Analytical validation of automatic speech recognition tools used for voice biomarker development

Rachel Kindellan, Mengdan Xu, Jennifer Ruan, Sasha Sirotkin, Jessica Robin

Winterlight Labs, Toronto, ON, Canada

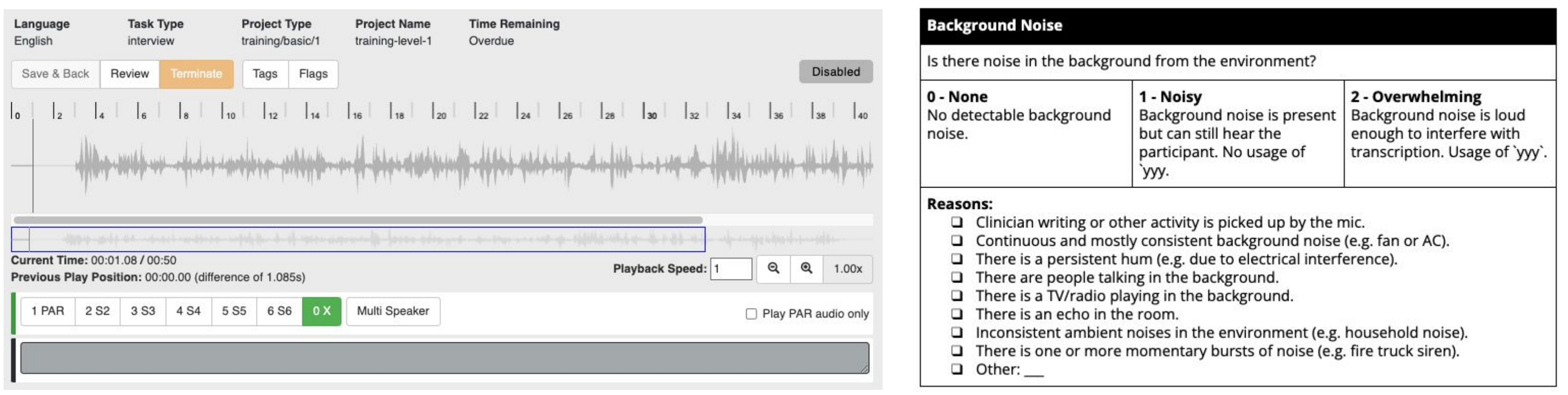
Background

- Speech-based biomarkers have the potential to offer scalable, automated solutions for disease detection and patient monitoring in clinical research.[1,2,3]
- Previous research has shown that changes to speech and language patterns occur in a variety of neurological and psychiatric diseases and disorders.[4,5,6]
- Linguistic changes, including variations in vocabulary, sentence structure and information content have been linked to neurodegenerative disease as well as psychiatric and mood disorders.[6,7]
- While acoustic properties of the voice can be directly computed from an audio recording, linguistic properties require transcription of speech to text.
- Automated speech recognition services (ASR) are playing a growing role in speech-based biomarker solutions due to the increased scalability, namely reduction in time and cost, compared to human transcription.
- It is important to validate the accuracy of ASR-generated text transcripts from speech recordings to understand their suitability for use in the computation of speech-based digital biomarkers.
- This project aims to investigate the accuracy of ASR in varying environments of (a) linguistic content and (b) audio quality to further evaluate the reliability of ASR according to content and context of the audio recording.

Methods

- In this validation study, we compared text transcripts generated by trained human raters to those generated by Amazon Web Services (AWS) ASR.
- 875 speech samples were collected from 359 English-speaking individuals, including healthy controls and individuals with Alzheimer's disease, aphasia, depression, schizophrenia and other conditions.
- Samples were elicited in a variety of speech tasks, including open-ended tasks (journalling, picture description) and structured speech tasks (fluency, naming).
- For each recording, two text transcripts were produced: one generated by a trained human rater on an in-house web-based transcription platform (Figure 1) and one generated by the Amazon Web Services (AWS) ASR service.
- The human raters were prompted to score/rate each recording on a scale of 0 (no issues) to 2 (serious issues) for audio quality (e.g. distortion), background noise (e.g. typing), participant clarity (e.g. stuttering), participant accent (by prevalence).
- The word error rate (WER) for each audio recording was calculated by comparing the number of deviations in the ASR transcript relative to the human-generated transcript (gold standard).
- WER is calculated by dividing the sum of all substitutions ("word" instead of "bird"), deletions ("all that" instead of "all of that"), and insertions ("all of that" instead of "all that") by the total number of words in the transcript.
- WER was compared across task types (only samples without quality issues) and audio quality ratings (only picture description task type) using 1-way ANOVAs and paired comparisons using Tukey's HSD test.

