Investigating Conversational Speech Latency as a Marker of Cognitive and Temporal Processing During Picture Description Tasks across CNS Conditions

Cathy Zhang[1], Hardik Kothare[1], Michael Neumann[1], Beverly Insel[2], Anzalee Khan[2], Jean-Pierre Lindenmayer[2] and Vikram Ramanarayanan[1,3]

[1] Modality.AI, Inc., [2] Nathan Kline Institute,[3] University of California, San Francisco, CA, USA

<hardik.lkothare, vikram>@modality.ai

Methodological Question and Introduction

- **Speech latency**—the time between the end of a conversational prompt and the onset of a verbal response—integrates **cognitive, linguistic, motivational and motor processes** involved in speech production and turn-taking.
- Altered latency has been associated with **schizophrenia, depression and MCI**, but its value as a **cross-condition digital biomarker** remains unclear.
- Manual latency annotation is time-intensive and difficult to scale. Automated speech recognition, including Whisper, enables scalable extraction.

Research Question 1: Analytical validation — Can an automated pipeline approximate manual annotation with sufficient accuracy across diverse clinical recordings?

Research Question 2: Clinical validation — Does automated latency discriminate patients from controls and track symptom severity across conditions?

Data and Methods

	SCZ	DEP	MCI
Participants	175	320	200
Controls	103	76[1]	100
Patients	72	244[1]	100
Sessions	334	320	362
Manual annotated	334	148	347
Age (controls)	36.4 (11.6)	48.8[1] (14.7[1])	71.3 (8.6)
Age (patients)	43.0 (11.7)	38.0[1] (12.8[1])	71.0 (9.1)
% Female	37%	51%	10%
Clinical scale	PANSS, BNSS	PHQ-8	—

Table 1: Participant Demographics. [1]Depression participants did not have clinical diagnoses. We label participants with a PHQ-8 score < than 5 as Controls (indicating minimal symptoms) and those scoring 5 or above as Patients, following standard cutpoints.

- Virtual guide (**Tina**) administered standardized picture description prompts; audio and video were recorded.
- Participants were instructed to respond immediately after the prompt, engaging **rapid turn-taking, response planning, and spontaneous speech initiation**.

Automated latency annotation

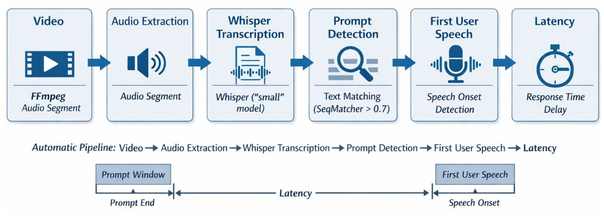

Figure 2. Automated latency pipeline

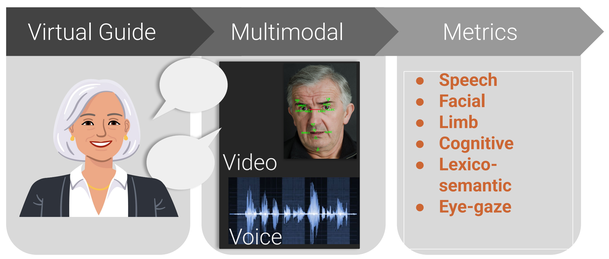

Figure 1. Schematic of the Modality dialogue platform

Manual latency annotation

- Praat + synchronized video were used to mark prompt offset and response onset, including fillers but excluding non-speech noises.
- The same trained annotator applied the protocol across all cohorts while **blinded** to **clinical status** and **symptom scores**.

Statistical analysis

- **Mann–Whitney U tests** and **Glass's Δ** assessed patient–control differences using manually-annotated latency values; **Spearman's ρ** tested associations with symptom severity.
- Manual vs. automated latency was evaluated using **MAE**, **MedAE**, and **Spearman's ρ**, with analyses repeated using automated measures.

