INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Assessing Neuromotor Coordination in Depression Using Inverted Vocal Tract Variables *

Carol Espy-Wilson[1], Adam Lammert[2], Nadee Seneviratne[1], Thomas Quatieri[2]

[1]Speech Communication Laboratory, Institute of Systems Research, and Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20742
[2] Bioengineering Systems & Technologies, MIT Lincoln Laboratory, Lexington, MA 02421
espy@umd.edu, adam.lammert@ll.mit.edu, nadee@umd.edu, quatieri@ll.mit.edu

Abstract

Speech articulation is a complex activity that requires finely timed coordination across articulators, i.e., tongue, jaw, lips, and velum. In a depressed state involving psychomotor retardation, this coordination changes and in turn modifies the perceived speech signal. In previous work, we used the correlation structure of formant trajectories as a proxy for articulatory coordination, from which features were derived for predicting the degree of depression. Ideally, however, we seek coordination of the actual articulators using characteristics such as the degree and place of tongue constriction, often referred to as a *tract variable* (TV). In this paper, applying a novel articulatory inversion process, we investigate the relation between correlation structure of formant tracks versus that of TVs. We show on a pilot depressed/control dataset that, with the same number of variables, TV coordination-based features, although with some characteristics similar to their counterpart, outperform the corresponding formant track correlation features in detection of the depressed state. We speculate on the latent information being captured by TVs that is not present in formants.

Index Terms: speech production, vocal tract variables, psychomotor retardation, neuromotor coordination, depression, mental health, biomedical applications

1. Introduction

According to the World Health Organization, 350 million people worldwide suffer from depression. Depression is the most common precursor to suicide and suicidality is the third leading cause of death in youth ages between 10 and 24. A goal of our research is to develop a home monitoring system for depression that can provide the patient and their therapists with information to assess their mental health. The work in this paper is a major step in that direction. We discuss a novel joining of articulatory parameters from a speech inversion system [1][2] and a method for quantifying changes in neuromotor coordination from behavioral signals [3][4] that allows for highly accurate classification of depression.

In individuals suffering from major depressive disorder (MDD,) neurophysiological changes often alter motor control and thus affect mechanisms controlling speech production and facial expression. Clinically, these changes are typically associated with psychomotor retardation, a condition of slowed neuromotor output manifested in altered coordination and timing across multiple observables of acoustics and facial movements during speech. Although there has been significant effort in studying vocal biomarkers for depression classification, study into changes in articulatory coordination has been limited to two approaches, both focused on the use of formant tracks. The first study uses correlation structure of the formants as a proxy for underlying articulatory coordination [3][4]. Though this approach has been very effective in predicting the severity of MDD, it does not use articulation parameters directly, such as place and manner, i.e., degree of constriction. The second approach gets at articulatory parameters through a neuro-computational model of articulation (DIVA) [5] with the Maeda-model basis of speech synthesis [6]. Though showing promise as a feature discriminant, this approach however is also constrained by formant inputs to the model as production targets, where the Maeda model in effect inverts formants to a higher articulatory space (specifically in [5][6], 3 formants to 13 articulatory parameters). The one-to-many mapping is also ambiguous in this inversion process.

In this paper, we leverage a novel speech-to-articular inversion system that is not constrained to a formant representation but rather uses the entire speech signal and its gestural representation and provides a physiologically plausible resolution to the one-to-many mapping. The resulting (articulatory) *tract variables* then provide a basis for a truer representation of articulatory coordination. We show on a pilot depressed/control dataset that, with the same number of variables, TV coordination-based features, although with some characteristics similar to their counterpart, outperform the corresponding formant tracks in detection of the depressed state.

In Section 2 of this paper, we describe the dataset, inversion technique, and correlation structure methodology. Section 3 describes our effect size and classification results, while Section 4 speculates on interpretability and next steps.

