A Prototype System with Speech Recognition Function for Practicing Speaking English

Harumi Kashiwagi, Kobe University, Japan
Min Kang, Kobe University, Japan
Kazuhiro Ohtsuki, Kobe University, Japan

The Asian Conference on Education 2020
Official Conference Proceedings

Abstract
The literature indicates that many Japanese students enrolled in English language classes do not use English outside of class. Therefore, we assume that speech recognition technology can help create opportunities to speak English. Although some speech recognition software is widely available, few studies have used software with a speech recognition function to investigate how we should adapt this technology to foreign language learners with various proficiency levels, including Japanese students. The authors, therefore, developed a prototype system with a speech recognition function to create opportunities for Japanese students practicing English. We tested the system in a pilot study with 17 Japanese university students. During the test, students were asked to use the correct English word for an image displayed by the system. Students' responses to the system were collected via a survey questionnaire. The pilot test indicated that most words were recognized accurately, and the students' speech was correctly recognized by the testing system to a large extent. In addition, 88% of the students expressed a positive attitude toward the system. These results suggest that speech recognition functions create opportunities for students to practice their English. They also suggest that we should consider the balance between a software's recognition rate and students' motivation for practicing English when using speech recognition software for language instruction, since students might be less confident if their pronunciation is repeatedly found to be incorrect.

Keywords: Language Learning System, Speech Recognition, Computer-Generated Characters, Practice Speaking in a Foreign Language

Introduction

In 2014, Japan's Ministry of Education, Culture, Sports, Science and Technology published its English Education Reform Plan. This plan placed a priority on improving Japanese students' oral English proficiency. However, many students in Japan lack the opportunity to use English outside of language classes and lack confidence in speaking English (Kashiwagi, Kang, & Ohtsuki, 2017). Thus, it is important that these students have a welcoming environment in which they can practice speaking English. In this paper, we suggest that speech recognition software can help create this environment for students who lack English speaking opportunities outside of class. Although some speech recognition software, such as Dragon Naturally Speaking, is widely available, it is not yet sophisticated enough to recognize the English speech of all levels of English language learners (Chapelle & Voss, 2016). According to Blake (2016), effectively utilizing this technology's limited range of operations, such as word recognition or short sentence repetition, is a desirable goal. By using the speech recognition function in this way, learners can identify which words or phrases deviate from the standard pronunciation that the engine recognizes and can adjust their speech accordingly. Unfortunately, few studies have used software with a speech recognition function to investigate how to adapt this technology to foreign language learners with various proficiency levels, including Japanese students.

To meet this challenge, we developed a prototype system with a speech recognition function to examine how to best adapt this technology for Japanese students practicing English. We conducted a pilot study with 17 Japanese university students and solicited their opinions of our prototype through a questionnaire. We aimed to investigate the following research questions:

1. How accurately does the prototype system recognize students' speech?
2. How do the students respond to the prototype system?
3. Are student's responses to the prototype system related to their nervousness when speaking English, or their negative attitudes toward English pronunciation?

Below, we first describe the prototype system, then lay out the experiment, deliver its results, and discuss its findings. We conclude by reflecting on the potential of our prototype.

The Prototype System

In our prototype system, SpeechRecog, recognizes speech inputs by referring to a registered dictionary. SpeechRecog uses Microsoft Windows' speech recognition function (see Windows Support Home Page, n.d.) to recognize students' pronunciation, following the process explained below. The system's dictionary is contained in a text file so that registered words and sentences can be edited with ease. When a word is recognized by the prototype system, it provides feedback by repeating the recognized word (or sentence) and a score from 0 to 1. This score indicates the degree to which the speech input matches the registered dictionary. This feedback comes with one of three letters: "H" indicates that the input has been recognized as an accurately pronounced registered word or sentence, "M" indicates that the input speech is hypothesized or assumed to be a registered word or sentence, and "L" indicates that the input speech has been detected but it is not clear whether the input

speech corresponds to a registered word or sentence. An example of such feedback would be: "[H] sea bream (0.9426834)."

Structure of the Prototype System

The structure of the prototype system is outlined in Figure 1. It consists of three different types of software: SpeechRecog, MINI BASIC, and AnimeViewer. SpeechRecog (Figure 3) initiates and concludes the recognition of input speech and receives and saves the results of the Windows' speech recognition function. SpeechRecog sends only "H" level results to MINI BASIC.

