the jaltcalljournal

ISSN 1832-4215
Vol. 13, No.3 Pages 229–239

The suitability of cloud-based speech recognition engines for language learning

Paul Daniels
Kochi University of Technology
daniels@kochi-tech.ac.jp

Koji Iwago

As online automatic speech recognition (ASR) engines become more accurate and more widely implemented with CALL software, it becomes important to evaluate the effectiveness and the accuracy of these recognition engines using authentic speech samples. This study investigates two of the most prominent cloud-based speech recognition engines-Apple's Siri and Google Speech Recognition (GSR) to determine which engine would be more accurate at transcribing L2 learners' speech. The average recognition accuracy of Siri and GSR is reported using language samples of Japanese learners speaking English. The study also presents a series of computerized speech assessment tasks that were developed by the researchers using a cloud-based speech recognition engine in conjunction with Moodle, a widely used course management system.

Background of speech recognition

Computerized speech recognition systems were being designed as far back as the early 1930s when Bell Labs began conducting research on computerized transcription of human speech. As personal computers became more widespread, speech recognition software, such as Dragon NaturallySpeaking, shifted to the desktop market. While speech recognition initially was lauded as an effective text input method, users unsurprisingly preferred keyboards to microphones for text input. Speech recognition technology has seen wider use in assisting people with text input who are not able to use traditional text input devices such as

keyboards. As the accuracy and the efficiency of speech recognition software improve, a wider range of user may embrace it.

It was not long before language educators and CALL developers became interested in integrating speech recognition technology with CALL activities, particularly with language production practice. Speech recognition software was utilized early on in *Dyned's* language learning software, in *Subarashii*, an interactive dialog system for learning Japanese and in ECHOS, a voice interactive French language training system. Voice recognition also was adopted by companies debuting automated speech assessment technology. *PhonePass*, now Pearson Versant, offered one of the first fully automated tests of spoken language.

With the growing popularity of mobile devices, speech recognition is now becoming a useful tool for mobile users as it enhances multitasking. It was initially used to assist users with hands-free searching for contacts and with dialing numbers, useful when driving. With early mobile devices, speech recognition was rudimentary since the recognition engine was installed on those mobile devices. Speech recognition changed significantly with the introduction of Smartphones. These new 'Smart' devices are typically bundled with data services allowing users to be connected to the Internet anytime, anywhere. With today's robust mobile networks, the speech recognition engines are able to processes speech on powerful cloud servers. The mobile device is simply acting as a microphone which sends the audio out over the Internet to a server which performs the CPU intensive processing of the speech and sends back the transcribed text to the mobile device. This kind of a software system, called a client-server model, has several advantages. One advantage is that applications that use speech recognition can be easily deployed on a mobile device without additional strain on the device's CPU or memory. Another advantage is that the speech recognition software is easy to update and maintain because it is installed on the server side. Today, cloud based speech recognition is embedded in almost all mobile operating systems.

Cloud-based speech recognition

Apple's Siri and Google Speech recognition (GSR) have evolved as two of the most promising cloud-based speech recognition technologies. When designing language learning tasks, it is possible to use either of these cloud-based recognition engines to analyze L2 pronunciation. While it is no surprise that past research (Ploger, 2015) concludes that human beings can understand accents and mispronunciation better than speech recognition software, Siri and GSR can handle accented and mispronounced speech to some extent. L2 speakers are often unaware of their pronunciation problems. However, by using a recognition engine, such as Siri or GSR, pronunciation problems can be instantly identified by the learner because the actual utterance is transcribed to text in real-time. When mistakes are identified, L2 speakers can become more aware of their pronunciation problems. With online speaking tasks, learners can practice and easily check their pronunciation again and again. However, improving pronunciation is not always obvious to a learner. The learner must first identify which syllables are mispronounced. Once problematic areas are identified, a specific remedy can be suggested. The accuracy of L2 speech transcription becomes an important element when employing speech recognition tools for language learning (Neri, et al., 2003). Therefore, the purpose of this study is to determine whether Siri or GSR is more accurate at transcribing L2 speech.

Background research

Previous research on ASR and language learning has focused predominantly on pronunciation training. Studies conducted by Neri, et al. (2002), Ploger (2015), Hincks (2002), and Elimat & AbuSeileek (2014) suggest that ASR holds potential benefits for language learners, particularly when coupled with self-study CALL activities that incorporate practical learner feedback. Neri, et al. (2002) observed that pronunciation training using ASR offered a valuable, stress-free learner experience, particularly when learners were provided verification of correct responses as well as effective remedies for their learning errors. Ploger (2015), reporting on a single learner in a case study, found that dramatic pronunciation improvements occurred when using dialogue practice along with ASR. Ploger (2015) also suggested that feedback was more helpful to the learner when a score or accuracy percentage was provided by the ASR application rather than a simple positive or negative response, although the researcher also pointed out that the ASR's false negatives posed a problem for the learner. The importance of immediate and useful feedback is a recurrent theme and therefore, a feature which needs to be given careful consideration when designing ASR activities for language learning purposes.

