Vocal Biomarkers of Depression Based on Motor Incoordination

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

Rachelle Horwitz
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
(781) 981-7964
rachelle.horwitz@ll.mit.edu

Bea Yu
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
(781) 981-8473
bea.yu@ll.mit.edu

Daryush D. Mehta
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02421
(781) 981-7480
daryush.mehta@ll.mit.edu

Abstract[1]

In Major Depressive Disorder (MDD), neurophysiologic changes can alter motor control [1, 2] and therefore alter speech production by influencing the characteristics of the vocal source, tract, and prosodics. Clinically, many of these characteristics are associated with psychomotor retardation, where a patient shows sluggishness and motor disorder in vocal articulation, affecting coordination across multiple aspects of production [3, 4]. In this paper, we exploit such effects by selecting features that reflect changes in coordination of vocal tract motion associated with MDD. Specifically, we investigate changes in correlation that occur at different time scales across formant frequencies and also across channels of the delta-mel-cepstrum. Both feature domains provide measures of coordination in vocal tract articulation while reducing effects of a slowly-varying linear channel, which can be introduced by time-varying microphone placements. With these two complementary feature sets, using the AVEC 2013 depression dataset, we design a novel Gaussian mixture model (GMM)-based multivariate regression scheme, referred to as Gaussian Staircase Regression, that provides a root-mean-squared-error (RMSE) of 7.42 and a mean-absolute-error (MAE) of 5.75 on the standard Beck depression rating scale. We are currently exploring coordination measures of other aspects of speech production, derived from both audio and video signals.

Keywords

major depressive disorder, motor control, clinical assessment, vocal biomarkers, incoordination, correlation structure, Gaussian mixture models

1. Introduction

Major Depressive Disorder (MDD) is the most prevalent mood disorder, with a lifetime risk observed to fall between 10-20% for women and 5-12% for men [5]. As the number of people suffering from MDD steadily increases, so too does the burden of accurate diagnosis. Currently, diagnosis of MDD requires clinical assessment by a professional with significant clinical experience. However, the inter-clinician variability of these assessments makes the tracking of medication efficacy in clinical trials difficult. The growing global burden of MDD suggests that a convenient and automated method to assess depression severity would both simplify and standardize the task of diagnosing and monitoring depression, allowing for greater availability and uniformity in assessing depression. An automated approach avoids in-office clinical visits, and thus may facilitate automated measurement and identification, as well as quicken evaluation of treatment.

Classes of potential biomarkers of growing interest are based on vocal characteristics, which have been shown to change with a patient's mental condition and emotional state [3, 6-14]. Although there has been significant effort in using potential vocal biomarkers for emotion classification, there has been little or no exploiting of the effects of incoordination in the vocal modality (or in other modalities such as facial expression) that result from a depressed state. Such changes, indicated by qualities such as monotony, slur, hoarseness, and breathiness in voices, reflect changes in interacting brain components [15-17] that are associated with the transition from a healthy to a depressed brain. Clinically, many of these characteristics are associated with psychomotor retardation, where a patient shows sluggishness and motor disorder in vocal articulation, affecting coordination across multiple aspects of production. Indeed, dynamic motor disturbances seen in depression, although lesser in extent, have been found to bear similarity to those in Parkinson's disease with perhaps similar root neural sources [3, 4].

In this paper, we exploit inter-relationships across aspects of speech production by selecting features that reflect dynamical

[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. The work of R. Horwitz was supported in part by The National Institutes of Deafness and Other Communicative Disorders under Grant 5 T 32 DC00038.

changes in coordination within two particular vocal tract representations: (1) formant-frequency tracks, capturing coordination across vocal tract resonant frequencies, and (2) temporal characteristics of mel-cepstral features, capturing coordination in vocal tract spectral shape dynamics. For both representations, coordination measures are obtained from auto- and cross-correlations of the multichannel vocal signals. Using the AVEC 2013 depression dataset, we show the complementary nature of these two feature sets and design a GMM-based multivariate regression scheme that provides a root-mean-squared-error (RMSE) of 7.42 and a mean-absolute-error (MAE) of 5.75 on the standard Beck depression rating scale. We are also exploring coordination of other aspects of speech production, both from audio and video signals.

Our paper is organized as follows. In Section 2, we briefly describe the 100-subject audio database collected by AVEC [18]. In Section 3, we describe our signal-processing methodologies for vocal-feature extraction, with emphasis on using multichannel correlations themselves as features. In Section 4 we describe our novel GMM-based approach for multivariate regression, and in Section 5 we describe the results of our correlation and regression analyses on the AVEC 2013 challenge. Finally, in Section 6 we provide conclusions and projections toward future work.

