the jaltcalljournal

ISSN 1832-4215
Vol. 13, No.1 Pages 59–76

How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production?

Tim Ashwell
Komazawa University
tashwell@komazawa-u.ac.jp

Jesse R. Elam
Tokyo Denki University
elamj@mail.dendai.ac.jp

The ultimate aim of our research project was to use the Google Web Speech API to automate scoring of elicited imitation (EI) tests. However, in order to achieve this goal, we had to take a number of preparatory steps. We needed to assess how accurate this speech recognition tool is in recognizing native speakers' production of the test items; we had to assess its accuracy with our Japanese EFL learners; and, on the basis of these trials, we needed to evaluate the potential for using the API for our purposes. Through comparing our own assessments of the learners' pronunciation with the system's ability to transcribe utterances, we were able to ascertain that the learners' pronunciation of certain sounds is probably the single biggest reason for a fall in recognition accuracy compared to native speaker input. However, we argue that pronunciation may not be an insurmountable barrier to using this speech recognition system for our EFL purposes. By going through this double screening process, we feel we have arrived at a set of items which can be used to assess student's grammatical ability in an EI test using a custom Google Web Speech system.

Keywords: Automated Speech Recognition (ASR), Elicited Imitation (EI) tests, Google Web Speech API, pronunciation, Sphinx

Introduction

Automated speech recognition (ASR) is revolutionizing the way humans interact with computers, hence, second language (L2) learning researchers are understandably excited about its potential to help revolutionize how EFL students acquire a second language. Unfortunately, even with

Forum

the hugely powerful systems that are used widely nowadays on smartphones and PCs, recognition accuracy is not perfect. In this study, we were interested in exploring whether the Google Web Speech API could be used to help automatically score L2 learners' performance on elicited imitation (EI) tests designed to measure spoken grammatical ability, but in order to reach that goal a number of exploratory, preparatory steps needed to be taken to assess the viability of using the API to create a custom Google Web Speech system. Therefore, the aim of this study was to ascertain how accurately our custom Google Web Speech system could recognize and transcribe both native speakers' and Japanese L2 English learners' oral production.

A brief history of ASR

Over the past two decades, the race for a reliable ASR has attracted a number of major tech-based companies with large players such as Apple, Amazon, Microsoft, Nuance, and Google all vying for space in the field; however, ASR has had a longer history than one might expect. "Research on speech recognition dates back to the 1930s when AT&T's Bell Labs began using [mechanical] computers to transcribe human speech" (Daniels, 2015, p. 177). This later inspired the development of a one-syllable ten-digit recognition system in the 1950s by Bell Labs, MIT, and NEC, which could identify numbers spoken by the users (Juang & Rabiner, 2005). Nevertheless, the early systems faced serious limitations, as they used transistors and frequency sensors to conduct speech recognition. Hence, they were only able to identify a restricted range of phonetic sounds (e.g., numbers 1–10), and they were typically only reliable with an acoustic model of the original speaker whose voice was recorded to establish the original waveform. As technology improved in the 1970s and 1980s, ASR systems could handle up to 1000 different words, by utilizing language models and more sophisticated algorithms for analyzing data (2005). During this period, researchers and programmers began to develop pattern clustering methods for speaker-independent recognizers and introduced dynamic programming methods for improving connected word recognition (2005). This meant that the ASR systems could distinguish between a wider range of speakers regardless of their inherently different pronunciations and could begin to recognize more than single words – as had been the case until that point.

Dragon was one of the first commercial companies to develop a product that brought voice recognition to personal computers using these new methods. "In the late 1990s Dragon Naturally Speaking [was] released . . . [and] was later purchased by Nuance which offered speech recognition applications for Windows and for mobile devices" (Daniels, 2015, pp. 177–178). Eventually, Nuance and other commercially available ASR systems were developed that used Hidden Markov Models (HMM) to analyze syntax and semantics, which in turn increased the recognition and accuracy of the programs. Even today the HMM is one of the most influential algorithms for automated speech recognition and, thanks to the tremendous advancements in the last decade, systems are now able to handle an almost unlimited vocabulary set integrated with text-to-speech processing (Juang & Rabiner, 2005). Some of these systems are even open-source and/or cloud-based, making it possible for consumers to have their first experience with the latest technology.

In 1986 Sphinx was launched and became one of the most successful open source ASR systems developed for research purposes using HMM (Juang & Rabiner, 2005). This was a major advancement in ASR because now anyone with programming skills could implement their own customized ASR system. Since its initial release, however, Sphinx has gone

through a number of different modifications. In the past, "the decoding strategy of these systems tended to be deeply entangled with the rest of the system. As a result of these constraints, the systems were difficult to modify for experiments in other areas" (Walker et al., 2004, p. 1). Nonetheless, the newest version, (Sphinx-4) which was released in 2010, "works with various kinds of language specifications such as grammars, statistical language models (SLMs), or blends of both" (Twiefel, Baumann, Heinrich, & Wermter, 2014, p. 1). This means that researchers are given more flexibility in the way they can incorporate acoustic models, which allows constraints to be imposed on the expected input (language models) from the user.

Thanks to the advancements in mobile technology, internet speed, and cloud-based computing, voice recognition systems like, Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google's Assistant are becoming ubiquitous in everyday life. Furthermore, these systems are continually improving on their respective accuracy rates by constantly gathering acoustic information and utilizing machine learning. According to Twiefel et al. (2014), Google's acoustic models were originally based on data collected from a free telephone service and had over 5,000 hours of training when version 2 was released in 2010. Through this endeavour, it was believed that, "distributed speech recognition systems [could] offer better recognition accuracy than local customization systems" (Twiefel et al., 2014, p. 2). That is, systems, like Google's ASR, no longer needed to be reliant on the data stored locally on the computer, as it had the ability to transcribe speech-to-text in real-time, making the number of identifiable words seemingly limitless. "While voice activity detection (VAD) and feature extraction may be performed on the client . . . the computationally expensive decoding step of speech recognition is performed on Google's servers," (p. 1) ultimately allowing third-party developers to easily integrate the Google Speech API into their own custom ASR systems. At the time of writing this research paper, Google Cloud Platform had just been released including Google Speech API alpha, which could potentially be used on virtually any platform including iOS (Google, 2016).

