Consensus automatic speech recognition (CASR) in the California Cognitive Assessment Battery (CCAB)

INTRODUCTION. Recent reports have investigated the use of automatic speech recognition (ASR) to analyze and score verbal responses in cognitive tests. ASR scoring is objective, permits the efficient computerized administration of verbal tests, and generates timestamps that enable the detailed temporal analysis of responses. However, ASR transcription accuracy varies by engine, task, and participant, and ASR can incorrectly score responses from participants with atypical speech patterns. Here we describe the speech-transcription pipeline of the California Cognitive Assessment Battery (CCAB), which incorporates consensus ASR (CASR) to produce more accurate transcripts than possible with any single ASR engine. We also developed a Transcript Review Tool (TRT) which facilitates the manual correction of mis-transcribed words in problem subjects.

METHODS. Figure 1 shows the CCAB speech transcription pipeline. Realtime ASR transcriptions are obtained along with the transcriptions of the digital recordings of responses using six cloud-based ASR engines (e.g., Google, etc.). Individual transcripts are then combined to produce a "consensus" transcript, and a transcription confidence measure based primarily on the agreement between ASR engines (Figure 2). If needed, "consensus" transcripts can be manually corrected using the Transcript Review Tool which enables the review of all words or just those words below a predefined CASR confidence threshold (Figure 3).

RESULTS. ASR transcriptions were obtained from 442 healthy adults (mean age = 65.1 ±14.4) who each underwent three days of cognitive testing that included 25 verbal tests. In all, approximately 276 hours of speech were transcribed. Preliminary analyses show that CASR transcription accuracy surpassed 99% for tests with limited response sets (e.g., digit span, verbal list learning, face-name binding, etc.) and exceeded 95% for discursive speech tests (e.g., picture description and logical memory).

DISCUSSION. CASR transcription is more accurate than that of any single ASR engine. When combined with the TRT, "consensus" ASR can produce error-free, timestamped transcripts that enable the detailed analysis of verbal responses from older individuals at risk of cognitive decline.

