Advances in automated speech transcription during cognitive testing

David Woods[1], Michael Blank[1], Kristin Geraci[1], Isabella Jaramillo[1], Evan Richards[1], Peter Pebler[1], David K Johnson[2], Timothy J. Herron[3], and Kathleen Hall[1]
[1]Neurobehavioral Systems, Inc, [2]UC Davis Alzheimer's Disease Center, [3]VA Medical Center, Martinez, CA

Introduction

Accurate automatic speech recognition (ASR) is essential to modernize computerized test batteries, enabling objective, automated scoring of discursive speech to replace traditional manual tools like the WAIS. We evaluated transcription accuracy of the California Cognitive Assessment Battery (CCAB) by comparing word error rates (WERs) of individual ASR engines and consensus ASR (CASR) for **~416,000** manually reviewed words.

Methods

Participants: 453 older healthy subjects

- Age: 63.7 y.o. sd = 13.5
- 71% female
- Mean education: 13.8 years
- Race: White 46%, Asian 16%, Black 21%, Other 17%

Procedure:

- Up to 6 high-fidelity recordings (24 bit, 48 kHz) were captured in participants' homes during two logical memory and a picture description task.

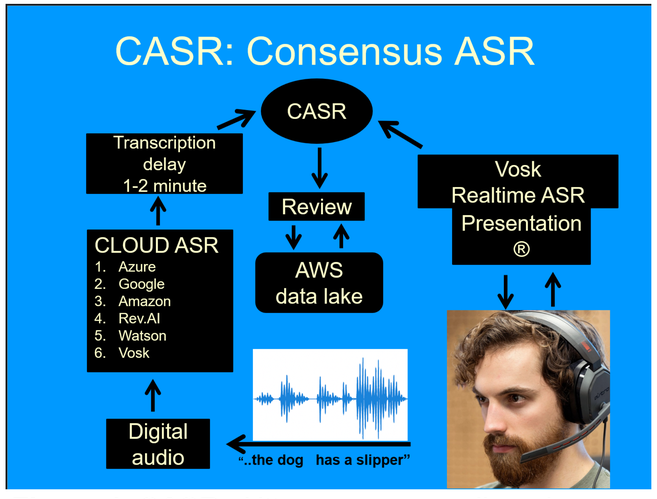

Figure 1. CASR. Utterances were aligned across engines using the Levenshtein algorithm. Word-level evidence was derived from lookup tables mapping each engine's confidence to its empirical accuracy. CASR selected the word with the highest aggregate evidence.

Results

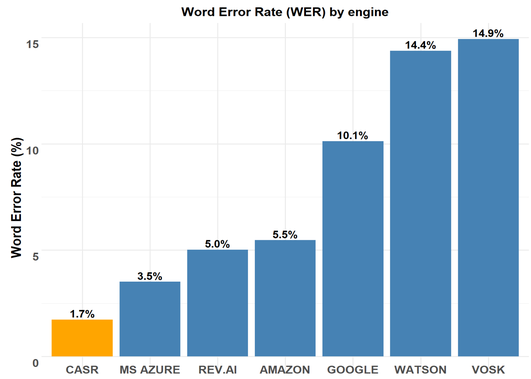

Figure 2. WER by engine

	Microsoft	Amazon	Google	Rev.AI	Vosk	Watson	CASR	age	education	gender	vocab_qsin
Microsoft	1.00	0.79	0.71	0.77	0.68	0.69	0.86	0.10	0.29	0.13	0.45
Amazon	1.00	0.68	0.86	0.67	0.69	0.83	0.07	0.31	0.11	0.49	
Google	1.00	0.68	0.77	0.78	0.71	0.10	0.28	0.25	0.46		
Rev.AI	1.00	0.68	0.69	0.87	0.03	0.29	0.04	0.46			
Vosk	1.00	0.94	0.67	0.15	0.39	0.19	0.65				
Watson	1.00	0.71	0.14	0.36	0.18	0.60					
CASR	1.00	0.07	0.30	0.08	0.45						
age	1.00	0.25	0.00	0.28							
education	1.00	0.01	0.53								
gender	1.00	0.01									
vocab_qsin	1.00										