One of the negative aspects of Google ASR is that the expected input cannot be controlled; nonetheless, Twiefel et al. (2014) believe that integrating Google ASR with Sphinx might alleviate this issue. This is because all of the computations are happening on Google's servers, and the output is a text string that best represents the audio input statistically, using machine learning. This means, Google ASR is quite accurate with the acoustic models it uses to decipher speech input, but it is a black box when it comes to analyzing the phonemes produced by the users. Hence, researchers at Brigham Young University have realized that although Sphinx is limited in its acoustic models, its strengths lie in its ability to break down input at the syllable level, input which can then be transcribed into phonemes (2014). To this end, Twiefel et al. have suggested using the benefits of each system by making a hybrid ASR which allows phonetic post-processing. This means the original input can be run through the Google ASR with the resulting text string sent to Sphinx to be deconstructed into the phonetic form. As a result, investigators can have access to more useful data for conducting analysis. Moreover, Sphinx can allow researchers to constrain the exact expected input, making it a great candidate for identifying particular phonetic sounds for set phrases. Thus, the resulting output can then be compared to an expected input at the syllable level. Such a combination would increase reliability and accuracy of voice recognition dramatically, therefore creating more opportunities to employ voice recognition for educational purposes.

Pedagogical applications of ASR

In the field of L2 learning, researchers have become interested in how ASR systems can be used to increase students' confidence, pronunciation, and motivation (see Chiu, Liou, & Yeh, 2007; Elimat & Abuseileek, 2014; Golonka, Bowles, Frank, Richardson, & Freynik, 2014; Kim, 2006; McCrocklin, 2016; Wang & Young, 2014). After analyzing 350 research articles in a historical account of the pedagogical use of technology for foreign language learning, Golonka et al. (2014) concluded, that Computer-Assisted Language Learning (CALL) and ASR systems have had a reasonable influence in increasing L2 students' motivation, been a useful aid in giving feedback, and helped learners develop metalinguistic skills.

Much of ASR research has focused on pronunciation in the past (Chiu et al., 2007; Elimat & Abuseileek, 2014; Golonka et al., 2014; Kim, 2006; Luo, 2016; McCrocklin, 2016)According to McCrocklin (2016), "Research in pronunciation learning strategies has struggled to provide methods for autonomous pronunciation practice in which students can also get clear feedback to help them improve" (McCrocklin, 2016, p. 26). Consequently, many educators and researchers have been turning to ASR systems – specifically Sphinx – to provide students with pronunciation feedback (Chiu et al., 2007; Golonka et al., 2014; Kim, 2006; McCrocklin, 2016). Recent studies have shown that pronunciation and intelligibility are connected to Japanese students' self-efficacy (Lear, 2013; Toyama, 2015) and these researchers have advocated pronunciation activities utilizing ASR systems as the most beneficial for increasing pronunciation accuracy due to the inherent nature of the immediate corrective feedback (CF) they provide (Golonka et al., 2014). In this way, De Vries, Cucchiarini, Bodnar, Strik, and Van Hout (2014) maintain that the current technological limitations of ASR systems suggest that they are more suitable for giving implicit forms of feedback in pronunciation, which require students to concentrate more on their errors, thus, promoting teachable moments.

Although the role of CF in language acquisition has a long-debated history (Russell, 2009), it is clear from many studies that some form of feedback is better than no feedback at all (Ellis, 2012; Lyster & Ranta, 1997). In one experimental study, for example, De Vries et al. (2014) used an ASR system called GREET (utilizing Sphinx) to determine the effectiveness of CF on grammar correction. When users in the experimental group made mistakes, the system would notify them of their errors in red, while the control group did not receive any feedback. After analyzing the pretest-posttest, student logs, and surveys, De Vries et al. found that students who received CF enjoyed the system more. Furthermore, in a similar Taiwanese study, Wang and Young (2014) explored the idea of using a Sphinx-based ASR system named iCASL to measure multiple levels of CF in pronunciation practice for self-paced learning. An experimental group received a three-step error correction that included implicit and explicit error correction while the control group only received implicit correction. After using t-tests to compare the two groups and analyzing the qualitative data, the researchers concluded that the experimental group largely improved; a combination of both implicit and explicit correction through the use of an ASR system enabled them to gain better pronunciation skills over an 8-week period (Wang & Young, 2014). These studies demonstrate how ASR systems can now be used to test CF in a more controlled environment.

ASR systems have also been the focus of research into motivation and learner autonomy recently (Chiu et al., 2007; Golonka et al., 2014; Kim, 2006; McCrocklin, 2016). For example, McCrocklin (2016), examined students' beliefs about autonomy using the feedback of different ASR systems. He found that some ESL students did not even want to experiment

with Dragon Dictation, and three students did not even attempt to use the system because they could not get the ASR to recognize their voices properly (2016). Overall, the students felt that "Dragon Dictation was too [inaccurate] for the program to be useful for pronunciation practice" (p. 31). In short, Dragon may be a suitable program for native English speakers to use, however, as McCrocklin points out, even when the Dragon's ASR is trained to a single speaker's voice, it lacks the acoustic models that would make it beneficial for EFL or ESL students.

