TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language

Auto-scoring of Student Speech: Proprietary vs. Open-source Solutions

*** * * On the Internet * * ***

November 2022 – Volume 26, Number 3

https://doi.org/10.55593/ej.26103int

Paul Daniels
Kochi University of Technology, Japan
<daniels@kochi-tech.ac.jp>

Abstract

This paper compares the speaking scores generated by two online systems that are designed to automatically grade student speech and provide personalized speaking feedback in an EFL context. The first system, *Speech Assessment for Moodle* (*SAM),* is an open-source solution developed by the author that makes use of Google's speech recognition engine to transcribe speech into text which is then automatically scored using a phoneme-based algorithm. *SAM* is designed as a custom quiz type for *Moodle,* a widely adopted open-source course management system. The second auto-scoring system, *EnglishCentral*, is a popular proprietary language learning solution which utilizes a trained intelligibility model to automatically score speech. Results of this study indicated a positive correlation between the speaking scores generated by both systems, meaning students who scored higher on the *SAM* speaking tasks also tended to score higher on the *EnglishCentral* speaking tasks and vice versa. In addition to comparing the scores generated from these two systems against each other, students' computer-scored speaking scores were compared to human-generated scores from small-group face-to-face speaking tasks. The results indicated that students who received higher scores with the online computer-graded speaking tasks tended to score higher on the human-graded small-group speaking tasks and vice versa.

***Keywords*:** Speech recognition, computer-scored speech, language learning

The increasing importance of speaking skills in Japan

Faced with a lack of authentic speaking opportunities and limited time allotted for English instruction in elementary school, Japanese students often struggle to make sufficient improvements with their speaking abilities. (Aoki, 2017). To overcome these challenges, the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) has been planning substantive reforms to English education from the elementary to the higher secondary school levels, specifically with an aim to improve English speaking skills (Nemoto, 2018). In addition to the MEXT educational reforms, the National Center for University Entrance Examinations is planning to recognize a number of standardized tests

such as Cambridge CAE, TOEFL, TOEIC, GTEC, TEAP, and IELTS as part of the new university admission exam system. In response to these changes, standardized test developers are striving to find more efficient methods to evaluate speaking abilities in a computer testing environment (Ockey, 2017; Saito, 2019). Student preparedness for these speaking exams is also essential. To better prepare students for MEXT's English language reforms and for the entrance exam speaking component, creative solutions need to be explored that can help to maximize speaking opportunities for language learners. As computerized testing environments become more common (Matthews et al., 2012; Ramesh & Sanampudi, 2022), learners will need to become familiar with the types of speaking tasks that they will encounter on the latest speaking exams in order to feel more relaxed when speaking during the actual test. Students also need to be exposed to extensive speaking tasks to improve their speaking fluency. In Japan, L2 speaking and communication skills are typically acquired in the language classroom and evaluated using standardized language tests. Speaking and communication skills are becoming increasingly more important for employers. Companies have expressed concerns that current language learning proficiency tests available on the market may not be the best indicators of English fluently levels (Murai, 2016). As companies demand better measures of speaking competencies, standardized language test companies are beginning to add speaking assessment components, which can be scored entirely by a computer or using a hybrid approach where speaking skills are assessed using humans and computers to rate the speech.

While there is a growing number of computer-scored speaking platforms available for language learning, there are few studies that evaluate the validity of the scoring methods, particularly with open-source speech assessment solutions. Open-source computer-scored speaking applications, such as *SAM*, could be a viable, inexpensive and flexible solution to provide extensive speaking opportunities in EFL environments and offer opportunities to better prepare students for computer-based speaking tests. In addition, open-source solutions typically support customization of the learning content, which may help to better align the speaking tasks to the curriculum goals.

Literature Review

CALL-based speaking tasks

Because of limited opportunities to practice speaking, the use of computer assisted language learning (CALL) together with automatic speech recognition (ASR) technology may provide learners with additional extensive speaking opportunities to help improve English communication skills. A number of proprietary speech evaluation systems have been reviewed, but little research exists on open-source speech recognition solutions for language learning. To better evaluate how cost-effective speech recognition solutions can foster speaking skills, two online speaking platforms were evaluated that are designed to automatically score student speech and provide individualized speaking feedback. The first system, an open-source speech assessment plugin for *Moodle* (*SAM*), employs *Google's* speech recognition engine to transcribe student speech into text which is then automatically scored. The second system, *EnglishCentral*, is a proprietary language learning platform that, like *SAM*, has the ability to automatically generate speaking scores from student speech. But unlike *SAM*, which uses speech recognition to generate text that is then scored, *EnglishCentral* relies on a proprietary algorithm to score speech based on a statistically trained model that estimates the intelligibility of speech (Lanting, 2015).