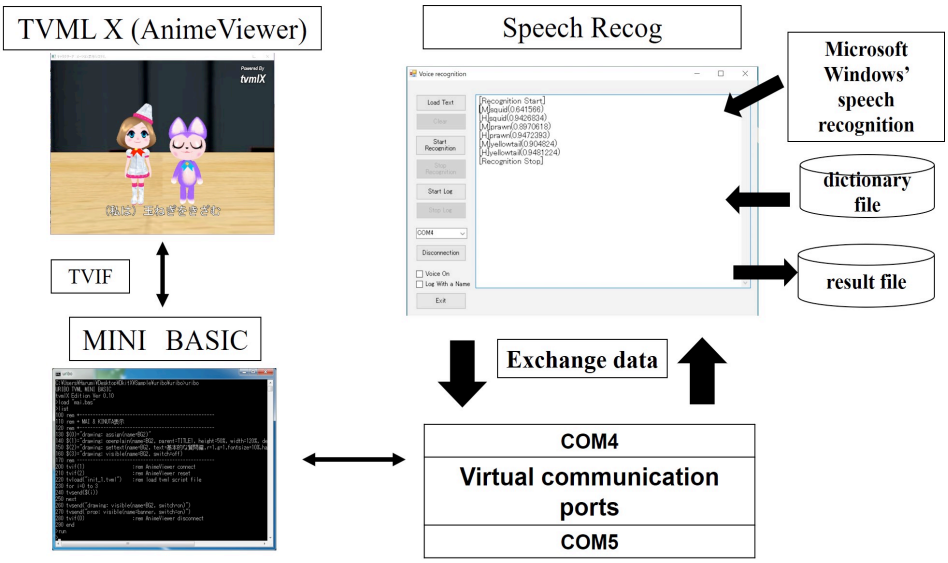

Figure 1: System overview.

MINI BASIC (Kashiwagi, Kang, & Ohtsuki, 2020) creates questions by automatically obtaining question text data and inputting this data into a TV Program Making Language (hereinafter called TVML, TVML home page. n.d.) file template. TVML is a text-based scripting language that automatically generates television programs (Hayashi, 1999). This file is sent to AnimeViewer, which is a viewer tool that displays computer-generated (CG) content from TVML scripts. AnimeViewer is prepared in TVML Player X, which produces CG content. MINI BASIC also checks whether the recognized data from SpeechRecog match the correct answer and creates feedback messages in a TVML file.

Experiment

Participants

This study had 17 participants, all students at a university in Japan, of whom 12 were second-year students, 2 were third-year students, and 3 were fourth-year students. Participants were asked to provide an English word to accompany a picture displayed by the testing system (see below) and answered a questionnaire after completing the experiment.

Procedures

First, we provided study participants with 10 English words (shown in Table 1, below) and instructed them to focus on remembering these words for 30 minutes. The words were the names of fish and shellfish. Next, we asked participants to provide the appropriate English word for an image displayed by the testing system. The system then recognized their answers, checked whether their answers were correct, and delivered feedback. The post-experiment questionnaire (Table 2) had three items, each scored on a 5-point Likert-type scale: strongly agree, moderately agree, neutral, moderately disagree, or strongly disagree. They were also encouraged to give unstructured, longer-running feedback about the experiment and the testing system.

The experiment had five steps detailed below to provide an idea of how our system might be applied in a classroom setting.

1. The experimenter boots up the MINI BASIC software, and a CG character appears on the AnimeViewer screen, gives instructions, and displays an image for a question (Figure 2).

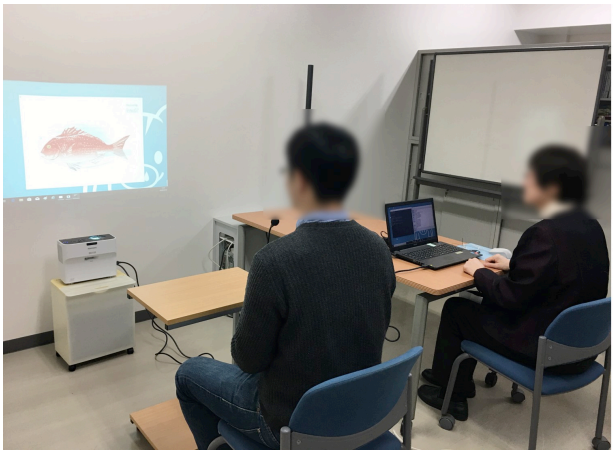

Figure 2: Experiment scene using the prototype system.

