Consensus Automatic Speech Recognition (CASR) in Cognitive Testing

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

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 consensus ASR (CASR) procedures that incorporate multiple ASR engines to increase transcription and timing accuracy and 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 of 41 subjects from the California Cognitive Assessment Battery (CCAB). A novel Recognizer Output Voting Error Reduction (ROVER) algorithm was used to mutually align the transcripts [3], and a Bayesian weighted voting algorithm [4] 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.

Results

Database sentence WERs from 1767 subjects ranged from a mean of 22% (Windows10 UWP) to 6% (Rev.ai) with CASR producing 5%, with no significant gender or age effects but better performance for native english speakers (Figure 1). In CCAB test responses, for limited word response tests CASR WERs ranged from 3% to less than 1% (Figure 2); for expansive word response tests CASR WERs ranged from 8% to 2% (Figure 3); and for discursive speech, CASR WERs ranged from 6% to 5%. Word start time ASR estimates for 594 database words in lists ranged in mean deviations from true times from 250ms std.dev. (Google) to 17ms std.dev. (Amazon) with CASR obtaining 14ms errors (Figure 4). Finally, consensus confidence scores from CCAB test responses, ranging from 0 to 1 (1=complete agreement across ASR engines), show that CASR words with consensus scores above 0.8 and 0.9 are correct >99% and >99.8% of the time, respectively (Figure 5).

Conclusion

CASR produces transcripts for verbal test responses accurate enough for estimating scores in most limited word response tests. 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 also indicate future substantial reductions to CASR WER on a per test basis.

References

[1] https://accent.gmu.edu

[2] https://inetapps.nus.edu.sg/aelp/

[3] S.Jalalvand, M.Negri, D.Falavigna, M.Matassoni & M.Turchi, Automatic quality estimation for ASR system combination, Computer Speech & Language, Vol. 47, January 2018, pp 214-239, DOI: 10.1016/j.csl.2017.06.003.

[4] L. Kuncheva & J.J. Rodríguez, A weighted voting framework for classifiers ensembles, Knowledge and Information Systems, 38:259–275, Feb 2014, DOI: 10.1007/s10115-012-0586-6

Figure 1: Transcription errors over spoken sentences from the GMU Accent Archive for various ASR engines for all speakers and only native English speakers. CASR: consensus ASR; amazon: Amazon transcription service; google: Google transcription service; ms: Microsoft Azure transcriptions; revai: Rev.ai transcriptions; uwp: Microsoft Windows 10 UWP realtime transcription; vosk: Vosk kaldi-based transcription; 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; vosk: Vosk kaldi-based transcription; watson: IBM Watson transcription service.

Figure 3: CASR transcription errors for tests that have limited response vocabularies. ASRnumbers: Automated Speech Recognition of numbers screen; ASRwords: Automated Speech Recognition of words screen; AudScreen: Auditory hearing screen using words; BAVLT: Bay area verbal learning test; PASAT: Paced auditory serial addition test; DigitSpan: DigitSpan forward and backward; 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 response 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 test; Verbal Fluency: category verbal fluency test.

Figure 5: Accuracy of CASR transcriptions vs. the CASR consensus confidence value indicating level of agreement across ASR engines. Values based upon all CCAB test transcripts used in Figure 3 and Figure 4.