Speech recognition technology

Speech recognition technology consists of software that detects the words and phrases in spoken language and converts the speech to a text format. Earlier speech recognition systems, developed by Bell Labs in the 1950's, were used to transcribe spoken numbers, for example voice to digits, rather than voice to text. Later IBM improved speech recognition capabilities to better recognize and respond to a limited set of spoken words. In the 1970's the US Department of Defense made further advances and was able to develop a system that was able to recognize over 1,000 words (Boyd, 2018). Over the past 10 years we have seen the major technology corporations introduce a speech engine- Apple Siri, Microsoft Cortana, Amazon Alexa, and Google Cloud Speech (Pinola, 2011). Advances in computing power and artificial intelligence have been the main drive behind the recent advances in speech technology, with Google claiming to have achieved over 95% accuracy in recognizing speech.

With the accuracy of ASR now equal to human accuracy, speech recognition technology is being rapidly deployed in a number of industries such healthcare, consumer, military, legal, education, automotive, banking, financial services, and insurance, and government. The speech and voice recognition market is expected to be worth over 26 billion USD by 2025 (Arora, 2020; Millward, 2020).

The applications of speech recognition in language learning

A review of previous studies shows that speech recognition is one of the most commonly used computer recognition technologies (Shadiev, et al., 2020). Speech recognition is increasingly being adopted when developing online language learning applications, and past research demonstrates that that speech recognition can have a positive impact on student learning (Shadiev, et al., 2014). Automatic speech recognition first appeared in computer assisted pronunciation training systems to assess phonetic segments such as vowel metrics, phonetic vowel reduction and sentence boundary detection, as well as prosodic features including stress, rhythm, and intonation. While these systems have potential to provide constructive feedback on language learners' pronunciation errors, challenges remain in providing feedback not only on pronunciation but also on grammar and word usage (Chen & Li, 2016). Research on ASR and pronunciation training systems suggest that vowel quality and stress reduction can be easier to assess, whereas intonation and rhythm assessment can become more complicated. (Graham et al., 2015). In addition to being a pronunciation aid tool, ASR is utilized in popular online language learning platforms such as *Rosetta Stone, Babbel, EnglishCentral* and *Rocket Languages*. ASR-enhanced speaking tasks can offer immediate and individualized feedback on speaking and pronunciation skills, providing language learners with immersive self-study speaking opportunities. These tools can be especially useful in EFL environments where access to the target language is limited. (McCrocklin, 2016). A number of studies have indicated that ASR technology together with language learning content can be effective in improving speaking skills and pronunciation (Huang et al., 2016; Shadiev et al., 2018).

Speech recognition and Language testing

In addition to autonomous learning advantages, speech recognition technology may prove useful when developing low-stakes speaking tests. The Educational Testing Service (ETS) developed an automated scoring system based on their *SpeechRater* system that identified and filtered problematic responses that could be human-scored, while the un-problematic responses were auto-scored. Using this method, an improved correlation was obtained between the automated scores and the human scores (Zechner et al., 2015). Wang et al.

(2018) evaluated ETS' *SpeechRater* system and found that automated scoring of speech by the *SpeechRater* computer system was more consistent than human scoring. The study also reported that computer generated speaking scores can be useful to identify overly strict or overly lenient human raters. Additional validity studies have shown that scores from computer-scored speaking tests were able to accurately predict scores from oral proficiency interviews (Bernstein et al., 2010). Automated scoring of speech is still in the development stage, and presently may be most useful in low-stakes testing environments using close-ended or predictable speaking tasks such as read-aloud and sentence completion tasks (Isaacs, 2018) which are described in more detail in the following section.

Computer-scored speaking task types

Computerized-scoring of speaking skills often involves highly constrained speaking tasks with a limited set of possible correct responses. Table 1 outlines typical speaking task types that can be automatically scored by computers.

Table 1: Computer-scored speaking task types and description

Task type	**Task description**
Shadowing (*imitative*)	Learner listens to and repeats, or reads aloud the target language, and then speaks.
Speak correct answer (*responsive*)	Learner listens or reads a prompt and speaks the correct answer from a list of possible choices.
Speak correct order (*intensive*)	Learner listens to or reads a series of words or phrases that are not in the correct order and then speaks the phrases in the correct order.
Retell a story (*responsive*)	Learner listen to a short story and then retells the story using vocabulary from the original story.
Free speaking activities (*extensive*)	Student speaks freely about a topic.

