Investigating Conversational Speech Latency As a Digital Biomarker of Schizophrenia: A Comparison of Manual and Automated Approaches

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

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

<cathy.zhang, hardik.kothare, vikram>@modality.ai

Methodological Question and Introduction

- Schizophrenia is characterized by negative symptoms, including reduced spontaneity and increased latency in conversational speech.
- Conversational **speech latency**—the interval between prompt completion and patient response — can serve as a remotely measurable digital biomarker.
- Manual annotation of latency from remote video assessments is labor-intensive and limits scalability.

Research Question 1 (*Clinical validity – manual*): *Does manually annotated conversational speech latency discriminate between patients with schizophrenia and healthy controls in remote picture description tasks?*

Research Question 2 (*Clinical validity – automated*): *Does automatically computed speech latency, using Whisper-based transcription and timing, retain discrimination between cohorts comparable to manual annotation?*

Research Question 3 (*Analytical validity*): *How accurately does the automated pipeline recover manual latency at the session level (e.g., MAE, correlation, and error thresholds)?*

Data and Methods

	Number of participants	Number of sessions	Mean age ± SD (years)
Healthy controls	64 (59 female)	200	37.6 ± 12
Schizophrenia	77 (22 female)	149	42.3 ± 11.6

Table 1: Participant Demographics

- Virtual guide (**Tina**) administered standardized picture description prompts; audio and video were recorded.

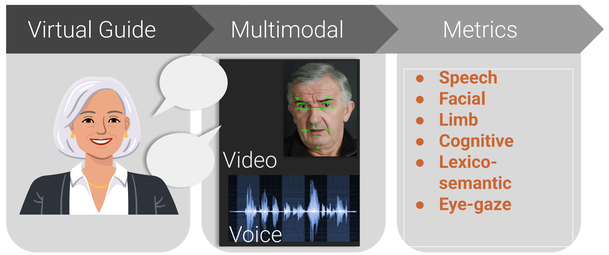


Figure 1. Schematic of the Modality dialogue platform

Manual latency annotation

- For each picture description session, annotators marked the end of the virtual agent's prompt and the onset of the participant's first speech response.
- Manual latency defined as latency = speech start time - prompt end time.
- Group differences between cohorts were evaluated using Mann–Whitney U tests, and effect size quantified using Glass's Delta.

Automated latency annotation

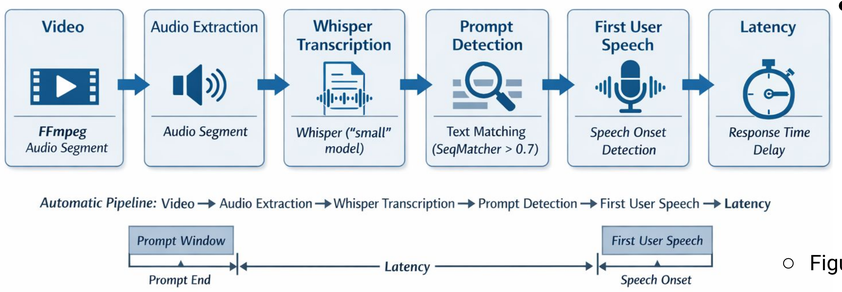


Figure 2. Automated latency pipeline

Evaluation metrics

- **Analytical validity (RQ3):**
  - Session-level comparison of automated vs manual latency.
  - Metrics: mean absolute error (MAE), Spearman correlation, and proportions of sessions with absolute error ≤ 0.5 s, 1.0 s, and 2.0 s, chosen to reflect increasingly permissive clinical tolerances (sub-second, ~1-s, and ~2-s deviation from manual latency)
- **Clinical validity (RQ1 & RQ2):**
  - For manual and automated latency, computed cohort-wise descriptive statistics for schizophrenia vs controls.
  - Used Mann–Whitney U tests and Glass's Delta to quantify discrimination

Results and Discussion

Clinical validity: latency by cohort

- Manually annotated latency was longer in schizophrenia than controls (3.68 s vs 2.92 s; Mann–Whitney $p < 0.01$; Glass's Delta = 0.55).
- Automated latency also differed between cohorts (5.48 s vs 3.67 s; $p < 10^{-6}$; Glass's Delta = 0.78).

