The Correlation Between Standardized Test Scores, ASR Scores and Oral Presentation Scores

Paul Daniels
Kochi University of Technology

Paul Raine
J. F. Obirin University

Recent advances in AI and neural machine learning have drastically improved the accuracy of automatic speech recognition (ASR) technologies. Consequently, ASR tools have been gaining headway in language learning environments. The present study seeks to investigate the relationship between students' standardized language test scores, their speaking scores derived from a series of computer-scored speech tasks, and their speaking scores from human-rated presentations. Data was gathered from 130 first year undergraduate students enrolled at a technology university in Japan. The results show a positive correlation between the standardized test scores and the computer-scored ASR speaking tasks. The pedagogical implications of the ASR and standardized test scores correlation are also discussed.

When it comes to evaluating language competence, testing instruments have historically focused on receptive (listening and reading) skills rather than productive (speaking and writing) skills. This has been largely due to the temporal, financial, and practical challenges of evaluating productive skills (Powers, 2010). However, in recent years there has been a move to try to incorporate evaluation of speaking competence into formerly receptive-only testing instruments (Pearson, 2017a; ETS, 2018). One way to achieve this while also overcoming time, money, and practical constraints is to develop and deploy automated or computer augmented speaking tests.

We are rapidly moving into an era where speaking with a computer is becoming as common and as important as communication with humans. An increasing number of organizations are using automated menu systems to replace or supplement call center workers (Morgan, 2016), and the rise of personal assistants such as Amazon Alexa, Google Assistant and Apple Siri (Verto Analytics, 2017) mean that the frequency of interactions with artificially intelligent interlocutors is rapidly increasing.

In light of the preceding factors, this paper describes and evaluates a Moodle speech assessment plugin jointly designed and developed by the authors. The speech assessment plugin was designed to automatically score speech using the Web Speech API (Mozilla Developer Network, 2018), an Application Programming Interface for recognizing speech via the web browser, which is currently fully implemented only in Google Chrome. In an attempt to establish validity and equivalence with other testing instruments and procedures, it also investigates the correlation of scores generated by the authors' speech assessment plugin with scores generated by the Computerized Assessment System for English Communication (CASEC), as well as scores derived from human-rated speaking assessments administered by the authors.

Existing automated speaking tests

There are several automated and computer-augmented speaking tests available in the language testing market. In this section, two well-established automated speaking tests (Pearson's *Versant* and ETS's *SpeechRater*) will be briefly examined, and the test makers' claims of equivalence with human-rated scoring methods and other standardized tests will be summarized. In doing so, it is hoped that the professional standard for automated and human rated speaking tests can be established, and thereby provide a point of reference for our own findings.

Pearson Versant

Pearson's *Versant* automated speaking test purports to "assign independent scores based on the content of what is spoken and on the manner in which it is said" (Pearson, 2017a). The *Versant* test includes a variety of diagnostic tasks, such as "reading aloud, repeating sentences, building sentences, giving short answers to

questions, retelling brief stories, response selection, conversations, and passage comprehension" (Pearson, 2017a). The makers of the *Versant* test claim that it is equivalent to "repeated independent human judgements" with a correlation coefficient of 0.97 (Pearson, 2017b).

ETS SpeechRater

ETS' *SpeechRater* purports to be "the world's most advanced spoken-response scoring application" and has been deployed in the evaluation of TOEFL online practice tests since 2006 (ETS, 2018). *SpeechRater* uses Natural Language Processing (NLP) techniques to analyse "fluency, pronunciation, vocabulary usage, grammatical complexity and prosody" (ETS, 2018). The NLP techniques utilized by SpeechRater relies on a linguistic model that is based on human-rated speech. This allows the system to automatically score speech based on fluency, pronunciation, vocabulary usage, grammatical complexity and prosody. A 2011 study published by ETS (Bridgeman et al, 2012) suggests a correlation coefficient of between 0.37 and 0.55 between *SpeechRater* and human-rated scores, which were intended to measure the comprehensibility of test takers' spoken responses. The authors note that some important components of communicative competence, such as the ability of test takers to comprehend spoken and written information, were clearly not being recognized or appropriately rewarded by *SpeechRater* (Bridgeman et al, 2012).