Figure 1: Transcription platform and rating scales



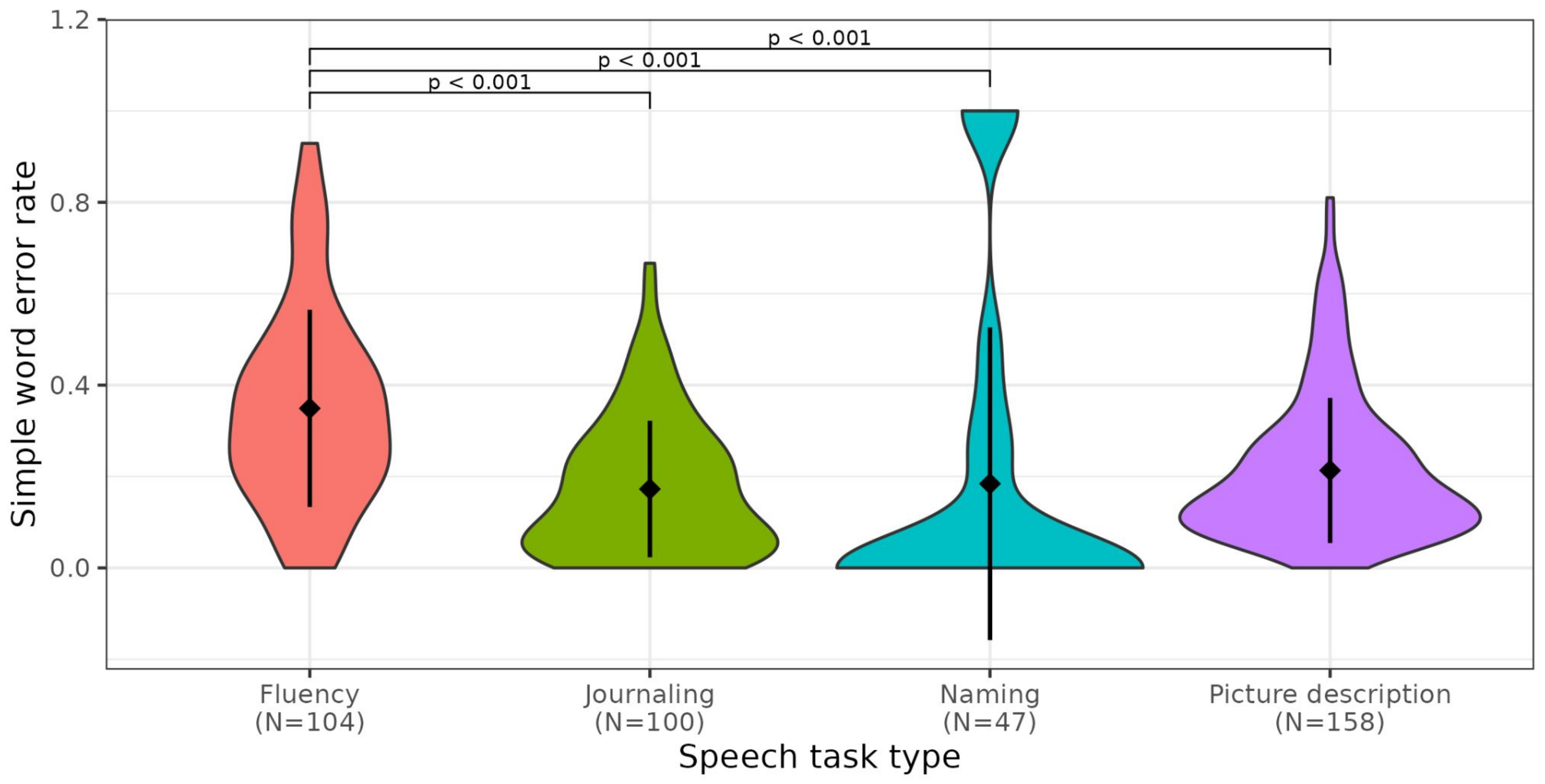
Language	Task Type	Project Type	Project Name	Time Remaining
English	interview	training/basic/1	training-level-1	Overdue

Current Time: 00:01.08 / 00:50
Previous Play Position: 00:00.00 (difference of 1.085s)
Playback Speed: 1

Background Noise		
Is there noise in the background from the environment?		
0 - None No detectable background noise.	**1 - Noisy** Background noise is present but can still hear the participant. No usage of `yyy.	**2 - Overwhelming** Background noise is loud enough to interfere with transcription. Usage of `yyy`.

