Towards an Interpretable Index Score for the Assessment of Schizophrenia based on Multimodal Speech and Facial Biomarkers

Michael Neumann[1], Hardik Kothare[1], Christian Yavorsky[2], Anzalee Khan[3], Jean-Pierre Lindenmayer[3,4], and Vikram Ramanarayanan[1,5]
[1]Modality.AI, Inc., [2]Valis Bioscience, [3]Nathan S. Kline Institute for Psychiatric Research,
[4]New York University, School of Medicine, [5]University of California, San Francisco
v@modality.ai

Motivation and Research Question

- **Schizophrenia** is a mental disease that causes hallucinations, delusions, and disordered thinking
- **Speech and oro-facial biomarkers** are promising for remote assessment and monitoring

Goal: combine speech and facial biomarkers into one composite index score

→ Useful as an endpoint in clinical practice & pharmaceutical trials

→ Better noise robustness and statistical power than multiple individual markers

→ Maintain interpretability of clinically meaningful metrics

Research Question:
Given a large, multicollinear feature set from remote audiovisual assessments, **how can we determine an interpretable composite index score for remote monitoring of Schizophrenia?**

Data and Feature Selection

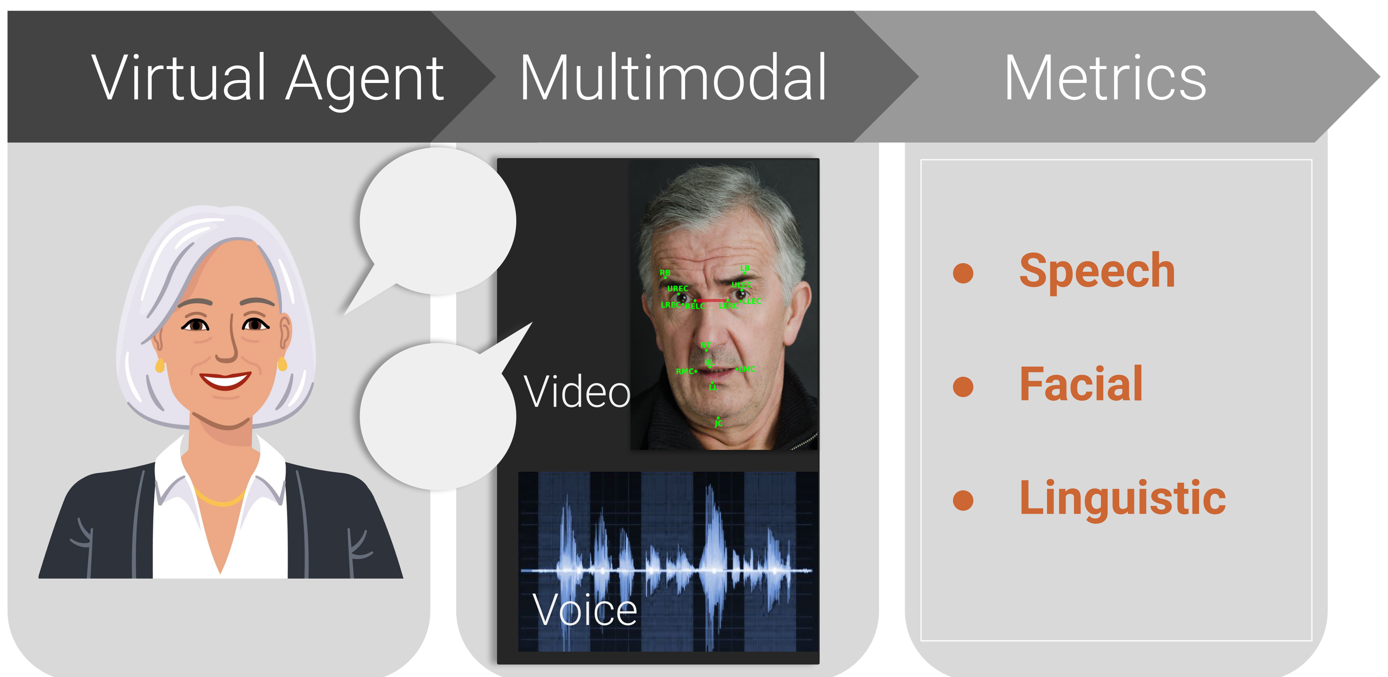

Figure 1. Modality.AI dialogue platform.