2. AVEC Database

The AVEC 2013 challenge uses a subset of the audio-visual depressive language corpus (AVDLC), which includes 340 video recordings of 292 subjects performing a human-computer interaction task while being recorded by a webcam and a microphone and wearing a headset [18]. The 16-bit audio was recorded using a laptop's sound card at a sampling rate of 41 KHz. The video was recorded using a variety of codecs and frame rates, and was resampled to a uniform 30 frames-per-second. For the challenge, the recording sessions were split into three partitions, with 50 sessions each: a Training, Development, and Test set.

Recording lengths fall between 20-50 minutes with a 25-minute mean value. The mean age is 31.5 years, with a standard deviation of 12.3 years over a range of 18 to 63 years. The recordings took place in a number of quiet environments and consisted of: sustained vowel phonation; speaking loud while solving a task; counting from 1 to 10; read speech; singing; telling a story from the subject's own past; and telling an imagined story. For our particular scenario, we use read speech only (the 3rd read passage).

3. Feature Construction

We selected two vocal feature domains in which to represent underlying changes in vocal tract shape and dynamics: Formant frequencies and delta-mel-cepstra. We hypothesize such changes occur with motor control aberations due to a depressed state. The auto- and cross-correlations among "channels" of each measurement domain become the basis for key depression features.

3.1 Data segmentation

The goal of data segmentation is to provide, from each session in the Training and Development sets, representative speech data segments with as much extraneous variation removed as possible. In previous work, we have found that vocal biomarkers for depression assessment are most reliable when comparing identical read passages [19]. We decided therefore to focus on the third read passage, which has sufficient duration to provide robust feature estimates (mean duration = 226 seconds, with standard deviation = 66 seconds), and which is also in the speakers' common native language (German). This passage was segmented using a semi-automated procedure.

To remove an additional source of extraneous variation, we detected all speech pauses greater than .75 seconds, and then removed these pause segments from both of the feature domains, stitching together the feature values across each removed pause segment. This was performed because the presence of long speech pauses provides an extraneous source of low frequency dynamics in the formant and delta-mel-cepstral features that are not necessarily related to depression level. Pause detection was performed using an automated procedure that detects local smooth periods in the formant frequency tracks. These smooth periods occur when the formant tracker (described below) coasts over non-speech regions.

3.2 Formant frequencies

We associate (loosely) vocal tract formant dynamics with vocal articulation as one means to represent articulatory changes in the depressed voice. There are a variety of approaches to the on-going challenge of formant estimation and tracking. We have selected an algorithm recently developed by Rudoy, Mehta, Spendley, and Wolfe [20, 21], based on the principle that formants are correlated with one another in both the frequency and time domains. Formant frequencies are computed at 10-ms data frames. Embedded in the algorithm is a voice-activity detector that allows a Kalman predictor to smoothly coast consistently through non-speech regions. Because we are using only formant frequencies, these features are approximately immune to slowly-varying linear channel effects.

3.3 Mel-cepstra

To introduce vocal tract spectral magnitude information, we use standard mel-cepstra (MFCCs), provided by the AVEC challenge, as a basis for a second feature set. Specifically, we use delta-mel-cepstra generated by differencing the first 16 mel-cepstra across consecutive 10-ms data frames, thus introducing a dynamic spectral component and also reducing slowly-varying channel effects through the cepstral difference operation.

3.4 Correlation structure features

We hypothesize that the structure of the correlations of formant frequencies and of delta-mel-cepstral coefficients reflects the physiological coordination of vocal tract trajectories, and thus reveals motor symptoms of depression. A multivariate feature construction approach, based on cross-correlation analysis, is used to characterize the correlation structure among the signals from these two speech feature domains. This multivariate feature construction approach was first introduced for analysis of EEG signals for epileptic seizure prediction [22, 23], and has since been successfully applied to detection of epileptic seizures [24], and to prediction of infant apneas from cardio-respiratory signals [25, 26]. A detailed description of this feature analysis approach is in [23].

In this approach, channel-delay correlation and covariance matrices are computed from multiple time series channels. Each matrix contains correlation or covariance coefficients between the channels at multiple relative time delays. The 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 signals. Changes over time in the eigenvalue spectra of these channel-delay matrices register the temporal changes in coupling strengths among the channels.

