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

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

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

1. INTRODUCTION

- Major Depressive Disorder (MDD)
  - Long-lasting depressed mood or loss of interest in activities
  - Monitoring and providing treatments heavily rely on human intervention
- Automated solutions can provide the patient and their therapists with timely information to assess their mental health
- Depression is associated with changes in speech
  - Features derived from speech are expected to capture information which can distinguish depressed speech from non-depressed speech
- Psychomotor Slowing (PMS) [1]
  - A condition of slowed neuromotor output that manifests changes in speech, ideation, and motility
  - A long-established necessary feature of MDD that can track its severity
  - Altered coordination and timing across articulators

2. OVERVIEW OF THE STUDY

- Articulatory coordination features:
  - Previously extensively applied to the first three formant time series as a proxy for vocal articulation [2]
- **Objective**: Use of direct articulatory parameters from a speech inversion system (Vocal Tract Variables - TVs) to quantify changes in the way speech is produced by depressed and non-depressed subjects
- A preliminary study showed that 3 TVs corresponding to constriction degree can outperform 3 formants in classifying depressed vs. not depressed speech [3]
- Extending the preliminary study by:
  - Including results from adding constriction location TVs
  - Using a wider range of coordination features as inputs to the classification model
  - Using data from additional subjects

3. ACOUSTIC-TO-ARTICULATORY SPEECH INVERSION SYSTEM

- Based on Articulatory Phonology

Constriction Organ	Tract Variable	Articulators
Lip	Lip Aperture (LA) Lip Protrusion (LP)	Upper Lip, Lower Lip, Jaw
Tongue Body	Tongue body constriction degree (TBCD) Tongue body constriction location (TBCL)	Tongue Body, Jaw
Tongue Tip	Tongue tip constriction degree (TTCD) Tongue tip constriction location (TTCL)	Tongue Body, Tip, Jaw
Velum	Velum (VEL)	Velum
Glottis	Glottis (GLO)	Glottis

- Feedforward Network trained on Wisconsin X-ray Microbeam database [4]

