Towards fully automated rating scale review: Identifying speech feature signature for administration variances in CDR interviews during Alzheimer's Disease clinical trials

Rachel N. Newsome[1], Rachel Kindellan[1], Celia Fidalgo[1], William Simpson[1,2]
1 Cambridge Cognition Toronto, Canada, 2 McMaster University, Hamilton, Canada

Author contact: rachelnewsome@winterlightlabs.com

Background

Fidelity of rating scale administration is crucial for valid clinical assessments. As rating scales are primary endpoints in clinical trials, Quality Assurance (QA) of scale administration is vital for data integrity and therapeutic signal detection. Building on our previous work showing non-expert reviewers can detect administration errors in the ADAS-Cog effectively[1], this study explores how speech biomarkers could be leveraged to automatically detect CDR administration variances in Alzheimer's disease (AD) clinical trials**.**

Objective

To investigate how administration variances in the CDR map to voice characteristics – thereby allowing for future automation of QA reviews

Methods

- We utilized the Winterlight speech platform to manually diarize, transcribe, split by subtask and annotate 236 recorded administrations of the Clinical Dementia Rating (CDR) interview.
- This analysis focused on rater interjections in the interview, which were rated as part of the QA process. More details are provided in **Table 1**.
- Speech features were extracted from participant speech for each subtask.
- Recordings comprised patients from cognitively unimpaired to mild cognitive impairment.
- We focused on a core set of speech features including timing, acoustic and lexical characteristics of speech.
- Ratings were made by trained, human transcribers.
- To examine how rater interjections may be reflected in the speech signal we completed ANOVA group comparisons using the level of rater interjection as the grouping variable.
- Differences between group levels were evaluated post-hoc using Tukey's HSD test.

Results

The distribution of rater interjections showed a similar pattern across multiple CDR tasks (**Figure 1**) with most receiving a level of "None".

These patterns were notably different for the "Memory Problem" and "Recent Experience" questions, suggesting more rater adaptation was required to elicit the required details.