Elicited imitation tests

The present investigation was aimed at assessing whether it may be possible to use an ASR system to score learner performance on an elicited imitation (EI) test. EI tests can be used to measure L2 learners' spoken grammatical ability (Purpura, 2004). The simplest format for an EI test is one in which the learner hears a sentence and then imitates the sentence, with the response recorded onto tape or computer (Bley-Vroman and Chaudron, 1994). Although the learner may have little difficulty imitating simple sentences perfectly, when the length of test sentences is increased, the load on working memory increases and the learner may begin to have difficulty reproducing certain parts of a sentence. The EI test is taken to be a measure of how well the grammatical structures contained in the sentence have been automatized as part of the learner's interlanguage system. It is believed that the ability to chunk information (Abney, 1991) allows a limited capacity working memory to cope with the demands of reproducing sentences of greater length, and that chunking into larger units at the phrasal and clausal level is what enables more expert speakers to produce longer and more complex utterances even under the demands of on-going interaction (DeKeyser, 2001). When an L2 learner has difficulty reproducing a grammatical feature contained in a stimulus sentence this is believed to be due to the feature still not being fully automatized as part of the learner's interlanguage knowledge. The EI procedure allows particular grammatical features to be elicited and tested and also allows for productive grammatical ability, as opposed to receptive grammatical ability, to be assessed. The tests are therefore an attractive way of measuring a learner's mastery over particular grammatical features.

It has been argued that by manipulating certain aspects of the test, it is possible to produce EI instruments that can measure underlying implicit as well as explicit knowledge of grammatical features (Erlam, 2006). For example, by introducing an intervening step between stimulus sentence and imitation it is possible to encourage the learner to focus on meaning rather than form. If the stimulus sentence is a question, the learner can be asked to provide a simple answer before imitating the question. Or if the stimulus sentence is a contentious statement, the learner can be asked to say "True" or "False" before imitating the sentence. Thus, the learner's attention can be diverted away from the accuracy of the form and performance can be claimed to be more likely based on implicit knowledge than explicit knowledge. This claim is further strengthened if a time limit is set for the response so that the learner is not given a chance to reflect on explicit knowledge before responding (Ellis, 2005). It is also possible to use pictures to make the context of the stimulus sentence clear and to link test items together thematically so as to induce a focus on meaning and away from form.

One major problem with EI tests is that scoring is extremely labor intensive. It can take nearly as long to score one test manually as it does for one individual to take the test,

making it impossible for a teacher to provide feedback to an entire class in a timely manner. If the results are to be used in a research project, this may not be such a problem, but when the results are to be used diagnostically in an on-going program, for example, or if tests are to be used in class as pedagogical tools in themselves, the scoring issue becomes a major hurdle to the usefulness of EI tests. If it were possible to replace manual scoring with automated scoring, EI tests could be used much more widely because immediate feedback would become a reality.

Research into automated scoring of EI tests

Interest in automated scoring of EI tests has grown with the development of the Sphinx ASR system (e.g. Christensen, Hendrickson, & Lonsdale, 2010; De Wet, Muller, Van der Walt, & Niesler, 2011; Lonsdale & Christiansen, 2011). Graham, Lonsdale, Kennington, Johnson, and McGhee (2008), for example, developed a computer-based EI test and subsequently attempted to score test performance by L2 learners from a variety of language backgrounds using Sphinx. They describe several reasons why they thought it might be difficult to apply ASR to scoring EI tests. Firstly, at that time, ASR was still an emerging technology and recognition accuracy – even of native speaker input – was variable. Secondly, using the Sphinx ASR system to automate scoring required integrating complex systems that were hard for non-computer specialists to manipulate. And finally, the speakers who took their EI tests were non-native speakers sometimes with heavy accents which an ASR system designed for use with native speakers might find difficult to recognize. However, Graham et al. (2008) also pointed out several reasons for optimism. One was that with an EI test the expected input is already known, so the ASR task is far more constrained than for systems designed to deal with unpredictable input of any kind. Over several trials, Graham et al. were finally able to score non-native EI data using the Sphinx open source ASR tool achieving good correlations with human scoring.

Nowadays with the availability of tools to develop ASR applications such as the Google Web Speech API, the possibility of automatically scoring tests for EFL has come a step closer as developers can access the open source code. Sphinx is a complex system to set up whereas the Google Web Speech API is already in use and is growing in power all the time. By using this API and its transcription function, and by matching this transcribed output string with the original string of the EI item stimulus, we believed that we could begin to develop a system that automatically scored EI tests. As an initial step, however, we needed to investigate how well the Google Web Speech API was at recognizing both native speaker and non-native speaker input.

Research questions

1. How accurate is the Google Web Speech API in recognizing and transcribing English native speakers' oral production? – Which words does the Google Web Speech API have difficulty recognizing?
2. How accurate is the Google Web Speech API in recognizing and transcribing Japanese L2 English learners' oral production? – Which words does the Google Web Speech API have difficulty recognizing?
3. To what extent does the learners' pronunciation affect the system's ability to recognize and transcribe words?

The EI test

The 13 EI test items used in this study formed the first section of a 39-item EI test designed to elicit performance on 13 grammatical features (possessive -s, plural -s, 3rd person -s, articles, question tags, comparative adjectives, relative clauses, conditionals, modal verbs, relative adverbs, verb complements, since/for, and direct/indirect objects) and four tenses (simple present, simple past, present perfect, and present perfect continuous). Multiple instances of each feature appeared in the test. Items ranged in length between 4 and 16 syllables. The 13 items in each section were arranged in increasing length order so that the first items would be relatively easy to imitate and later items would become progressively more difficult to imitate. Each item was in the form of a question and was accompanied by a slide displayed on the computer to help provide context. It is hoped that this test can eventually be used to assess grammatical ability under varying conditions (timed/untimed; requiring an intervening answer or not), but in this study, we were testing to see how well our Google Web Speech API-based system could recognize native and non-native speaker input using just the items in the first section of the test.