2. Methods

2.1 Data

The database [7] used for this work contains speech collected using interactive voice response (IVR) technology. Thirty-five physician-referred patients who recently started on pharmacotherapy and/or psychotherapy treatment for depression, participated in an observational study that spanned six weeks. Using standard depression severity measures, the subjects were assessed weekly by clinicians and the patients themselves. Speech data used in this analysis include both read speech (the Grandfather passage) and spontaneous speech where patients describe how they feel emotionally, physically

* Distribution Statement & Disclaimer in Section 5

http://dx.doi.org/10.21437/Interspeech.2019-1815

and their ability to function in each week. In addition, subjects rapidly repeated the syllables /pa ta ka/ for five seconds.

For this study, we used the Hamilton Depression Rating Scale (HAMD) provided by the clinicians to choose subjects who transitioned from a depressed state to a not-depressed state. The distribution of HAMD scores on Day-00 for all 35 patients is given in Table 1. If the HAMD score was 17 or greater, the subject sessions were labeled as depressed and if the score was 7 or lower, the sessions were marked as not depressed. Scores between 8 and 16 were not assigned a label due to the ambiguity of their depression status. According to this criteria, six patients were found to transition from being depressed to not depressed as shown in Table 2. In addition to these six subjects, Subject 121 was included in the analysis since her psychomotor-retardation score changed from 2 to 0 (higher score means more psychomotor retardation). This was the largest change in a psychomotor retardation score observed across the subjects.

Table 1: *HAMD score distribution on Day-00.*

HAMD Score	State of depression	No. of patients
0-7	Normal	1 (Score is 7)
8-13	Mild depression	5
14-18	Moderate depression	12
19-22	Severe depression	11
>=23	Very severe depression	6

Table 2: *Days and corresponding HAMD scores of patients who transitioned from Depressed to Not Depressed.*

Subject	Depressed	HAMD Score	Not Depressed	HAMD Score
101	Day-14	19	Day-42	3
111	Day-00	21	Day 14, 28, 47	7, 3, 5
119	Day-00	17	Day-31	3
121	Day-00	14	Day-42	3
123	Day-00	22	Day-42	3
127	Day-00	24	Day-42	7
128	Day-00	20	Day-27	7

2.2 Speech Inversion based on Articulatory Phonology

We developed a Speech Inversion (SI) system [1][2] based on Articulatory Phonology (AP) [8] that maps the acoustic signal to vocal tract variables. AP is a phonological theory of speech as a constellation of overlapping articulatory gestures, and gestures are defined as discrete action units whose activation results in constriction formation or release by five distinct constrictors (lips, tongue tip, tongue body, velum, and glottis) along the vocal tract. The kinematic state of each constrictor is defined by its corresponding constriction degree and location coordinates, which are called vocal tract constriction variables or, in short, tract variables (TVs) (see Table 3 and Figure 1 for more details). Fig. 2 shows the TVs related to constriction degree along with a portion of the waveform and spectrogram of one of the rapidly spoken /pa ta ka/ tokens when a subject is not depressed. Observe that the TVs (sampled at 100 Hz) accurately capture the lip closure for /p/, the tongue tip closure for /t/ and the tongue body closure for /k/.

Table 3: *Constrictors and TVs.*

Constrictors	Vocal Tract (VT) Variables
Lip	Lip aperture (LA)
	Lip protrusion (LP)
Tongue Tip	Tongue tip constriction degree (TTCD)
	Tongue tip constriction location (TTCL)
Tongue Body	Tongue body constriction degree (TBCD)
	Tongue body constriction location (TBCL)
Velum	Velum (VEL)
Glottis	Glottis (GLO)

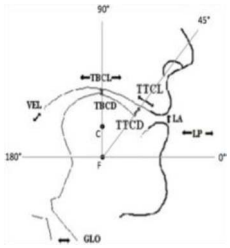

Figure 1: *Visual representation of the vocal tract variables at five distinct constriction organs (taken from Saltzman & Munhall [10]), along with a listing of constrictors and their vocal tract variables. See Table 2 for TV labels.*

2.3 Formant Estimation

The first three formant frequencies were extracted using Praat [9]. Settings were: tracking 5 formants, 5500 Hz maximum formant, window length of 25 milliseconds. Formants tracks were sampled at a rate of 160 Hz.

