Multimodal Approach for Assessing Neuromotor Coordination in Schizophrenia Using Convolutional Neural Networks

Yashish M. Siriwardena[1], Chris Kitchen[2], Deanna L. Kelly[2], Carol Espy-Wilson[1]
[1]University of Maryland College park, MD, USA, [2]University of Maryland School of Medicine, Baltimore, MD, USA

ACM ICMI 2021 MONTREAL

1. INTRODUCTION

- Schizophrenia is a chronic mental disorder with heterogeneous presentations
- Symptoms of schizophrenia are broadly categorized as,
  - Positive (e.g. hallucination, delusions)
  - Negative (e.g. blunted effect, alogia)
  - Cognitive (e.g. disorganized thinking, slow thinking)
- Previous studies in Major Depressive Disorder (MDD) support affects of neurophysiological changes to speech production and facial movements
- Capturing these neurophysiological changes by coordination features based on the correlation structure of the movements of various articulators
- Study focuses on Schizophrenic patients who are ***markedly ill*** and exhibit strong ***positive symptoms in schizophrenia***

2. Dataset

• Details of the database collected for the collaborative observational study

Longitudinal	5 weeks
Number of Subjects	31 M, 30 F
31 M, 30 F	26 African American, 28 Caucasian, 5 Asian
Assessment	HDRS, MADRS, BPRS, CAPE-42 (Weeks 1,3,5)
Recording Type	Video and Audio
Session Length	10-50 mins

• Details of the subject data used for the study

	SZ	HC
No of Subjects	7	11
BPRS range	45<score<=62	18<score<=23
HAMD range	0<score<14	0<score<7
Mean session duration	35 min	18 min
No of Utterances	1208	1132
Hours of speech	10.0	9.43

3. Audio Feature 1: Vocal Tract Variables

Constricted 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

Speech Inversion System

Aperiodicity, Periodicity and Pitch (APP) detector

4. Audio Feature 2 : MFCCs

- 13 MFCCs from the librosa python library
- Analysis window of 20 ms with a 10 ms frame shift
- Only 12 MFFCs were used for analysis by discarding the 1st coefficient

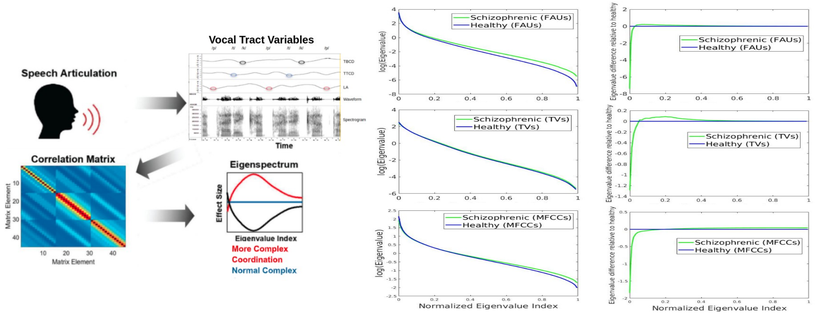
5. Video Features : Facial Action Units (FAUs)

- List of 17 FAUs extracted from the Openface 2.0 Facial Behavior Analysis toolkit

FAU No	FAU Name	FAU No	FAU Name
1	Inner brow raiser	14	Dimpler*
2	Outer brow raiser	15	Lip corner depressor*
4	Brow raiser	17	Chin raiser*
5	Upper lid raiser	20	Lip stretcher*
6	Cheek raiser*	23	Lip tightner*
7	Lid tightner*	25	Lips part
9	Nose wrinkler*	26	Jaw drop
10	Upper lip raiser*	45	Blink
12	Lip corner puller*		

*FAUs used in the TDEC method

6. Coordination feature 1 : Time delay embedded correlation analysis (TDEC)



7. Coordination feature 2 : Full Vocal Tract Coordination (FVTC)

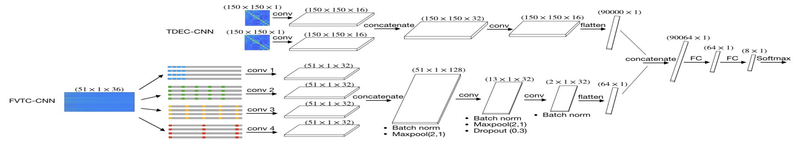
- Huang et al. in a recent study with MDD introduces a new channel delay correlation method inspired by TDEC
- FVTC includes every correlation within the considered D (design choice) frames
- Avoids the repetitive use of same correlations as in the TDEC correlation matrix

8. Unimodal Systems

- Model 1 : Parallel delay scale TDEC-CNN model (TDEC-CNN)
  - 2 coordination matrices with 2 delay scales as parallel inputs for two 2D-CNN layers
- Model 2 : FVTC CNN model (FVTC-CNN)
  - FVTC coordination matrix as the input to a Dilated-CNN layer
- Leave one subject out cross validation for model training
- Subject level classification derived from segment level predictions

	TDEC-CNN (Model 1)		FVTC-CNN (Model 2)	
	Accuracy (%)	F1(SZ)/F1(HC)	Accuracy (%)	F1(SZ)/F1(HC)
FAU	**83.33**	**0.80/0.86**	83.33	0.77/0.89
TV (8 TVs)	66.67	0.57/0.73	**72.22**	**0.62/0.78**
MFCC	61.11	0.46/0.70	**72.22**	**0.55/0.80**
MFCC + Glottal TVs	60.05	0.45/0.69	**72.22**	**0.55/0.80**

9. Multimodal Systems



Models	Accuracy (%)	F1(SZ)/F1(HC)
FAU (Model2)+TV(Model2)	66.67	0.67/0.67
FAU (Model1)+TV(Model1)	72.22	0.67/0.76
FAU (Model2)+MFCC(Model2)	72.22	0.62/0.78
FAU (Model1)+MFCC(Model2)	77.78	0.67/0.83
FAU (Model1)+(MFCC+Glottal TVs)(Model2)	83.33	0.73/0.88
FAU (Model1)+TV(Model2)	**88.89**	**0.86/0.91**

10. Summary on Key findings

- Subjects with schizophrenia who exhibit strong positive symptoms follow a pattern which suggests **higher articulatory coordination complexity** compared to healthy controls
- FAUs outperform TVs and MFCCs in classification metrics
- TVs perform better than MFCCs in both coordination feature types
- Models with heterogeneous architectures perform the best when fused
- Multimodal fusion of features significantly improve classification performance