The custom ASR system for EI

The conceptualization of the custom ASR system used in this research project, originated in a prior study conducted by the content specialist (Author1), who aspired to digitize EI tests to help automate the process of assessing students' grammatical ability in future studies. Although Google has now discontinued the use of external queries, the initial program was designed by the technology specialist (Author2), who used Linux shell scripting to send .flac formatted audio files to Google's server which were transcribed and returned back as simple text strings. After analyzing multiple audio file transcriptions of sample EI inputs (varying in syllable length) it appeared that Google's ASR system was accurate enough to warrant further development. Therefore, an initial plan for a customized Google Web Speech API program was drawn up (Figure 1) that would be capable of presenting EI test items and analyzing L2 learners' oral reproductions of EI items.

Although the Google Cloud Platform (included in Google Speech alpha which can run on any operating system) was released in March, 1st 2016 (Google, 2016), the design of the system used in Figure 1 was built on Google Web Speech API beta which required the use of the Google Chrome browser to operate. The object of the system was very straightforward: register users, show instructions, display prompts, capture inputs, compare text strings, and process/export results (Figure 2). The system used .mp3 files for the audio, .png files for the images, and .csv files for the data which were imported into the system for the prompts using JavaScript and PHP (Figure 1). Nevertheless, access to the Google Web Speech API using HTML5 was discontinued at the time of creating, so the technology specialist used JavaScript, PHP and AJAX to import and export external files instead.

Initially, each student heard a prompt (EI test item), clicked the record button, and then imitated it. After they spoke into the microphone, they would push the record button again, the Google ASR would capture the students' input, send it to Google's sever for decoding, and finally returned a text string of the closest match to each student's utterances. The resulting string was analysed against the original prompt string imported from the .csv file to see if the sentences perfectly matched. However, if the strings did not match, the system would report which words were regarded as missing, as well the student's accuracy

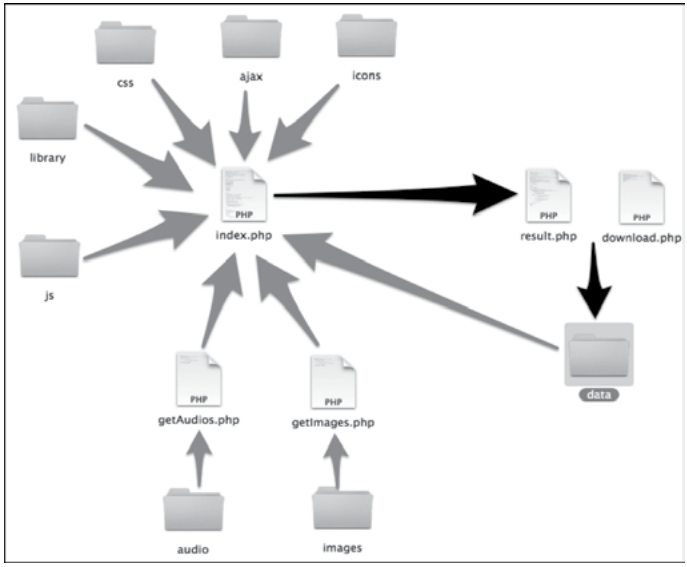


Figure 1. Custom ASR design using Google Web Speech API

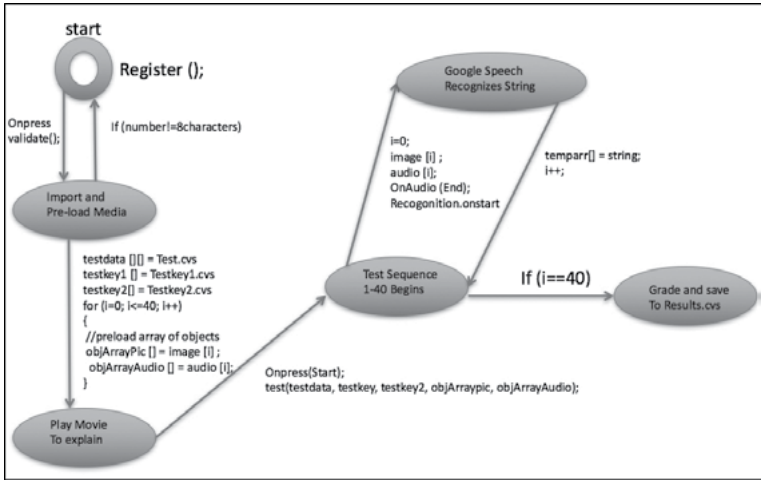
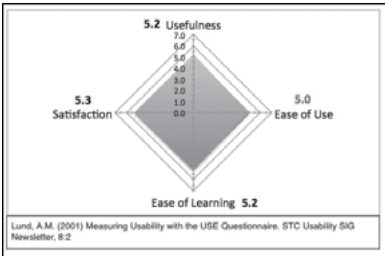


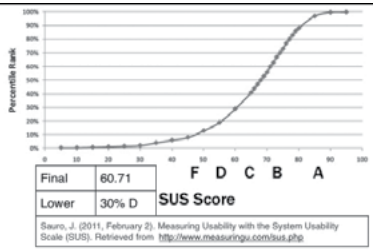
Figure 2. Custom ASR flow diagram using Google Web Speech API

in terms of the number of matching words/total words in the prompt. All of the data was processed using an algorithm in the results.php (Figure 1), which exported and amended the .csv file into the data folder.