2. SpeechRecog software is booted up and the dictionary file is read (Figure 3). The experimenter clicks Start Log to initiate recording data and Start Recognition to initiate speech recognition.

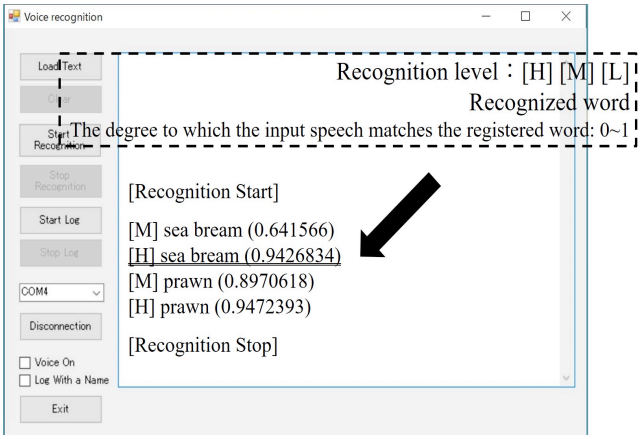

Figure 3: Example screen from SpeechRecog.

3. The student orally answers the question.
4. The speech recognition results appear in SpeechRecog (Figure 3).
5. SpeechRecog sends only the "H" level results to MINI BASIC to check whether the recognized data match the correct answer. If so, a TVML file is sent to AnimeViewer and a CG character appears to tell the user that their answer is correct. If not, the same happens but the CG character says "Once again, please" to tell the user that their answer was incorrect.
6. At the end of the experiment, the experimenter clicks Stop Recognition to stop the software's speech recognition activity and Stop Log to stop recording data. The results are then saved.

Results and Discussion

RQ1: How accurately does the prototype system recognize students' speech?

We determined the prototype system's accuracy by calculating the percentage of participants' spoken words that were recognized as correct answers at the "H" level (Table 1). Every participant's use of the words "prawn," "sea bream," "mackerel," "yellowtail," and "saury" was recognized accurately by our prototype, and 16 participants' (94%) uses of the words "globefish" and "eel" were recognized accurately, while 13 of the 17 participants' (76%) uses of the words "squid," "crab," and "turban shell" were recognized accurately. These results indicate that most of the words were recognized accurately, and the prototype system usually reflected students' speech correctly.

Table 1: System recognition of participants' use of key words.

English words	Percentage of spoken words recognized as correct at the "H" level
prawn, sea bream, mackerel, yellowtail, saury	100%
globefish, eel	94%
squid, crab, turban shell	76%

RQ2: How do students respond to the prototype system?

Participants' responses to the post-experiment questionnaire are listed in Table 3. The responses to Q1 indicate that 15 of the 17 participants (88%) agreed strongly or moderately that they had a positive attitude toward the prototype system. Participants gave some positive feedback, including: "The speech recognition software is more accurate than I expected, and it might be useful for confirming our English pronunciation," and "This system might be useful when we are more familiar with speaking English."

Two outlying responses were neutral, though the participants who gave those responses also commented that the prototype system's "use of multimedia helps us remember new words" and that "With this system, we will be able to practice answering the appropriate English word reflectively."

There were some negative comments as well, most concerning the fact that the prototype system only tells students when, not where or how, they made errors in

pronunciation. This feedback suggests that speech recognition software does not necessarily improve students' English pronunciation, even though it gives them the opportunity to practice.

Table 2: Questionnaire items.

No	Questionnaire items
Q1	I have a positive view of this software and system.
Q2	I feel nervous when I use English during face-to-face communication.
Q3	I am not good at English pronunciation.

Table 3: Questionnaire results (n = 17).

	Strongly Agree	Moderately Agree	Neutral	Moderately Disagree	Strongly Disagree
Q1	7	8	2	0	0
Q2	6	9	0	1	1
Q3	5	5	4	2	1

In this study, we have not conclusively determined how we can effectively introduce speech recognition into language learning activities. However, this type of practice with pictures and a speech recognition function can serve for oral practice with multimedia information that is not character-based and will provide students chances to say what they want to say in English quickly. One participant commented that this prototype system would be "useful for preschool and elementary school children" due to its ease of use and the way in which feedback is delivered.

