Developing and Evaluating an E-learning Material for Speaking Practice with the Latest AI Technology

Makoto Shishido, Tokyo Denki University, Japan

The IAFOR International Conference on Education – Hawaii 2021
Official Conference Proceedings

Abstract
Recent years have seen remarkable advancements in information and communication technology (ICT) and artificial intelligence (AI), and many e-learning materials that use ICT and AI have become available on the market. However, none of them has been proven effective at improving speaking skills in English as a Second Language (ESL) and English as a Foreign Language (EFL) classrooms. The author's previous work suggested that speaking practice with voice recognition and an AI scoring system is an effective method because it allows students to practice speaking anytime and anywhere, without a partner. Furthermore, learners' psychological anxiety may be alleviated, eliminating the resistance to speaking English. With AI, learners do not have to feel embarrassed even if they make mistakes, and they can repeat the practice materials many times. This study describes a newly developed e-learning material that makes the best use of advanced AI technology. It includes activities such as shadowing, role-playing conversations, and speech delivery practice with automatic evaluation and scoring. Also reported are the results of comparisons between pre- and post-test scores on the Test of English for International Communication (TOEIC) and the Oral Proficiency Interview – computer (OPIc) test to evaluate the effectiveness of speaking practices and the improvement of oral skills with the proposed system. Finally, a questionnaire is used to examine learners' opinions on learning English with AI. Possible further developments of the e-learning system using AI technology include improving the accuracy of speech recognition for utterances by non-native English speakers, updates to reflect the latest evolution of AI, and the addition of a capability for free conversations instead of just set phrases.

Keywords: Voice Recognition, Artificial Intelligence, Shadowing, Pronunciation, TOEIC, Opic

1. Previous Studies

There has been much discussion about the problems in teaching English speaking skills in Japan. One issue is that many Japanese EFL teachers are not good at or do not have confidence in their own pronunciation, and as a result, speaking classes are often left to be taught by native speakers. However, those native speakers are sometimes not qualified to teach ESL or lack appropriate strategies for teaching non-native speakers.

Another problem is class size. Many EFL classes in Japan have about 30 to 40 students, which makes it difficult to teach speaking skills. Other problems are a lack of appropriate teaching materials and Japanese students' characteristics such as shyness and hesitance to speak in public.

Yet despite the various problems, there have been growing expectations for improving Japanese students' EFL speaking skills. For instance, Japan's Ministry of Education has recently emphasized a balance of the four skills of English, especially speaking. Globalization is another key issue, and, though it is doubtful, the Olympics and Paralympics might take place in Tokyo in 2021.

1.1. English Central

As a Japanese EFL teacher, I have tried various materials and methods for teaching speaking over the past few years. I describe some of my trials here. In 2017, I asked students to use the English Central website, which provides short video clips for listening practice. On the site, students repeat phrases from the video clips, and an AI technology called Intellispeech recognizes and evaluates their pronunciation. The results are shown in three colors. However, Intellispeech uses a somewhat primitive voice recognition system, and its evaluation is vague. Also, it only recognizes certain set phrases.

Alongside the use of English Central in 2017, I arranged for the students to practice speaking with Filipino teachers through Skype. This combined method seemed to be a better approach; however, the limited class budget allowed only for two Skype sessions per year (i.e., once a semester). The students needed more interaction with native English speakers.

1.2. Filipino Teachers

Another trial was done with a Filipino teacher (Sakamoto et al., 2016). This time, I organized one-to-one speaking practice through Skype. This trial was made possible by financial support from a company. Without the financial support, the students would have had to pay their own cost, which would have posed an extra burden for them and would not have been an ideal practice for a classroom setting.

Then, I had another opportunity to work with Filipino teachers (Sakamoto et al., 2017). This time, I received financial support and invited Filipino teachers to Japan. I organized small group lessons with them, and they used the Callan Method. The lessons were popular and received a good reputation. However, it was difficult to obtain approval for the lessons to be recognized as a formal class where students

could earn credits because the Filipino teachers were not approved by the ministry even though they held Teaching English as a Second Language (TESL) certificates in the Philippines.

2. The First Trial

2.1. Shikibu App

After the previous trials, I created my own e-learning materials with voice recognition and AI. I implemented the materials using three different platforms: Shikibu, SpeakBuddy, and the English Listening and Speaking Test (ELST). Below, I describe the three trials in which students used the materials on these platforms. There are several technologies related to voice recognition such as Siri, Google Assistant, and many AI speakers. Also, there advancements have been made in AI technologies like IBM's Watson.