ASR and spontaneous speech

As speech recognition technology advances, it will gradually become a more reliable tool for scoring spontaneous speech. ETS is currently experimenting with the scoring of spontaneous speech in their TOEFL iBT Speaking section (Zechner, 2019). ASR-aided spoken dialogue systems are another promising interface that can provide more realistic and engaging language learning experiences by allowing students to hold spoken conversations with a computer in the target language (Litman et al., 2018). Another study, which employed both computers and humans to score a simulated job interview conversational task, indicated that computer-generated speaking scores could accurately predict human scores of the same speech task. (Ramanarayanan et al., 2017). Personal robots, such as Amazon Echo and Google Home are also making their debut into language classrooms providing additional opportunities for speaking practice (Moussalli & Cardoso, 2016)

Open-source speech assessment - SAM

Moodle is one of the most popular open-source learning management systems worldwide and is typically used as a platform for distributing educational content and deploying online practice activities (Young, 2018). Perhaps its most distinctive advantage is that it is open-source, allowing educators and developers to add tailored plugins that can expand

functionality to better meet learners' needs. In addition to plugin flexibility, the activity content is generated by the instructor. Because of Moodle's ability to allow educators to customize both content and activity types, it was chosen as a platform for the development of an open-source speech plugin named Speech Assessment for Moodle, or *SAM*.

SAM was designed as an open-source module that can be administered within an institution's existing Moodle course management system. The *SAM* quiz question-type plugin is intended to assist EFL learners with their speaking skills and to better prepare learners for the speaking component of the next-generation of entrance exams in Japan which include a speaking component. *SAM* has the ability to provide extensive online speaking opportunities for language learners to complement in-class face-to-face speaking tasks. *SAM* allows educators to easily create customized speaking tasks that can supplement current course content, that can be automatically graded, and that offer individualized speaking feedback to the learner. *SAM* incorporates a phoneme-sequence matching algorithm to improve scoring accuracy by assigning points for each correctly spoken phoneme in a string of words. The phoneme-sequence check is able to provide more detailed feedback to the learner on specific words or phrases that the speaker may have difficulty speaking.

Proprietary speech assessment – English Central

EnglishCentral is a popular proprietary English language learning platform used primarily in Asia. It is based on a self-learning concept where learners select their own study content based on their proficiency levels and interests. The platform steps learners through three core tasks where they listen to video content, select and study vocabulary items from the video content and finally speak chosen lines from the video content. *EnglishCentral's* key feature is its ability to automatically score student speech and provide personalized speech feedback to the learner. Studies (Dixon, 2015; Robb, 2016) have suggested that *EnglishCentral* can be an effective tool to improve speaking skills as it is able to provide greater opportunities for language practice than a traditional teacher-centered language instruction.

EnglishCentral uses a proprietary system called *Intellispeech* to evaluate a speaker's accent and then scores the 'sound' of the learner's speech by comparing it with speech samples gathered from native speakers. The *EnglishCentral* system also checks for fluency by determining the number and the length of pauses. Finally, it checks whether the learner spoke all of the words that were expected.

Key research questions

The purpose of this research is to advance the development of open-source automatic speech recognition and evaluation platforms in order to efficiently and reliably evaluate speaking abilities. In order to investigate how ASR along with the automatic grading and feedback of speech can be best implemented to improve oral production skills, the following research questions are investigated.

A. To what extent do the speaking scores of shadowing tasks that were derived from *SAM* correlate with the speaking scores of the same shadowing tasks within *EnglishCentral* in an EFL context?

B. To what extent do computer-graded speaking task scores correlate with human-graded speaking task scores in an EFL context?

Methodology

Participants

20 Japanese undergraduate engineering students participated in the study- 8 female and 12 male students. Participants were in their third and fourth year of undergraduate study, and were all enrolled in an English language elective course. The speaking activities used in this study were designed to be integrated into the course content and were requirements for the final course evaluation, therefore participants were encouraged to complete all of the speaking activities to the best of their abilities to receive credit for the course. The participants' TOEIC scores ranged from 450 to 550 and participants had lower-intermediate speaking skills. Most never studied abroad and had limited L2 speaking experiences both inside and outside of the classroom.

Design & procedure

The focus of the speaking and communication elective course was to actively engage students in various communicative activities, typically in groups of 3 or 4. For the course evaluation, students completed two types of speaking activities. The first activity, which was the main activity of the course, involved 4 small-group mini-presentations and weekly discussion sessions on topics chosen by the learners. In addition to the group speaking tasks, students were asked to complete online speaking tasks which were automatically scored. These online tasks involved watching a video based on a technology topic and speaking or 'shadowing' 5 to 10 lines, or sentences, of the video. Each spoken line was automatically scored by the computer. The students completed the identical 'shadowing' tasks using two different systems. Figure 1 shows the open-source *SAM* speaking task and Figure 2 shows the proprietary *EnglishCentral* speaking task. While additional speaking tasks could have been investigated in this study, the shadowing activity was chosen as it was the central speaking task within the *EnglishCentral* system and therefore identical speaking tasks could be compared between the open-source and the propriety systems. With the shadowing tasks, the students would listen and read one line of the video and then speak the line. Shadowing tasks have been found to be associated with effective listening comprehension and fluency skills (Hamada, 2019).