Figure 3. Engine hit correlations

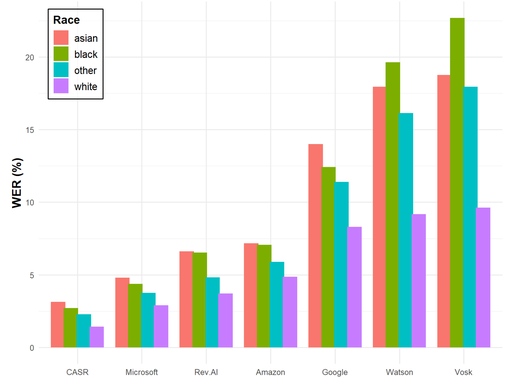

Figure 4. Race effects[1]

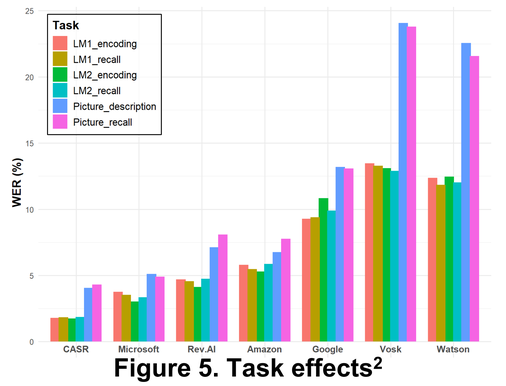

Figure 5. Task effects[2]

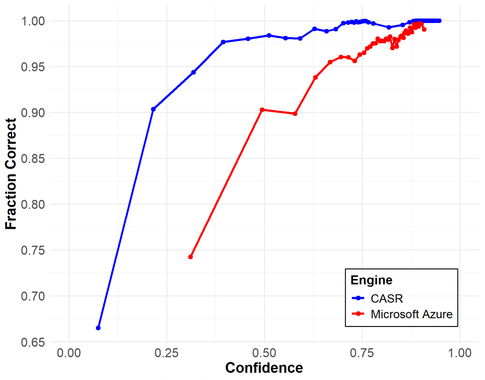

Figure 6. Confidence vs. accuracy

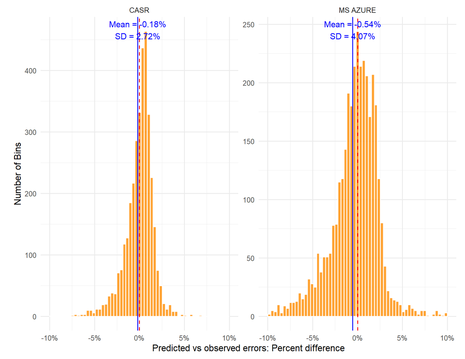

Figure 7. Detecting transcript errors[3]

Summary

- CASR achieved higher transcription accuracy than any individual engine..
- Vocabulary and education were the strongest predictors of ASR accuracy; gender and age had smaller effects.
- Race had a modest impact on CASR accuracy but more pronounced effects on some individual engines.
- Accuracy was higher for Logical Memory than for picture description.
- CASR confidence was a strong predictor of transcription errors, enabling realtime quality assessment.

Discussion

- CASR WERs were generally too low to significantly affect keyword recall z-scores.
- ASR accuracy correlated with keyword recall, reflecting the influence of vocabulary and education on both measures.
- The impact of CASR WERs and timing precision on speech and language biomarkers (SLBs) warrants further investigation.

Footnotes

(1) Elevated WERs in Asian participants reflected the presence of strong accents in some subjects.
(2) Custom grammars containing LM story words contributed to higher accuracy on LM tasks.
(3) Confidence/accuracy slopes were used to estimate per-word error probabilities, which were aggregated and compared to observed errors.

Contact us

drdlwoods@neurobs.com
ccabresearch.com
neurobs.com

Visit us at booth 646!

Supported by NIA R44AG080951