Methodology

Participants

The participants in this study were selected from a convenience sample of 130 first year Japanese university engineering students (4 classes of 32 students) who were enrolled in a semester long general English course.

Procedure

Participants were required to complete weekly speaking tasks, hereby referred to as "the ASR tasks", that were graded by the authors' ASR Moodle speech assessment plugin, as well as complete human-rated speaking tasks that were graded by the instructor. At the end of the 16-week course, data was collated and grouped into three sets: CASEC test scores, five ASR task score averages, and two presentation score

averages. The researchers were interested in finding out whether the students who obtained high scores on the human-rated speaking tasks also scored highly on the Moodle ASR speech assessment plugin tasks. A similar question was also examined in relation to whether students who performed well on the CASEC English language proficiency test also performed well on the Moodle ASR speech assessment tasks.

Standardized English language test

Before the start of the English course, all students completed the CASEC English language proficiency test. CASEC is an adaptive test that adjusts the difficulty of test items according to the proficiency level of the learner, and the CASEC score can be converted to a comparable TOEIC test score. The test items include listening tasks, reading tasks, and grammar tasks. Since the test focused on the learners' receptive skills only, no speaking or writing was required of the test-takers. The test took about an hour to complete, and scores were available immediately after completion of the test.

Moodle speech assessment ASR tasks

Participants were assigned weekly speaking tasks that were automatically graded by the authors' Moodle ASR speech assessment plugin.

The ASR tasks included "Read Aloud", "Random Word Order", and "Speak the Best Answer" activities. Each of the weekly tasks consisted of 6 to 10 items and were completed by students either in the CALL classroom or outside class at the learners' discretion.

"Read Aloud" ASR task

One of the ASR tasks that students completed throughout the semester was the "Read Aloud" task (Figure 1). During this task, the participant listens to a question with optional text support, and then selects the record button while reading the provided response.

After completing the "Read Aloud" task (Figure 1), the participant is provided with feedback which shows how well the ASR engine was able to "understand" and transcribe their spoken words. In lieu of a numerical score, participants can receive general non-numeric feedback (Figure 2) such as "Excellent" or "Good start, but try again to improve". Under the ASR transcription, the target answer

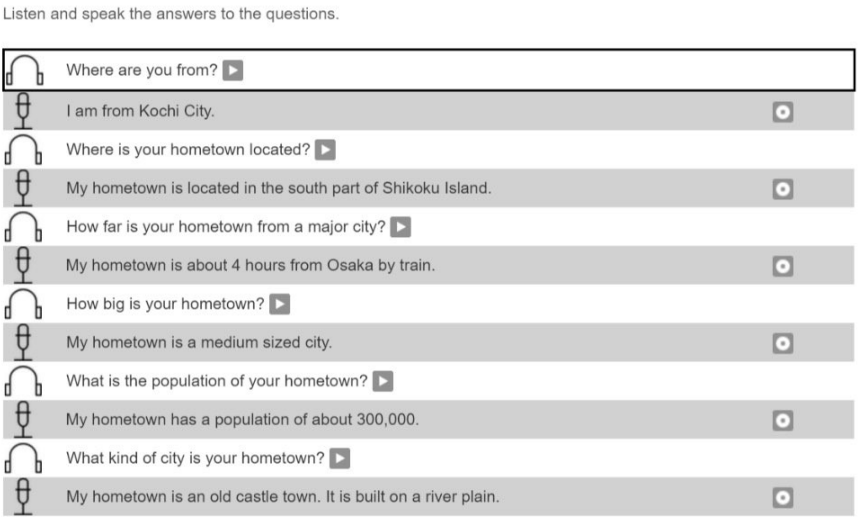

Figure 1. 'Read Aloud' ASR task with listening prompt and written response.

text is displayed, and the audio of the recorded speech can be captured using a popular JavaScript library called 'Recorder.js'. The captured audio enables both the learner and the instructor to evaluate the spoken responses.