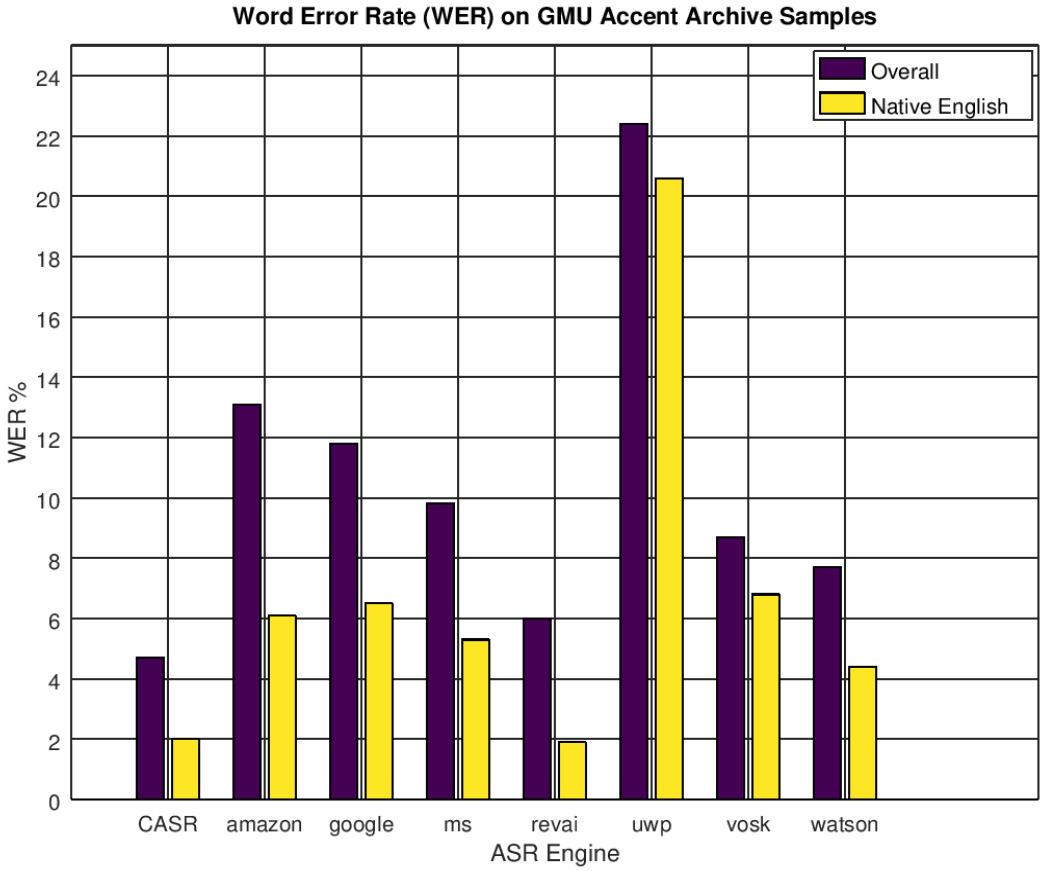

Figure 1: Transcription errors over spoken sentences from the GMU Accent Archive for various ASR engines for all speakers and only native English speakers. CASR: consensus ASR; amazon: Amazon transcription service; google: Google transcription service; ms: Microsoft Azure transcriptions; revai: Rev.ai transcriptions; uwp: Microsoft Windows 10 UWP realtime transcription; vosk: Vosk kaldi-based transcription; watson: IBM watson transcription service.