The first trial was carried out using the Shikibu app, with support from NTT Learning Systems. NTT Learning Systems developed a speaking practice system using Apple's Siri and IBM's Watson. The system was designed for Japanese students of EFL to practice English utterances. It uses Siri for voice recognition and Watson for evaluation and giving responses. Since it uses Siri, it is available only on iPhones.

The study for the first trial was done with 26 students at my university, and all of them were computer science majors. The experiment was carried out over approximately a three-month period from July to September 2017, during their summer break. Students used the app to practice English on a self-study basis.

Fig. 1 Shikibu's Main Menu

Fig. 2 Lesson Menu on Shikibu

Fig. 3 Example of a Speaking Exercise on Shikibu

Fig. 1 shows the main menu of the Shikibu on the iPhone. Fig. 2 shows a lesson menu. All the exercises are designed based on the questions for TOEIC speaking test. There are a read-aloud practice, a word and phrase exercise with multiple choices for fill in the blank, and a picture description practice. In the picture description practice (Fig. 3), Siri recognizes the student's utterance and Watson evaluates whether the answer is correct or not. In the question practice, the conversation is documented as a chat record for students to refer back to.

2.2. Questionnaire

After the experimental practice, I analyzed the effectiveness of the system by asking the students about their impressions and reactions using a Google questionnaire form (Shishido, 2018: 1). The questionnaire items were divided into five major categories based on whether they were related to 1) English learning, such as the student's English ability, 2) Shikibu's functions, including the app's effectiveness, 3) Shikibu's contents, 4) benefits of using the app, and 5) future use.

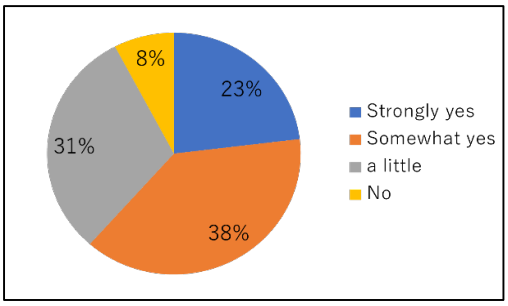

Fig. 4 Interest in Shikibu

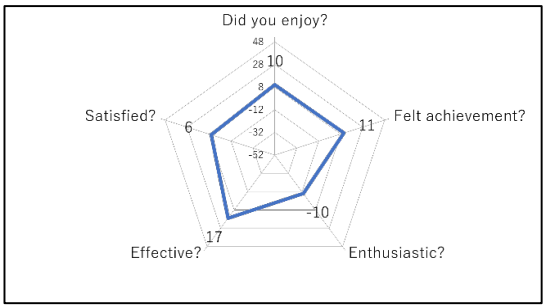

Fig. 5 Radar chart

In response to the question "Were you interested in the app with AI?," 23% of the students answered "yes," while 38% said "somewhat yes" and 31% said "a little." Thus, a total of 92% responded that they were interested in the app (Fig. 4). This is a very high percentages of students indicating interest.

Fig. 5 summarizes the results for the responses to the five questions about the students' impressions of Shikibu, which related to enjoyment, achievement, enthusiasm, effectiveness, and satisfaction. The total scores were calculated based on a score scale with four categories: 2 points for "yes," 1 for "somewhat yes," -1 for "somewhat not," and -2 for "no." The scores for each question were as follows: enjoyment, 10 points; achievement, 11 points; effectiveness, 17 points; satisfaction, 6 points; and enthusiasm, -10 points (Fig. 5). Only the score for enthusiasm was negative, while those for the other questions were affirmative and very high. This seems to indicate that the students found Shikibu very effective and fun to study with, and they felt a sense of achievement and satisfaction when they had completed their studies. Some of the reasons for the low score for enthusiasm may lie in the unfamiliarity of TOEIC among the students and the difficulty level of the questions. It may be possible to improve the enthusiasm scores by changing the contents of future lessons.

In answer to the question "Did you find the English spoken in Shikibu natural?," 84% of the students reported that it was natural even though Shikibu uses Siri and its voice is artificially composed (Fig. 6). The majority of the students' responses were

positive and affirmative. Japanese learners often have negative feelings about speech practice in English because they feel embarrassed about speaking in English and are not confident. However, when asked about their experience using the app, 85% answered that they "felt less resistance to speaking English" and 92% said they were "not embarrassed to make mistakes." Thus, practicing English conversation with e-learning materials using AI allowed the students to overcome negative feelings, make progress, and gain confidence in English conversation with autonomy (Fig. 7).