"Random Word Order" ASR task

Using the "Random Word Order" ASR task, (Figure 3) the participant listens to a question and then attempts to speak the answer to the question using a series

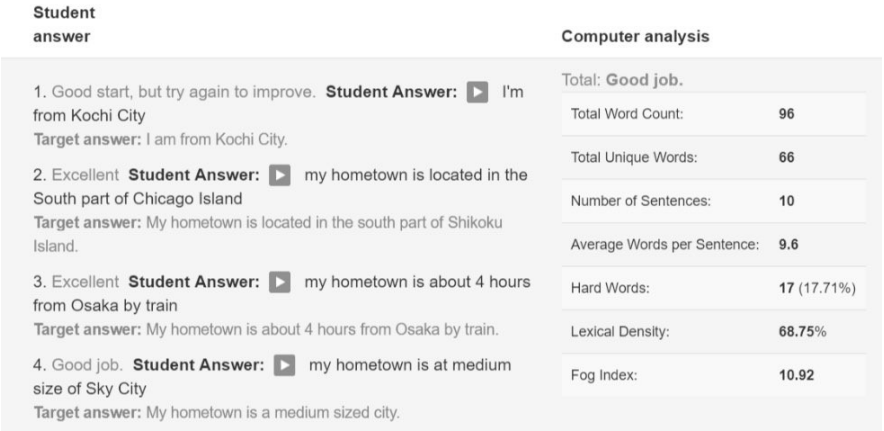

Figure 2. Results of the "Read Aloud" speaking task.

of given words or phrases that are not in the correct order. In the example below, the participant listens to the question "What is the temperature of ice?" and is prompted with the following jumbled text: [is] [the temperature] [0 °C] [of ice]. The learner then needs to speak the words or phrases in the correct order, in this case, "The temperature of ice is 0 °C."

"Speak the Best Answer" ASR task

Another ASR task administered to the participants consists of an audio prompt followed by three possible responses from which the participant must choose the most appropriate one. The participant speaks one of the three choices. For example, the audio prompt could be "How often do you go back to your hometown" and the three possible responses could be "once a month", "for three hours" and "only on weekends". The participant would then speak the best response: "once a month".

As with the other speaking tasks, the audio is captured to allow both the students and the instructor to review the student-produced speech. The captured audio was used as a reflective activity for the students, and as a way for the instructor to ensure that the learners were on task. However, the recorded speech was not used to manually calculate the speaking task scores.

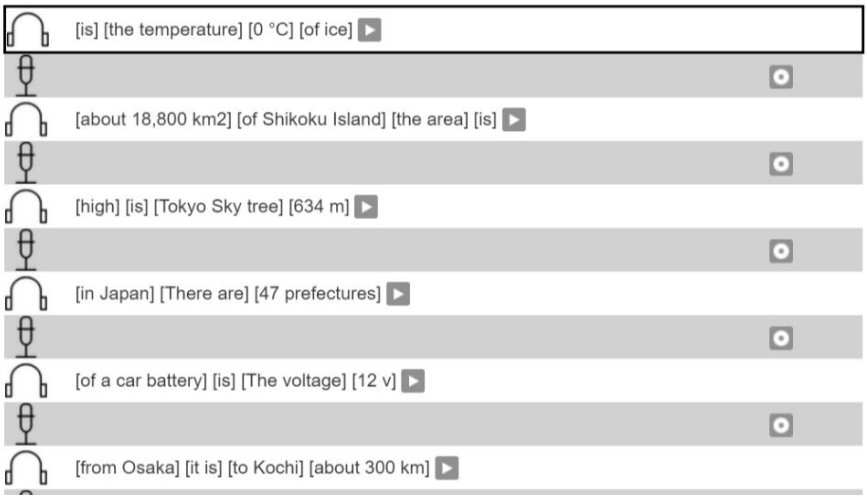

Figure 3. "Random Word Order" ASR task.

The ASR scoring algorithm

To transcribe participants' speech, the Web Speech API (Mozilla Developer Network, 2018) was used. The Web Speech API provides access to a speaker-independent continuous speech recognizer (Lumenvox, 2018) with an unrestricted grammar that allows for full dictation of most conceivable utterances. Although it is not optimized for non-native speakers (NNS) of English, a previous study has shown that the Web Speech API is able to recognize NNS speech with a high degree of accuracy (Ashwell & Elam, 2017), and that the API is amenable to digital English language learning tasks (Daniels & Iwago, 2017).

