Consensus Automatic Speech Recognition (CASR) in Cognitive Testing

Timothy Herron[1], Kathleen Hall[2], Gabriel Sucich[2], Mike Blank[2], Kristi Geraci[2], Juliana Baldo[1] and David Woods[2]

[1]VA Northern California Healthcare System, [2]Neurobehavioral Systems, Inc.

Introduction

Scoring verbal cognitive tests with automatic speech recognition (ASR) engines increases the efficiency of scoring and provides word timestamps that enable detailed temporal analyses of spoken responses. Here, we describe novel consensus ASR (CASR) procedures that incorporate multiple ASR engines to increase transcription and timing accuracy and to generate CASR transcript confidence scores.

Methods

- Seven ASR engines produced automatic transcriptions of both speech database samples (GMU Speech Accent Archive [1] and NUS Auditory English Lexicon Project [2]) and verbal test responses from 41 subjects (Age 19-84, mn 49 std 20; Edu 12-18, mn 16 std 2; 52% female; 69% White; 12% non-native English) using the California Cognitive Assessment Battery (CCAB) [3].
- A novel Recognizer Output Voting Error Reduction (ROVER) algorithm was used to mutually align the transcripts [4], and a Bayesian weighted voting algorithm [5] produced the best CASR transcript, mean word timestamps, and consensus scores.
- Word error rates (WER) gauged CASR accuracy against either predetermined or manually corrected transcripts.

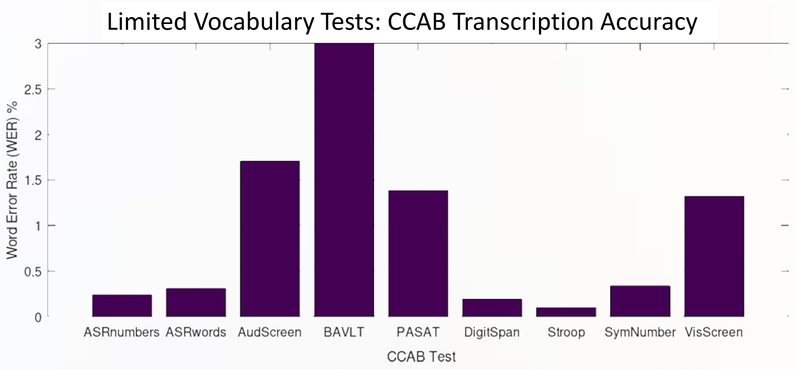

Figure 3: CASR transcription errors for tests that have limited response vocabularies. ASRnumbers: Automated Speech Recognition (ASR) of numbers screening test; ASRwords: ASR of words screening test; AudScreen: Auditory hearing screening using words; BAVLT: Bay Area verbal learning test; PASAT: Paced auditory serial addition test; DigitSpan: DigitSpan forwards and backwards; Stroop: Stroop color naming test; SymNumber: Symbol-Number test; VisScreen: Visual acuity test using words.

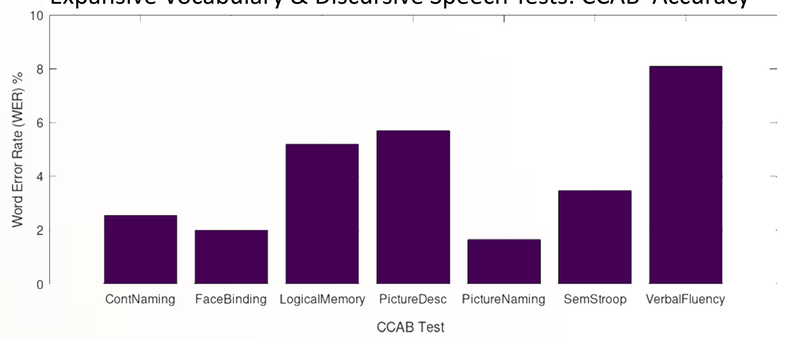

Figure 4: CASR transcription errors for tests that have expansive vocabularies or discursive responses. ContNaming: Continuous picture naming; FaceBinding: Face binding memory test; LogicalMemory: Logical memory test; PictureDesc: Picture Description test; PictureNaming: Single picture naming test; SemStroop: Semantic Stroop synonym-antonym test; VerbalFluency: category verbal fluency test.

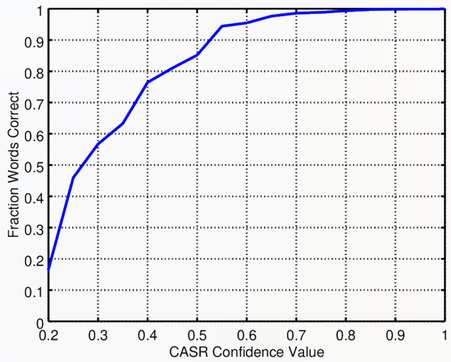

Figure 5 Accuracy of CASR transcripts (y-axis) vs. CASR consensus confidence value indicating level of agreement across ASR engines. Values based on test transcripts used in Figures 3 and 4

Results

- Figure 1 shows that CASR WER is lower than that for any individual ASR. Similarly, Figure 2 shows that CASR start timestamps are more accurate than those for any individual ASR. There were no gender or age effects in CASR WER.
- Figures 3 and 4 show that CASR performance on CCAB tests is better for limited vocabulary tests (worst is 3%).
- Figure 5 shows that CASR confidence values can be used to predict which words require manual transcription.