Our two feature sets consist of the first 3 formants and the 16 delta-mel-cepstra, both of which are provided at 10-ms frame intervals. The cross-correlation analysis of these time series is conducted at four different time delay scales. These scales involve computing correlations among time series that are shifted in time relative to each other at four different sample delay spacings: 1, 2, 4, and 8. These spacings correspond to time delays in increments of 10-ms, 20-ms, 40-ms, and 80-ms.

A multi-scale approach is used to characterize the coupling patterns among the signals over different ranges of delays. For the formant frequency feature set, 30 time delays are used per delay scale, and for the delta-mel-cepstral feature set, 10 time delays are used per delay scale. The formant features are analyzed using a single feature frame that spans the entire data segment, whereas the delta-mel-cepstral features are analyzed using a sliding 60s feature frame, applied at 30s intervals.

For example, Figure 1 shows channel-delay correlation matrices (3rd delay scale, with time delays in increments of 40-ms) constructed from the formant tracks of two different subjects. These matrices each contain nine 30 × 30 blocks, each block consisting of the within- or cross-channel correlation coefficients for a pair of formant tracks. These coefficients are computed using all possible pairwise combinations of the 30 time-delayed versions of each channel. The 30 × 30 blocks along the main diagonal contain the within-channel correlations and the 30 × 30 off-diagonal blocks contain the cross-channel correlations. The matrix on the top is from a healthy subject (Beck score 0), and the matrix on the bottom is from a severely depressed subject (Beck score 44). Note that the healthy-subject matrix has a more vivid appearance, containing auto- and cross-correlation patterns that look sharper and more complex.

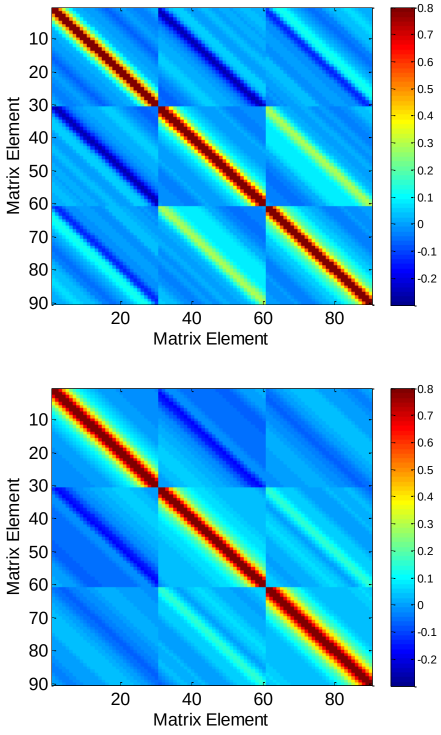

Figure 1. Channel-delay correlation matrix computed from formant tracks for healthy subject (top) and from severely depressed subject (bottom). Red denotes high and blue low (auto-)cross-correlation value.

These qualitative differences between the correlation matrices can be quantified using the matrix eigenspectra, which are the rank-ordered eigenvalues. These features are invariant to the underlying ordering of the channels (randomly permuting them will produce identical eigenspectra), capturing instead the levels of correlation among all the channels. The eigenspectra from the two matrices are plotted in Figure 2, with the eigenvalues from the healthy subject in blue and from the depressed subject in red. The eigenspectra from the depressed patient contain a greater fraction of power in the first few eigenvalues, indicating reduced complexity and independent variation in this subject's formant tracks. The divergence in eigenspectra between the healthy and the depressed subject suggests that this technique could provide an effective basis for estimating depression levels.

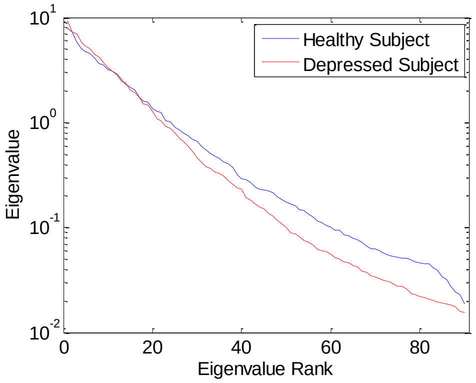

Figure 2. Eigenspectra from formant channel-delay matrices of a healthy and a depressed subject.

Additionally, eigenspectra from channel-delay covariance (as opposed to correlation) matrices at each delay scale are also used in order to characterize signal magnitude information. From each covariance eigenspectrum, two summary statistics are computed that capture the overall covariance power and entropy.