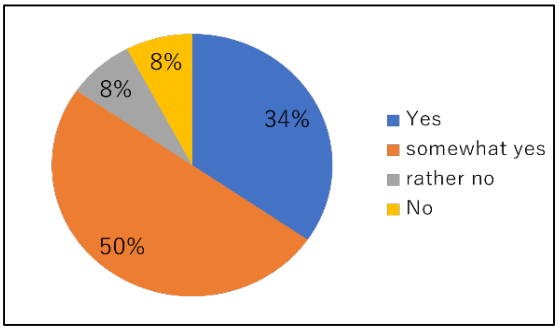

Fig. 6 Is the AI English natural?

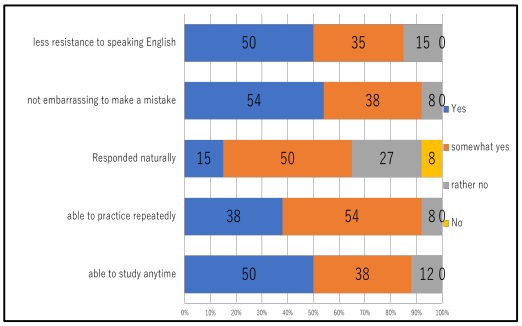

Fig. 7 Benefits of Shikibu

When asked if the English conversation practice with AI matched their learning style, 19% of the respondents said "yes" and 62% said "somewhat yes." Thus, a total of 81% thought the conversation practice with AI somehow matched their learning style and wanted to engage in this kind of English learning exercise.

3. The Second Trial

3.1. SpeakBuddy App

For the second trial, I received support from a company called AppArray. They created a smartphone app named SpeakBuddy with Google voice recognition and AI. The second trial was conducted from April to December 2018 in regular university classes on English speaking skills. The students were 209 first-year students in six different classes, all majoring in information and computer science. The students studied with the smartphone app during the class for nine months and took the questionnaire in the last week of May.

Using smartphones is usually prohibited during the class, but for the experiment, the students used their smartphones to practice English conversation with AI. Fig. 8 is a screenshot of the smartphone app SpeakBuddy. Once the user taps the app icon, the menu screen appears. The contents of SpeakBuddy focus on business English conversation. There are main lesson contents and several exercises in a lesson. The vocabulary practice consists of multiple-choice questions for matching vocabulary words and their meanings in Japanese. Then, in the listening practice section, a model conversation is played. After the students have listened to the model conversation, they proceed to the speaking practice section, where they repeat certain phrases they heard. The AI evaluates their pronunciation and provides feedback using three colors: red, yellow, and green (Fig. 9).

There is another speaking practice with blank spaces (Fig. 10). The students guess the words to fill in the blanks and pronounce the full phrases. The AI judges if their

answers are correct and evaluates their pronunciation at the same time.

Fig. 8 SpeakBuddy lesson menu

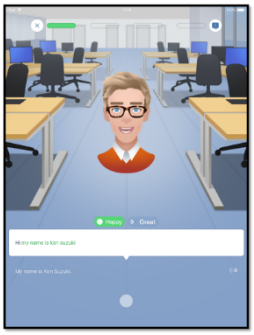

Fig. 9 Screenshot of the speaking exercise in SpeakBuddy

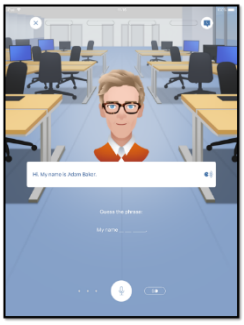

Fig.10 Screenshot of the exercise on speaking to fill in the blank in SpeakBuddy

3.2. Questionnaire

In the last week of May, about two months after the students had studied with SpeakBuddy, they completed a questionnaire (Shishido, 2018: 2). The questionnaire items were classified into five categories: English ability, SpeakBuddy functions and effectiveness, lesson contents, benefits of the English conversation app using AI, and future use.

The radar chart in Fig. 11 summarizes the results for the students' responses to five questions about enjoyment, achievement, enthusiasm, effectiveness, and satisfaction, respectively. The total scores were calculated based on a score scale with four categories: 2 points for "yes," 1 for somewhat "yes," -1 for "somewhat not," and -2 for "no." The scores for each question were as follows: enjoyment, 27 points; achievement, 55 points; effectiveness, 64 points; satisfaction, 7 points; and enthusiasm, -76 points. Only the score for enthusiasm was negative, while those for the other questions were positive and relatively high. This seems to indicate that the students found SpeakBuddy very effective and fun to study with, and they felt a sense of achievement and satisfaction when they had completed their studies. Some of the reasons for the low score for enthusiasm may lie in the unfamiliarity of business English conversation among the university students and the difficulties of the lesson contents. It may be possible to improve the enthusiasm scores by changing the contents of future lessons.

