Evaluating a method for automatic and objective scoring of verbal response for the Montreal Cognitive Assessment (MoCA)

Liam D. Kaufman, MSc [1], Aparna Balagopalan, MSc [1], Jekaterina Novikova, PhD [1], and Fariya Mostafa, MSc [1]

(1) Winterlight Labs, Toronto, ON, Canada

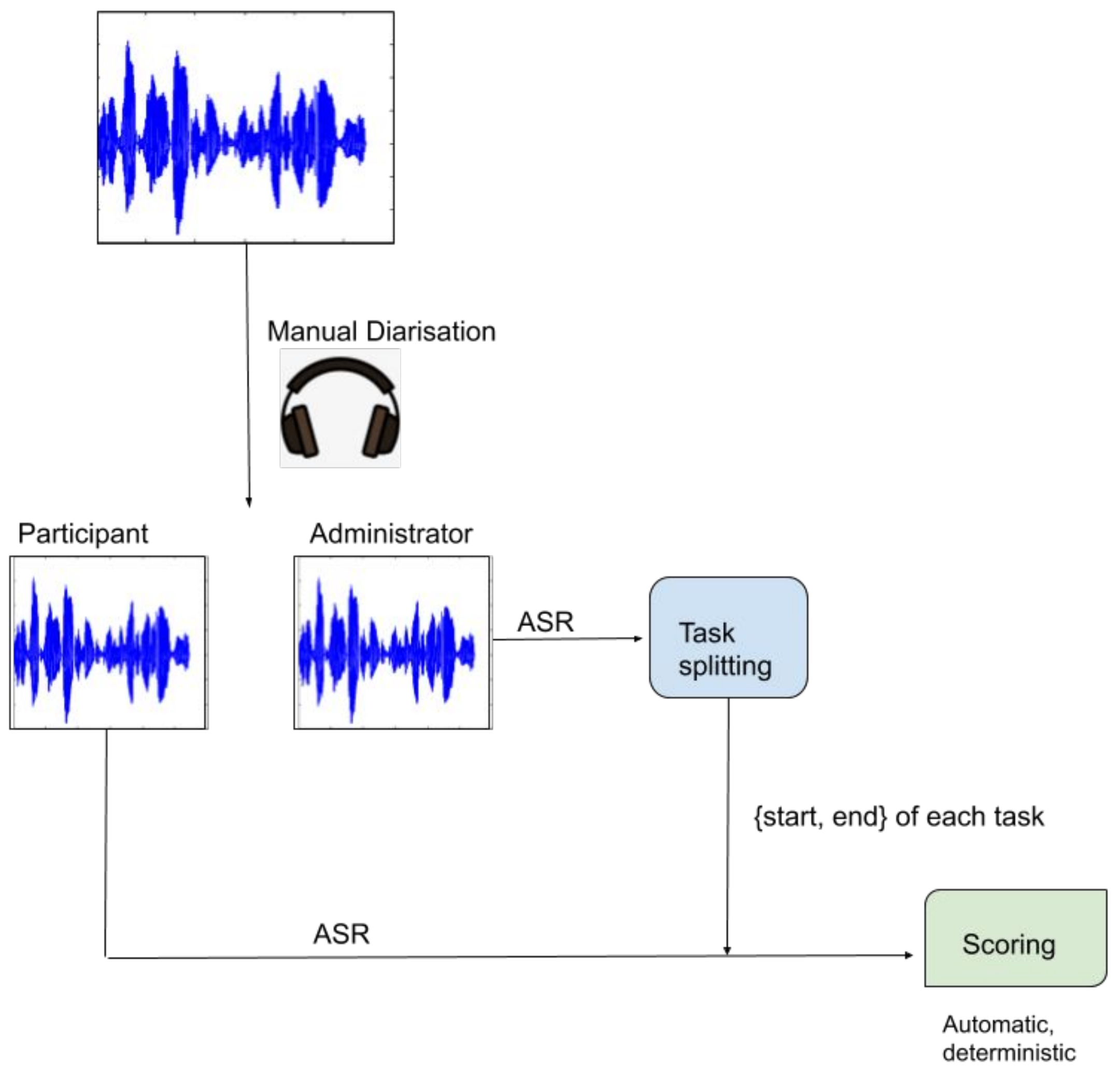
Background

Current cost estimates for bringing an Alzheimer's Disease (AD) therapeutic to market amount to $5.6 billion, nearly 3 times higher than the cost of other therapies. Existing outcome measures are partially responsible for this expense. These measures require a significant amount of time, expertise and rater training to administer correctly. Deviations in administration and errors in scoring are very common. These errors reduce the overall precision of measurement, putting upward pressure on the AD trial sample sizes. To improve precision, many trials rely on recording the administration of all primary cognitive outcomes so that they can be reviewed for errors by independent, often PhD level raters. This process is both time consuming and costly. Recent advances in speech recognition and Natural Language Processing (NLP) could make identifying errors and automatically scoring individual tasks significantly more efficient. Automatic scoring and error identification could open the door to streamlined quality assurance evaluation and dramatically reduce the cost of deploying these measures in a clinical trial. In this proof of concept study, we explored the feasibility of automatically scoring the delayed recall task from audio recordings of the Montreal Cognitive Assessment (MoCA).