In order to derive a score for the three ASR tasks described above, a multistep process is followed for each participant utterance. First, the audio recording of the utterance is sent to the speech recognizer, which returns a transcript of the utterance. The transcript is then broken down into individual words, which are transliterated into their ARPABET phonetic equivalents. ARPABET was developed by the US Department of Defense in the 1970s, and provides a way to phonetically represent any English word using standard letters of the English alphabet. Using the ARPABET system helped ensure that homophones such as "which" and "witch" were not rejected by the scoring algorithm. The ARPABET tokens are then compared to the equivalent tokens in the target utterance, and scores and calculated based on how many of the tokens are exact matches between the participant utterance and the target utterance.

Oral presentations

Participants also performed two small group oral presentations during the semester, which were scored by a human rater. In the first presentation, students introduced data about their hometown. In the second presentation, participants reported the results of a previously undertaken design project. Both presentations were about 3 minutes in length, and included PowerPoint or Google Slides visuals. The presentation guidelines incorporated language structures that were introduced in the course textbook and in the online ASR tasks. Both presentations were recorded, and the videos were later evaluated by the course instructor using a rubric (Figure 4).

Speaker used a clear, audible voice.
Speaker used inflection, i.e. not a monotone voice.
Speaking was poised, controlled, and smooth. Had few pauses.
Pronunciation didn't interfere with understanding of language.
Speaker used complete grammatical sentences, not speaking only titles & keywords on slides.
Score 4: Excellent. 3: Good. 2: Fair. 1: Needs improvement.

Figure 4. Oral Presentation Grading Rubric.

Both the human-rated presentation tasks and the ASR tasks were designed using phrases and vocabulary that were introduced via in-class speaking activities such as pair work, and via language that students encountered while completing listening and reading activities in the course textbook.

Results

In order to establish whether the ASR task scores correlated with either the oral presentation scores or the CASEC scores, both parametric (Pearson) and non-parametric (Spearman) statistical tests were applied to the score data. The values for asymmetry and kurtosis (Table 1) fell between -2 and +2, so normal distribution of the proficiency, ASR and presentation scores was considered acceptable (George & Mallery, 2010).

A moderate positive relationship was observed between the ASR scores and the CASEC scores. The r value of Spearman's correlation coefficient was .37, and Pearson's correlation coefficient was .33, with a p value of $p < .001$ suggesting that the result is significant at $p < .05$.

A weaker relationship emerged between the ASR task scores and the presentation speaking scores. Again using the Pearson correlation coefficient test, the value of r was .18, showing a positive correlation, and the p value was .039, significant at $p < .05$.

The scatter chart of the ASR scores and CASEC test scores (Figure 5) reveals the moderate correlation between the two variables. The ASR scores and oral presentation scores scatter chart (Figure 6) reveals the weaker correlation between the two.

Table 1
Descriptive Statistics of Proficiency Test Scores, ASR Scores and Oral Presentation Scores

Proficiency test scores		*ASR scores*		*Oral presentation scores*	
Mean	519.085	Mean	76.308	Mean	82.019
Standard Deviation	69.295	Standard Deviation	13.972	Standard Deviation	6.224
Kurtosis	-0.828	Kurtosis	0.213	Kurtosis	1.291
Skewness	0.105	Skewness	-0.814	Skewness	-0.606
Count	130	Count	130	Count	130

Discussion

Our results suggest that speaking ability can to some extent be inferred from the results of receptive testing instruments such as the CASEC test, as there is a moderate positive correlation between the two. One could argue that this fact negates the need to assess speaking skills separately to receptive skills. Notwithstanding this argument, it is clear that ASR speaking activities are