The majority of the research on ASR was conducted before cloud-based speech recognition tools were readily available to the public sector. Older ASR systems often provided pronunciation feedback using speech waveforms that illustrate air movement of fricative consonants or aspiration of stops. Hincks (2002) reported that learners found these waveforms to be ineffective. This may explain why the results of this study suggest that the ASR software and activities employed did not discernibly improve pronunciation. Newer cloud-based ASR engines, such as Siri and GSR, which convert L2 utterances into text, can help improve learner feedback by returning a transcription of a spoken utterance to the user. Therefore, locating errors in pronunciation using a real-time transcription may be easier for the learner to interpret compared to waveforms and spectrograms which tend to be difficult for L2 learners to effectively utilize.

Feedback on pronunciation needs to be accurate to ensure that the correct pronunciation is not mistakenly modified and that poor pronunciation is not reinforced. Although the quality of pronunciation cannot be accurately analyzed, cloud-based speech recognition engines are far less complicated and less expensive to deploy compared to traditional ASR engines which typically need to be installed and maintained on a local server. The ease of use makes cloud-based ASR suitable for quick self-pronunciation practice. Additionally, since instructors often don't have enough time to constantly monitor and provide feedback to individual learners in a large class, cloud-based ASR language tasks can be both effective and motivating. L2 learners who are afraid of making mistakes in public can comfortably practice speaking in a private setting.

In addition to feedback, a wider range of ASR tasks need to be employed when designing ASR systems for language learning. While most ASR research focuses on pronunciation training, a few studies suggest other innovative uses of ASR for language learning. Cai, et al. (2013) report on a study on how ASR could be used to apply gamification theory to a word/picture matching task. The researchers claimed that by relaxing the constraints of ASR or making it more lenient, users became more engaged in the activity. Because false negatives are common with non-native speakers using cloud-based ASR systems designed for native speakers, learning engagement can be negatively affected.

ASR has been shown to be effective as a language learning tool in language games and

pronunciation practice. It is able to provide learners with greater opportunities to practice language. ASR appears to offer numerous advantages for oral practice, and further research needs to be conducted on its effectiveness on improving the accuracy of the language through pronunciation training but also on improving oral fluency. ASR can easily be implemented in tasks that encourage extensive speaking. Using ASR, speech reports can be compiled for evaluative purposes which summarize, for example, word counts of spoken utterances, length of utterance, and lexical density of the language produced. As speech recognition applications are becoming more popular in CALL, educators often question the effectiveness of the speech technology, particularly as CALL developers continue to add additional features to their applications. To date there has not been a tremendous amount of studies conducted on the effectiveness of language learning activities that incorporate speech recognition technology. One can look at the motivational aspects of using speech recognition technology in EFL settings where limited speaking opportunities exist.

Since popular speech recognition engines such as Siri and GSR are developed specifically for L1 speakers, it is important to verify if these tools can adequately transcribe L2 speech in order for the output to be meaningfully applied to language learning activities. Both Apple and Google's speech engines rely on the context of the utterance in order to 'guess' the meaning of the phrase when transcribing speech. Siri and GSR more accurately transcribe strings of speech that occur more frequently, such as "have you ever" or "went to the." Therefore, we can assume that if an L2 speaker leaves out an article or uses a preposition incorrectly, the software may run into difficulty with the transcription. This grouping of language may play an important role in both authentic listening activities as well as in speech recognition accuracy. With this in mind, it seems appropriate that before speech recognition activities could be adequately assessed as to how well they can aid in language instruction, the performance of the speech recognition engines need to be assessed to determine how well they deal with L2 speech. Since Siri and GSR are the most pervasive engines available on mobile devices, with iOS devices using Siri and Android devices using GSR, the researchers set out to determine how accurate these two tools are at transcribing L2 speech.

Research questions

Which online speech recognition tool is more accurate at transcribing L2 spoken utterances?

For L2 learners which tool could be used more effectively for designing online speaking activities for learners for English study?