- **Multimodal dialogue platform** used to collect audiovisual data (illustration Fig. 1); sessions were overseen by a psychiatrist
- Speech assessments included: **diadochokinesis (DDK), reading passage (RP), picture description (PD), spontaneous speech (S)**
- Clinician administered rating scales for people with Schizophrenia (pSz): PANSS, BNSS, CDSS, CGI-S, AIMS, SAS, BARS
- **Patient eligibility:** Inpatients with diagnosis of schizophrenia, age 18-60, English speaking, WRAT-IV Reading Score ≥ 8th grade, Negative symptoms as evidenced by score of ≥ 18 on PANSS Marder Negative Symptom Factor
- **Healthy control eligibility:** Individuals with no prior history of mental illness, age 18-60, English speaking

	Number of participants	Mean age ± SD (years)	Median BNSS and PANSS ± SD*
People with Schizophrenia	48 (12 female)	39.2 ± 10.9	BNSS: 38.0 ± 9.4 PANSS P: 16.0 ± 4.6 PANSS N: 25.0 ± 2.6
Healthy controls (HC)	63 (29 female)	39.2 ± 11.0	-

Table 1: Demographics. BNSS ranges from 0 to 78, PANSS Positive & Negative range from 7 to 49. * at first visit

- Extracted **Acoustic**, **visual (facial)**, and **linguistic** features
- Identify multicollinear features: **Hierarchical clustering** on Spearman rank-order correlations between features
- **Select one feature** for each cluster based on ROC analysis

Feature cluster	Selected representative
Voice quality	CPP (RP) Shimmer (PD)
Timing	CTA (RP)
Jaw movement	avg. JC speed (PD)
Mouth measures	max. mouth surface area (PD)
Lip movement	avg. LL jerk (RP) max. LL velocity down (S)
Eyes	avg. eye opening (DDK)
Lexico-semantic	noun-to-pronoun ratio (PD) word count (S)

Table 2: Selected features. CTA: canonical timing alignment, CPP: cepstral peak prominence, JC: jaw center, LL: lower lip.

Index Score Computation

Method	Description
Baseline	Linear combination with **equal weights**
LDA	Linear combination that **maximizes area under ROC curve** (AUC) Caveat: assumption of Gaussian distributions
Logistic regression	**Logistic regression coefficients as weights** for linear combination; L1 regularization enforces sparse weight vector
Constrained log. regr.	**Logistic regression coefficients** as weights, constrained to be **non-negative**

Table 3: Methods to compute an index score as (weighted) linear combination of features. Features were inverted by taking (1 - scaled feature) when median value was smaller in Control cohort in the train set. LDA: Linear Discriminant Analysis

Results

- **Index scores yield better test results** than individual metrics with **reduced variability**
- All methods >80% UAR
- Weak to moderate negative **correlations between index scores and negative symptoms** (-0.35 (p=0.001) b/w cLogReg and BNSS total, -0.37 (p=0.001) b/w cLogReg and PANSS Negative total)