The cross-correlation analysis produces, from each feature frame, a high dimensional feature vector of correlation matrix eigenvalues and covariance matrix power and entropy values. These feature vectors consist of 368 elements in the formant domain (3 formant channels, 4 delay scales, 30 delays per scale, and 2 covariance features per scale), and 648 elements in the delta-mel-cepstral domain (16 delta-mel-cepstral channels, 4 delay scales, 10 delays per scale, and 2 covariance features per scale). The features within each domain are highly correlated, and so the final stage of feature construction is dimensionality reduction using principal component analysis (PCA) into a smaller set of uncorrelated features. This is done independently within each feature domain. A critical step is to first normalize each of the features into standard units (zero mean unit variance), which allows the variation of each feature to be considered relative to its baseline variation across the feature frames in all the sessions in the Training set. The top ***n*** principal components are input from each domain into the machine learning algorithm for depression estimation, which is described in Section 4. The appropriate ***n*** is empirically determined independently for each feature domain.

4. GMM-Based Regression Analysis

The feature construction approach described in Section 3.4 may produce multiple principal component features that are each weakly correlated with the Beck score. In addition, the patterns of correlation between features and Beck score may differ from one subject to the next. Therefore, we desire a multivariate fusion and regression approach that can effectively combine the information from multiple input features and also take advantage of contextual information such as subject identity (or potentially gender) in making its depression predictions. We propose to use for this purpose Gaussian Mixture Models (GMMs), which have been successfully applied for classification in the domains of automatic speaker recognition [27] and, more recently, speech-based depression classification [13, 14]. Below, we propose a novel approach for training GMMs for multivariate regression.

4.1 Multivariate fusion

Our regression approach accomplishes fusion of the multivariate feature density for non-depressed (Class 1) and depressed (Class 2) subjects using a novel approach that we call Gaussian Staircase Regression. This approach creates a GMM for Class 1 and for Class 2 based on multiple data partitions. The GMMs produce likelihoods for Class 1 and Class 2 on the multiple data frames for each session. The GMM test statistic for a session is the log likelihood ratio of the mean Class 2 likelihoods and mean Class 1 likelihoods. A univariate regression function is then created from the GMM test statistics on the (AVEC) Training set and the corresponding Beck scores. This regression function, when applied to the GMM test statistic from a (AVEC) Development session, is used to produce a Beck score prediction.

Gaussian Staircase Regression uses multiple partitions of the Training feature vectors. In each partition, vectors are assigned to the two classes by comparing their Beck scores to a different Beck score threshold. We use 8 partitions, corresponding to Beck score thresholds of 5, 10, …, 40. Therefore, rather than the standard approach of training a GMM using Expectation-Maximization from a fixed data partition between depressed and non-depressed subjects (e.g., [14]), the GMM is formed directly from an ensemble of Gaussian classifiers that are trained from multiple data partitions. This partitioning approach thereby creates a "staircase" of increasing Gaussian density support in the feature space for Class 1 along the continuum of lower Beck scores, and for Class 2 along the continuum of higher Beck scores. The Gaussian densities use full covariance matrices, with a constant value of 0.2 added to the diagonal terms for improved regularization.

This approach results in a test statistic that tends to smoothly increase with increasing depression, providing a strong basis for subsequent univariate regression. In addition, by using explicit Gaussian densities, it allows the use of Bayesian adaptation of the Gaussian densities from contextual information such as subject identity (and potentially gender), using an approach similar to the widely used technique developed in [27]. In the original technique, the GMM components of a background speaker model are moved toward the training data from a target speaker by amounts proportional to the posterior probabilities of the GMM components given the target speaker data, resulting in a speaker-adapted GMM. In our current work, the Gaussian means are adapted independently in each data partition based on subject identity, using mixing weights computed as $n/(.5+n)$, where n is the number of sessions from the currently evaluated Development subject that are in the Training set.

The frame rates for correlation structure features are different for the two feature domains, and so multivariate fusion of the principal component features from the two domains requires frame registration. The formant-based feature vector is computed using a single frame for each session, whereas the delta-mel-cepstral-based feature vectors are computed using 60s frames with 30s overlap. Frame registration is done by duplicating the single formant feature vector from each session, and pairing it (via vector concatenation) with the 6-dimensional delta-mel-cepstral feature vector from each frame, thereby creating the 11-dimensional fused feature vectors. When evaluating the formant features by themselves, these duplicated formant feature vectors are also used, in order to make comparisons over different feature combinations consistent. Using features extracted at fixed time intervals (60 second frames, with 30 second overlap) causes longer duration read passages to produce a larger number of feature vectors, thereby causing these passages to be slightly over-represented in the Training set.