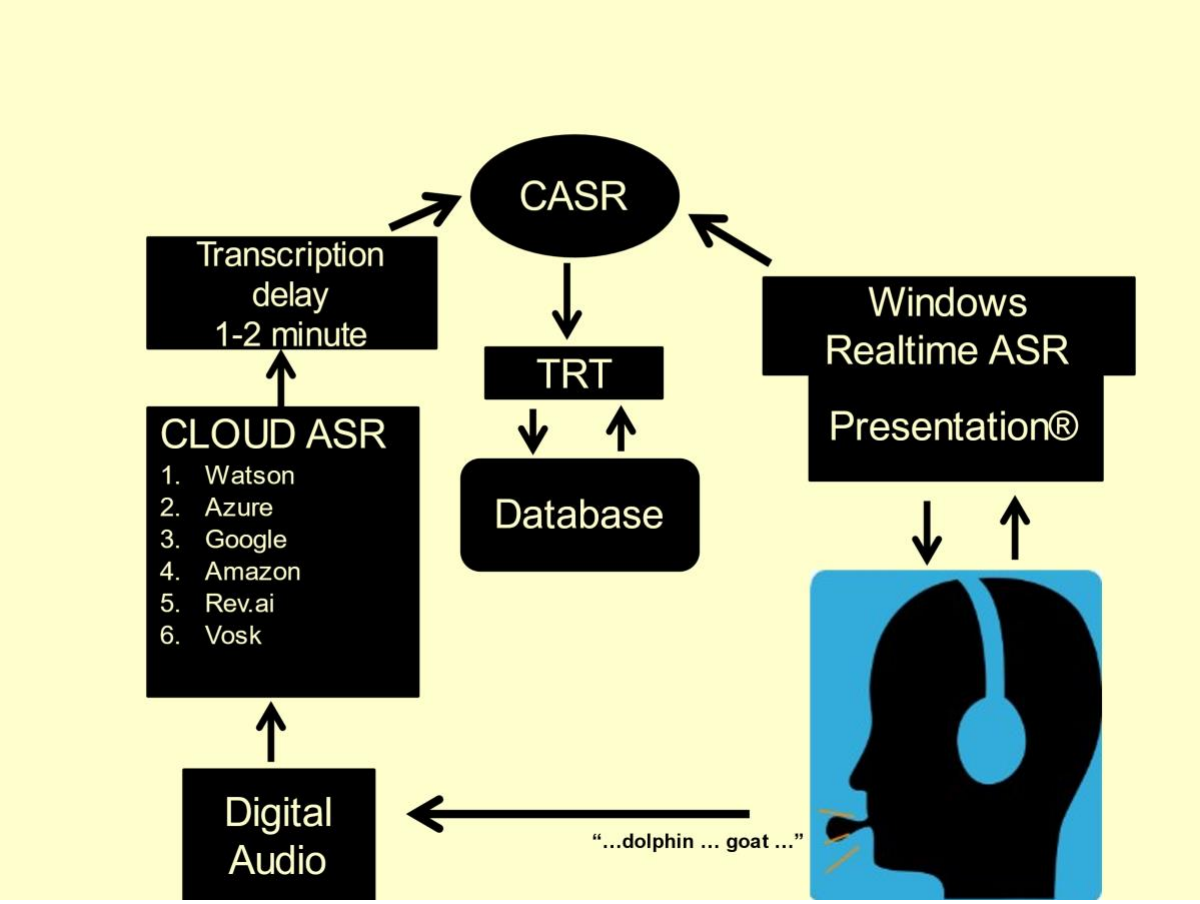

Figure 1. The CCAB speech-analysis pipeline. Spoken responses were digitally recorded through a head-mounted microphone and simultaneously analyzed in realtime with Windows ASR. Recordings were then sent to six different ASR engines whose transcripts were combined in consensus ASR (CASR) analysis. Transcripts with low-confidence words can then be reviewed with the Transcript Review Tool (TRT) prior to storage in the database.

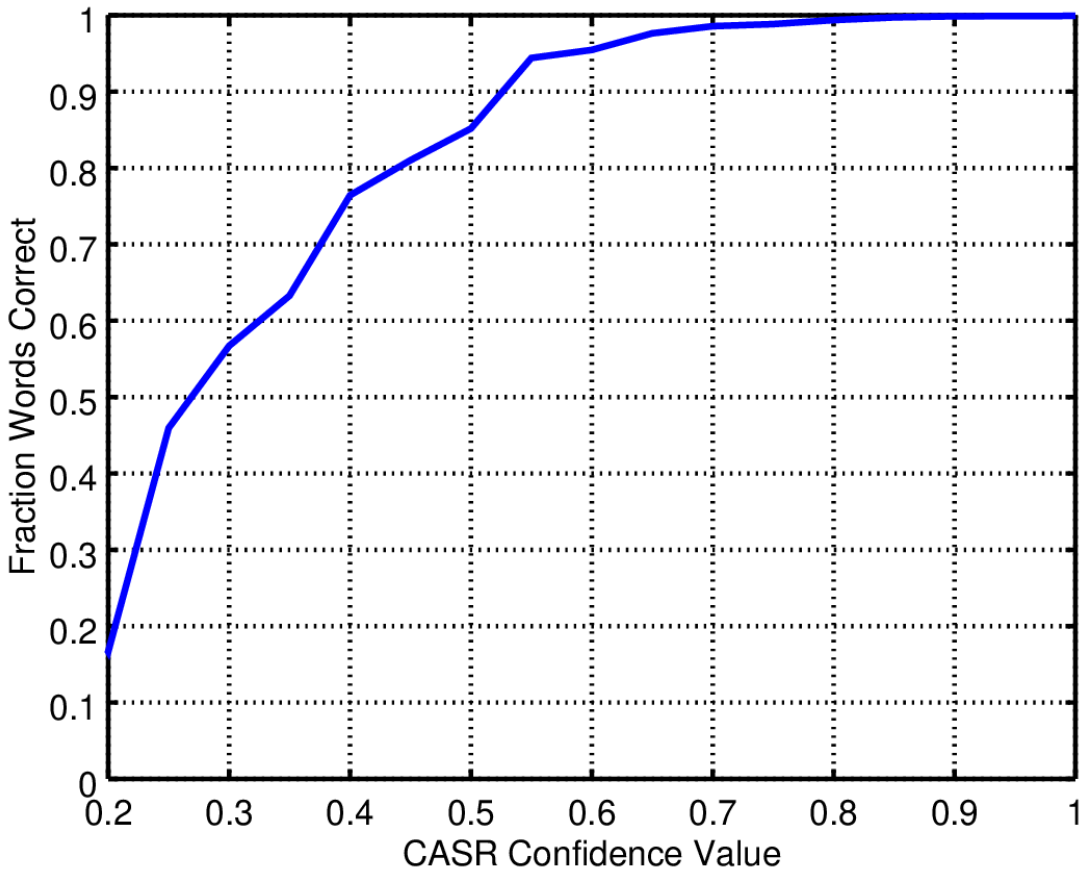

Figure 3. CASR confidence values and transcription accuracies, showing the fraction of words correctly transcribed as a function of CASR confidence. At CASR confidence values of 0.7, 0.8, and 0.9 transcription accuracies respectively exceeded 98.0%, 99%, and 99.8%.