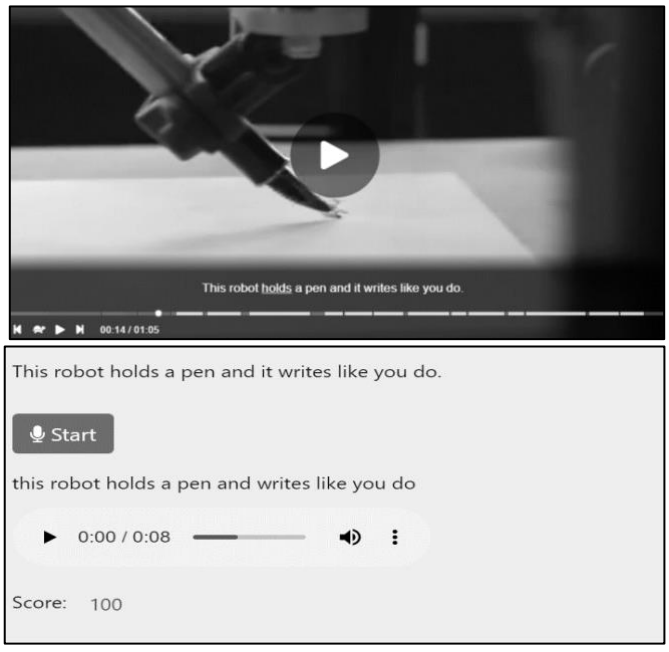

Figure 1. *SAM* speaking task

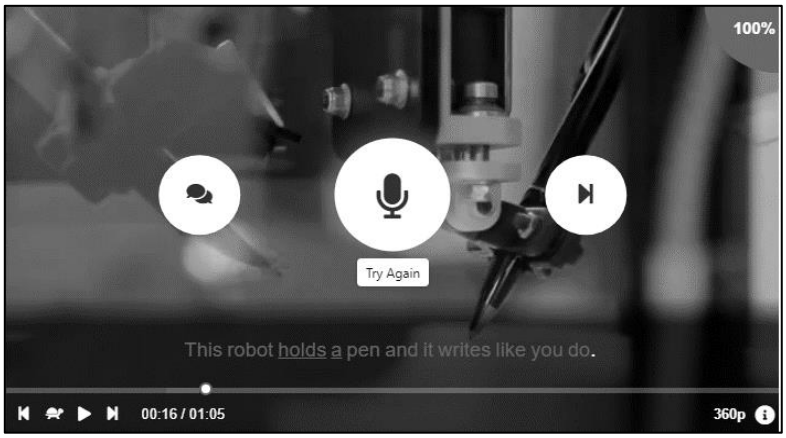

Figure 2. *EnglishCentral* speaking task

Speaking tasks

The participants' speech for the *SAM* auto-graded speaking tasks was transcribed using Google's Web Speech API and scored using *SAM*'s custom phoneme scoring algorithm. Although Google's Web Speech API is not specifically trained to transcribe speech from non-native speakers of English, it was found to be accurate when used with EFL learners (Ashwell & Elam, 2017), and has effectively been used to support online speaking practice in the language classroom (Daniels & Iwago, 2017). *SAM*'s speaking task scoring algorithm uses Google's text transcription to first break down the speech into individual words as well as their ARPABET phonetic equivalents, which are then compared to target phonetic equivalents based on the speech task. A score is then calculated using a percentage match between the phonemes from the students' transcription and the phonemes from the target text.

All participants completed the identical online speaking tasks using both *EnglishCentral* and *SAM*. Participants completed the online speaking tasks a total of 4 times, and each online speaking task took 30 minutes on average to complete. *EnglishCentral's* computerized scoring algorithm is somewhat different from *SAM's* scoring algorithm. *EnglishCentral* employs a statistically trained intelligibility model (Gokgoz-Kurt, 2017), rather than transcribing the speech, which is the method employed by *SAM*. The *EnglishCentral* scores are based on a phonetic match of the 'sounds' of the speech, whereas *SAM*'s scores are based on a 'text' match of the transcribed speech and target text.

In addition to the online speaking tasks, all participants took part in 4 small-group speaking tasks during the semester, in which each participant was instructed to speak for approximately three minutes on a prepared topic. Each of the 4 small-group speaking tasks lasted 30 minutes. The first speaking task topic focused on 'summer breaks', the second task involved 'explaining a process', and the third task was a 'cause & effect' topic. All three speaking tasks were both peer-scored and instructor-scored using a scoring rubric shown in Table 2, which was created by the course instructor. The in-class speaking tasks were not intended to replicate the shadowing tasks, but rather to identify if any relationships exist between computer-scored shadowing tasks and teacher-scored live speaking tasks. Although, both the shadowing tasks and small group discussions focused on improving student's oral fluency skills via extensive speaking tasks, rather than a focus on vocabulary and grammar use. The main reason for this approach is that Japanese junior and senior high school English language programs typically focus on grammar and vocabulary skills, and therefore this undergraduate course was primarily centered on improving speaking fluency.