Methods

- Scoring a multi-component verbal cognitive assessment from an audio recording requires differentiating the administrator from the participant (diarization), generating a transcript (automatic speech recognition; ASR), identifying the boundaries of each subtask in the transcript (task splitting), using the transcript of participant verbal responses to score the subtask (automatic scoring), and combining the subtasks into a total score.
- For this study, we chose to focus on evaluating automatic task segmentation approaches and the downstream scoring accuracy of the delayed recall subtask of the MoCA.
- MoCA recordings from 50 individuals were taken from a longitudinal natural history study of older adults (aged 55-90), recruited from the community and independent living facilities in Canada and the US.
- Recordings were manually diarized, transcribed, segmented and scored to produce a gold standard reference dataset.
- To segment tasks, ASR-produced transcripts from the rater were matched to the standard administration script for the MoCA. The following alignment algorithms were evaluated:
  - Phonemic Alignment
  - Keyword matching
  - Keywords matching with timestamp information
- We tested a variety of transcription and task splitting algorithm combinations to determine the upper and lower bound of performance.
- Delayed recall score was calculated (max = 5) based on the ASR of participant's responses.
- Task segmentation performance was measured by examining the proportion of correctly identified delayed recall task boundaries.
- Mean absolute error (MAE) in delayed recall score for a given strategy was also tested.

Figure 1: Schematic of MoCA audio transcription in preparation for automatic scoring



More errors in ASR transcription are related to higher errors in delayed recall scoring

Correct delayed recall score:
Mhm says likely to tell me as many words you can remember that were from that list. <participant speaks> *Okay I'll give you a chance to guess, one was the type of fruits*

Error = 1 point out of 5:
I read some xxx earlier and I asked you to trying to remember them. I'd like to try and help me as many of those words <participant speaks> *tell you about one of the other words with the type of flower*

Error = 3 points out of 5:
Great so a few minutes ago we had learned over sports so what I like to do is tell me as many words as you can remember that world that <participant speaks> *okay I'll give you hit it was a musical instrument*

Text Delayed Recall
Text Delayed Recall with cues

Algorithmic approaches to task splitting

Option A: Phonemic Alignment

Actual: "Please draw a line, going from a number **to a letter in ascending order**."

ASR: "so pleased draw a line going from a number **two a letter and they're sending order**

```
// T UW1 // AH0 // L EH1 T ER0 // IH0 N // AH0                     S EH1 N D IH0 NG // AO1 R D ER0
// T UW1 // AH0 // L EH1 T ER0           // AH0 N D // DH EH1 R // S EH1 N D IH0 NG // AO1 R D ER0
```

Option B: Keywords with and without timestamps

Match keywords based on ASR transcript and task-specific reference list

Delayed recall task e.g.: **remember, list, tell, words**

Timestamps provide structure for keyword *ordering*

Results

Scenario	Transcription for participant	Algorithm	Accuracy	Delayed Recall MAE
1	Manual	Phonemic	84.0%	0.76
2	ASR	Phonemic	84.0%	1.87
3	ASR	Keyword	78.0%	2.20
4	ASR	Keyword + timestamps	82.0%	2.05

- Using manual segmentation and transcripts, automated scoring of the delayed recall task (checking reported words against the word list) was 100% accurate
- Phonemic alignment was the the most accurate task splitting algorithm (84.0%)
- The addition of automated task segmentation alone imparted a mean error of 0.76/5 on the delayed recall task (Scenario 1)
- Adding in ASR for the participant responses (i.e. words recalled) increased the error to 1.87/5 (Scenario 2)
- This suggests that errors in ASR impart more error into the final delayed recall score than automatic segmentation of task boundaries.
- Additional combinations using Keyword based algorithms (Scenario 3 & 4), did not surpass phonemic alignment.

Conclusions

The results of this proof of concept study show that transcribed audio recordings can be used to automatically calculate Delayed Recall scores on the MoCA. Using an ASR-based algorithm to automatically segment MoCA tasks resulted in a mean error of 1.87/5 pts in Delayed Recall scores. Together these results suggest that automatic segmentation and scoring of audio recordings of cognitive assessments is feasible and further work using larger datasets is needed to fine tune the algorithms and improve scoring accuracy.

References

Scott, T. J., O'Connor, A. C., Link, A. N. & Beaulieu, T. J. Economic analysis of opportunities to accelerate Alzheimer's disease research and development. Ann. N. Y. Acad. Sci. 1313, 17–34 (2014).

Cano, S. J. et al. The ADAS-cog in Alzheimer's disease clinical trials: psychometric evaluation of the sum and its parts. J. Neurol. Neurosurg. Psychiatry 81, 1363–1368 (2010).

Nieuwenhuis-Mark, R. E. The death knoll for the MMSE: has it outlived its purpose? J. Geriatr. Psychiatry Neurol. 23, 151–157 (2010).

Connor, D. J. & Sabbagh, M. N. Administration and scoring variance on the ADAS-Cog. J. Alzheimers. Dis. 15, 461–464 (2008). Salloway, S. et al. Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. N. Engl. J. Med. 370, 322–333 (2014).

Smith, P. J., Need, A. C., Cirulli, E. T., Chiba-Falek, O. & Attix, D. K. A comparison of the Cambridge Automated Neuropsychological Test Battery (CANTAB) with 'traditional' neuropsychological testing instruments. J. Clin. Exp. Neuropsychol. 35, 319–328 (2013).

Presented at Clinical Trials for Alzheimer's Disease (CTAD), San Diego, CA, USA. Dec 4-7, 2019