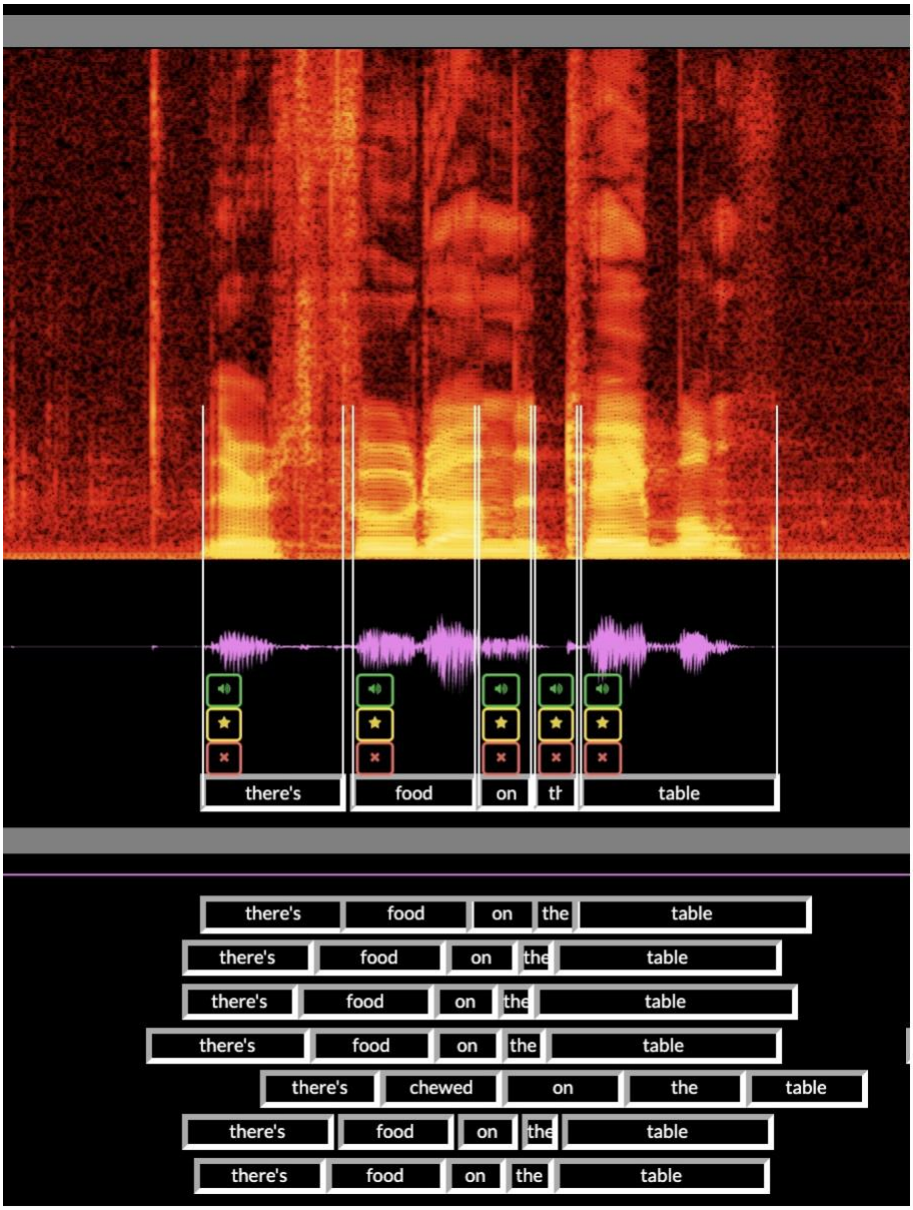

Figure 3. The transcript review tool (TRT). The TRT retrieves recordings of verbal response and displays their spectrograms (top) and waveforms (top center), a copy of the CASR transcript for editing, and estimates of word onsets and offsets,. Below the gray line is the CASR transcript followed by the individual transcripts from Rev AI, IBM Watson, Amazon, Google, Microsoft Azure, and Vosk. In the current example, Google mis-transcribed the sentence “There’s food on the table”.