In response to questions about the benefits of practicing English speaking skills with SpeakBuddy as an AI app, 28% chose that they could study anytime (Fig. 12).

Meanwhile, 26% responded that they did not feel embarrassed to make mistakes, 20% said that they could practice repeatedly, and 19% selected that they felt less resistance in speaking English.

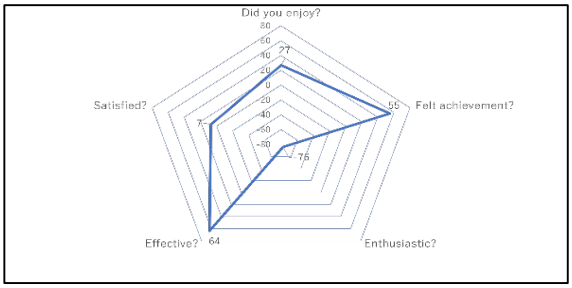

Fig. 11 Student impressions of SpeakBuddy

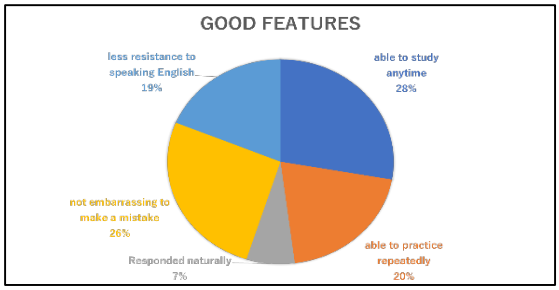

Fig. 12 Benefits of SpeakBuddy

Even though the questionnaire results showed many positive reactions from the students, as the class teacher, I needed to create supplemental materials for as exercises. Since the functions of the smartphone app are very limited, I created vocabulary exercises with Quizlet. I also used Hot Potatoes to create a variety of other grammatical and content-based exercises such as mixed up sentences and cloze tests. These exercises were available on the students' computers, so that they used a combination of the smartphone and computer during class.

In summary, in the second trial, the students practiced their English speaking skills on an individual basis at a low cost, anytime and anywhere. They felt less resistance to speaking, did not feel embarrassed about making mistakes, and were able to repeat the exercises many times. However, the contents of the lessons should be matched to the students' interests and needs. In addition, the voice recognition system needs to be developed further to enhance the recognition of non-native speakers' English.

4. The Third Trial

4.1 ELST Program with Textbooks

For the third trial, I worked with a Japanese company called Sinewave, which created the ELST. This system uses the Chinese voice recognition and AI of iFLYTEC, which was developed to recognize Asian speakers' English. The students were 25 first-year students at the Japanese university. The study was conducted from April to December 2019. The materials were presented in a regular English class.

In addition to the technological support from Sinewave, I also received support from Seibido, a Japanese publisher, to publish a matching textbook. The previous two smartphone apps did not have accompanying textbooks, and I believed this to be a strong necessity. Also, for this product, I created original contents on the topic of "Introducing Japan in English." The contents consisted of all four skills: reading, listening, speaking, and writing.

Figs. 13 to 16 are samples of the textbook. Fig. 13 shows a vocabulary exercise followed by the reading section. The comprehension questions, a note-taking exercise, and grammar practices come after the essay (Fig. 14). The second section is

the listening practice (Fig. 15). After listening to the model conversation, the students work on a fill-in-blank exercise, vocabulary, and comprehension exercise. Speaking and writing sections follow (Fig. 16).

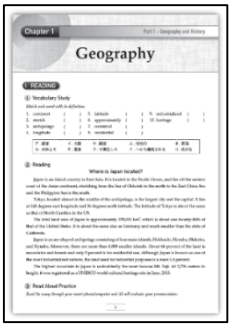

Fig. 13 Reading lesson in the textbook accompanying ELST

Fig. 14 Note-taking lesson in the textbook accompanying ELST

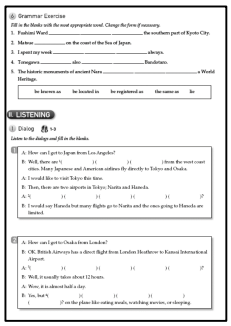

Fig. 15 Listening lesson in the textbook accompanying ELST

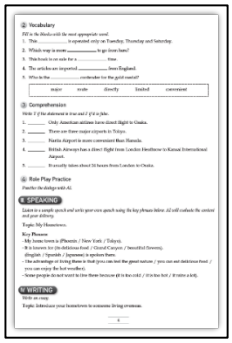

Fig. 16 Speaking lesson in the textbook accompanying ELST

When the computer program opens, the main menu appears. A vocabulary exercise is presented first. The students can listen to a model pronunciation, look at the phonetic