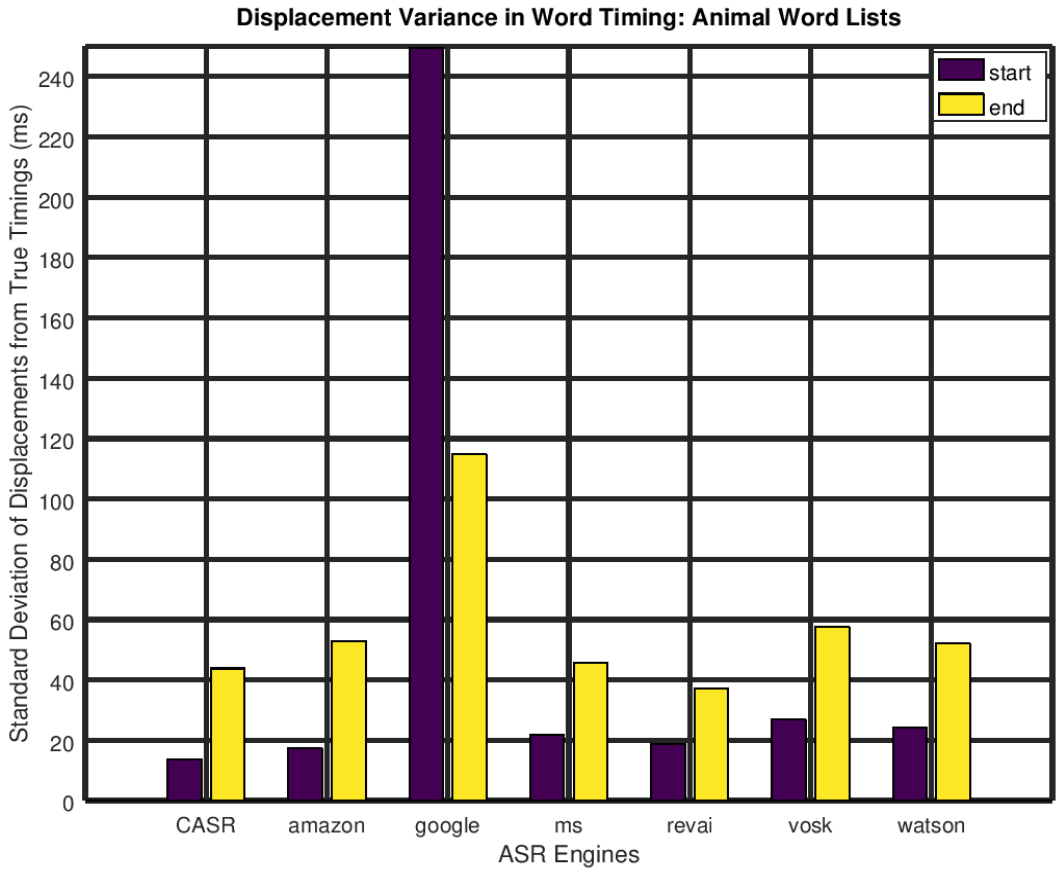

Figure 2: Variance from true timestamps, the start and end of individual words, estimated by ASR engines for artificial lists of spoken words from the NUS word database. CASR: consensus ASR; amazon: Amazon transcription service; google: Google transcription service; ms: Microsoft Azure transcriptions;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

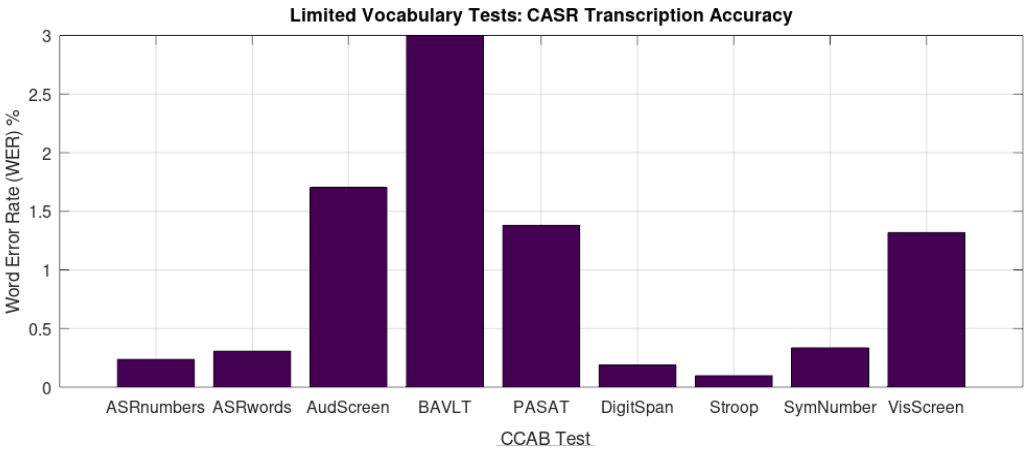

Figure 3: CASR transcription errors for tests that have limited response vocabularies. ASRnumbers: Automated Speech Recognition of numbers screen; ASRwords: Automated Speech Recognition of words screen; AudScreen: Auditory hearing screen using words; BAVLT: Bay area verbal learning test; PASAT: Paced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