4.2 Training and Testing Procedure

The Beck score predictions are made for each Development session based on parameters estimated from the 50 sessions in the Training set. The Beck score predictions are generated as follows. The high-dimensional correlation structure features from the Training feature frames are normalized to have zero variance and unit standard deviation, and these normalization coefficients are then applied to the high-dimensional correlation structure features from the Development feature frames. Next, PCA is applied independently to each feature domain, generating the following number of components per feature domain: 5 principal components for the formant domain, and 6 principal components for the delta-mel-cepstral domain. As with the feature normalization procedure, the PCA transformation coefficients are determined from the Training features and then applied to the Development features. The principal component features are subsequently normalized to zero mean, unit standard deviation (again, with normalizing coefficients obtained from the Training set only, and applied to the Development set) prior to the GMM-based multivariate regression, described in Section 4.1.

The following procedure is repeated for all of the 50 sessions in the Development set to obtain the 50 Beck score predictions. Given the subject identity for each Development session, subject-adaptation of the Training set GMMs is performed, and test statistics for the 50 Training set sessions are produced. Because GMM likelihoods are produced at each of multiple feature frames per session, the single test statistic per session is the log likelihood ratio of the mean of the GMM likelihoods for Class 2 and for Class 1. The 50 Training set test statistics are used to create a 2nd order regression with the corresponding Training set Beck scores. This regression equation is then applied to the single Development test statistic value to obtain a predicted Beck score for that session. Because negative Beck scores are impossible, negative predictions are set to zero.

5. Prediction Results

The feature extraction and regression approach described above was applied to the 3rd read AVEC passage from each session as a basis for predicting depression. Figure 3 shows scatter plots, with the GMM Development set test statistics on the x-axis and the Development set Beck scores on the y-axis. These plots are shown given three different combinations of the two feature sets: (1) formant features only, (2) delta-mel-cepstral features only, and (3) both feature domains combined. Plotted in red are 2nd-order regressions fit to these Development test statistics. Observe that these are not the regressions used to generate the Beck predictions. As described in Section 4.2, those predictions are made using different regressions for each Development set subject, based on subject-adapted Training set GMMs.

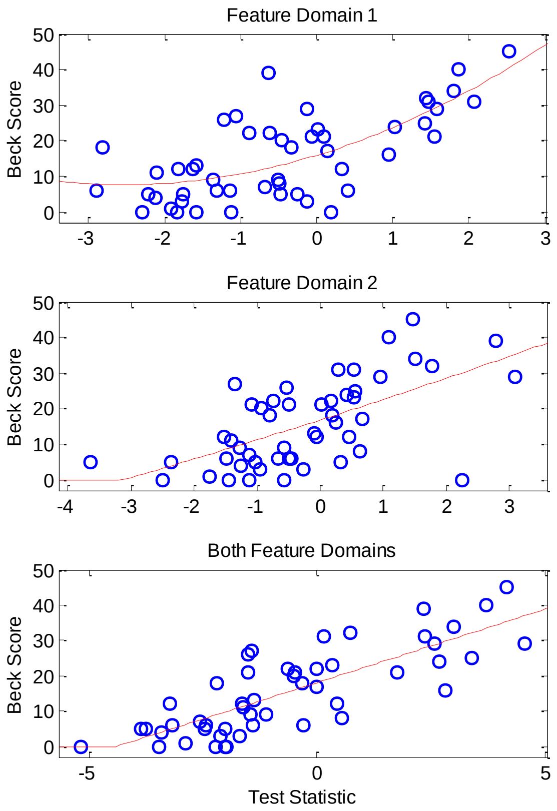

Figure 3. Scatter plots relating GMM test statistic to Beck score on Development set for three feature domain combinations. 2nd-order regression lines are shown in red. Top: formant features only. Middle: delta-mel-cepstral features only. Bottom: Both feature domains combined. These plots match the results shown in Table 1.

Table 1 shows the error metrics and Pearson correlations for the three feature set combinations., The best Beck score predictions by far are obtained using the combined feature sets (in which the feature vector consists of 11 principal components), thereby demonstrating their complementary nature. These results are a root-mean-squared-error (RMSE) of 7.42 and a mean-absolute-error (MAE) of 5.75 on the standard Beck depression rating scale. These results also demonstrate large performance improvements compared to the AVEC baseline audio prediction scores, which are RMSE = 10.75 and MAE = 8.66 [18].