symbols, and read about the meaning. They can also practice their pronunciation and receive an evaluation of it from the AI (i.e., red, yellow, or green). The score appears on the top right of the screen. At the end of the section, the students take a quiz to check their understanding of the vocabulary meanings. There is also a fill in the blank grammar exercise.

The shadowing exercise is the main feature of ELST. In this exercise, the students listen to a model pronunciation and repeat what they hear sentence by sentence. The AI system recognizes their pronunciation and assigns each word a color based on how well it is pronounced: green, yellow, red, or gray. The top left of the screen also shows total evaluations for accuracy, fluency, and integrity. At the same time, a total overall score appears on the top right (Fig. 17). This is also a notable feature of the ELST system.

In the listening section, the model conversation is played, and the students can practice the conversation. The AI system evaluates their pronunciation and generates evaluation feedback in the same way as in the reading section (Fig. 18).

In the speaking section, the students select words and generate their own utterances based on key phrases. The system provides an evaluation of their utterances on the screen. Figs. 17 and 18 show screenshots of the ELST system. The students demonstrated very concentrated and enthusiastic study attitudes in the class.

Fig. 17 Shadowing exercise in ELST

Fig. 18 Dialog practice in ELST

4.2 Comparison of the Pre- and Post-test Results

This section compares the pre- and post-test results of OPIc and TOEIC. The students in the ELST class were requested to take both OPIc and TOEIC at the beginning and end of their academic year. The pre-test OPIc and TOEIC were given in May, and the post-test ones were done in December.

The Oral proficiency Interview-computer test (OPIc) is a speaking proficiency test organized by the American Council on the Teaching of Foreign Languages (ACTFL), and the scales are from 1 the lowest to 8 the highest. In Fig. 19, the blue bar indicates the pre-test results, and the orange bar shows the post-test results. Both the pre- and post-test results have the same lowest and highest level, namely, 1 and 4. In terms of the lowest and highest, there were no differences in improvement. However, a comparison of the mean shows that there was a slight improvement: the mean of the pre-test was 2.04, while that of the post-test was 2.28.

A closer comparison of the OPIc results focuses on the number of students at each

level (Fig. 20). The blue bars are for the pre-test, and the orange bars are for the post-test. As one can see, there was an improvement. Nine students were at level 1 on the pre-test, while this number decreased to four on the post-test. Meanwhile, the number of students increased from 8 to 10 for level 2 and from 6 to 10 for level 3. Even though there was a decrease in the number from 2 to 1 for level 4, most students improved, and the results showed a positive effect of the system overall.