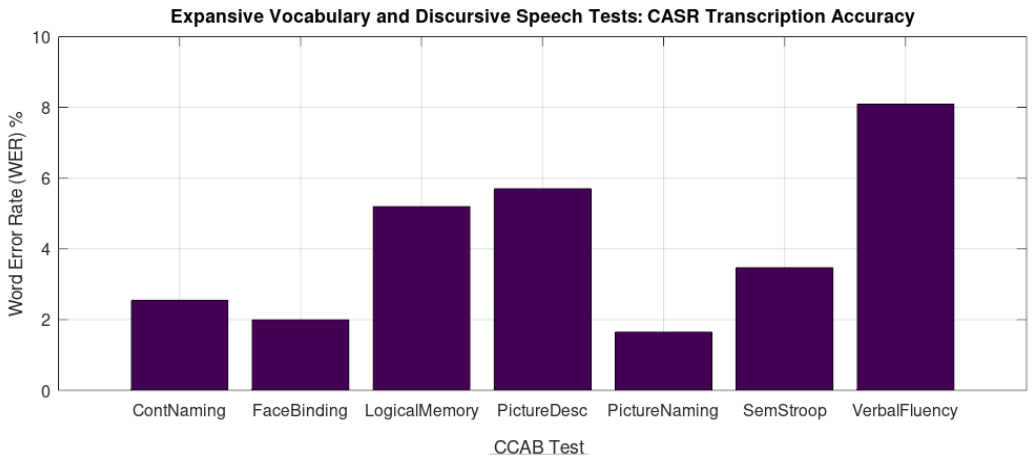

Figure 4: CASR transcription errors for tests that have expansive response vocabularies or discursive responses. ContNaming: Continuous picture naming; FaceBinding: Face binding memory test; LogicalMemory: logical memory test; PictureDesc: Picture description test; PictureNaming: Single picture naming test; SemStroop: Semantic stroop test; Verbal Fluency: category verbal fluency test.

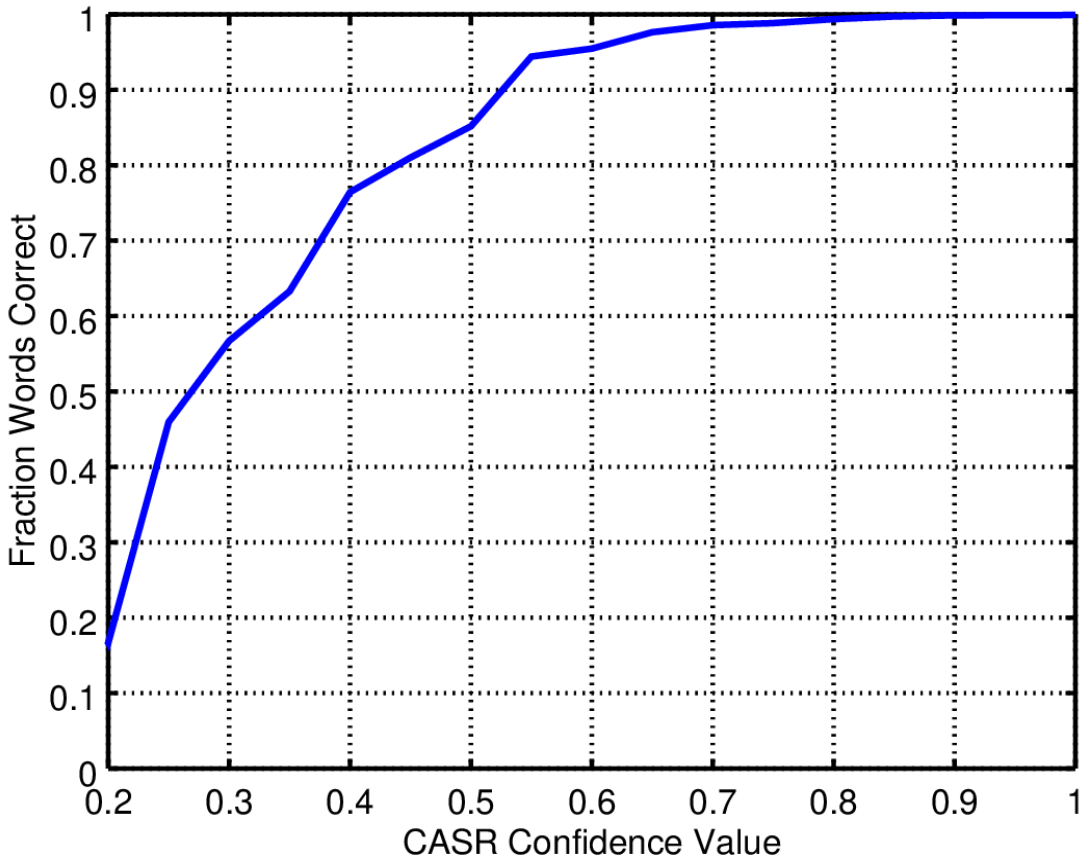

Figure 5: Accuracy of CASR transcriptions vs. the CASR consensus confidence value indicating level of agreement across ASR engines. Values based upon all CCAB test transcripts used in Figure 3 and Figure 4.