Classification of Depression by Quantifying Neuromotor Coordination Using Inverted Vocal Tract Variables

Nadee Seneviratne[1], Carol Espy-Wilson[1], James Williamson[2], Adam C. Lammert[3], Thomas F. Quatieri[2]

[1] *University of Maryland College Park,* [2] *MIT Lincoln Laboratory,* [3] *Worcester Polytechnic Institute*

Major Depressive Disorder (MDD) is known to affect speech. Psychomotor retardation is a long-established key feature of major depressive disorder (MDD) [1], and results in speech that is slower in rate and contains many more pauses, that are also longer in duration than pauses of speech from a non-depressed person. Speech changes of this kind are being developed to noninvasively diagnose and monitor MDD. Recently, articulatory coordination features have been developed that show promise for depression severity classification [2]. These coordination features are based on quantifying the correlation structure among speech-relevant time series. They have been extensively applied to the first three formant time series as a proxy for vocal articulation [2], but have been much less extensively validated in their application to articulatory speech features.

To explore coordination closer to actual articulation, a preliminary study was presented in [3], which uses speech-inverted vocal tract variables (TVs) as a direct measure of articulation to quantify changes in the articulatory coordination of depressed and non-depressed speech. TVs are derived from a speech inversion (SI) system [4] that maps the acoustic signal to vocal tract variables. TVs define the constriction degree and location of five distinct constrictors located along the vocal tract: lips, tongue tip, tongue body, velum, and glottis. The SI system in [4] estimates six TVs namely, Lip Aperture (LA), Lip Protrusion (LP), Tongue tip constriction degree (TTCD), Tongue tip constriction location (TTCL), Tongue body constriction degree (TBCD), Tongue body constriction location (TBCL).

In this preliminary study [3] experiments were conducted to determine the extent to which the coordination features proposed by [2] computed over three formants and three TVs (TTCD, TBCD, and LA) could be used as the basis for building a model to classify depressed vs. not depressed speech. Recordings of 7 subjects from the Mundt database [5] who transitioned from a depressed state (Hamilton Depression Rating Score (HAMD)$\geq$17) to a non-depressed state (HAMD$\leq$7) were used for this study. Altogether, there were 14 read utterances and 14 spontaneously spoken utterances. To compute the coordination features, a time-delay correlation matrix for each utterance is computed as an intermediate representation of the complexity of speech coordination as proposed in [2]. Each correlation matrix has dimensionality *(MN* $\times$ *MN)*, based on M TV/formant channels and N time delays per channel. This representation provides details about which TV/formant is correlated with which, and at which time delays, and is therefore rich with information about the mechanisms underlying coordination level. An eigenspectrum is computed from the correlation matrix, taking the form of an MN-dimensional feature vector. This is used to characterize the within-channel and cross-channel distributional properties of the multivariate TV time series. Using only two eigenspectrum features at 0.2 and 0.95 and a linear classifier, the classification results for the TVs were better than those for the formants by 7% for read speech and 28% for free speech.

The **present work** extends the preliminary study by (1) including results from 6 TVs (adding LP, TTCL, TBCL) (2) using data from more subjects in the Mundt database, (3) utilizing more eigenspectrum features computed from the correlation matrix to do the classification and (4) using a non-linear classifier (Support Vector Machine with Radial Basis Function kernel). In the case of read speech, we use all speech when subjects are depressed (HAMD$\geq$20) and all speech when subjects are not depressed (HAMD$\leq$ 7). In the case of free speech, we use the same HAMD threshold, but use only those utterances that are less than 30 sec in duration to obtain a balanced distribution of two classes. In the case of free speech there were 51 utterances for depressed speech and 66 utterances for non-depressed speech. For read speech, there were 33 and 20 utterances for depressed and non-depressed speech, respectively.

Consistent with the methodology reported in [2], the eigenspectrum for each utterance was computed using 3 or 6 channels and 15 time delays with the delay scaling of 7 samples yielding a 45 or 90-dimensional eigenspectrum respectively. In Figure 1, the eigenspectra are plotted as standardized feature-wise mean values as a function of the normalized eigenvalue feature index. The effect sizes are plotted relative to depressed state. It can be seen that low-rank eigenvalues from the low HAMD sessions

(not-depressed) are smaller than those from the high HAMD sessions (depressed). This difference means that higher depression is associated with simpler neuromotor coordination as described in [2].