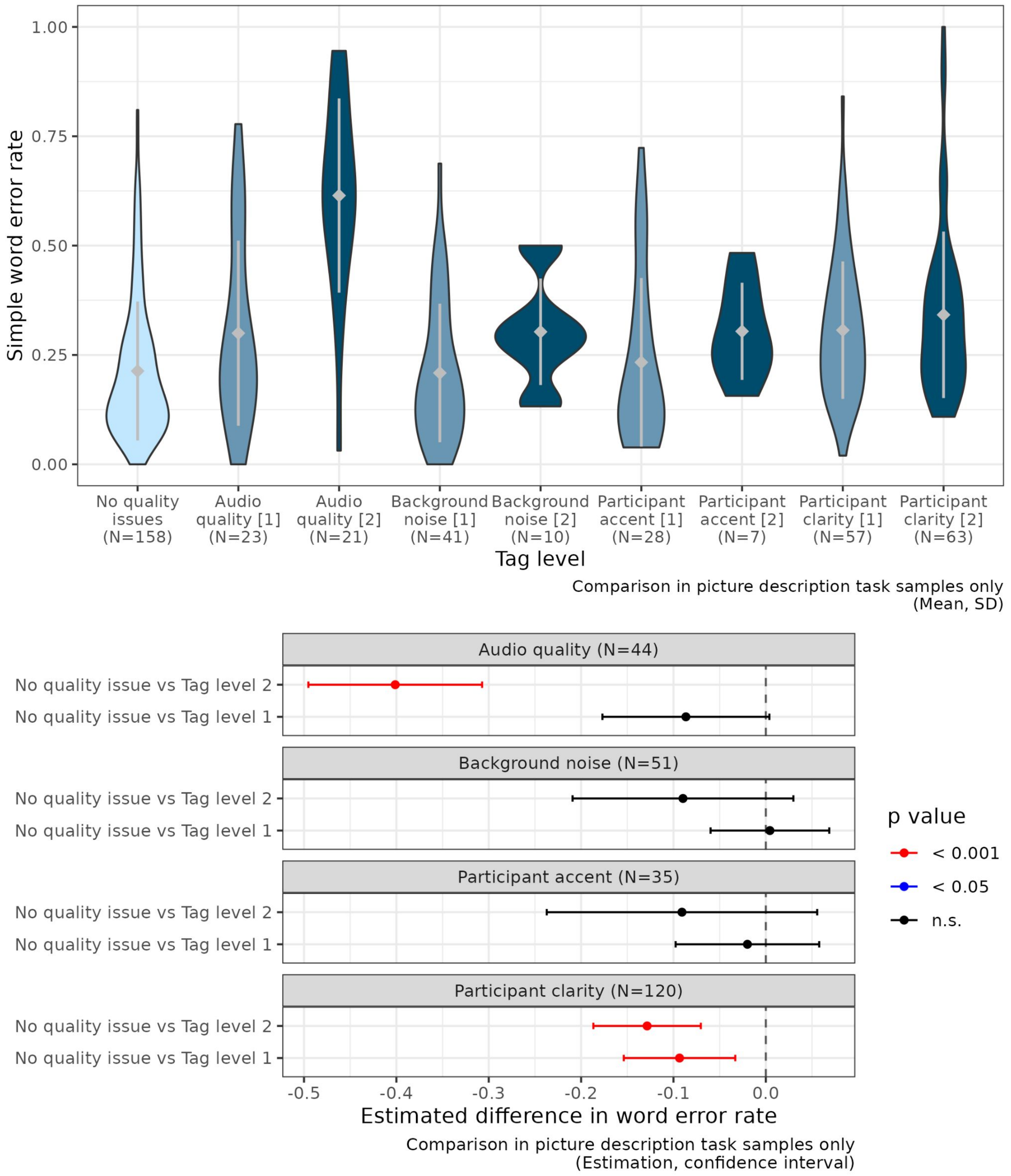
Reasons:
- ❑ Clinician writing or other activity is picked up by the mic.
- ❑ Continuous and mostly consistent background noise (e.g. fan or AC).
- ❑ There is a persistent hum (e.g. due to electrical interference).
- ❑ There are people talking in the background.
- ❑ There is a TV/radio playing in the background.
- ❑ There is an echo in the room.
- ❑ Inconsistent ambient noises in the environment (e.g. household noise).
- ❑ There is one or more momentary bursts of noise (e.g. fire truck siren).
- ❑ Other: ___