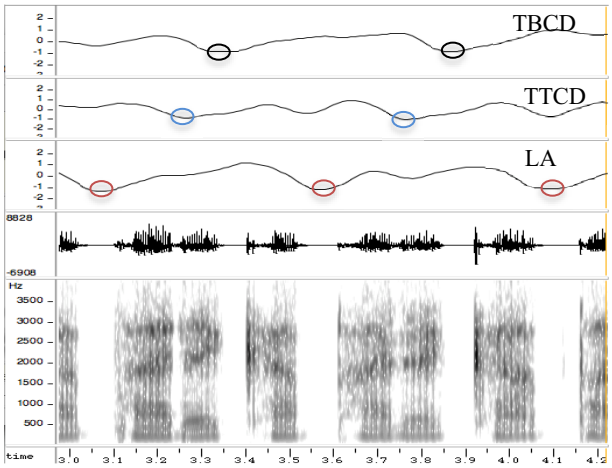

Figure 2: *Spectrogram, waveform and TVs for two productions of /pa-ta-ka/ when subject 119 is not depressed. Ovals encircle minima related to lip closure in LA for /p/ (red), tongue tip constriction in TTCD for /t/ (blue) and velar constriction in TBCD of /k/ (black).*

2.4 Coordination Features

Coordination among the three formant tracks (F1-3) and among three TVs (LA, TTCD, TBCD) was estimated using correlation structure features. The number of TVs was constrained to three to facilitate direct comparison between the formant- and TV-based coordination features.

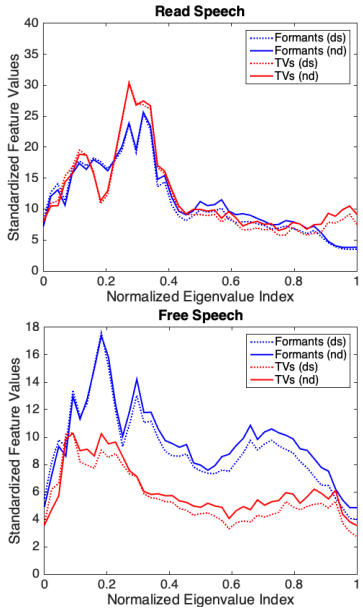

Figure 3: *Standardized feature values of coordination features in the not-depressed speech samples relative to those in the depressed speech samples.*

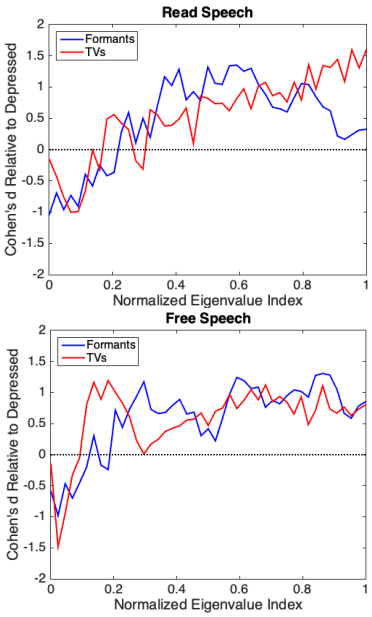

Figure 4: *Effect sizes between the feature-wise means (Cohen's d) of coordination features in the not-depressed speech samples relative to those in the depressed speech samples.*

The correlation structure features are estimated by first computing a channel-delay correlation matrix, using time-delay embedding at a fixed delay scale [3][4]. The delay scale chosen was 7 samples, which corresponds to 7/160 = 44 ms for the formants tracks and 7/100 = 70 ms for the TVs. Each correlation matrix has dimensionality (45 x 45), based on 3 formant or 3 TV channels and 15 time delays per channel. Changes over time in the coupling strengths among the channel signals cause changes in the eigenvalue spectra of the channel-delay matrices.