Table 2. Scoring rubric for small-group speaking tasks

Eye contact	Energy	Speaking	Content	English only
- Looked at & communicated with classmates when speaking.	- Was excited about topic and motivated audience**.**	- Spoke clearly and at a speed that was easy to understand.	- Gave enough information about topic.	- Used only English when speaking.

4: Very good, **3**: Good, **2**: So-so, **1**: Need more effort

Results

Descriptive Statistics

To determine whether the speaking scores of shadowing tasks derived from *SAM* correlate with the speaking scores of the same shadowing tasks within *EnglishCentral*, the Pearson Correlation Coefficient test was used to observe any linear relationship between the two variables, For the Pearson Correlation Coefficient test, variables should be normally distributed, or form a bell curve. To check for a normal distribution, or normality, of the scores derived from *EnglishCentral* and *SAM*, Kurtosis and Skewness values, shown in Table 3, were generated. Both values fell between -2 and 2, which are acceptable. The Pearson Correlation Coefficient test also assumes that the variables have equal variances. Levene's test for equality of variance was used to determine whether the scores from the auto-scored speaking tasks have equal variances. The requirement of homogeneity, an assumption underlying both t tests and F tests, was met with a non-significant result of ($p=0.17$) between the auto-scored speaking scores obtained from *SAM* and from *EnglishCentral*. The f-ratio value was 1.99, and the p-value was 0.17, therefore the result is not significant at $p < .05$.

With the small-group speaking tasks, which were human-scored, the Skewness was acceptable at 1.7, but the Kurtosis value was 4.9 indicating that the scores were not normally distributed. In addition, when examining equality of variance between the *SAM* scores and small-group speaking scores, the requirement of homogeneity was not met, as the obtained p-value was 0.000438.

Table 3. Descriptive statistics of the independent variables

EnglishCentral speaking scores		*SAM speaking scores*		*Small-group speaking scores*	
Mean	56.99	Mean	48.57	Mean	85.82
Standard Deviation	20.00	Standard Deviation	13.55	Standard Deviation	3.02
Kurtosis	-0.37	Kurtosis	1.38	Kurtosis	4.94
Skewness	0.58	Skewness	0.95	Skewness	1.70
Range	68.00	Range	54.70	Range	13.77

Statistical data of SAM and EnglishCentral

To address the first research question, the Pearson Correlation Coefficient test was used to determine if the scores from an open-source auto-graded speaking task, correlate with computer-generated scores from an identical task using *EnglishCentral*. The results in Figure 3 indicate a moderate positive correlation (The P-Value was 0.004 and the value of

R, or the coefficient of determination, was 0.43) suggesting that there was a tendency that students who scored high with *SAM* also scored high with *EnglishCentral*, and vice versa.