Pilot study

A pilot research project was conducted during the spring semester of 2015 at Meiji Gakuin University in a TOEFL IBT preparation course for determining the usability, functionality, and limitations of the custom ASR system. In a controlled environment using microphones in a CALL laboratory located on the campus, seven participants were asked to take three different, five minute tests: single words, phrases and sentences. Additionally, the overall accuracy of the system was tested to see if the students' pronunciation affected the system's ability to transcribe their inputs. To achieve this, each input was compared word-by-word against students' actual audio recordings to determine the accuracy of the custom ASR system; initial findings suggested an accuracy rate of nearly 70%.





Student Positive Qualitative Feedback (n=7)
I could know my weak points which are difficult to know by myself
I think my English skill is improved before
It is easy to use
Able to do it when I only have a little time
It takes my voice correctly

Student Negative Qualitative Feedback (n=7)
Pushing "allow" every was time consuming
Sometimes I can't listen to the speaking
There should be more conversational tests or questions
More long sentences
The page should be more fashionable

Figure 3. Usability results from pilot research

After using the ASR system in the pilot study, the students were asked to take two different surveys to identify any difficulties (Figure 3). Through both surveys, it was obvious that the system was not as easy to use as initially thought. Responses to the Usability Questionnaire (USE) showed that the students recognized inconsistencies in the design, and they also noted that the system would not be useable without directions. As can be seen in the System Usability Scale (SUS) in Figure 3, the system used in the pilot study was in the lower 30th percentile of acceptability. Furthermore, qualitative feedback from the users showed there were some other negative aspects of the system that made it difficult for them to utilize it properly. All of the feedback was used to redesign the system for a better user experience, ensuring that the custom ASR system usability had no influence on the students' input (see Figure 4).

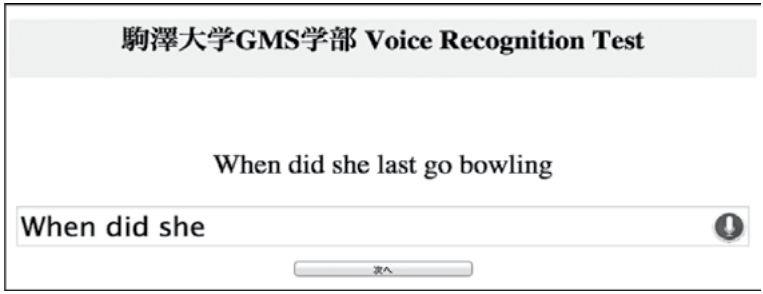


Figure 4. Custom ASR user interface

Procedure

Ultimately, we would like to use a custom ASR system to score an EI test, but first we needed to check to see how accurate the Google Web Speech API-based system designed for this study was at recognizing, transcribing and scoring input. We reasoned that, if the custom ASR system did not work well even with native speaker input, we were unlikely to have much success with non-native speaker input. However, if the system performed reasonably well with native speaker input, we could then go on to see how it performed with non-native speaker input. We assumed that getting one American English (AE) and one British English (GB) speaker to repeat 13 EI items 40 times each would give us a large enough sample to work with for native speaker input; we did not feel it necessary to find 40 AE and 40 GB native speakers for this purpose. However, we did choose to see how setting the system to expect four different varieties of English might affect recognition for native speaker input. Due to time constraints, we did not feel that we could impose on the students who participated in the study more than to have them record each item once in a language lab at the end of a regular lesson. As we were not intending to make a principled, detailed comparison between accuracy rates for native and non-native speakers, we did not feel the need to have identical conditions for the two groups. The next step in the process will be to compare automated and manual scoring accuracy of non-native speaker input. For now, we were concerned to see whether the system recognized native and non-native input accurately at all and to see what features of the input might affect recognition.

Results of the present investigation

Results pertaining to RQ1: NS input

Each of the 13 EI items in Trial 1 was recorded 40 times by a British English (BE) native speaker and 40 times by an American English (AE) native speaker. The custom ASR system was set to expect American English (US) for the first 10 recordings, Australian English (AU) for the next 10, British English (GB) for the next 10, and Canadian English (CA) for the final 10 for each speaker. The custom ASR system scanned the transcribed output and gave an accuracy score for each utterance based on the number of words the system recognized divided by the total number of words in the item. Thus, for example with Item 1, "Who is

taller", the system would give a score of 100% if all three words were fully transcribed, and a score of 66.7% if, for example, the word "taller" was not transcribed. Table 1 shows the mean accuracy scores for 20 recordings of each item (10 by the BE speaker and 10 by the AE speaker) with the system set to expect input in the four varieties of English, and an overall accuracy score based on all 80 recordings for each item for the two speakers.

Table 1. Mean accuracy scores (%) for NS input in four input modes and overall

Item no.	Item	US	AU	GB	CA	Overall	Below 90%
1	Who is taller	96.7	96.7	96.7	91.7	95.4	
2	Does Sue like swimming	85.0	87.5	82.5	91.3	86.6	*
3	When did she last go bowling	99.2	95.0	97.5	97.5	97.3	
4	Does Robert enjoy reading	100.0	97.5	100.0	98.8	99.1	
5	Mike and Sue have a bike do they	66.9	78.1	76.3	77.5	74.7	*
6	How old was Tom when he started work	98.8	100.0	99.4	99.4	99.4	
7	This week has she done any washing	93.6	97.1	99.3	97.9	97.0	
8	Is this the first coat she ever bought	73.1	75.0	83.8	81.9	78.5	*
9	Would Emma study abroad if she could	67.1	88.6	87.9	80.7	81.1	*
10	After dinner did they both wash the dishes	70.6	73.8	83.1	87.5	78.7	*
11	Is there a reason why he does not study	88.9	97.2	92.2	93.3	92.9	
12	If he practiced more would he be a better golfer	79.0	86.5	92.0	91.5	87.3	*
13	Does Alison like listening to her iPod when she goes to school	87.1	99.2	95.0	95.0	94.1	
	Overall	85.1	90.2	91.2	91.1	**89.4**	

The words in two of the items (4 and 6) were recognized and transcribed with an overall accuracy score of over 99% and five other items (1, 3, 7, 11 and 13) had mean accuracy scores of 90% or over. Item 12 was over 90% for the BE speaker, but below 90% for the AE native speaker. The remaining 5 items (2, 5, 8, 9 and 10) were not recognized and transcribed so effectively for either speaker, with the scores for Items 5, 8 and 10 being particularly low for the BE native speaker.