It is useful to understand the relative importance of different elements of the prediction system. One element is subject-based adaptation of the Gaussian components. Table 2 illustrates the importance of this step, showing that prediction accuracy is degraded if this step is removed. Another element is the use of multiple data partitions in the Gaussian staircase regression technique. The results shown in Figure 3 and Tables 1 and 2 were obtained with 8 data partitions, corresponding to Beck score thresholds of 5, 10, …, 40. We have investigated the effect of varying the number of partitions, and thereby obtained the results shown in Figure 4. The MAE values from the combined features are plotted as a function of the number of data partitions. For multiple partitions, the outside partition threshold values of 5 and 40 were kept fixed, and intermediate threshold values spaced at equal intervals were used. For the single partition case, the mid-point threshold value of 22.5 was used. As Figure 4 shows, the algorithm is relatively insensitive to the number of partitions, provided there are at least four of them. The number of partitions corresponds to the number of Gaussian components in the Class 1 GMM and Class 2 GMM created by the Gaussian staircase technique. An alternative method of training the GMMs using expectation-maximization from a single fixed data partition was also attempted, but produced inferior results compared to Gaussian staircase regression.

Table 1. Prediction results for three feature domain combinations, with speaker-based adaptation. AVEC baseline audio prediction scores are RMSE = 10.75 and MAE = 8.66 [18].

Feature Domain	**RMSE**	**MAE**	***R***
Formant only	8.50	6.87	0.68
Delta-mel-cepstral	9.66	7.92	0.61
Combined	7.42	5.75	0.80

Table 2. Prediction results for three feature domain combinations, without speaker-based adaptation. AVEC baseline audio prediction scores are RMSE = 10.75 and MAE = 8.66 [18].

Feature Domain	**RMSE**	**MAE**	***R***
Formant only	9.97	7.92	0.56
Delta-mel-cepstral	10.05	8.24	0.53
Combined	8.68	7.12	0.70

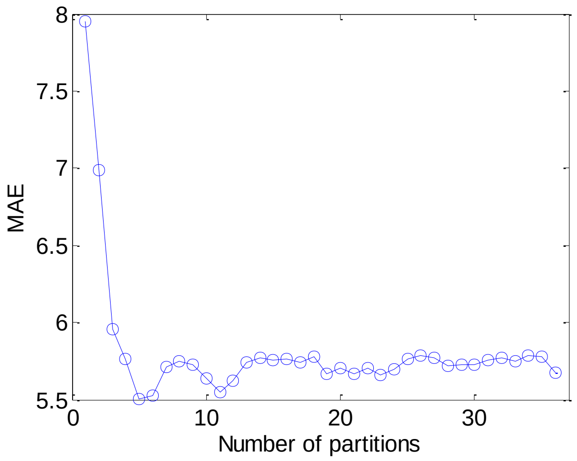

Figure 4. MAE plotted as a function of the number of data partitions used in Gaussian staircase regression. Results obtained using combined features and subject-based adaptation.

Another interesting data comparison concerns the relative usefulness of mel-cepstral versus delta-mel-cepstral features as input to the cross-correlation analysis technique. Both features are useful, but we have found better performance using the delta-mel-cepstral features, obtaining MAE = 7.92 as opposed to MAE = 8.52 for the mel-cepstral features (processed with smallest delay scale only, and 3 principal components). While using these two cepstral feature sets in conjunction improves performance compared to either one alone (MAE = 7.32), we have found that adding the mel-cepstral features to the combined formant and delta-mel-cepstral features slightly degrades performance (MAE = 5.92 vs. MAE = 5.75).

Insight into these cepstral variants can be obtained by viewing examples of their channel-delay correlation matrices. Figure 5 shows the delta-mel-cepstral matrices at the smallest delay scale for the same healthy and depressed sessions that are also illustrated in Figure 1 for the formant-based correlation matrices. Observe that the healthy subject shows sharper and less erratic cross-correlation patterns compared to the depressed subject. Figure 6, on the other hand, shows the corresponding correlation matrices for these sessions from the original mel-cepstral features. These matrices show greater differentiation in the correlations between different channel pairs, which is due to slowly varying channel effects. This results in lower relative differentiation within the same channel pair across time delays. Further investigation is needed to discover the best combinations of cepstral-based features, as well as potentially additional cepstral feature variants.