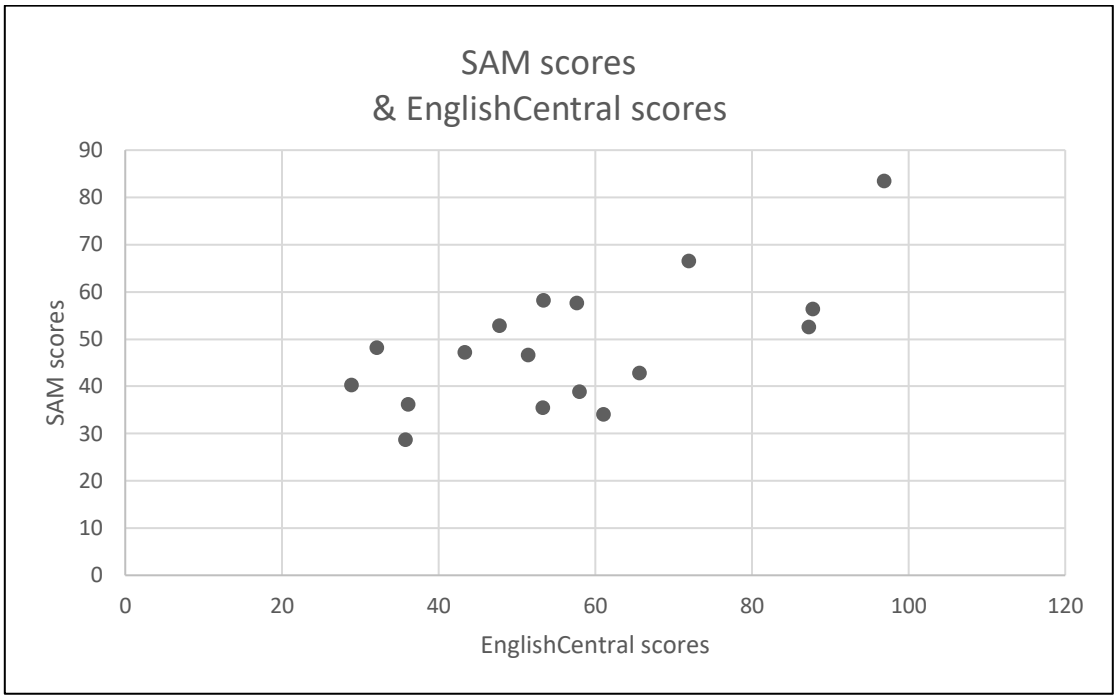

Figure 3. Correlation between *SAM* & *EnglishCentral* speaking task scores

Auto-graded and human-scored speaking task scores

The second research question investigates the relationship between the computer-scored and the human-scored speaking tasks. Again the Pearson Correlation Coefficient test was adopted and the results shown in Figure 4 indicated a moderate positive correlation between the two variables (The P-Value is .033, The value of R is 0.52), suggesting that students who scored high on the auto-graded speaking tasks with *SAM* also scored high on the human-scored small-group speaking tasks, and vice versa. The Pearson Correlation Coefficient test, used to compare auto-scored grades and live speaking task scores, assumes normality of data and equal variance among other factors. The normality values of the *SAM* and small-group speaking scores were acceptable as they fell between -2 and 2, although the equal variance value of the live speaking task scores was not acceptable. The human-scored speaking task scores ranged from 80% to 95%, with very little variance. On the other hand, the computer generated scores generated from *SAM* had a much greater range from 28.7% to 83.4%.

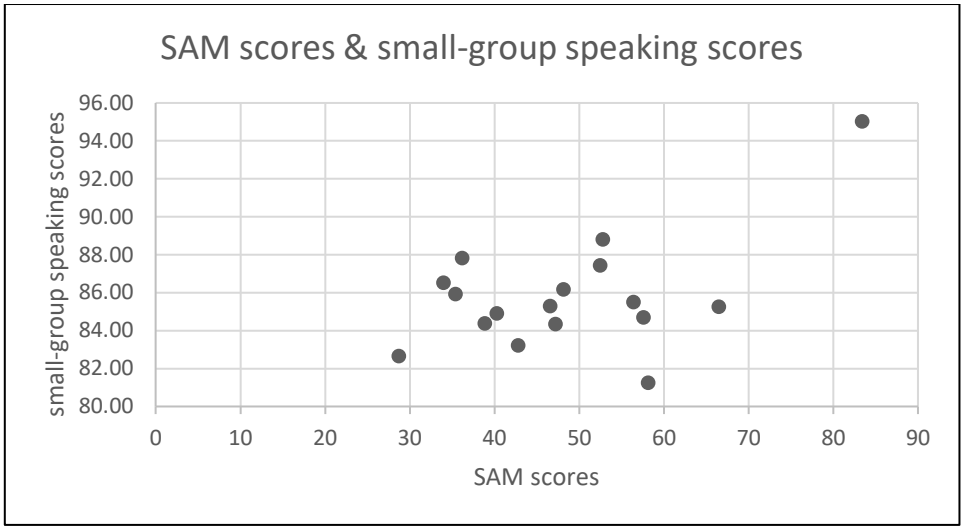

Figure 4. Correlation between *SAM* scores & small-group speaking task scores

Discussion

From the data obtained from this study, the most significant result was the positive relationship between the speaking scores generated by English Central and *SAM*- the two auto-scoring speech systems that were being compared in this study. The validity of the speaking scores between the systems which employ different scoring algorithms was confirmed. These findings suggest that an open source auto-scoring system can be deployed inexpensively, which can automatically generate speaking scores that are consistent with speaking scores generated from a proprietary language learning system that has been previously tested and confirmed to have improved students speaking proficiency (Kimura, 2013; Dixon, 2015; Robb, 2016). Along with the cost-saving benefits, the open source *SAM* system, allows the language teacher to add customized content that can be tailored to suit the individual goals of a particular language learning curriculum or learner group, whereas with the proprietary system, the content already exists and cannot be easily modified or added to by the language teacher. The key takeaway of this study is that with the recent improvements of web-based speech recognition, it is becoming easier and less expensive to deploy speech recognition that can support a variety of speaking tasks in the language classroom.