Looking at the 5 common problematic items in more detail, it was possible to identify particular words, collocations and word order issues that seem to have caused recognition difficulties. Item 2 contains four words and was recorded 80 times making a total of 320 words for the system to transcribe. Of these 320 words, 43 were not transcribed and were then regarded as being "missing" giving the overall mean accuracy score of 86.6%. Thirty-five of the missing words were "Sue" suggesting that the system had particular difficulty recognizing this proper noun. Recognition of the name "Emma" in Item 9 was also particularly poor. It was subsequently found that by replacing these proper nouns with "Tom" or "Robert", recognition accuracy improved dramatically. For the BE native speaker, recognition of "Sue" and "Emma" was particularly poor in the US English input mode, but recognition of "Sue" was also poor in the GB English mode, which was his own variety.

For Item 5, 86 of the 162 missing words were either "do" or "they", and "do they" was

overwhelmingly the most commonly missed collocation. This suggests that the system is not good at recognizing question tags. Subsequently, replacing "Mike and Sue" with "Tom and Robert" did not have any impact on recognition of "do they".

For Item 8, 128 of the 138 missing words were in the second half of the sentence. Subsequently, by replacing "coat" with "time" and "bought" with "sang", recognition accuracy improved dramatically not only for these two words but also for the intervening words "she" and "ever". This suggests that the highly frequent collocation "first time" helps the system to recognize the following words more effectively than a highly unusual collocation like "first coat".

In Item 9, "would" accounted for 46 of the 106 missing words and "Emma" 39. Afterwards, by replacing "Emma" with "Robert", the recognition of "would" improved dramatically suggesting that the ASR system is good at back-forming for initial auxiliary verbs when it recognizes the proper noun that appears in second place in the sentence.

Finally, for Item 10, "after", "did", "they" and "both" accounted for 115 of the 136 missing words and "did they both" was the only collocation found to be commonly missing. Subsequently, it was found that by moving "after dinner" to the end of the sentence, recognition of "after" and "did" improved noticeably.

Results pertaining to RQ2: NNS Input

Table 2 shows the mean accuracy scores for 44 university students recording the same 13 EI items as above. Two of the members from this student cohort were Chinese nationals. The other 42 were Japanese. Table 2 also shows the two words the SR system judged were missing most often.

Most obviously, and as might be expected, the mean accuracy scores for the NNS input are generally much lower than those for the NS input (65.7% overall accuracy). Only on Item 11 is there close parity between the NNS and NS mean scores. Naturally, one assumes that the discrepancy must be caused by pronunciation issues. It will be noted, however, that the two proper nouns, "Sue" and "Emma", which were problematic for the system with native speaker input were also among the words which were most problematic for the system with non-native speaker input.

Results pertaining to RQ3: NNS pronunciation

Table 3 shows which word was judged to be missing most often by the custom ASR system, to what extent these words were judged to be mispronounced by a rater (one of the authors), and the word in each item which was judged to be mispronounced most often. For example, for Item 1, "taller" was the word that the system judged to be missing most often. It represented 72% of all missing words. "Taller" also represented 82% of the words that the rater thought were mispronounced. Overall, it was the word judged to be most often mispronounced. In Item 2, "Sue" was the word the system failed to recognize most often and was regarded as missing. It accounted for 55% of all missing words. According to the rater, however, "Sue" only accounted for 19% of the mispronounced words. The word "swimming" was, according to the rater, the word most often mispronounced accounting for 51% of mispronounced words. A word was judged to be mispronounced if it was thought that someone not familiar with Japanese English would have difficulty catching the word, even in context.

Table 2. Mean accuracy scores for NNS input and most problematic words

Item no.	Item	NNS mean accuracy (US)	No. 1 missing word	As % of missing w.	No. 2 missing word	As % of missing w.
1	Who is taller?	67.4	taller	72	who/is*	14
2	Does Sue like swimming?	58.0	Sue	55	does/ swimming*	19
3	When did she last go bowling?	70.5	she	33	last	24
4	Does Robert enjoy reading?	53.4	Robert	32	enjoy/ reading*	22
5	Mike and Sue have a bike do they?	60.5	they	28	do	25
6	How old was Tom when he started work?	84.9	Tom	34	work	21
7	This week has she done any washing?	59.4	washing	20	she/done*	19
8	Is this the first coat she ever bought?	57.1	coat	22	bought	19
9	Would Emma study abroad if she could?	48.1	would	26	Emma	24
10	After dinner did they both wash the dishes?	51.1	both	17	they	16
11	Is there a reason why he does not study?	94.4	is/there/ not*	23	reason	14
12	If he practiced more would he be a better golfer?	68.9	golfer	21	practiced	18
13	Does Alison like listening to her iPod when she goes to school?	80.7	Alison	30	does	14
	Overall	65.7				

Note: * tied for same place

It is clear that the word which was most often judged to be mispronounced in each item was not always the word most often regarded as missing by the system. In six cases the word most often mispronounced coincided with the word recorded as most often missing, but in seven cases there was no correspondence.