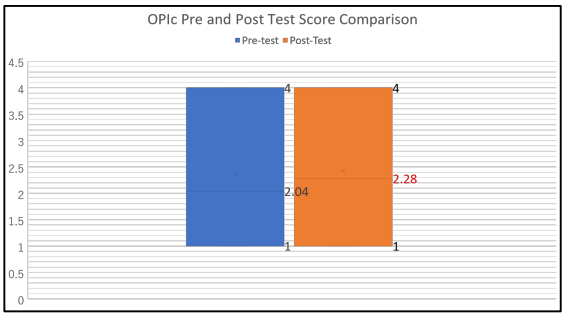

Fig. 19 OPIc pre & post-test results 1

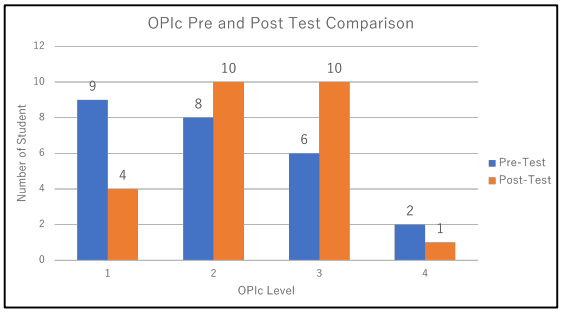

Fig. 20 OPIc pre & post-test results 2

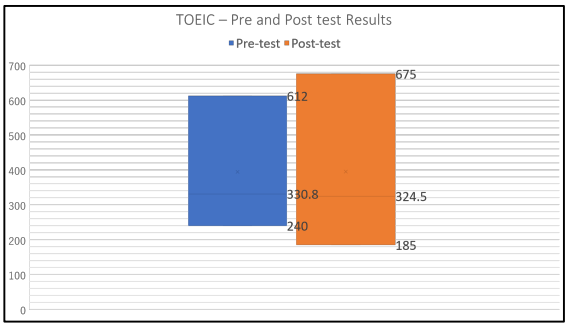

Fig. 21 TOEIC pre- and post-test 1

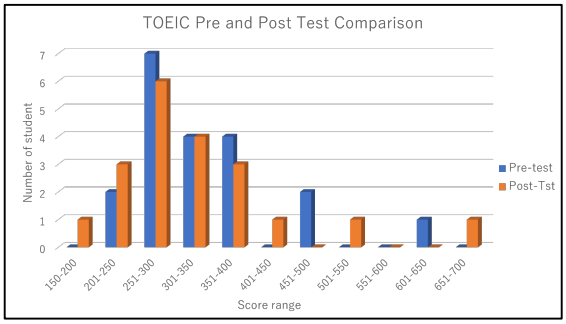

Fig. 22 TOEIC pre- and post-test 2

Next is a comparison of the TOEIC pre- and post-test results. The highest score showed an improvement from 612 to 675, while the lowest score declined from 240 to 185. The mean showed a slight decrease of 6.3 points, from 330.8 to 324.5. A comparison of the number of students in each 50-point range shows some improvement in the higher scores and a decrease in the lower scores.

To analyze and compare the TOEIC scores, I developed the idea of Score Development Rate (SDR), which indicates a potential increase of scores of a test-taker. The concept is intended to rectify the slight unfairness of comparing just the increases or decreases in raw scores. For example, if a student with 490 on the pre-test scores 590 on the post-test, the increase is 100 points. In the same way, if a student with 890 on the pre-test scores 990 on the post-test, the increase is also 100 points. However, these increases may have different meanings. When the increases are calculated with SDR, there is a distinctive difference between the two examples. The SDR of the former student is 20%, while that of the latter is 100%. The SDR formula is given below. The difference is significant, and so the SDR was used to analyze the increases of test scores for the third trial.

SDR = (Post-Test Score–Pre-Test Score) / (Full Mark–Pre-Test Score) × 100
Example 1: TOEIC score increase 490 to 590
(590-490) / (990-490) × 100 = 20%

Example 2: TOEIC score increase 890 to 990

$$(990-890) / (990-890) \times 100 = 100\%$$

The highest SDR was 16.67, and the lowest was -10.56.

The mean SDR was -1.49. Therefore, the SDR overall showed a negative result (Fig. 23).

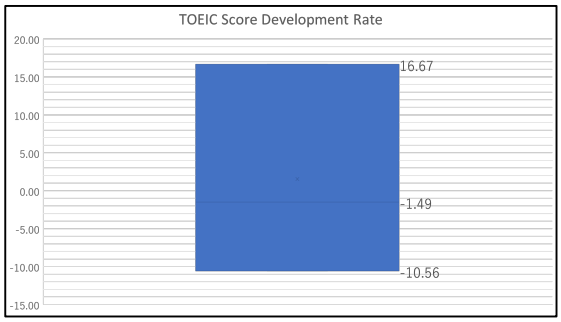

Fig. 23 SDR of TOEIC

4.3 Questionnaire

The questionnaire were administered in December around the end of the study. The answers were scaled from 5 to 1, positive to negative. In response to the question "Were you interested in the ELST?," 23% of the students responded with 5 "very much," 50% with 4 "yes," 23% with 3 "a little," 4% with 2 "not much," and 0% with 1 "no." The average score for interest was 3.93 out of 5. This is a very positive result (Fig. 24).

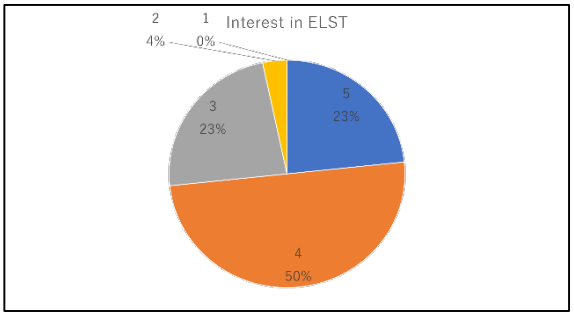

Fig. 24 Student interest in ELST

Fig. 25 Student impressions of ELST

Fig. 25 summarizes the results for six questions concerning the students' impressions of the ELST system: enjoyment, achievement, enthusiasm, effectiveness, satisfaction, and improvement. The students were asked to assign number scores from 5 "the best" to 1 "the worst." The total scores were calculated as follows: 2 points for a score of "5," 1 for "4," 0 for "3," -1 for "2," and -2 for "1." For enjoyment, achievement, and satisfaction, the scores were very high (6, 5.4, and 4.4, respectively). Enthusiasm and improvement also showed relatively high scores of 3.6 and 3.2, respectively, while effectiveness did not receive a very positive result with only 2 points.