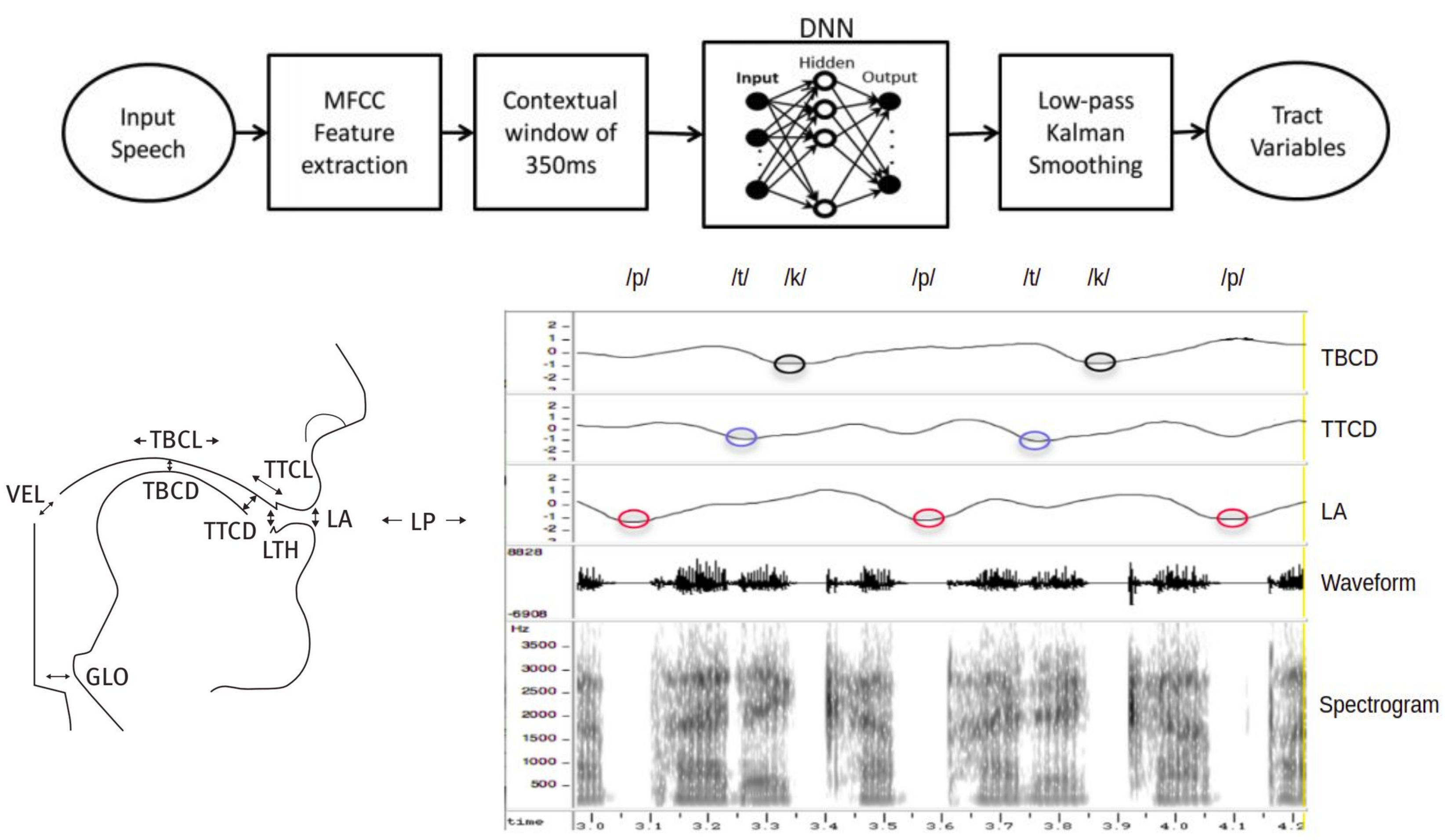

Figure 1: The schematic of the DNN based speech inversion system and an example of estimated TVs.

4. ARTICULATORY COORDINATION FEATURES

To characterize the level of articulatory coordination and timing.

Step 1:

A channel-delay correlation matrix is computed from feature vectors at a specified delay scale (Eg: 7 samples = 70ms)

- Each time-series signal is shifted by multiples of the delay scale (7 samples) up to 15
- Auto- and cross- correlations are computed among these shifted time series signals

Each correlation matrix R_j has dimensionality (MN x MN), based on M channels and N time delays per channel:

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

Step 2:

An eigenspectrum is computed from the correlation matrix, taking the form of an *MN*-dimensional feature vector

- Magnitude of eigenvalues represent the average correlation in the direction of corresponding eigenvectors
  - Depressed speech has few eigenvalues with significant magnitudes
  - Thus, depressed speech can be represented using a few independent dimensions compared to non-depressed speech

5. DATASET & EXPERIMENTAL SET-UP

- Mundt Database [5]:
  - Data collected from 35 physician-referred patients over a six week period
  - Hamilton Depression (HAMD) Rating Scale used for assessment
  - Speech types used - free speech (FS), read speech (RS)

	Not Depressed	Excluded from the study	Depressed
HAMD Score	0-7	8-19	20-52

	# Dep Segments	Dep Mean Duration	# Ndep Segments	Ndep Mean Duration
Free Speech	51	17.47 s	66	48.62 s
Read Speech	33	52.27 s	20	45.84 s

Comparison among coordination features derived from:

3 TVs (constriction degree TVs only - LA, TTCD, TBCD), 6 TVs (location constriction degree TVs), and first 3 formants

6. ANALYSIS OF COORDINATION FEATURES

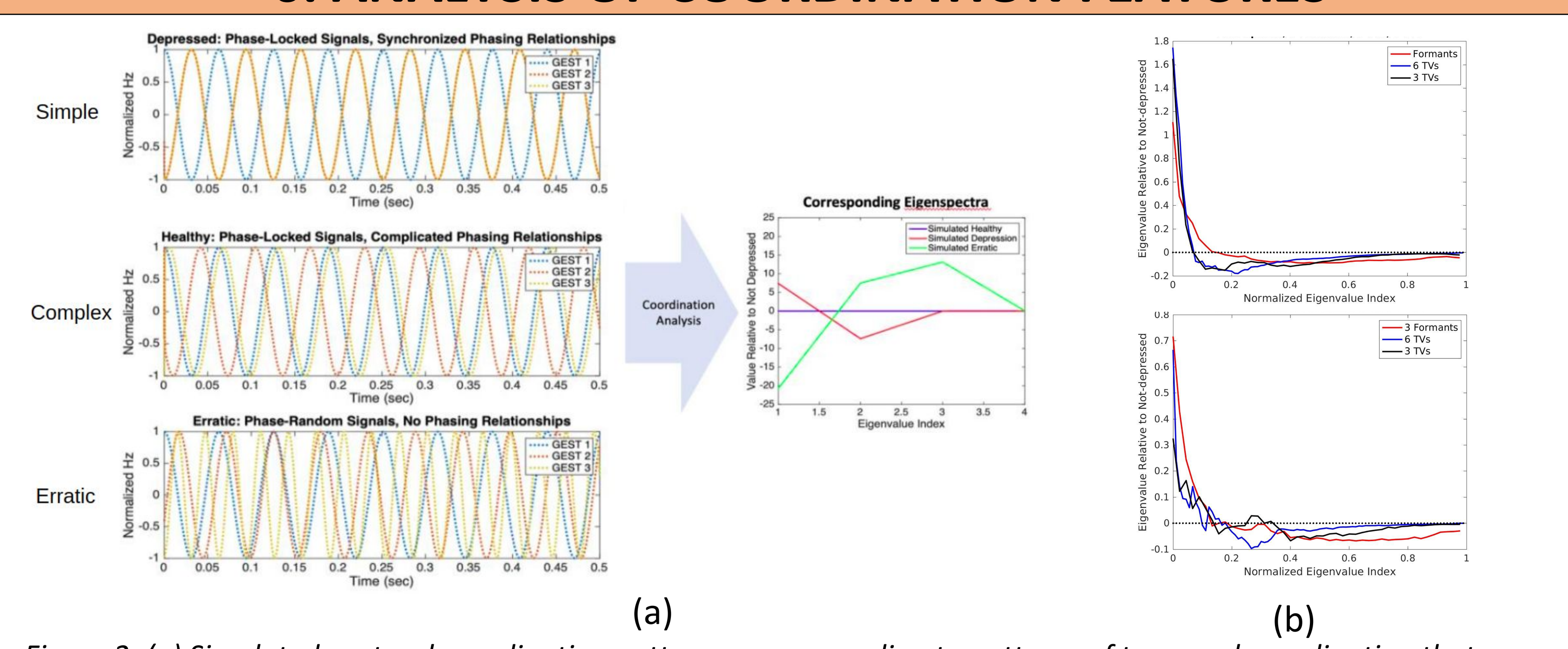

Figure 2: (a) Simulated gestural coordination patterns, corresponding to patterns of temporal coordination that are either oversimplified, speech-appropriate, or erratic. Associated eigenspectra show differences resulting from these different coordination patterns. (b) Relative differences plotted using free speech (top) and read speech (bottom).

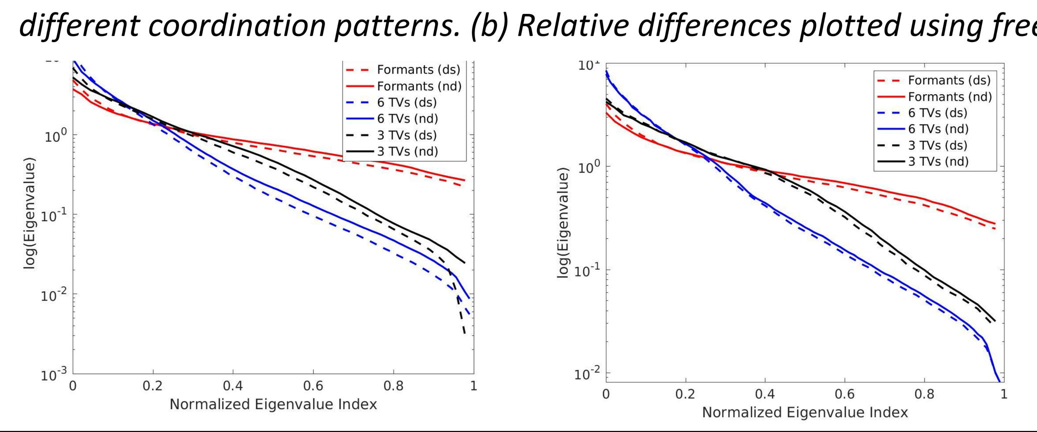

Figure 3: Eigenspectra of free speech (left) and read speech (right) obtained from 3 TVs, 6 TVs and 3 Formants

7. CLASSIFICATION EXPERIMENTS

- Leave-one-subject-out cross-validation scheme using an **SVM Classifier**
  - The features were individually standardized (i.e., z-scored) across all instances prior to model training and testing
  - Averaged the eigenspectrum features in different ranges to obtain a low-dimensional representation of the high dimensional eigenspectrum feature vector

8. EXTENDED WORK BASED ON THIS STUDY

Based on the findings of this study we made several improvements to the classification model over the past few months.