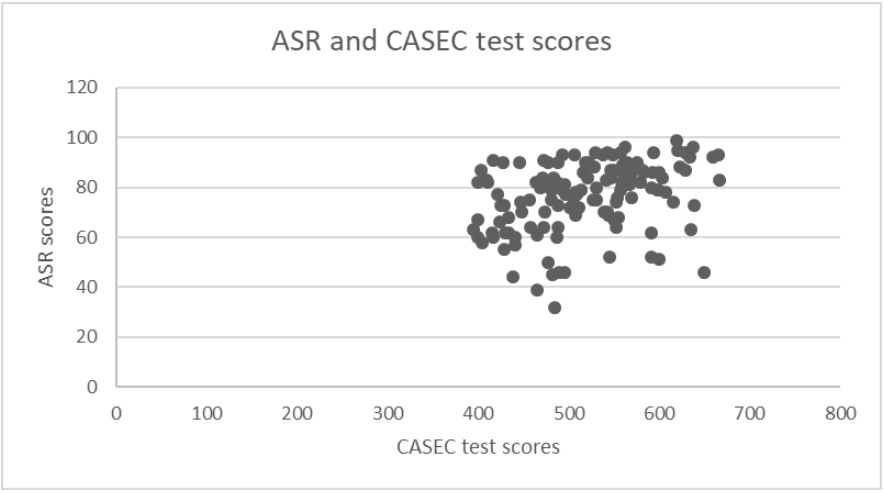

Figure 5. Correlation between ASR scores and CASEC test scores.

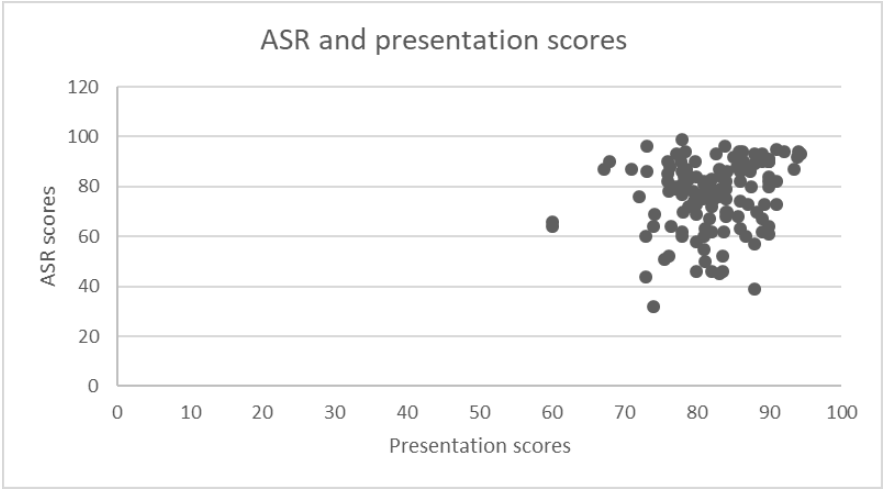

Figure 6. Correlation between ASR scores and oral presentation scores.

very useful for automating speaking practice both inside and outside language classrooms, providing an interlocutor where one would otherwise be unavailable, and delivering instant feedback on the comprehensibility of students' speech.

Furthermore, in the near future, Japanese students will be facing the prospect of more spoken English tests due to entrance exam reforms (McCrostie, 2017). ASR activities such as those introduced in this paper will be of use to educators seeking to prepare their students for such tests, especially given the fact that such activities are a statistically valid type of formative assessment. We predict the move toward four-skill tests from 2020 in Japan will promote a "positive washback" (Saito, 2019) and have a beneficial impact on learners' spoken skills as they focus on practicing for such exams, using both ASR and traditional face-to-face speaking activities.

The weaker positive correlation between the ASR task scores and the human-graded oral presentation scores suggests that our ASR activities were suited to scoring the "Read Aloud" tasks, while humans were better able to determine whether test takers could competently "go beyond" simply repeating words written on cards and use inflection effectively (Figure 4).

The fact that the positive correlation between the ASR task scores and the human-rated oral presentation scores is not stronger than expected is perhaps

not surprising given the differences in skills required by the learners to complete the ASR tasks versus the presentation tasks. As opposed to simply repeating text from a computer screen, presenting in front of classmates, as the subjects did in this study, requires extra-linguistic strategies, such as gestures, facial expressions, tone of voice, eye contact, body language and posture. Some students possess stronger extrovert characteristics or are skilled at conveying ideas using extra-linguistic strategies. Such strategies would be spotted by a human rater, but ASR systems would fail to recognize them.

Finally, several outliers appear in the data (Figures 5 and 6). There are more outliers in relation to the ASR scores, suggesting that some students were not able to properly complete the ASR tasks. Feedback from students indicates that technical and time management difficulties may have been factors in this regard.