Procedure

The participants consisted of 41 undergraduate students at two separate Japanese universities who were enrolled in general English language courses. The students' majors ranged from science to humanities, however none of the majors were related to English or language studies. Each participant was instructed to speak a total of 8 sentences into a microphone one sentence at a time. The transcription of each student recording was then entered into a spreadsheet and compared to the target sentence to determine the accuracy of the transcription. The vocabulary and grammar of the target sentences that were used in the task were at a similar level to the language being introduced in the English course in which they were enrolled. Each of the 8 sentences was spoken by the participants and

transcribed a total of four times- two times using Siri and two times using GSR. To ensure a more objective evaluation of the two transcription engines, half of the students started the speaking task using GSR while the other half of the students started with Siri. This was an effort to ensure that the attempt at speaking the sentence and the order of the recognition engine being used were equal – neither Siri nor GSR had an advantage of transcribing speech that the participant had practiced more.

Data analysis

Table 1 provides a summary of the accuracy of the transcribed data by both the GSR and Siri transcription engines. The columns correspond to the target sentences, and the accuracy of the transcription of each speech recognition engine is listed in the corresponding column. The accuracy of the transcriptions was determined using a string comparison tool that calculates a similarity coefficient between two texts (Oliver, 1993). For example, if all of the words in the transcribed text matched the target sentence and were in the same order, a score of 100% was assigned.

Table 1. Data analysis from the string comparison tool

Average recognition accuracy of GSR and Siri for 8 spoken sentences									
Sentence #:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Ave**
GSR (%):	95.7	77.6	96.8	86.6	84.4	69.2	85.0	60.5	82.0
Siri (%):	96.9	59.5	75.9	72.6	72.7	51.1	60.5	46.2	66.9

n=41

As seen in the above table, the data reveal that the average score of GSR's accuracy is considerably higher than that of Siri for seven sentences out of eight. The overall averages were 82.0% for GSR and 66.9% for Siri. When each transcription is analyzed, it becomes apparent that Siri sometimes missed words as if they were not pronounced at all. For instance, the first sentence "Where are you from" was transcribed as "Where are you." Siri may not recognize some sounds such as a weakly pronounced 'R'. For instance, GSR transcribed 'earth' correctly. On the other hand, Siri transcribed it as 'us.'

GSR appears to make use of contextual clues to make corrections as a sentence is being transcribed. On the other hand, Siri did not appear to make corrections as intelligently as GSR while transcribing based on the context. Furthermore, after one word is transcribed incorrectly, the remaining words in the sentence were sometimes transcribed incorrectly. Siri appears as though it was confused by a single word at which point it was not able to process the rest of sentence.

It is observed that the average accuracy of the first sentence is very high because it is relatively short and easy to pronounce. The second sentence is longer and more difficult to pronounce. The Japanese language does not have the R sound which may be the reason many participants have a problem pronouncing it correctly. Having said that, not all words with the R have the same degree of difficulty. When the R is at the beginning of a word, it is typically easier for a Japanese speaker to pronounce. However, if it is in the middle or at the end of a word, it may be dropped or mispronounced. For example, 'born' is sometimes wrongly transcribed as 'bone.'

It may have been also useful to look at phoneme matches since quite often Siri and GSR would transcribe the student's speech with part of a word matching, for example if the target word is 'there' and the student's speech is transcribed as 'they', partial credit should be given for the correctly matched 'th' or ð phoneme.

Implications and ASR activities

Not only does GSR appear to be more accurate at recognizing L2 speech than Siri, it is also relatively easy to integrate into web-based language-learning apps. Apple only allows developers to make use of Siri via a native app. GSR, on the other hand, offers a web-based API for voice technology, allowing developers to add voice recognition capabilities to ordinary HTML web pages as well as web-based apps. Because of the numerous advantages of GSR technology, the researchers decided to employ the GSR API with an automated speech assessment plugin for Moodle to allow teachers to administer a number of online speaking tasks which incorporate automated scoring and feedback. The following section provides a description of the types of speaking tasks that can be administered online.

Using the speech assessment activity, tasks can be administered online to capture audio, transcribe this captured audio, and perform basic text analysis of the transcription. Depending on the assessment algorithm, a speaking score can be automatically generated by comparing the transcribed text to the model answers. This automated assessment is typically beneficial with closed-ended questions that have a limited or restricted number of responses, for example, if a learner is asked to respond to a question while looking at an illustration which provides a clue to the correct answer. An example of a closed-ended question might be "What is the circumference of the circle?" Possible correct answers may include "The circumference of the circle is 10 centimeters" or "The circle has a circumference of 10 centimeters." Dictation tasks can also be set up to be closed-ended. As seen in Figure 1, the learner is able to listen and participate in conversational dialogues, which learners typically encounter in language textbooks. In this example, each active line of the dialogue is highlighted. The user selects the play icon to listen to that particular line. After listening to one line of the dialogue, the user can then select the record button and repeat that line of the dialogue. The learner is then presented with a score as well as the transcribed text, which appears to the right of the target phrase. The score is generated by comparing the target text to the transcribed text using a 'similar_text' PHP function [3]. The score, the transcript and the captured audio are saved to the Moodle course for both the learner and instructor to access.