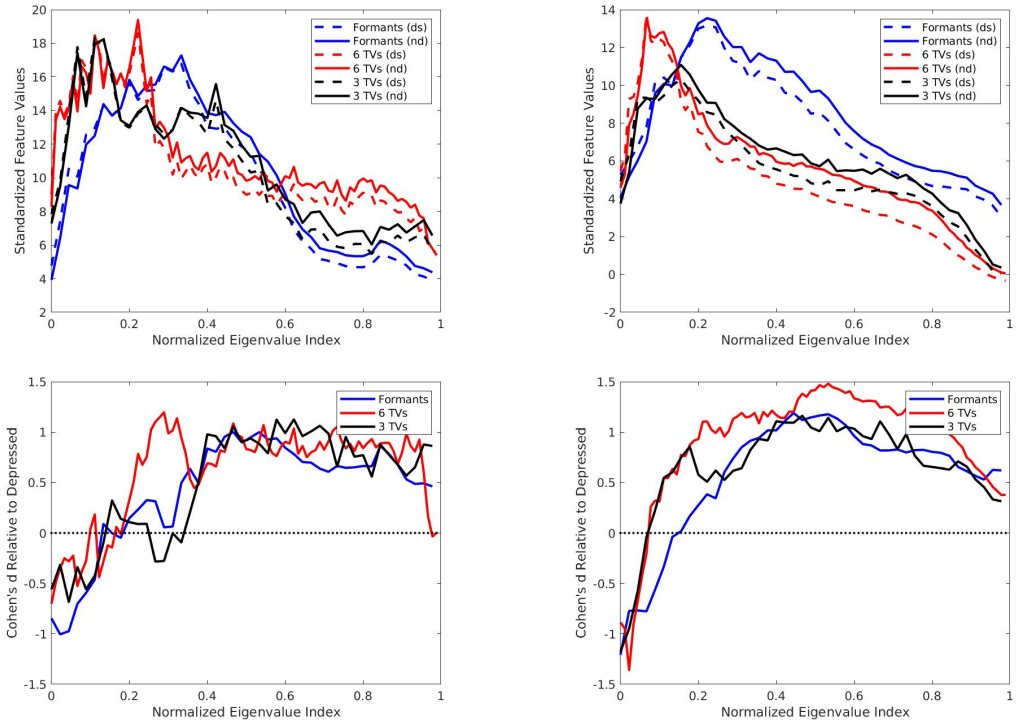

Figure 1: *Eigenvalues of 3 formants, 3 TVs and 6 TVs and effect sizes between the feature-wise means (Cohen's d) of coordination features in the not-depressed speech relative to those in the depressed speech for read speech (left) and free speech (right).*

Model training and testing were carried out within a leave-one-subject-out cross-validation scheme. At each fold, a classifier was trained on and used as the basis for estimating a label for the test utterances from the remaining subject. Classification accuracy of these estimated labels was calculated across all folds. We averaged the eigenspectrum features in many different ways and best results are shown in Table 1. "Index Range" specifies which features of the eigenspectrum that were averaged.

Table 1: *Classification accuracies for extended experiments*

	3 TVs	**6 TVs**	**3 Formants**
Read speech	70.56%	73.06%	66.39%
Index Range	<0.3, >0.8	<0.3, >0.8	<0.2, [0.2-0.8], >0.8

	3 TVs	**6 TVs**	**3 Formants**
Free speech	69.09%	72.06%	68.98%
Index Range	<0.2, [0.2-0.8], >0.8	<0.2, >0.8	<0.3, [0.3-0.7], >0.7

Results show that adding constriction location TVs increases the accuracy compared to the case of 3 TVs and 3 formants. It is also observed that read speech accuracy is higher compared to free speech when TVs are used. In [6], the authors argue for depression classification based on read speech since a stable, repeatable protocol can be used across speakers for measuring speech disfluencies (expected to be considerably higher when the speaker is depressed relative to when not depressed) based on the location of affective keywords in sentences.

[1] S. Kennedy, "Core symptoms of major depressive disorder: Relevance to diagnosis and treatment," *Dialogues in clinical neuroscience*, vol. 10, pp. 271–7, 02 2008.

[2] J. R. Williamson, D. Young, A. A. Nierenberg, J. Niemi, B. S. Helfer, and T. F. Quatieri, "Tracking depression severity from audio and video based on speech articulatory coordination," *Computer Speech & Language*, vol. 55, pp. 40 – 56, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230817303510

[3] C. Espy-Wilson, A. C. Lammert, N. Seneviratne, and T. F. Quatieri, "Assessing neuromotor coordination in depression using inverted vocal tract variables," *Proc. Interspeech 2019*, pp. 1448–1452, 2019.

[4] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 316–329, jul 2019.

[5] J. C. Mundt, P. J. Snyder, M. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology," *Journal of Neurolinguistics*, vol. 20, pp. 50–64, 2007.

[6] B. Stasak, J. Epps, and R. Goecke, "Automatic depression classification based on affective read sentences: Opportunities for text-dependent analysis," *Speech Communication*, vol. 115, pp. 1 – 14, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639318304266