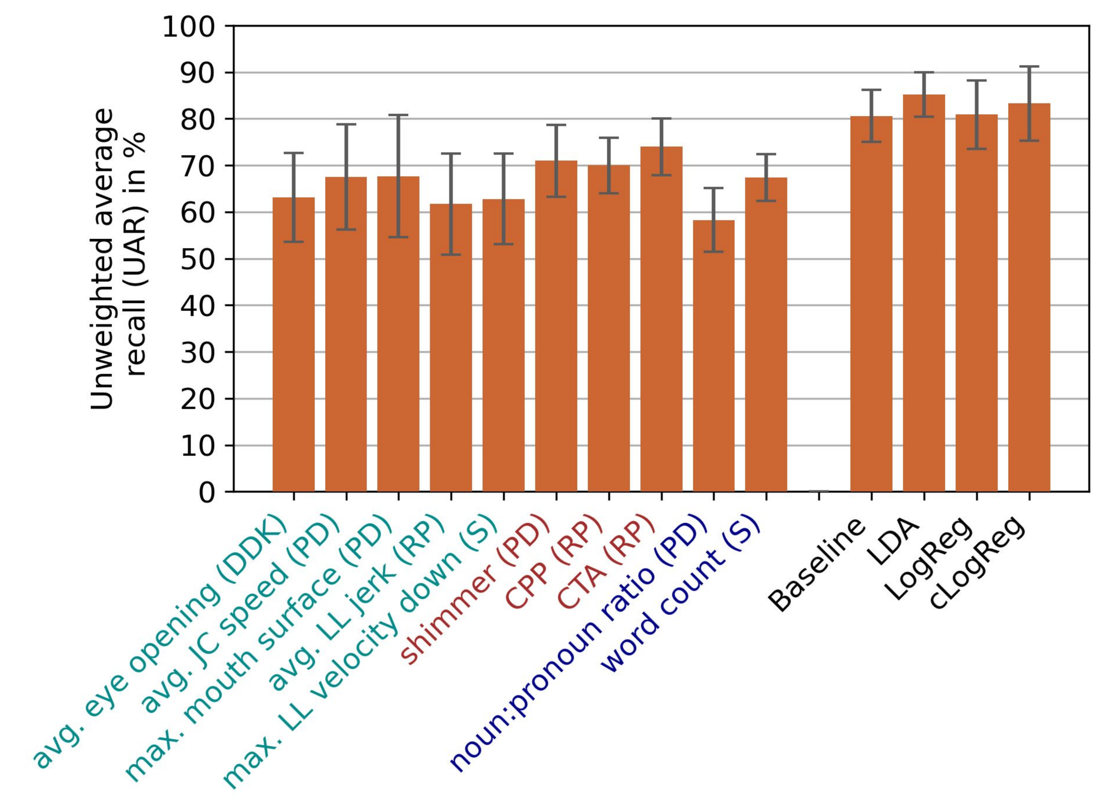

Figure 2. Classification accuracy for individual features and for the different index scores, for the binary classification pSz vs. controls. Error bars represent standard deviation across validation folds for 5-fold cross validation. LogReg: logistic regression, cLogReg: constrained logistic regression.

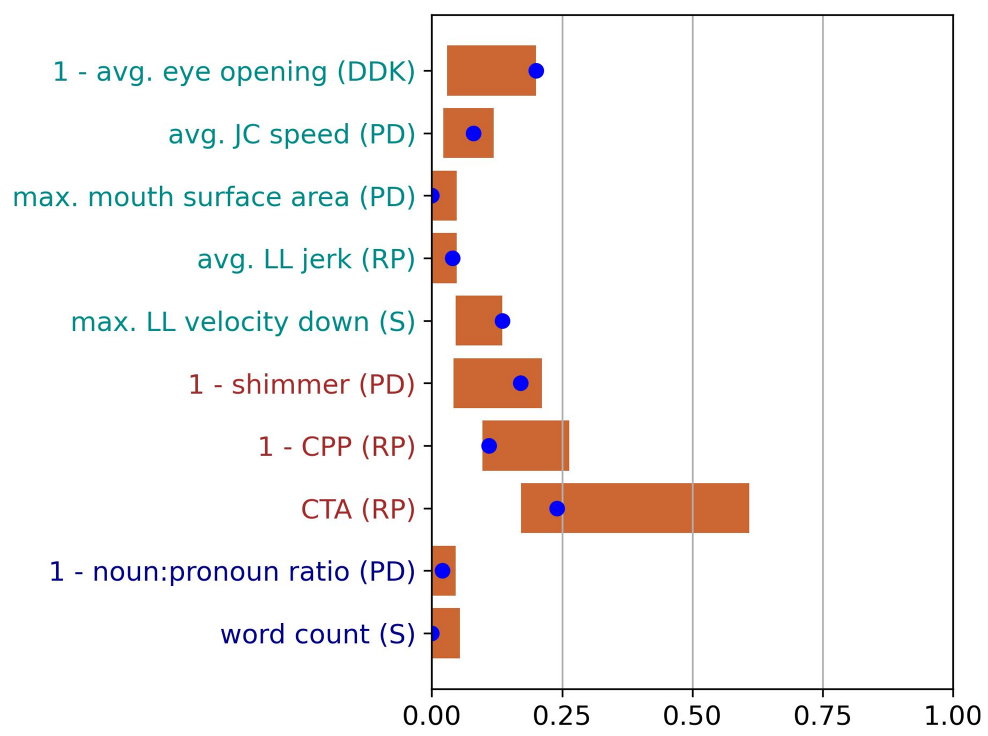

Figure 3. Normalized feature weights (constrained log. regr.) expressed as ranges that result from the variation across validation folds. Blue dots represent the weights from one representative validation fold.

- **Interpretation of index:** score decreases with increasing severity of neg. symptoms
- Normalized weights reveal **contribution of each component metric**
- Speech metric CTA is assigned highest weights
- Facial metrics add valuable information
- Linguistic features not given as much weight

Conclusions

Key findings:

- Proposed a method to combine speech, oro-facial, and linguistic features into **one composite index → potential endpoint for trials**
- Index scores **improve classification accuracy** and **reduce variation** within cross validation
- Weighted linear combinations **maintain interpretability**
- For differentiating pSz from HC, speech features are most dominant

Limitations and future work:

- Variation in feature weights depending on training data
- Index tailored to specific task (classify pSz and healthy controls) – future work will explore broader use case

modality.ai nki NYU School of Medicine UCSF