Figure 1. A conversational dialogue using speech recognition

Figure 1 illustrates an online speaking task that can be automatically scored where the learner reads or listens to a series of words that are not in correct order and then attempts to speak the phrase in the correct order. In the example illustrated in Figure 2, the learner is prompted with: [you] [did] [What] [arrive] [time], and needs to speak: 'What time did you arrive?'.

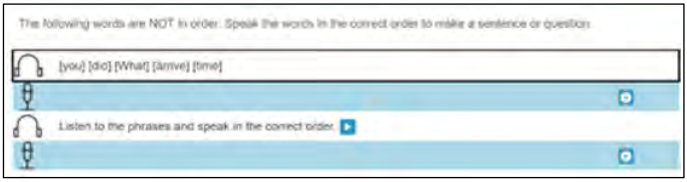

Figure 2: A scrambled word task using speech recognition

Figure 3 illustrates an online speaking task where the learner listens to an audio prompt, for example, "How often do you study in the library?" and is then shown three possible responses. The learner should then speak the best response from the following:

[everyday] ——— [for 3 hours] ——— [in between classes].

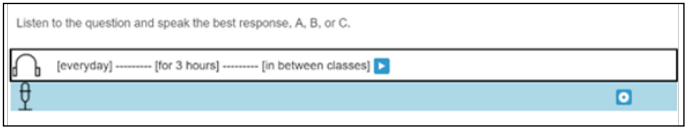

Figure 3: Speaking the best response task using speech recognition

Open-ended speaking tasks can also be administered online and, to some extent, automatically scored. One such example is a task where the learner listens to a short story and then attempts to retell the story. With this task an automated text comparison can be performed to match words or phrases from the target story with the student's transcribed text of the story. The transcription can also be automatically analyzed for word count, number of sentences, average words per sentence, and lexical density. In addition, the student audio is

captured at the same time for self, peer, or instructor assessment. A completely open-ended speaking task might be a simple prompt such as "Speak for 1 minute about your weekend." Like the open-ended story retelling task, the audio, transcript, and analyzed text data can be saved to the course.

Figure 4: Text analysis of transcribed speech

Conclusion

From the analysis of this study, the researchers determined that GSR was both more accurate at transcribing L2 speech and easier to deploy than Siri. Therefore, GSR was chosen as the recognition engine for a series of speech assessment plugins that are currently being developed by the researchers for Moodle. Several types of online speaking tasks were illustrated which can be used to automatically score L2 speech. These speaking tasks will be employed in the next stage of this research project where the reliability of the scoring algorithm and student responses to the use of online speaking tasks will be evaluated. It will be important to determine the motivational aspects of online speaking activities as well as the importance of reaching out to different learner types. As students learn in different ways, online speaking activities should be administered as supplemental practice activities. Ideally, learners should be able to make choices as to how they practice speaking, with automated online speaking tasks as one of the options. In addition, GSR should not be used as an assessment tool as its assessment algorithm cannot be verified. Both Siri and GSR are closed source, and educators do not have access to the recognition algorithms that are employed. GSR, for example, has an option for different types of native English input, such as American, British, or Australian, but no options exist to instruct GSR that the language input is from a L2 speaker, which may offer non-native speakers inaccurate speech recognition results. Finally, educators and learners need to be aware of privacy concerns of these cloud-based services. The audio as well as the transcription are captured on Google's servers with little knowledge of how this user data will be used.

Appendix 1

Materials provided to participants for the study

音声認識・音声入力ソフトの精度を比較する為にボランティアを募集しています。ボランティアにはいくつかの英語のフレーズや文章を読んでもらいます。回数は1度で、時間は15分から20分くらいになると思います。

[Warm-up]

At https://www.google.com/intl/en/chrome/demos/speech.html, please speak the text below:

"Hello. Today I will practice speaking English using a computer. I am speaking into the microphone now. The words that I speak appear on the screen as text. It is difficult but the computer understands some of my words."