From the correlation matrix, R_i, associated with sample i, the eigenspectrum is computed. The eigenspectrum takes the form of a 45-dimensional vector, λ_i containing eigenvalues of the correlation matrix, rank-ordered from index $j = 1, \ldots, 45$.

The eigenspectrum can be considered a feature vector (of features j) that can be used to characterize the within-channel and cross-channel distributional properties of the multivariate formant or TV time series. Features of this kind have previously been used for estimating formant coordination changes in the context of depression severity, Parkinson's disease, age-related cognitive decline, mild TBI, and cognitive load [3-6, 11-16]. We have previously found on the Mundt dataset (and other depression datasets) that low-rank eigenvalues from the low HAM-D session are smaller than those from the high HAM-D session. This means that higher depression is associated with a time-embedded formant scatter distribution that is less isotropic, which can be thought of as less "complex." Though still coordinated, less complexity corresponds to more highly coupled movements.

2.5 Classification Experiments

Experiments were conducted to determine the extent to which coordination features, computed over formants and TVs, could be used as the basis for building a model to classify depressed vs. not depressed speech. In previous applications of these features, the eigenspectrum vector has been projected, using principal component analysis (PCA), into a lower-dimensional representation. In the present work, a simpler, less data-intensive approach was employed in which two features were selected from the full eigenspectrum, corresponding to 0.20 and 0.95 normalized eigenvalue index, to act as input features to the model. Indices at these approximate locations have been shown to provide good discrimination between neurological states in prior work [3-6, 11-16]. These indices have been considered previously because they provide a simple, low-dimensional indication of the eigenspectrum shape in terms of linearity and slope. The features were individually standardized (i.e., z-scored) across all instances prior to model training and testing.

Model training and testing was carried out within a leave-one-subject-out cross-validation scheme, with a total of seven (7) total folds. At each fold, a linear classifier was trained on data from six (6) subjects – a total of 12 data points – and used as the basis for estimating a label on the two (2) data points from the remaining subject. Classification accuracy of these estimated labels was calculated across all folds.

3. Results

Figure 3 shows the eigenspectrum features associated with not-depressed speech samples and depressed speech samples. Eigenspectra are plotted as standardized feature-wise means as a function of the normalized eigenvalue feature index $(j-1)/45$. More precisely, for a given feature j, the values of the curves plotted in Fig 3 were calculated according to:

$$\varepsilon_j = \frac{\mu_j^{\gamma}}{s_j} \tag{1}$$

where μ_j^{γ} is the mean feature value, $(1/n_{\gamma})\sum_{i\in\gamma}\lambda_{i,j}$, for all samples taken in the state $\gamma \in$ {not depressed (nd), depressed

state (ds)}. The quantity s_j is the pooled standard deviation, defined as:

$$s_j = \sqrt{\frac{(n_{ds}-1)s_j^{ds} + (n_{nd}-1)s_j^{nd}}{n_{ds}+n_{nd}-2}} \quad (2)$$

Eigenspectrum features are shown for read and free speech, and for both formants and TVs. In comparing eigenspectrum features in the depressed vs. not-depressed case, but under corresponding conditions, the distance between the lines is equivalent to a Cohen's d measure of effect size. This can be seen by considering the definition of Cohen's d:

$$d_j = \frac{\mu_j^{nd} - \mu_j^{ds}}{s_j} \quad (3)$$

Effect sizes are plotted in Figure 4. The largest magnitude Cohen's d value on the read speech using the formants was 1.35, with a mean absolute magnitude of 0.75 across all features, whereas Cohen's d on the TVs went as high as 1.60, and a mean absolute magnitude of 0.77 across all features. The largest magnitude Cohen's d value on the free speech using the formants was 1.31, with a mean absolute magnitude of 0.77 across all features, whereas Cohen's d on the TVs went as high as -1.48, with a mean absolute magnitude of 0.70 across all features.

