Facial expression and emotional face recognition in schizophrenia and depression. Eur. Arch. Psychiatry Clin. Neurosci. 242 (1), 46–52.

Greden, J.F., Carroll, B.J., 1981. Psychomotor function in affective disorders: an overview of new monitoring techniques. Am. J. Psychiatry 138 (11), 1441.

Hedlund, J.L., Vieweg, B.W., 1979. The hamilton rating scale for depression: a comprehensive review. J. Oper. Psychiatry 10 (2), 149–165.

Helfer, B.S., Quatieri, T.F., Williamson, J.R., Keyes, L., Evans, B., Greene, W.N., Vian, T., Lacirignola, J., Shenk, T., Talavage, T., et al., 2014. Articulatory dynamics and coordination in classifying cognitive change with preclinical MTBI. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M., 2011. The computer expression recognition toolbox (cert). In: Proceedings of 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011). IEEE, pp. 298–305.

Low, L.-S. A., Maddage, N.C., Lech, M., Sheeber, L., Allen, N., 2010. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In: Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). IEEE, pp. 5154–5157.

Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended Cohn–Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 94–101.

Mehta, D.D., Rudoy, D., Wolfe, P.J., 2012. Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. J. Acoust. Soc. Am. 132 (3), 1732–1746.

Moore, E., Clements, M., Peifer, J., Weisser, L., 2003. Analysis of prosodic variation in speech for clinical depression. In: Proceedings of the 25th IEEE Annual International Conference of the Engineering in Medicine and Biology Society, 2003, 3. IEEE, pp. 2925–2928.

Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S., 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J. Neurolinguist. 20 (1), 50–64.

Nilsonne, Å., Sundberg, J., Ternström, S., Askenfelt, A., 1988. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. J. Acoust. Soc. Am. 83 (2), 716–728.

Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K., Wilkes, D.M., 2004. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE Trans. Biomed. Eng. 51 (9), 1530–1540.

Quatieri, T.F., Malyska, N., 2012. Vocal-source biomarkers for depression: a link to psychomotor activity. In: Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association.

Quatieri, T.F., Williamson, J.R., Smalt, C.J., Perricone, J., Patel, T., Brattain, L., Helfer, B., Mehta, D., Palmer, J., Heaton, K., et al., 2017. Multimodal biomarkers to discriminate cognitive state. Role Technol. Clin. Neuropsychol. 409.

Sahu, S., Espy-Wilson, C., 2015. Effect of depression on syllabic rate of speech. J. Acoust. Soc. Am. 138 (3), 1781.

Scherer, S., Lucas, G.M., Gratch, J., Rizzo, A.S., Morency, L.-P., 2016. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. IEEE Trans. Affect. Comput. 7 (1), 59–73.

Scherer, S., Morency, L.-P., Gratch, J., Pestian, J., 2015. Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4789–4793.

Schwartz, G.E., Fair, P.L., Salt, P., 1976. Facial expression and imagery in depression: an electromyographic study. Psychosom. Med. 38 (5), 337–347.

Smith, K.M., Williamson, J.R., Quatieri, T.F., 2017. Vocal markers of motor, cognitive, and depressive symptoms in parkinson's disease. In: Proceedings of the , 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 71–78.

Sung, M., Marci, C., Pentland, A.,. Objective physiological and behavioral measures for identifying and tracking depression state in clinically depressed patients. Massachusetts Institute of Technology Media Laboratory, Cambridge, MA, Technical Report TR 595, 2005.

Trevino, A.C., Quatieri, T.F., Malyska, N., 2011. Phonologically-based biomarkers for major depressive disorder. EURASIP J. Adv. Signal Process. 2011 (1), 42.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M., 2014. Avec 2014: 3D dimensional affect and depression recognition challenge. In: Proceedings of the Fourth International Workshop on Audio/Visual Emotion Challenge. ACM, pp. 3–10.

Williamson, J.R., Bliss, D.W., Browne, D.W., 2011. Epileptic seizure prediction using the spatiotemporal correlation structure of intracranial eeg. In: Proceedings of the , 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 665–668.

Williamson, J.R., Bliss, D.W., Browne, D.W., Narayanan, J.T., 2012. Seizure prediction using EEG spatiotemporal correlation structure. Epilepsy Behav. 25 (2), 230–238.

Williamson, J.R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., Kung, H.-T., Dagli, C., Quatieri, T.F., 2016. Detecting depression using vocal, facial and semantic communication cues. In: Proceedings of the Sixth International Workshop on Audio/Visual Emotion Challenge. ACM, pp. 11–18.

Williamson, J.R., Quatieri, T.F., Helfer, B.S., Ciccarelli, G., Mehta, D.D., 2014. Vocal and facial biomarkers of depression based on motor incoordination and timing. In: Proceedings of the Fourth International Workshop on Audio/Visual Emotion Challenge. ACM, pp. 65–72.

Williamson, J.R., Quatieri, T.F., Helfer, B.S., Horwitz, R., Yu, B., Mehta, D.D., 2013. Vocal biomarkers of depression based on motor incoordination. In: Proceedings of the Third ACM International Workshop on Audio/Visual Emotion Challenge. ACM, pp. 41–48.

Williamson, J.R., Quatieri, T.F., Helfer, B.S., Perricone, J., Ghosh, S.S., Ciccarelli, G., Mehta, D.D., 2015. Segment-dependent dynamics in predicting parkinson's disease. In: Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association.

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: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association.