In the same way, Fig. 26 shows the benefits of practicing speaking with the ELST system. Questions were asked for five categories: if the students felt less resistance to speaking English, they were not embarrassed to make mistakes, ELST responded

naturally, they were able to practice repeatedly, and they were able to study anytime. All five questions received very high scores and very positive responses.

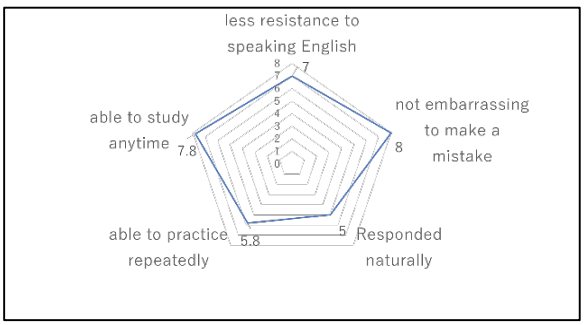

Fig. 26 Benefits of ELST

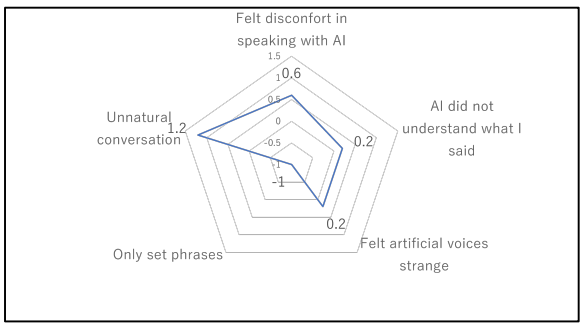

Fig. 27 Negative feelings about ELST

The radar chart in Fig. 27 shows the drawbacks of the ELST system. The lower the scores, the positive or the better the responses were. Questions were asked in five categories: if the students felt discomfort in speaking with AI, the AI did not understand what the students said, the students thought the artificial voice sounded strange, the students were able to practice only set phrases, and the conversations seemed unnatural.

Answers were given on a 5-point scale: 5 “strongly agree,” 4 “agree,” 3 “neutral,” 2 “disagree,” and 1 “strongly disagree.” The scores were calculated in the same way as in the previous radar charts: 2 points for a score of “5,” 1 for “4,” 0 for “3,” -1 for “2,” and -2 for “1.” The question about whether the conversation seemed unnatural received a mean score of 1.2, but the other categories had mean scores of 0.6 and 0.2. The question about practicing only set phrases received a mean score of -1, a very negative response, which means very good. Therefore, the students did not have any negative feelings about using the ELST system.

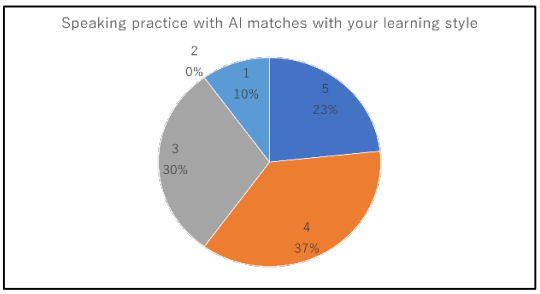

Fig. 28 Learning styl

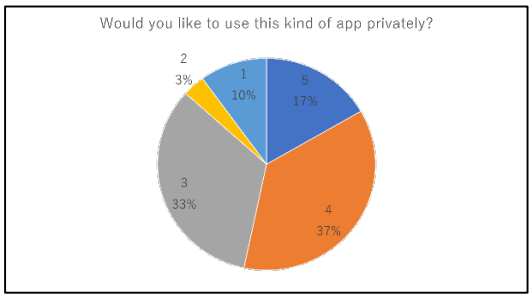

Fig. 29 Students who would like to use the app privately

In response to the question “Do you think speaking practice with AI matches with your learning style?” 23% of the students gave a rating of 5 points, 37% assigned 4 points, and 30% said 3 points. In total, 90% answered with positive responses, and the average score was 3.63 (Fig. 28). Therefore, the majority of students showed very positive reactions. For the question, “Would you like to use this kind of app privately?,” 17% of the students chose 5 points, 37% 4 points, and 33% with 3 points. The negative responses were 3% for 2 points and 10% for 1 point. A total of 87% was positive, and most students said they would like to use the app even outside the class (Fig. 29).