Table 3 shows accuracies (%) in labeling speech samples as depressed or not depressed, resulting from our classification experiments. Accuracies of 57.1% and 42.9% were observed when formant-based coordination features were used on read and free speech, respectively. Accuracies of 64.3% and 71.43% were seen when using TV-based coordination features on read and free speech, respectively.

Table 3: *Classification. Accuracy (%).*

	Formants	**TVs**
Read Speech	57.1	64.3
Free Speech	42.9	71.4

4. Discussion

Looking across features, effect sizes in Fig. 4 follow the same general pattern in formants and TVs – i.e., negative effect sizes for the low index eigenvalues, and positive effect sizes for the high index eigenvalues – but the TV-related effect sizes are much larger in magnitude toward either end of the eigenspectrum. Classification accuracies are considerably higher when using TVs, as compared to formants. This is consistent with the above larger effect sizes observed in TVs.

The flatter TV-based eigenspectrum for not depressed speech as compared to depressed speech shows that there is a strong difference in articulatory coordination. One interpretation of this difference in the eigenspectrum is that articulators of depressed speech have less complex coordination associated with more coupled movements. We have observed this phenomenon when studying other conditions [3, 4, 11, 12] and future work will focus on whether it holds more generally for the TVs.

A second interpretation for less complexity in depressed speech is reduced coarticulation and lenition. Correspondingly, many studies have cited slower speech rate with more and longer pauses as biomarkers for psychomotor In Fig. 5, we compare productions by patient 127 of "Well, he

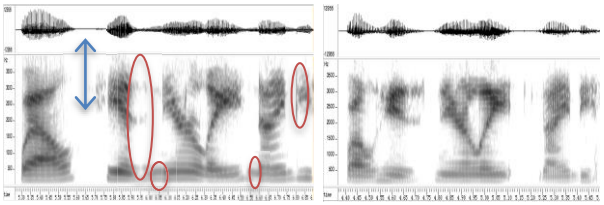

Figure 5: *Spectrogram of "Well, he is nearly ninety" spoken on day zero (left) and day 42 (right) by patient 127. Red ovals highlight the strong voicing during the /z/ is "is", the nasal murmur for the /n/s at the beginning of "nearly" and "ninety" and the /d/ burst in "ninety" The arrow points to the 100 ms pause when she is depressed.*

is ninety-three" excised from the Grandfather Passage. We see that she speaks more slowly when she is depressed (phrase of 1.6s vs. 1.1s). In part. The longer duration is due to the pause after "well" on the day she is depressed that is absent on the day she is not depressed. However, there are also differences related to the degree of coarticulation and lenition. The first oval highlights the /z/ which has significantly more voicing energy at low frequencies and less frication energy at high frequencies which occurs because of a weakened alveolar constriction. On the day she is depressed, she produces nasal murmurs for the word-initial /n/'s in "nearly"(around 92 ms in duration) and "ninety" (67 ms), whereas the word-initial /n/'s are completely coarticulated with the following vowel on the day she is not depressed. Additionally, on the day she is depressed, the /d/ in "ninety" is produced with a closure of 50 ms (spectrogram shows very low-frequency energy due to the vibration of the vocal cords that radiate through the neck) followed by a burst release around 6.8 s. On the other hand, when she is not depressed, the /d/ is produced as a flap (30 ms in duration).

Another view of interpretability involves the inversion method capturing more than 3 formants via a more complete waveform representation, a hypothesis to be addressed by comparing effect size with more than 3 formants. However, one could also argue that the formants contain about place and manner of articulation. Regardless of the precision of our interpretations, based on our findings thus far, this paper presents the first evidence that psychomotor retardation results in less complex coordinated movement of the articulators and, moreover, that direct involvement of articulator parameters as a feature set provides a potentially powerful alternative as a basis for detection of the depressed state. Validation of our preliminary results will require larger and more diverse datasets.

5. Distribution Statement & Disclaimer

Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. 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 Under Secretary of Defense for Research and Engineering.