Discussion

In answer to the first research question, it is possible from the results presented in Table 1 to conclude that, although the overall recognition accuracy of the system was 89.4%, the recognition part of the system had difficulty with particular words, collocations and word orders even when the input was from what we take to be typical BE and AE native speakers.

We are certain the system faithfully transcribed what was recognized and found no malfunction with the part of the custom ASR system which matched the transcription with the original EI Test item letter string, so we are led to believe that it was particular

Table 3. Words judged to be missing by the system and mispronounced by a rater

Item no.	Item	No. 1 missing word	As % of missing w.	As % of mis-pronounced words	Word most often mis-pronounced	As % of mis-pronounced words
1	Who is taller?	taller	72	82	taller	-
2	Does Sue like swimming?	Sue	55	19	swimming	51
3	When did she last go bowling?	she	33	17	bowling	34
4	Does Robert enjoy reading?	Robert	32	53	Robert	-
5	Mike and Sue have a bike do they?	they	28	35	they	-
6	How old was Tom when he started work?	Tom	34	0	old	42
7	This week has she done any washing?	washing	20	25	this	37
8	Is this the first coat she ever bought?	coat	22	28	this/coat	28
9	Would Emma study abroad if she could?	would	26	45	would	-
10	After dinner did they both wash the dishes?	both	17	14	dishes	23
11	Is there a reason why he does not study?	is/there/ not	23	6/36/4	there	-
12	If he practiced more would he be a better golfer?	golfer	21	25	golfer	-
13	Does Alison like listening to her iPod when she goes to school?	Alison	30	12	school	18

features that caused the system difficulty. Certain proper nouns such as "Sue" and "Emma", certain collocations such as "first coat", certain structures such as question tags ("do they" at the end of the question), and certain word orders such as starting a question with an adverbial clause before the question word ("After dinner did they…") caused problems for the system even with native speaker input. Setting the input mode to one of four different varieties of native speaker English did not appear to strongly affect recognition accuracy except in the case of certain proper nouns.

In answer to the second research question, we know from the evidence presented in Table 2 that the Google Web Speech API incorporated into the custom ASR system judged these Japanese L2 English learners' oral production to be 65.7% accurate overall. Although direct comparison is unwise due to the different conditions set for native and non-native speakers, this accuracy rate is much lower than the overall accuracy assigned to NS input. It was noticeable, however, that some of the words, collocations, and word orders that seemed to cause recognition problems with NS input were also responsible for the low

accuracy score given for NNS input. Some of the same proper nouns ("Sue" and "Emma"), the same question tag ending ("do they"), and the same unusual word order putting an adverbial phrase in front of the question word ("After dinner did they both…") were partly responsible for poor recognition of NNS input just as they were for NS input. It was also noticeable that for items which avoid these particular features and follow a more canonical word order, recognition accuracy can climb to be on a par with that for NS input.

By investigating the third research question, we found that while the learners' pronunciation obviously does affect the system's ability to recognize and transcribe words, there are often other factors affecting recognition that outweigh non-target-like pronunciation. In Table 3, it is clear that certain pronunciations that are typically difficult for Japanese speakers caused the recognition problem. The medial lateral in "taller" is difficult and means the word is often pronounced more like "tora". This is similar to a typical problem with the word "golfer" so that it is pronounced "gorufaa", a problem exacerbated by the existence of this as a loanword in Japanese. The initial "r" in "Robert" and the second vowel sound mean it is often pronounced "lobaat". The initial "th" (/ð/) in "they" and "there" is often pronounced as "z" (/z/). And the initial glide in "would" is often approximated with "oo", so that "would" becomes "ood". These undoubtedly caused many of the recognition difficulties for the system. However, it is also clear from the data in Table 3 that NNS pronunciation was not the principle problem in around half the items. Words judged to be mispronounced most often were not always the ones the system failed to recognize most often. For example, in Item 6, the word "Tom" was the word the system did not recognize most often but it was judged to be pronounced perfectly well by all 44 NNS. Yet the word "old", which was judged to be mispronounced most often, was recognized by the system on 41 out of 44 occasions.

The conclusion we draw from this is that because the system does not rely solely on acoustic information but is also drawing on predictive syntax algorithms (SLM) and vast amounts of stored collocational data (HHM), it tends to overlook local pronunciation problems when the input conforms to canonical order and typical sentence patterns due to the machine learning algorithms in play.

Recommendations and next steps

One recommendation we would like to make to people considering using the Google Web Speech API for L2learning and testing purposes is that, even for pronunciation practice, it is advisable to check that the items learners are being asked to imitate are not difficult for the system to recognize. This may sound rather counter-intuitive. One might expect the system to be rather good at showing learners what they need to work on, especially in terms of pronunciation. To some extent this is true, but, as we have shown from the native speaker screening process, some items are not recognized well because they contain certain problematic words, collocations or word orders. It would therefore seem inadvisable to use these "faulty" items with L2 learners because they will be either disheartened or marked down when they get the item "wrong". What we are advocating is adapting to the system's strengths to some extent by constraining the input in some ways so that learners are judged fairly according to their true abilities.

The second recommendation is that it makes sense to tailor the system to the learners who will use it, keeping in mind the ultimate purpose. In our case, we are hoping to develop a system that will allow us to automatically score an EI test for Japanese learners of English to provide information on learners' productive grammatical ability. It therefore follows that

we should screen out features in items which cause the system to fail to recognize input even though it may be grammatically correct. For people who want to use the Google Web Speech API for pronunciation practice and training, this may not apply, but for our purposes (EI tests) we need items that will allow learners to display their grammatical ability and allow the system to recognize what is being said. Thus, we intend to constrain the input even further by developing items that avoid the word, collocation, and word order issues and the typical pronunciation pitfalls for Japanese speakers identified above. Using these items, we intend to test whether the system can score target grammatical features as well as a human scorer. We are also hoping to add new features to the system to make it more flexible and adaptable for use by other teachers and researchers.