1. ***Using a more complete representation of TVs by adding glottal parameters (8 TVs in total) [6]***
   - Comparison with MFCCs showed a relative classification accuracy improvement of 8%
2. ***A deep learning based model was developed using a modified correlation matrix as the inputs [7]***
   - Use of dilated CNNs to incorporate multiple delay scales
   - More data points were created by segmenting longer segments with overlaps
   - Depressed class: HAMD > 7, Non-depressed class: HAMD <= 7
   - Heavily imbalanced

	# Dep Segments	Dep Mean Duration	# Ndep Segments	Ndep Mean Duration
Free Speech	2131	19.91 s	528	16.79 s
Read Speech	730	20 s	123	20 s

-Assigned class weights for the minority class

-AUC-ROC reported in addition to accuracy

- TVs show promise as a robust feature for depression classification task

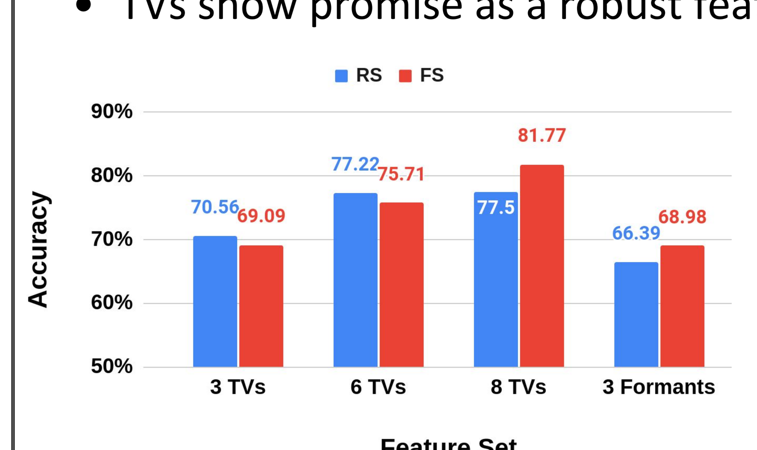

Figure 4: SVM Classification Accuracies of Free Speech (FS) and Read speech (RS)

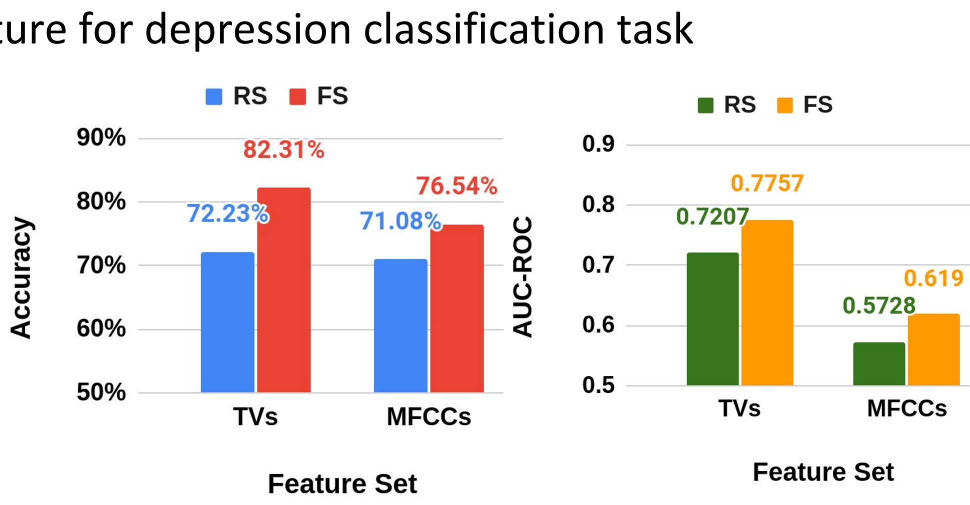

Figure 5: Dilated CNN Classification Accuracies and AUC-ROCs of Free Speech (FS) and Read speech (RS)

9. REFERENCES

[1] Christina Sobin and Harold Sackeim. "Psychomotor symptoms of depression". In: The American journal of psychiatry (1997)

[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

[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] Seneviratne, N., Williamson, J.R., Lammert, A.C., Quatieri, T.F., Espy-Wilson, C. (2020) Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression. Proc. Interspeech 2020, 4551-4555

[7] Seneviratne, N., Espy-Wilson, C., (2020) Deep Learning Based Generalized Models for Depression Classification. Submitted to ICASSP 2021 [arxiv]