Figure 2: Word error rate by speech task



Comparison in samples without quality issues only
(Mean, SD, significant p values from Tukey's HSD test)

Figure 3: Word error rate by audio quality



Comparison in picture description task samples only
(Estimation, confidence interval)

Results

- We found that the average WER for ASR transcripts significantly differed both by speech task (Figure 2) and by audio quality (Figure 3; p's < 0.001).
- Specifically, WER was lowest (indicating highest agreement with manual transcripts) for open-ended speech tasks, including picture description (WER=21%, SD=16%) and journaling (WER=17%, SD=15%), in which participants speak in full, natural sentences. Only samples with no audio quality issues were included in this comparison.
- WER was higher or more variable for structured speech tasks, like naming (WER=18%, SD=34%) and fluency tests (WER=35%, SD=22%), which elicit single words or sequences of words.
- As expected, WER was higher for picture description speech samples rated as having severe quality issues, including low overall audio quality (WER=45%, SD=27%) or participant clarity (WER=33%, SD=18%).
- WER was not significantly different between samples rated as having no quality issues (WER=21%, SD=16%), and those with quality issues relating to background noise or participant accent suggesting these have less of an impact on ASR accuracy.
- Even minor or intermediate issues with participant clarity led to significant differences in WER compared to samples with no quality issues.

Discussion

- This study found the highest analytical validity for ASR in the context of open-ended, naturalistic speech tasks with good quality audio.
- Structured speech tasks and lower audio quality (relating to audio quality or participant clarity) may result in lower accuracy transcripts.
- Interestingly, transcript accuracy was not affected by minor background noise or mild participant accents, suggesting that these samples are better suited for ASR analysis than those with overall quality issues or poor participant clarity.
- These findings imply that audio quality and linguistic content significantly impact the accuracy of ASR and thus the suitability of this transcription method in the computation of speech-based digital biomarkers.
- The lower accuracy of ASR in structured speech tasks may be a result of disagreement between the linguistic content of those tasks and the training dataset of the AWS ASR service.
- Future directions for this research include examination of ASR accuracy in the context of custom ASR language models whose training dataset accords with the linguistic content of the speech tasks in question as well as an exploration of options to enhance audio quality either pre-processing (during recording) or post-processing.

References

1. Coravos, A., Khozin, S. & Mandl, K. D. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. Npj Digit. Med. 2, 14 (2019).
2. Kourtis, L. C., Regele, O. B., Wright, J. M. & Jones, G. B. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. Npj Digit. Med. 2, 9 (2019).
3. Robin, J. et al. Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations. Digit. Biomark. 99–108 (2020) doi:10.1159/000510820.
4. de la Fuente Garcia, S., Ritchie, C. W. & Luz, S. Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review. J. Alzheimers Dis. 78, 1547–1574 (2020).
5. Low, D. M., Bentley, K. H. & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investig. Otolaryngol. 5, 96–116 (2020).
6. Boschi, V. et al. Connected Speech in Neurodegenerative Language Disorders: A Review. Front. Psychol. 8, (2017).
7. Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F. & Novikova, J. Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech. Front. Aging Neurosci. 13, 189 (2021).
8. Aloshban, N., Esposito, A. & Vinciarelli, A. What You Say or How You Say It? Depression Detection Through Joint Modeling of Linguistic and Acoustic Aspects of Speech. Cogn. Comput. (2021) doi:10.1007/s12559-020-09808-3.
9. Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G. & Naylor, M. Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine 28, 100583 (2020).

Disclosures: All authors are full-time employees of Winterlight Labs, Inc.

WINTERLIGHT

Presented at the ISCTM Autumn Conference, 2022