6. Conclusions and Future Work

Our ability to achieve good prediction accuracy of depression using only two vocal feature domains, and only a single, roughly 4-minute long read passage, demonstrates that we have achieved a solid foundation for depression estimation from vocal biomarkers. We will expand on this foundation in many different ways, continuing our exploration of correlation structure techniques applied to both audio and video signals.

It is also important to note that the array of speech features we have selected, as well as their manner of computation, need to be expanded and refined in future work, and that the initial correlation analysis of this paper is currently serving as a guide to our on-going work in designing automatic classifiers of depression severity. Such severity is reflected in the perception of qualities such as monotony, slur, hoarseness, and breathiness in voices and sluggishness and strain in facial expression of depressed individuals, reflecting a change in the interacting brain components in moving from a healthy to a depressed brain [15-17], and in a variety of other coordinated speech characteristics.

For video-based face analysis, the coordination across mirror components of the mouth may serve as a basis for novel features. Specifically, we are exploring the facial action coding system (FACS), which allows for the coding of distinct movements in the face, referred to as facial action units (FAUs). Prior work has shown the ability of a manually extracted FAU associated with the buccinator muscle to distinguish between depressed and non-depressed individuals with an accuracy of 88% [28]. In particular, we will focus on mouth-based FAUs that reflect incoordination of lip and near-lip movements during voice activity. Furthermore, we are considering the coordination between voice and face modalities, which reflects the coupling of articulation with facial motion.

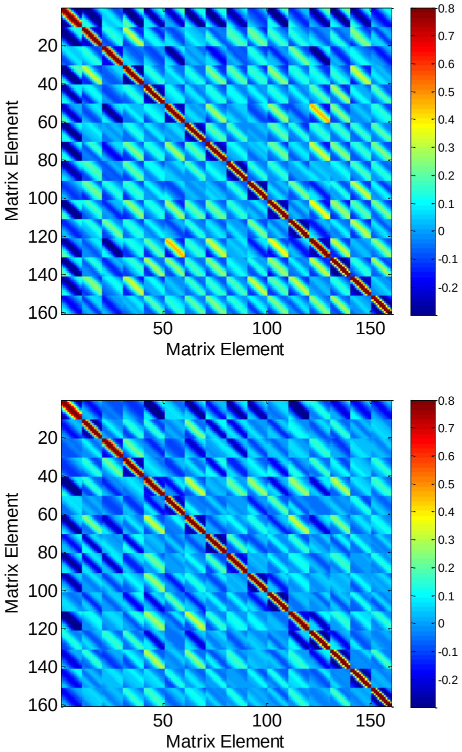

Figure 5. Channel-delay correlation matrices from delta-mel-cepstral features for healthy subject (top) and depressed subject (bottom).

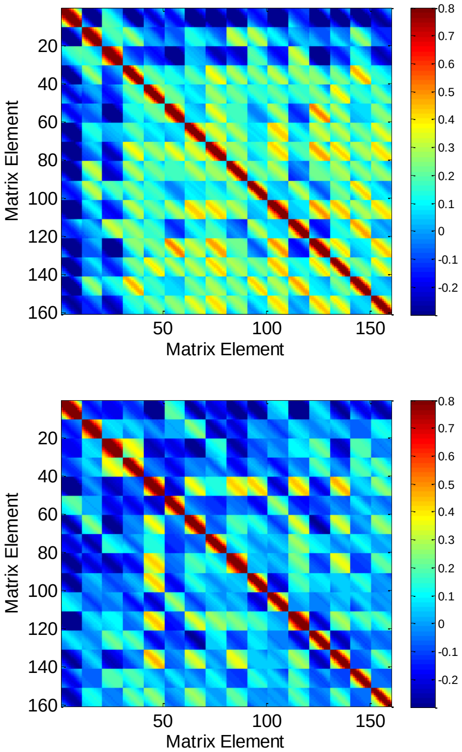

Figure 6. Channel-delay correlation matrices from mel-cepstral features for healthy subject (top) and depressed subject (bottom).

References

[1] C. Sabin, and C. Sackeim, H.A. Psychomotor Symptoms of Depression. *Am J Psychiatry*, 154:4-17, 1997

[2] J.S. Buyukdura, S.M. McClintock, P.E. Psychomotor retardation in depression: biological underpinnings, measurement, and treatment. *Prog Neuropsychopharmacol Biol Psychiatry*. 28(2):395-409, 2011.