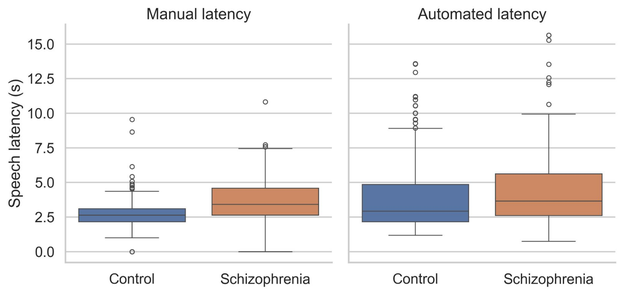


Figure 3. Speech latency by cohort and method

Error distribution by cohort and thresholds

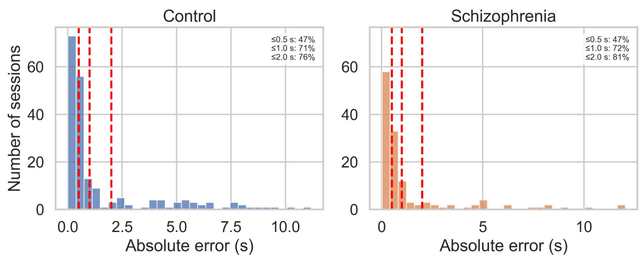
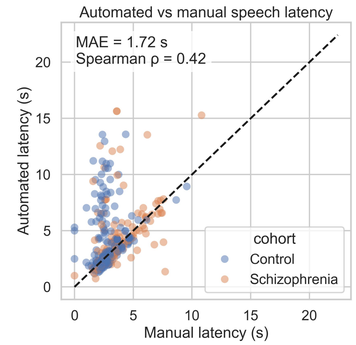


Figure 5. Histogram with dashed lines at 0.5, 1.0, and 2.0 s and a single inset box listing the three percentages.

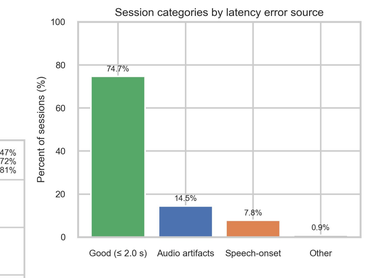
Analytical validity: automated vs manual latency



- The automated measure showed moderate agreement with manual latency

Figure 4. Automated vs manual latency scatterplot

Effect of prompt alignment quality



- Prompt alignment quality was generally high. Errors were driven primarily by speech-onset detection and audio artifacts.

Figure 6. Figure 6. Proportion of sessions falling into four latency-error categories. 'Good' sessions have absolute latency error ≤ 2.0 s. Among sessions with error > 2.0 s, remaining cases are attributed to audio artifacts (e.g., network-related distortions or missing audio), speech-onset issues (ambiguous or delayed onset), or other rare causes.

Conclusions

- **Manual conversational speech latency robustly differentiates patients with schizophrenia from healthy controls in remote picture description tasks.**
- **A fully automated Whisper-based pipeline can estimate speech latency with moderate accuracy relative to manual annotation, while maintaining or enhancing cohort discrimination.**

References

- Neumann, M., Kothare, H., Insel, B., Khan, A., Nadim, D., Lindenmayer, J.-P., & Ramanarayanan, V. (2025). Multimodal speech, language, and orofacial analysis for remote assessment of positive, negative, and cognitive symptoms in schizophrenia. Interspeech 2025.
- Sand, S. G., Kothare, H., Neumann, M., Insel, B., Nadim, D., Khan, A., Lindenmayer, J.-P., & Ramanarayanan, V. (2025). The advantage of combining health information from different modalities in objective remote assessment of schizophrenia. Proceedings of the International Society for CNS Clinical Trials and Methodology (ISCTM) 2025 / Ametris Digital Data Summit 2025.

Acknowledgements

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

modality.ai

UCSF