Results and Discussion

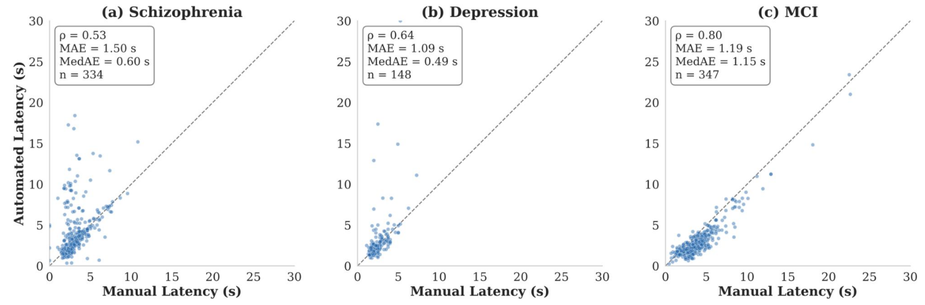

Figure 3. Manual vs. automated speech latency for (a) schizophrenia (n = 334), (b) depression (n = 148), and (c) MCI (n = 347).

	Manual			Automated		
Comparison	Mean (SD)	Δ	*p*	Mean (SD)	Δ	*p*
SCZ: Ctrl vs Pat	2.79 (1.12) / 3.80 (1.71)	**0.90**	<.001	3.39 (2.49) / 4.91 (3.31)	**0.61**	<.001
DEP: PHQ <5 vs ≥5	2.87 (1.11) / 2.61 (1.01)	−0.23	.119	3.42 (2.79) / 3.18 (3.28)	−0.07	.757
DEP: PHQ <5 vs ≥20	2.87 (1.11) / 2.46 (0.77)	−0.37	.180	3.42 (2.79) / 3.13 (2.84)	−0.10	.848
PHQ≥20: Med Off vs On	2.03 (0.69) / 2.71 (0.68)	**0.98**	.013	2.51 (1.09) / 3.54 (3.65)	0.94	.397
MCI: Ctrl vs MCI	4.40 (3.61) / 4.03 (1.98)	−0.10	.795	4.68 (4.22) / 4.15 (2.68)	−0.11	.717

Table 2: Group comparisons and automated–manual agreement. Δ refers to Glass's Delta.

	Manual		Automated	
Scale	*ρ*	*p*	*ρ*	*p*
Schizophrenia (n = 119)				
PANSS Negative	**0.31**	<.001	0.10	.280
PANSS Positive	−0.07	.460	−0.13	.151
PANSS General	0.10	.301	−0.12	.179
PANSS Exp. Deficit	**0.20**	.030	0.08	.400
BNSS Total	**0.22**	.019	−0.003	.975
BNSS Expressive	**0.20**	.028	−0.02	.860
BNSS Experiential	0.16	.093	−0.08	.389
Depression: continuous PHQ-8				
148-session subset	−0.08	.34	0.03	.73
320-session expanded	—	—	0.007	.91

Table 3: Symptom and severity correlations with speech latency (Spearman ρ). Schizophrenia correlations are computed among patients only (n = 119). The MCI cohort is not included because no clinical diagnostic test scores or continuous symptom severity ratings were available for correlation analysis.

- Speech latency was **most strongly associated with schizophrenia and expressive negative symptoms**, while depression and MCI showed weaker or non-significant effects, suggesting latency captures specific cognitive-motor dimensions rather than overall disease severity.

Conclusions

- **Speech latency may not be a universal psychiatric biomarker but may provide a condition-sensitive measure of cognitive-temporal coordination during conversation.**
- **Automated extraction is feasible, but clinical utility will require improved temporal precision and population-specific refinement of ASR-based methods.**

References

- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492-28518). PMLR.
- Ramanarayanan, V., Pautler, D., Arbatti, L., Hosamath, A., Neumann, M., Kothare, H., Roesler, O., Liscombe, J., Cornish, A., Habberstad, D. and Richter, V., 2023. When words speak just as loudly as actions: Virtual agent based remote health assessment integrating what patients say with what they do. In Proc. Interspeech 2023 (pp. 678-679).
- Zhang, C., Kothare, H., Neumann, M., Insel, B., Khan, A., Lindenmayer, J.-P., & Ramanarayanan, V. (2026). 16th International Conference on Auditory-Visual Speech Processing (AVSP 2026)

Acknowledgements

We thank all study participants, whose time and effort made this research possible.