RQ3: Are students' responses to the prototype system related to nervousness when speaking English or negative attitudes toward English pronunciation?

We also used the post-experiment questionnaires to examine whether participants' responses to our prototype were related to their nervousness in using English or their negative attitudes toward English pronunciation. To analyze the relationships between variables, we calculated Spearman's rank-order correlation coefficients on the data from the questionnaire shown in Table 3.

The results for the correlation coefficients between Q1 and Q2 ($r_{Q1Q2} = -0.26$) show no significant relationship between participants' responses to our prototype and their nervousness using English. Furthermore, the results for the correlation coefficients between Q1 and Q3 ($r_{Q1Q3} = 0.05$) show that there was no significant relationship between participants' responses to our prototype and their negative attitudes toward English pronunciation.

However, the results for the correlation coefficients between Q1 and Q2 ($r_{Q1Q2} = -0.26$) indicated that participants who feel nervous while using English in face-to-face communication might hold negative attitudes toward practicing their vocabulary with speech recognition software and that they might be less confident if their pronunciation is repeatedly found to be incorrect. This suggests that the software's recognition rate and students' motivation for practicing English must be considered when using speech recognition software for language instruction.

Limitations and Recommendations

Certain limitations of the current study should be mentioned. The primary aim of this study was to verify the operation of the prototype system's speech recognition function in a pilot study. More words and more participants would need to be tested before relying fully on this prototype system for English speaking practice. Moreover, it appears that students feel less confident if their speech is repeatedly recognized as incorrect. Thus, using this system's speech recognition function could actually decrease some students' English speaking if they shy away from using it in order to avoid negative feelings associated with doing something incorrectly. For future studies, we hope to further examine how we can incorporate speech recognition functions into language learning systems, considering both the recognition rate and the students' motivation to practice English.

Conclusion

We developed a prototype system that used speech recognition software to create opportunities for students to speak English. The results of our pilot study indicated that most speech was recognized accurately, and that 88% of participants viewed the prototype system positively. Our results also indicate that participants' responses to the system's use of speech recognition software was not related to their nervousness using English or their negative attitudes toward English pronunciation. It did, however, find that participants who felt nervous using English in face-to-face communication might have negative attitudes toward the use of speech recognition software, especially if their speech is repeatedly found to be incorrect. Together, our results suggest that the software's recognition rate and students' motivation for practicing English must be considered when speech recognition software is used for language instruction.

This study has several limitations. For instance, we had only 17 participants and provided them with 10 keywords. Future studies could examine more words and more participants. Future studies could also examine the utility of speech recognition in different cultural linguistic contexts to improve the generalizability and adaptability of systems such as ours.

Acknowledgement

This work was supported by JSPS Grant-in-Aid for Scientific Research Number JP18K02822.

References

Blake, R. (2016). Technology and the four skills. *Language Learning & Technology*, *20*(2), 129–142.

Chapelle, C. A. & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology*, *20*(2), 116–128.

Hayashi, M. (1999). TVML (TV program Making Language)—Automatic TV program generation from text-based script. *Proceedings of Imagina '99*, 119–133. doi:10.11485/itetr.23.4.0_1

Japan's Ministry of Education, Culture, Sports, Science and Technology. (2014). *English Education Reform Plan corresponding to globalization*. http://www.mext.go.jp/en/news/topics/detail/__icsFiles/afieldfile/2014/01/23/1343591_1.pdf

Kashiwagi, H., Kang, M., & Ohtsuki, K. (2017). A study on the affective factors of Japanese university students in language communication activities. *Proceedings of the 4th Symposium on Language and Sustainability in Asia*, 17–22.

Kashiwagi, H., Kang, M., & Ohtsuki, K. (2020). A study on a method for dynamic control and use of TVML contents. *Educational Technology Research*, 43(Suppl.), 9–12 (in Japanese).

TVML home page. (n.d.). Retrieved from http://www.nhk.or.jp/strl/tvml/index.html (in Japanese)

Contact email: kasiwagi@kobe-u.ac.jp