Task Matters: A Methodological Comparison of Speech Elicitation Types for Machine Learning Classification of Psychiatric Disorders and Controls

Felix Menne[1], Felix Dörr[1], Johannes Tröger[1], Alexandra König[1,2,3], Diana Immel[4], Simon Barton[4], René Hurlemann[4]

[1]ki:elements GmbH, Saarbrücken, Germany; [2]Cobtek (Cognition-Behaviour- Technology) Lab, University Côte d'azur, Nice, France; [3]Université Côte d'Azur, Centre Hospitalier et Universitaire, Clinique Gériatrique du Cerveau et du Mouvement, Centre Mémoire de Ressources et de Recherche, Nice, France; [4]Dept. of Psychiatry at Karl-Jaspers Clinic, School of Medicine & Health Sciences, Carl von Ossietzky University of Oldenburg

Methodological Issue

Digital speech biomarkers are increasingly investigated as objective measures of psychiatric disorders such as major depressive disorder (MDD) and schizophrenia (SZ). However, it remains unclear which types of speech elicitation tasks best capture the disorder-relevant signal needed for diagnostic differentiation. Because task demands can systematically shape speech output, task selection is a key methodological factor that can affect robustness and comparability across studies. Identifying the most informative task type is therefore essential for optimizing data collection while minimizing participant burden in digital assessments.

Background/Aims

Advances in computational linguistics and acoustic signal analysis have made it possible to quantify subtle speech abnormalities associated with psychiatric disorders. Building on these developments, this study compares four types of speech tasks: positive, neutral, and negative autobiographical recall, and a structured picture description (Boston Cookie Theft). The aim is to determine which task provides the strongest discriminative signal for classifying individuals with MDD, SZ, and healthy controls (HC) to inform future methodological standards in speech-based biomarker research.

Table 1: Demographic and clinical information of the sample

	HC	MDD	SZ	Group difference *p*-value
n	22	22	22	-
Age	41.09 (11.36)	39.73 (14.19)	40.77 (13.37)	0.91
Sex	m: 11; f: 11	m: 11; f: 11	m: 12; f: 10	-
Years of Education	13.26 (2.0)	11.0 (1.64)	10.86 (1.83)	<0.001
BDI-II	2.73 (2.43)	23.32 (10.47)	18.9 (11.69)	<0.001
MADRS	-	17.62 (6.959	-	-
PANSS	-	-	53.45 (15.85)	-

Disclosure

FM, FD, JT, and AK are employed by the speech biomarker company ki:elements. JT holds shares in ki:elements. The remaining authors have nothing to disclose.

Methods

A total of 66 participants were included: 22 with MDD, 22 with SZ, and 22 healthy controls recruited from the Karl-Jaspers Clinic of Psychiatry, University Hospital Oldenburg, Germany. Each participant completed four elicitation tasks: description of a positive, a neutral, and a negative autobiographical event, and a description of the Boston Cookie Theft picture. Speech recordings were processed to extract 92 acoustic (e.g., prosody, formant, spectral) and linguistic (e.g., syntactic complexity, lexical richness, sentiment) features. Multiple machine learning classifiers (Decision Trees, Extra Trees, Support Vector Machines, Linear Models) were trained in pairwise diagnostic comparisons. Classification performance was evaluated via cross-validation, and receiver operating characteristic (ROC) curves were generated for each task to compare discriminative ability across conditions.

Table 2: Classification performance by speech elicitation task and diagnostic contrast. Values report AUC (ROC), sensitivity, specificity, best-performing classifier, and number of selected features (n). Results are from cross-validation. PD = picture description; DT = decision tree; LM = linear model; RF = random forests.

	AUC	Sensitivity	Specificity	Model used	k features used
HC vs. MDD					
PD	0.96	0.91	1.00	DT	10
Pos. Story	0.96	0.91	1.00	DT	10
Neg. Story	0.96	0.91	1.00	DT	10
Neutr. Story	0.96	0.91	1.00	DT	10
HC vs. SZ					
PD	0.74	0.64	0.64	LM	44
Pos. Story	0.76	0.59	0.77	LM	10
Neg. Story	0.83	0.77	0.82	LM	44
Neutr. Story	0.80	0.68	0.73	LM	30
MD vs. SZ					
PD	0.91	0.77	0.96	RF	10
Pos. Story	0.96	0.91	1.00	DT	10
Neg. Story	0.96	0.91	1.00	DT	10
Neutr. Story	0.96	0.91	1.00	DT	10

Results

Across diagnostic group contrasts, performance patterns were consistent across tasks (Table 2). For HC vs. MDD, all four tasks yielded identical results (AUC = 0.96; sensitivity = 0.91; specificity = 1.00; 10 features). Similarly, SZ vs. MDD showed comparable performance for autobiographical recall (AUC = 0.96; sensitivity = 0.91; specificity = 1.00), with slightly lower discrimination for picture description (AUC = 0.91). For HC vs. SZ, discrimination was moderate and more variable across tasks (AUC = 0.74–0.83), with sensitivity, specificity, and selected features varying more widely.

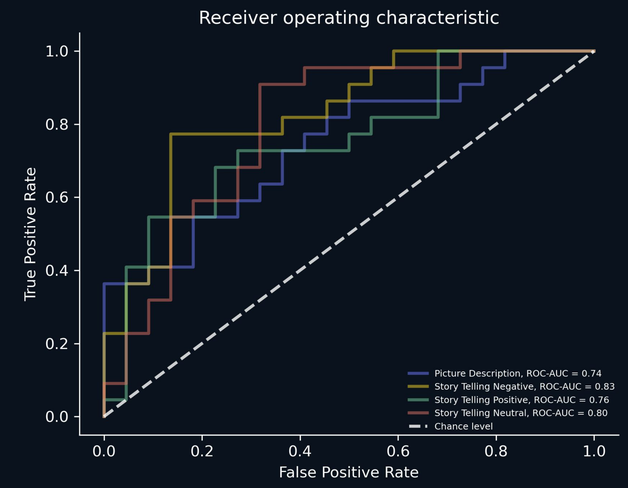

Figure 1: Receiver operating characteristic (ROC) curves for the schizophrenia (SZ) vs. healthy control (HC) classification across four speech elicitation tasks. Curves show task-specific performance (AUC) for positive, neutral, and negative autobiographical recall, and picture description (PD).

Conclusion

For MDD-related comparisons, consistent classification performance across autobiographical recall tasks indicates a stable, largely task-invariant diagnostic speech signal. In contrast, lower and more variable performance in the schizophrenia comparisons suggests greater task dependence. Methodologically, these results show that high discriminative accuracy can be achieved with lesser task standardization for depression, while task optimization may be more critical for schizophrenia and picture description tasks. These findings refine methodological approaches for speech-based digital phenotyping in psychiatric research.