Fig. 30 shows the co-occurrence network of the words in written comments, in other words, how each word in the comments is related. There is a strong connection among the words “English,” “fun,” and “study” and among the words “AI,” “interesting,” “study,” and “fresh.” Therefore, the students indicated that studying English with AI is interesting and fun, and it is a fresh experience for them.

Fig. 31 shows another co-occurrence relation. There is a strong relationship between the words “pronunciation,” “improvement,” “content,” “score,” and “change.” Relationships are also found between “pronunciation” and “good” and between “pronunciation” and “enhanced.” Thus, the students felt their pronunciation had improved and their abilities had increased.

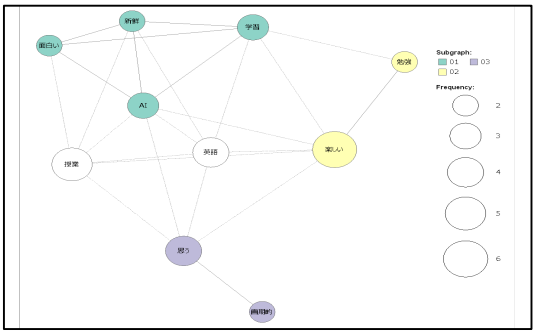

Fig. 30 Co-occurrence network 1

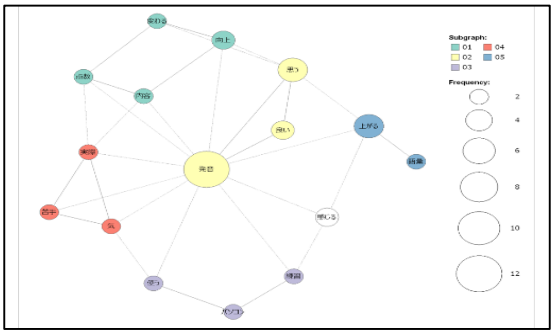

Fig. 31 Co-occurrence network 2

In summary, the results of the third trial indicate that the speaking practice with e-learning material using voice recognition and AI had a positive outcome. The students enjoyed studying with AI and saw it as a fresh experience. They believed they had improved their speaking skills, and their abilities and OPIc scores rose.

5. Future Development

In the future, I plan to improve the ELST system by adding a multiple-choice chatbot so that the students can have rather free conversations with AI. In the current development stage, the chatbot offers three multiple choices for six turns, yielding 27 possible conversations that begin with the same starter question.

Fig. 32 shows the structure of a possible conversation with the multiple-choice chatbot. With this improvement, the students can enjoy free conversation, and I expect the system will become more effective. The new system will become available in 2022.

AI	Student	AI	Student
			4.1.1. I live alone.
	2.1. I’m going home.	3.1. Who is waiting?	4.1.2. My mom.
			4.1.3. My dog.
			4.2.1.I study science.
1. Where are you going?	2.2. I’m going to school.	3.2. What do you study?	4.2.2.I study economics.
			4.2.3.I study math.
			4.3.1. I want a jacket.
	2.3. I’m going shopping.	3.3. What will you buy?	4.3.2.I will get a shirt.
			4.3.3.I am not sure yet.

Fig. 32 Example conversation with the multiple-choice chatbot

References

Sakamoto, Y. et al. “A Study of Learners’ Self-reflection in Oral Communication Practice.” *Language Learning and Educational Linguistics 2015-2016*. pp. 1-11. (2016)

Sakamoto, Y. et al. “Professors’ Opinion about Introducing Online-One-to-one Lessons with Filipino Instructors into University English Courses.” Language *Learning and Educational Linguistics 2016-2017*. pp. 17-24. (2017)

Shishido, M. “Developing e-learning system for English conversation practice using speech recognition and artificial intelligence.” *The EdMedia World Conference on Educational Media and Technology 2018*. pp. 226–231. Amsterdam, Netherland. Association for the Advancement of Computing in Education (AACE). (2018-1)

Shishido, M. “Practicing speaking through app with artificial intelligence on smartphone.” *The European Conference on Language Learning 2018*. Brighton, UK. The International Academic Forum (IAFSOR). (2018-2)