The second part of this study examined the correlation between the computer-generated scores and the human-generated scores for the speaking tasks. Again, for this part of the study, possible relationships were examined between the computer-score speaking tasks and human-rated speaking tasks, although the two tasks were not similar. The computer-scored speaking tasks were closed-ended shadowing tasks, while the human-scored activities were open-ended semi-prepared speaking tasks. A positive relationship emerged between the auto-scored *SAM* tasks and human-scored tasks, but these findings were not considered significant as the equal variance criteria of the live speaking task scores was not met. This could be due to the narrow bands of the scoring rubric (1, 2, 3, or 4) which may have contributed to the small spread in scores from the live speaking tasks. These results may also suggest that the computer-generated scores may be more reliable than the human-generated scores for certain types of speech tasks as computers leave less room for bias in the grading process compared to humans. Therefore, by implementing computer scoring systems, such as *SAM*, biased human raters may be more easily identified. In addition, based on the findings of this study, language instructors in an EFL setting may want to consider employing computer-scored speaking activities that reinforce course content and provide

additional speaking practice outside of the classroom. Computer-based speaking tasks should also be considered for learners who need additional practice for standardized speaking assessment tests that have an online speaking component. Finally, it would be worth investigating the differences and correlations between speaking scores produced by human raters and by computers for a parallel speaking task.

Limitations

Extraneous variables may have influenced the results of this study, for example, issues with background noise when completing the computer-graded speaking tasks, audio recording quality, and network robustness. In addition, the study did not employ the same type of language tasks when comparing the closed-ended *SAM* task scores and open-ended classroom speaking task scores. It cannot be assumed that close-ended speaking task scores accurately predict students' abilities when completing open-ended speaking tasks. Finally, this study on computer-scored speaking tasks employed a small sample size, therefore the results of this study cannot be generalized to a wider population.

Conclusion

From a review of recent literature on computer-scored speaking assessment, and from the results of this study, educators can better understand how speech recognition can potentially be adopted as a viable self-study tool for general speaking practice as well as for speaking test practice. Systems for auto-scoring speech remain in their infancy but initial data from this study and from other studies suggest that these systems can be used in conjunction with live-speaking tasks for language practice and for low-stakes speaking evaluations. With the ability to provide immediate feedback and recommend areas for improvement to the learner, these systems have the potential to increase speaking opportunities for EFL learners.

About the Author

Paul Daniels is a professor of English language at Kochi University of Technology in Japan. He has been using technology for over 30 years to enrich students' language learning experiences and he actively leads international workshops on computer-assisted language learning themes. His current research involves speech recognition and how it can be used to engage learners in speech practice. ORCID ID: 0000-0002-3552-0642

To Cite this Article

Daniels, P. (2022). Auto-scoring of student speech: proprietary vs. open-source solutions. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 26* (3).
https://doi.org/10.55593/ej.26103int

References

Aoki, M. (2017, April 6). Japan's latest English-proficiency scores disappoint. *The Japan Times*, https://www.japantimes.co.jp/news/2017/04/06/national/japans-latest-english-proficiency-scores-disappoint/

Arora, A. (2020, December 8). Global speech and voice recognition market size & share, future growth, trends evaluation, demands, regional analysis and forecast to 2026.

MarketWatch, https://www.marketsandmarkets.com/Market-Reports/speech-voice-recognition-market-202401714.html

Ashwell, T., & Elam, J. R. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *The JALT CALL Journal*, *13*(1), 59–76. https://doi.org/10.29140/jaltcall.v13n1.212

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Boyd, C. (2018). Speech recognition software: History, present & future. *Medium*, https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf

Chen, N.F., & Li, H. (2016). Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 1-7. http://www.apsipa.org/proceedings_2016/HTML/paper2016/227.pdf

Daniels, P., & Iwago, K. (2017). The suitability of cloud-based speech recognition engines for language learning. *The JALT CALL Journal*, *13*(3), 211-221. https://doi.org/10.29140/jaltcall.v13n3.220

Dixon, S. (2015). Evaluating the impact of an online English language tool's ability to improve users' speaking proficiency under learner- and shared-control conditions. [Unpublished doctoral dissertation]. Arizona State University. https://repository.asu.edu/attachments/150602/content/Dixon_asu_0010E_14813.pdf

Gokgoz-Kurt, B. (2017). EnglishCentral as a Tool to Improve Pronunciation. *TESOL Journal*, 8(4), 894-898. https://doi.org/10.1002/tesj.351

Graham, C., Caines, A., & Buttery, P. (2015). Phonetic and prosodic features in automated spoken language assessment. *Proceedings from the Workshop on Phonetic Learner Corpora, International Congress of the Phonetic Sciences*, 37–40. http://www.ifcasl.org/docs/Graham_final.pdf

Hamada, Y. (2019). Shadowing: What is It? How to Use It. Where Will It Go? *RELC Journal*, 50(3), 386–393. https://doi.org/10.1177/0033688218771380

Huang, Y., Shadiev, R., & Hwang, W. (2016). Investigating the effectiveness of speech-to-text recognition applications on learning performance and cognitive load. *Computers & Education*, *101*, 15-28. https://doi.org/10.1016/j.compedu.2016.05.011