Presented at ISCTM 20th Annual meeting, Washington DC. February 21-23-2023

Figure 1: Distribution of rater interjection level by CDR subtask

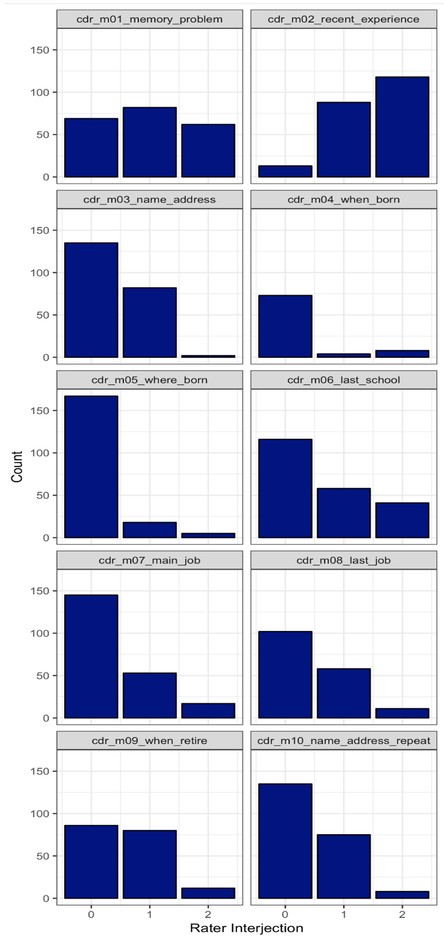

Figure 2: Group mean comparisons for speech features of interest

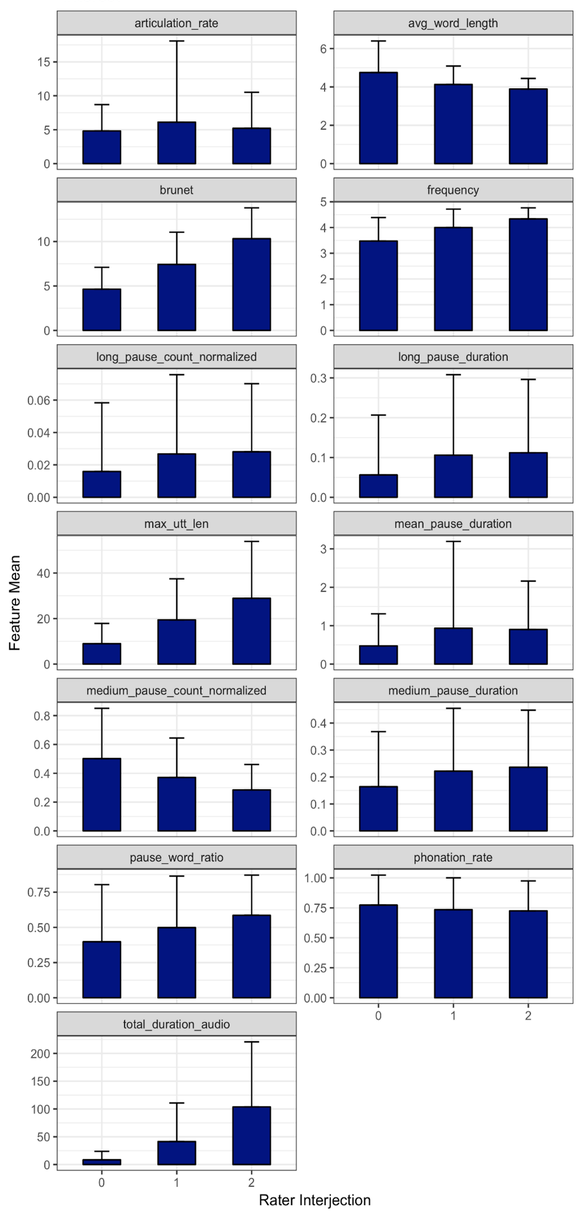

Table 1: Levels and example of rater interjection in current study

Levels of Rater Interjection	Examples
0 – None	Rater did not interject
1 – Minor interjection (likely does not affect performance)	Rater provided acknowledgement ("uh-huh") Rater provided encouragement ("good job") Rater elaborated on a question due to patient request
2 – Major interjection (likely does affect performance)	Rater has unrelated conversation with patient Rater provided hints or answers Rater interrupted patient or deviated from task entirely

Results con't

Statistically significant, stepwise group differences in participant speech characteristics (**Figure 2**) were seen for several features including: average word length, brunet's statistic, word frequency, maximum utterance length, medium pause count and audio duration (all main effect $p<0.001$, all pairwise, between group comparisons $p<0.03$).

Table 2: Description of expanded quality metrics in active development

Metric	Description
Rater interference [None, Minor, Major]	Does rater or other speaker interfere with task?
Rater clarity [Excellent, Somewhat unclear, Often unclear]	Is rater's voice hard to hear or understand?
Speaker number [Expected, More than expected]	Are there more speakers than required?
Skipped prompt or task [Not skipped, Skipped]	Is the prompt or task skipped?
Out of order prompt or task [Not out of order, Out of order]	Is the prompt or task out of order?
Repeated prompt [Not repeated, Repeated]	Is the prompt repeated?
Reworded prompt [No deviation from script, Minor deviation, Major deviation]	What is the level of rewording?
Disjointed administration [None, Disjointed]	Is the task administration disjointed (i.e., partially completed and revisited after another task was started)?

Conclusions

- These data suggest that rater behavior is captured within several participant speech parameters.
- Spurred by these findings, we are currently analyzing a similar CDR dataset using an expanded set of metrics (see **Table 2**).
- With this dataset we plan to:
  - Extract features for both rater and participant speech.
  - Generate a novel quality composite by weighting metrics based on their importance.
  - Correlate this composite metric with rater and participant features, as well as traditional clinical outcome assessments.
  - Conduct a series of human reviews with external quality experts to validate the performance of our quality composite against gold standard practices.

References

1. Kindellan, Newsome, Fidalgo, Robin. (2023) ISCTM. Assessments of ADAS-Cog administration are comparable across expert and non-expert reviewers.