Conclusion

This investigation has alerted us to some of the problems with using the Google Web Speech API to score EI test responses. By working with the strengths of the application we believe we will be able to develop a system that can reliably and efficiently score Japanese L2 learners' spoken production for grammatical accuracy. Because we are heavily constraining what the system needs to be able to recognize, the system does not need to work perfectly for any possible input. Through trial and error, we can find a set of items that are recognized with a high degree of accuracy with our learners. We can then test the system for how well it measures performance. At the same time, we can develop the ASR system so that it has greater functionality making it more useful to other researchers who may wish to employ it for other purposes.

References

Abney, S. P. (1991). Parsing by chunks. In R. C. Berwick, S. P. Abney, & C. Tenny (Eds.), *Principle-based parsing: Computation and psycholinguistics* (pp. 257–278). Dordrecht, NL: Kluwer Academic Publishers. http://doi.org/10.1007/978-94-011-3474-3_10

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 245–261). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chiu, T.-L., Liou, H.-C., & Yeh, Y. (2007). A study of web-based oral activities enhanced by automatic speech recognition for EFL college learning. *Computer Assisted Language Learning*, *20*(3), 209–233. http://doi.org/10.1080/09588220701489374

Christensen, C., Hendrickson, R., & Lonsdale, D. (2010). Principled construction of elicited imitation tests. *Proceedings of the 7th Conference on International Language Resources and Evaluation* (LREC'*10*), 233–238.

Daniels, P. (2015). Using web speech technology with language learning applications. *The JALT CALL Journal*, *11*(2), 177–187.

De Vries, B. P., Cucchiarini, C., Bodnar, S., Strik, H., & Van Hout, R. (2015). Spoken grammar practice and feedback in an ASR-based CALL system. *Computer Assisted Language Learning*, *28*(6), 550–576. http://doi.org/10.1080/09588221.2014.889713

De Wet, F., Muller, P., Van der Walt, C., & Niesler, T. (2011). Readability index as a design criterion for elicited imitation tasks in automatic oral proficiency assessment. In *ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2011)* (pp. 24–26). Venice, Italy.

DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). New York: Cambridge University Press.

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.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*(2), 141–172. http://doi.org/10.1017/S0272263105050096

Ellis, R. (2012). *Language teaching research & language pedagogy*. Malden, MA: Wiley-Blackwell.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*(3), 464–491. http://doi.org/10.1093/applin/aml001

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105. http://doi.org/10.1080/09588221.2012.700315

Google. (2016). Google Cloud Platform. Retrieved June 7, 2016, from https://cloud.google.com/speech/

Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 1604–1610. Retrieved from http://repository.dlsi.ua.es/242/1/pdf/409_paper.pdf

Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition – A brief history of the technology development. In *Elsevier Encyclopedia of Language and Linguistics* (2nd ed., pp. 806–819). Elsevier Science. http://doi.org/10.1016/B0-08-044854-2/00906-8

Kim, I. S. (2006). Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation. *Educational Technology and Society*, *9*(1), 322–334.

Lear, E. (2013). Using technology to improve society. *New Zealand Studies in Applied Linguistics*, *19*(1), 49–63.

Lonsdale, D., & Christiansen, C. (2011). Automating the scoring of Elicited Imitation Tests. *Symposium on Machine Learning in Speech and Language Processing*.

Luo, B. (2016). Evaluating a computer-assisted pronunciation training (CAPT) technique for efficient classroom instruction. *Computer Assisted Language Learning*, *29*(3), 451–476. http://doi.org/10.1080/09588221.2014.963123

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake; Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, *19*(1), 37–66. Retrieved from http://journals.cambridge.org/abstract_S0272263197001034

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System*, *57*, 25–42. http://doi.org/10.1016/j.system.2015.12.013

Purpura, J. E. (2004). *Assessing grammar*. New York: Cambridge University Press.

Russell, V. (2009). Corrective feedback, over a decade of research since Lyster and Ranta (1997): Where do we stand today ? *Electronic Journal of Foreign Language Teaching*, *6*(1), 21–31. Retrieved from http://e-flt.nus.edu.sg/

Toyama, M. (2015). Japanese EFL learners' beliefs about pronunciation learning and their pronunciation skills. *Bunkyo University Journal of Language and Culture*, *May*, 92–114.

Twiefel, J., Baumann, T., Heinrich, S., & Wermter, S. (2014). Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing. In *Proceedings of the 28th* AAAI *Conference on Artificial Intelligence* (AAAI-*14)* (pp. 1–7).

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P, & Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition. *Smli*, (TR-2004-139), 1–9. Retrieved from http://egouvea.users.sourceforge.net/paper/smli_tr-2004-139.pdf

Wang, Y. H., & Young, S. S. C. (2014). A study of the design and implementation of the ASR-based iCASL system with corrective feedback to facilitate English learning. *Educational Technology and Society*, *17*(2), 219–233.

Author biodata

Tim Ashwell taught in the UK and Thailand before becoming a full-time professor at Komazawa University in the Department of Global Media Studies. His interests are in grammar instruction, collaborative learning, and learner and teacher development.

Jesse R. Elam is a full-time lecturer at Tokyo Denki University and a doctoral candidate at the University of South Carolina focusing on curriculum and instruction with a cognate in TESOL. His research is related to the use of eLearning and instructional design to extend EFL classrooms. He also has a strong interest in the application of constructivism with educational technology to create blended- and flipped classrooms.