Isaacs, T. (2018) Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, *15*(3), 273-293. https://doi.org/10.1080/15434303.2018.1472264

Kimura, T. (2013). Improvement of EFL learners' speaking proficiency with a web-based CALL system. Glasgow, 10-13 July 2013 Papers, 141. https://www.scribd.com/document/317752775/WorldCALL2013-papers-pdf

Lanting, P. (2020, September 22). Speaking archives. *EnglishCentral Solutions*. https://solutions.englishcentral.com/category/speaking/

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, *15*(3), 294-309. https://doi.org/10.1080/15434303.2018.1472265

Matthews, K., Janicki, T., He, L., & Patterson, L. (2012). Implementation of an automated grading system with an adaptive learning component to affect student feedback and response time. *Journal of Information Systems Education*, *23*(1), 71-84. https://aisel.aisnet.org/jise/vol23/iss1/7

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *System*, *57*(0), 25-42. https://doi.org/10.1016/j.system.2015.12.013

Millward, W. T. (2020, February 25). Education robot maker Roybi boosts kids speech recognition technology with acquisition. *EdSurge*. https://www.edsurge.com/news/

Moussalli, S., & Cardoso, W. (2016). Are commercial 'personal robots' ready for language learning? Focus on second language speech. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), CALL communities and culture – short papers from EUROCALL 2016 (pp. 325-329). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.583

Murai, S. (2016, January 25). Changes in store for TOEIC, but test still not total Gauge of fluency: Experts. *The Japan Times*. https://www.japantimes.co.jp/news/2016/01/25/reference/changes-store-toeic-test-still-not-total-gauge-fluency-experts/#.Xmhol6j7QuU

Nemoto, A. (2018). Getting ready for 2020: Changes and challenges for English education in public primary schools in Japan. *The Language Teacher*, *42*(4). https://jalt-publications.org/articles/24344-getting-ready-2020-changes-and-challenges-english-education-public-primary-schools

Ockey, G. (2017). Approaches and challenges to assessing oral communication on Japanese entrance exams. *JLTA Journal*, *20*(0), 3-14. https://doi.org/10.20622/jltajournal.20.0_3

Pinola, M. (2011, November 2). Speech recognition through the decades: How we ended up with Siri. *PCWorld*. https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html

Ramanarayanan, V., Lange, P. L., Evanini, K., Molloy, H. R., & Suendermann-Oeft, D. (2017). Human and automated scoring of fluency, pronunciation and intonation during human–machine spoken dialog interactions. *Interspeech 2017*, 1711-1715. http://dx.doi.org/10.21437/Interspeech.2017-1213

Ramesh, D., & Sanampudi, S.K. (2022) An automated essay scoring systems: a systematic literature review. *Artif Intell*, 55, 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Robb, T. (2016). EnglishCentral improves test scores at Japanese university. *EnglishCentral Solutions*. https://solutions.englishcentral.com/2016/11/03/englishcentral-improves-test-scores-at-japanese-university/

Saito, Y. (2019). Impacts of introducing four-skill English tests into University entrance examinations. *The Language Teacher*, *43*(2), 9. https://doi.org/10.37546/jalttlt43.2-2

Shadiev, R., Hwang, W.-Y., Chen, N.-S., & Huang, Y.-M. (2014). Review of Speech-to-Text Recognition Technology for Enhancing Learning. *Educational Technology & Society*, 17(4), 65–84. http://www.jstor.org/stable/jeductechsoci.17.4.65

Shadiev, R., Sun, A., & Huang, Y. (2018). A study of the facilitation of cross-cultural understanding and intercultural sensitivity using speech-enabled language translation technology. *British Journal of Educational Technology*, *50*(3), 1415-1433. https://doi.org/10.1111/bjet.12648

Shadiev, R., Zhang Z. H., Wu, T.-T., & Huang, Y. M. (2020). Review of Studies on Recognition Technologies and Their Applications Used to Assist Learning and Instruction. *Educational Technology & Society*, 23(4), 59–74. https://www.jstor.org/stable/26981744

Wang, Z., Zechner, K., & Sun, Y. (2016). Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, *35*(1), 101-120. https://doi.org/10.1177/0265532216679451

Young, B. (2018, April 29). Top 3 advantages of Moodle. *eLearning Industry*. https://elearningindustry.com/advantages-of-moodle-top-3

Zechner, K. (2019). Summary and outlook on automated speech scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment* (pp. 192-204). Routledge. https://doi.org/10.4324/9781315165103-12

Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., Wang, X., & Yoon, S. (2015). Automated scoring of speaking tasks in the test of English-for-Teaching (TEFT™). *ETS Research Report Series*, *2015*(2), 1-17. https://doi.org/10.1002/ets2.12080