Conclusion

The positive correlation between the ASR task scores and the CASEC scores suggests that ASR activities provide valid and reliable evaluations of spoken language that are in line with standardized proficiency test scores. ASR tasks are useful for augmenting human-rated speaking tasks, particularly in contexts where there are a limited number of human raters available. However, Web Speech API based ASR activities may not be able to completely replace human graders as they are not optimized for assessing whether speakers can "go beyond" simply repeating or reordering utterances and use extra-linguistic techniques effectively. Web Speech API based ASR activities can nevertheless provide the opportunity for extensive speaking practice with instant feedback on comprehensibility, even when a human interlocutor is unavailable.

It could be argued that the correlation between ASR scores and standardized test scores in this study negates the need to specifically evaluate speaking ability, since speaking ability may be inferred from a standardized test score. However, the positive relationship between ASR tasks and standardized test scores could also help to expand the acceptance and commitment of automatically scored speaking evaluations and potentially impact curriculum design, teaching practices, and learning behaviors.

Finally, there are valid arguments for promoting a wider use of automated formative speaking assessments. More and more international employers are seeking graduates with good communication skills (Stevens, 2005), and more institutions are adopting four-skill standardized language tests (Saito, 2019). It is clear that ASR activities will have a major role to play in the future, with regards to both the evaluation and elicitation of learner English speech.

References

Ashwell, T., & Elam, J. R. (2017). How Accurately Can the Google Web Speech API Recognize and Transcribe Japanese L2 English Learners' Oral Production?. *JALT CALL Journal*, *13*(1), 59-76.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing*, *29*(1), 91-108.

Daniels, P., & Iwago, K. (2017). The suitability of cloud-based speech recognition engines for language learning. *JALT CALL Journal*, *13*(3), 229-239.

ETS. (2018). *Automated scoring of speech*. ETS Research. Retrieved from https://www.ets.org/research/topics/as_nlp/speech

George, D. and Mallery, M. (2010) *SPSS for Windows step by step: A simple guide and reference*, 17.0 Update, 10th Edition, Pearson, Boston.

Lumenvox. (2018). *Types of speech recognition*. Retrieved from https://www.lumenvox.com/resources/tips/types-of-speech-recognition.aspx

McCrostie, J. (2017) Spoken English tests among entrance exam reforms Japan's students will face in 2020, *The Japan Times*. Retrieved from https://www.japantimes.co.jp/community/2017/07/05/issues/spoken-english-tests-among-entrance-exam-reforms-japans-students-will-face-2020

Morgan, B. (2016). *The Economist Predicts robots will replace contact centers*. Retrieved from https://www.forbes.com/sites/blakemorgan/2016/02/16/the-economist-predicts-robots-will-replace-contact-centers/#78d363571e74

Mozilla Developer Network. (2018). *Web Speech API*. Retrieved November

23, 2018, from https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API

Pearson. (2017a). *How the Versant Testing System works*. Retrieved from https://www.versanttests.com/technology/scoring

Pearson. (2017b). *Developing tests to the highest standards*. Retrieved from https://www.versanttests.com/technology/validation

Powers, D. E. (2010). *The Case for a comprehensive, four-skills assessment of English language proficiency*. TOEIC Compendium. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.599.5802&rep=rep1&type=pdf

Saito, Y. (2019). Impacts of introducing four-skill English tests into university entrance exams, *The Language Teacher*, *43*(2), 9-14.

Stevens, B. (2005). What communication skills do employers want? Silicon Valley recruiters respond. J*ournal of Employment Counseling*, *42*, 2-9.

Verto Analytics. (2017). *Rise of the machines: How AI-driven personal assistant apps are shaping digital consumer habits*. Retrieved from https://insights.vertoanalytics.com/how-ai-driven-personal-assistant-apps-are-shaping-digital-habits

Author bios

Paul Daniels *is a Professor of English at Kochi University of Technology in Japan. His research interests include CALL, ESP and Project-based instruction. daniels@kochi-tech.ac.jp*

Paul Raine *(M.A. TESOL) is a teacher, presenter, and author. He is particularly interested in Computer Assisted Language Learning (CALL), and teaches at two universities in the Tokyo area. paul.raine@gmail.com*

Received: March 15, 2019
Accepted: May 21, 2019