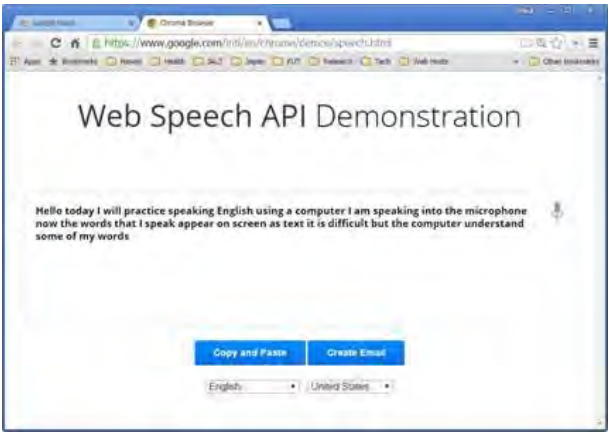

Figure 5: Speech transcription practice page

[Speaking task]

Using GSR: https://www.google.com/intl/en/chrome/demos/speech.html

I. Please speak the 8 sentences below clearly, one at a time.

II. After you speak each sentence, please take a photo or screenshot of the results that appear on your screen after each time.

III. Please speak each sentence a second time, and take a photo or screenshot of the results after each time.

Using Apple Siri on an iPad

I. Please speak the 8 sentences below clearly, one at a time.

II. After you speak each sentence, please take a photo or screenshot of the results that appear on your screen after each time.

III. Please speak each sentence a second time, and take a photo or screenshot of the results after each time.

1. Where are you from?
2. I was born and grew up in a small town in western Japan.
3. How long does it take to go from your home to school?
4. It takes about thirty minutes to walk from my home to school.
5. How many people are living on our earth?
6. There are over seven billion people living on our earth.
7. What is the diameter of the earth?
8. Earth has a diameter of about twelve thousand seven hundred kilometers.

Appendix 2

Notes for educators and developers interested in using speech recognition & audio capture.

1. Transcribe audio:

Using the HTML5 Speech Recognition API, JavaScript has access to a browser's audio stream which is converted to text using Google's speech recognition engine and returned to the browser as raw text. Tools: *webkitSpeechRecognition* API

2. Capture audio:

Recorder.js Javascript library can be used to capture audio from any input device. The audio stream is saved as a .wav file using getUserMedia. The .wav file can then be converted to an .mp3 file in real-time within the browser using libmp3lame.js. Tools*: getUserMedia* API, *record_wav.js* and *libmp3lame.js* JavaScript libraries

3. Capture & transcribe audio:

Audio capture is performed using Recorder.js as outlined in the previous example. The audio is then transcribed using Google's webkitSpeechRecognition API. The trick is that a python proxy is required to convert the captured wav audio file to flac – mono 22Hz, which is the format that Google's speech recognition engine requires. The transcribed text reply from Google's server then needs to be parsed. Tools: *speech_recognition* module written in Python.

References

Cai, C. J., Miller, R., & Seneff, S. (2013). Enhancing speech recognition in fast-paced educational games using contextual cues. *Speech and Language Technology in Education Proceedings*. Grenoble, France – August, 2013. Retrieved from http://www.slate2013.org/images/slate2013_proc_light_v4.pdf

Elimat, A. K., & AbuSeileek, A. F. (2014). Automatic speech recognition technology as an effective means for teaching pronunciation. *The JALT CALL Journal, 10* (1), 21–47.

Hincks, R. (2002). Speech recognition for language teaching and evaluating: A study of existing software. In ICSLP 2002 – interspeech 2002. *Proceedings of the 7th international conference on spoken language processing.* (pp. 733–6). Denver, Colorado, USA. September 16–20.

Neri, A. Cucchiarini, C., & Strik, H. (2002) Feedback in Computer-Assisted Pronunciation Training: technology push or demand pull?, in *ICSLP 2002, Proceedings of the International Conference on Spoken Language Processing.* Denver, USA. pp. 1209–1212.

Neri, A. Cucchiarini, C., & Strik, H. (2003) "Automatic Speech Recognition for second language learning: How and why it actually works", in *Proceedings of 15th International Congress of Phonetic Sciences.* Barcelona, Spain. pp. 1157–1160.

Oliver, I. (1993). *Programming classics: Implementing the world's best algorithms.* New York: Prentice Hall.

Ploger, D. (2015) Computer Speech Recognition and Language Learning: A Case Study. *Proceedings of the 2015 Conference for Industry and Education Collaboration,* Copyright ©2015, American Society for Engineering Education, Washington DC. Retrieved from http://www.indiana.edu/~ciec/Proceedings_2015/ETD/ETD315_Ploger.pdf