[3] J. K. Darby, N. Simmons, and P. A. Bergcr. Speech and voice parameters of depression: a pilot study. *Journal of Communication Disorders*, 11, 75 85, 1984.

[4] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of Psychiatric Research*, 27(3):309-319, 1993.

[5] M. Fava, K. Kendler. Major depressive disorder. *Neuron* 28(2), 335-341, 2000.

[6] J. Mundt, P. Snyder, M. S. Cannizaro, K. Chappie, and D. S. Geralts. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J. Neurolinguistics*, 20(1): 50-64, 2007.

[7] D. France, R. Shiavi, et al. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7): 829, 2000.

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

[9] L. A. Low, T. Maddage, M. Lech, L. Sheeber, and N. Allen. Influence of acoustic low-level descriptors in the detection of clinical depression in adults. *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010*.

[10] E. Moore II, M. Clements, J. Peifer, and L. Weisser. Analysis of prosodic variation in speech for clinical depression. *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, 2003.

[11] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, and D. Mitchell. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, 51(9), 2004.

[12] A. Trevino, T. F. Quatieri, and N. Malyska., Phonologically-Based Biomarkers for Major Depressive Disorder. *EURASIP Journal on Advances in Signal Processing: Special Issue on Emotion and Mental State Recognition from Speech, 2011(1),* 1-18, 2011.

[13] D. Sturim, P. Torres-Carrasquillo, T. F. Quatieri, N. Malyska, and A. McCree. Automatic Detection of Depression in Speech using Gaussian Mixture Modeling with Factor Analysis. *Interspeech,* 2011.

[14] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu. Classification of depression state based on articulatory precision. *Interspeech*, 2013.

[15] E. Tognoli and J. A. Scott Kelso. Brain coordination dynamics: True and false faces of phase synchrony and metastability. *Prog Neurobiol.* 87(1): 31–40, 2009.

[16] M. D. Greicius, B. H. Flores, V. Menon, G. H. Glover, H. B. Solvason, H. Kenna, A. L. Reiss, and A. F. Schatzberg. Resting-State Functional Connectivity in Major Depression: Abnormally Increased Contributions from Subgenual Cingulate Cortex and Thalamus. *Biol Psychiatry*, 62(5): 429–437, 2007.

[17] S. L. Bressler, E. Tognoli, Operational principles of neurocognitive networks. *International Journal of Psychophysiology* 60:139-148, 2006.

[18] M. Valstar, B. Schuller, K. Smith, et al. AVEC 2013 - The Continuous Audio/Visual Emotion and Depression Recognition Challenge. *AVEC 2013*.

[19] R. Horwitz, T. F. Quatieri, B. Helfer, B. Yu, J. R. Williamson, and J. Mundt. On the Relative Importance of Vocal Source, System, and Prosody in Human Depression. *IEEE Body Sensor Network Conference,* Cambridge, MA, May 2013.

[20] D. Rudoy, D. N. Spendley, and P. Wolfe. Conditionally linear Gaussian models for estimating vocal tract resonances, *Proc. Interspeech*, 526–529, 2007.

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

[22] J. R. Williamson, D. W. Bliss, and D. W. Browne. Epileptic seizure prediction using the spatiotemporal correlation structure of intracranial EEG. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (pp. 665-668). IEEE, 2011.

[23] J. R. Williamson, D. W. Bliss, D. W. Browne, and J. T. Narayanan. Seizure prediction using EEG spatiotemporal correlation structure. *Epilepsy & Behavior*, 25(2), 230-238, 2012.

[24] Y. Pan, C. Guan, K. K. Ang, K. S. Phua, H. Yang, D. Huang, and S. H. Lim. Seizure detection based on spatiotemporal correlation and frequency regularity of scalp EEG. In *Neural Networks (IJCNN), The 2012 International Joint Conference on* (pp. 1-7). IEEE, 2012.

[25] J. R. Williamson, D. W. Bliss, D. W. Browne, P. Indic, P., E. Bloch-Salisbury, and D. Paydarfar. Using physiological signals to predict apnea in preterm infants. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on* (pp. 1098-1102). IEEE, 2011.

[26] J. R. Williamson, D. W. Bliss, D. W. Browne, P. Indic, E. Bloch-Salisbury, and D. Paydarfar. Individualized apnea prediction in preterm infants using cardio-respiratory and movement signals. *IEEE Body Sensor Networks,* 2013.

[27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*(1), 19-41, 2000.

[28] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. Padilla, ... and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on* (pp. 1-7). IEEE, 2009.