6. Acknowledgements

We thank Dr. James Mundt for the depression database and Dr. James Williamson for the Matlab software to compute correlation structure.

7. References

[1] G. Sivaraman, V.Mitra, H. Nam, M.K. Tiede, C. Espy-Wilson (2016) Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion, Proc. of INTERSPEECH 2016.

[2] G. Sivaraman, "Articulatory representations to address acoustic variability in speech," Ph.D. dissertation, University of Maryland
College Park, 2017.

[3] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Ciccarelli, G., & Mehta, D. D. (2014, November). Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* (pp. 65-72). ACM.

[4] T. F. Quatieri, J. R. Williamson, C. J. Smalt, J. Perricone, T. Patel,
L. Brattain, B. Helfer, D. Mehta, J. Palmer, K. Heaton, M.EWddy, J. Moran, "Multi-modal Biomarkers to Discriminate Cognitive State", book chapter in "The role of technology in clinical neuropsychology," (March 2017), Oxford.

[5] F. H. Guenther. Neural Control of Speech. Mit Press, 2016.

[6] Ciccarelli, G., Quatieri, T.F., and Ghosh, S., Neurophysiological Vocal Source Modeling for Biomarkers of Disease, 2016 Annual Conference of the International Speech Communication Association.

[7] J. Mundt, P. Snyder, M. S. Cannizaro, K. Chappie, 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.

[8] C. P. Browman and L. Goldstein, "Articulatory Phonology : An Overview *," Phonetica, vol. 49, pp. 155-180, 1992.

[9] Boersma, P and Weenink, D, (2018): Praat: doing phonetics by computer [Computer program].=Version 6.0.37, retrieved 14 March 2018 from http://www.praat.org/

[10] Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*(4), 333-382.

[11] Quatieri, T. F., Williamson, J. R., Smalt, C. J., Patel, T., Perricone, J., Mehta, D. D., & Palmer, J. (2015). Vocal biomarkers to discriminate cognitive load in a working memory task. In *16 Annual Conf. of the Inter. Speech Comm. Assoc*.

[12] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., & Mehta, D. D. (2013, October). Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM Inter. workshop on Audio/visual emotion challenge* (pp. 41-48). ACM.

[13] Williamson, J. R., Quatieri, T. F., Helfer, B. S., Perricone, J., Ghosh, S. S., Ciccarelli, G., & Mehta, D. D. (2015, September). Segment-dependent dynamics in predicting Parkinson's disease. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[14] Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., ... & Quatieri, T. F. (2016, October). Detecting Depression using Vocal, Facial and Semantic Communication Cues. In *Proceedings of the 6th Int. Workshop on Audio/Visual Emotion Challenge* (pp. 11-18). ACM.

[15] 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 *INTERSPEECH* (pp. 1038-1042).

[16] Horwitz R, Quatieri TF, Helfer BS, Yu B, Williamson JR, Mundt J. 2013. "On the Relative Importance of Vocal Source, System, and Prosody in Human Depression." In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, 1–6.

[17] Helfer, B. S., Quatieri, T. F., Williamson, J. R., Keyes, L., Evans, B., Greene, W. N., Vian, T., Lacrignola, J., Shenk, T., Talavage, T., Palmer, J., & Heaton, K. (2014). Articulatory dynamics and coordination in classifying cognitive change with preclinical mTBI. In *INTERSPEECH* (pp. 485-489).

[18] Lammert, A. C., Williamson, J. R., Hess, A., Patel, T., Quatieri, T. F., Liao, H. J., ... & Heaton, K. J. (2017, October). Noninvasive estimation of cognitive status in mild traumatic brain injury using speech production and facial expression. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 105-110). IEEE.

[19] Bennabi, B., Vandal, P., Papaxanthis, C., Pozzo, T., and Haffen, E., "Psychomotor Retardation in Depression: A Systematic Review of Diagnostic, Pathophysiologic, and Therapeutic Implications, Biomed Research International, doi: 2013 Oct 30.