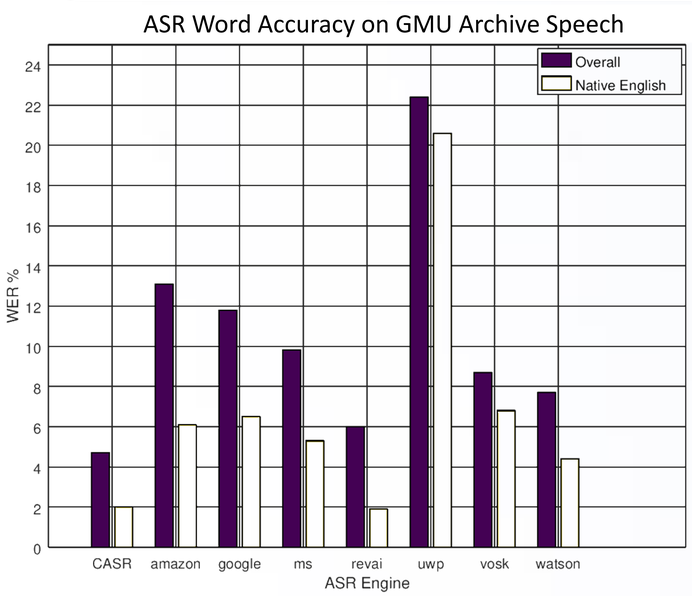

Figure 1: Transcription errors over spoken sentences from the GMU Accent Archive for CASR and 7 ASR engines for both native English speakers (34%) and all speakers. CASR: consensus ASR; amazon: Amazon transcription service; google: Google transcription service; ms: Microsoft Azure transcriptions; revai: Rev.ai transcriptions; uwp: Miscrosoft Windows 10 UWP real-time transcriptions; vosk: Vosk Kaldi-based transcriptions; Watson: IBM Watson transcription service.

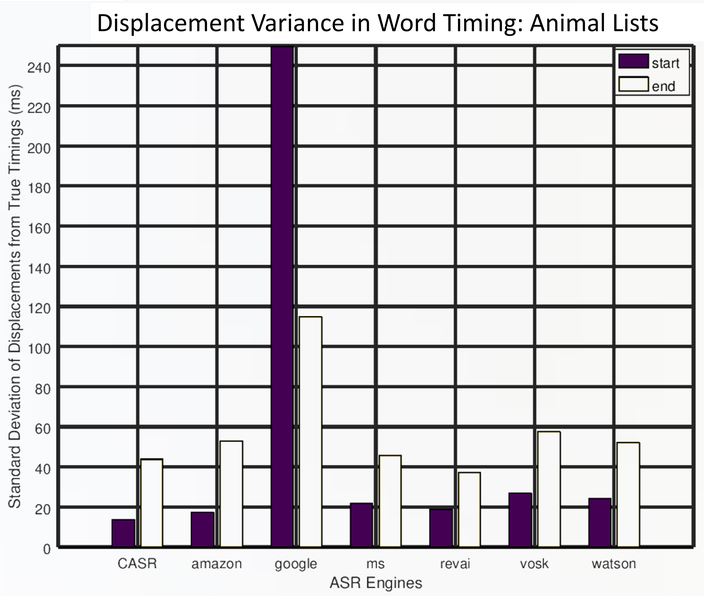

Figure 2: Variance from true timestamps, the start and end of individual words, estimated by ASR engines for artificial lists of spoken words from the NUS word database. CASR: consensus ASR; amazon: Amazon transcription service; google: Google transcription service; ms: Microsoft Azure transcriptions; revai: Rev.ai transcriptions; uwp: Miscrosoft Windows 10 UWP real-time transcriptions; vosk: Vosk Kaldi-based transcriptions; Watson: IBM Watson transcription service.

Discussion

- CASR produces transcripts for verbal test responses accurate enough for estimating scores in most limited word response tests and some tests with more expansive vocabulary.
- In large vocabulary response tests, CASR transcripts facilitate quick manual correction, and confidence values can identify transcript words needing manual correction.
- Patterns in CASR errors within each test also indicate further algorithm improvements that could reduce CASR WER.
- A version of CASR for US Spanish is being developed.

References

[1] https://accent.gmu.edu
[2] https://inetapps.nus.edu.sg/aelp/
[3] https://www.ccabresearch.com
[4] S.Jalalvan, M.Negri, D.Falavigna, M.Matassoni & M.Turchi, Computer Speech & Language, Vol. 47, January 2018, pp 214-239,
[5] L. Kuncheva & J.J. Rodríguez, Knowledge and Information Systems, 38:259–275, Feb 2014,

CLIMB
Center for Language, Imaging, Mind, and Brain