Utilizing Automatic Speech Recognition for English Pronunciation Practice and Analyzing its Impact

Katsuyuki Umezawa
Department of Informatics,
Shonan Institute of Technology
Kanagawa, Japan
umezawa@info.shonan-it.ac.jp

Makoto Nakazawa
Department of Industrial Information Science,
Junior College of Aizu
Fukushima, Japan
nakazawa@jc.u-aizu.ac.jp

Michiko Nakano
Faculty of Education and Integrated Arts and Sciences,
Waseda University
Tokyo, Japan
nakanom@waseda.jp

Shigeichi Hirasawa
Data Science Center,
Waseda University
Tokyo, Japan
hira@waseda.jp

***Abstract*—The advancement of AI in recent years has been remarkable, along with the widespread use of speech recognition functions. In addition, an increasing number of people are self-studying in their fields of interest. In a world considered a global society, many people in countries where English is not their first language are learning English. Therefore, in this study, the authors focus on self-study English learning. The pronunciation is assumed to be correct if the speech recognition feature correctly identifies the pronounced English words and sentences. It has been recognized that feedback plays a crucial role in English pronunciation practice. If the speech recognition function can be employed to provide feedback, it would alleviate the burden on teachers and enable students to practice pronunciation independently. In this research, the authors propose practicing pronunciation using each browser's built-in speech recognition features, including Google Chrome, iOS, and Microsoft Edge, and evaluating if one's English pronunciation is correctly recognized. Through a 7-day experiment, the authors clarify the speech recognition function suitable for self-studying English pronunciation. Through this experiment, it was observed that the speech recognition function on iOS (version 15.0) outperforms compared to Google Chrome (version 107.0.5304.63) and Microsoft Edge (version 107.0.1418.24) in accurately understanding speech, even when the pronunciation is incorrect. However, it became apparent that such highly accurate speech recognition capabilities may not be suitable for self-study pronunciation practice.**

***Index Terms*—Speech recognition function, Pronunciation practice，Google Chrome，iOS， Microsoft Edge**

I. Introduction

The advancement of AI in recent years has been remarkable, and speech recognition functions have come to be used daily. Moreover, more and more people are self-studying in their fields of interest. In a world considered a global society, many people are learning English. In this study, the authors focus on learning English by self-study. The pronunciation is regarded to be correct if the pronounced words and sentences are correctly identified using the speech recognition function, which is growing increasingly popular daily. In other words, using the speech recognition function would allow us to practice English pronunciation.

In previous research [1], in addition to the evaluator's check, researcher employed the speech recognition function (Siri 12.0) when testing pronunciation to confirm whether the pronounced words and sentences were accurately recognized. As a result, the speech recognition function is thought to be accurately recognized, and it is decided that using the speech recognition function for pronunciation practice is worthwhile. However, in a previous study [1], only Siri was used for the speech recognition function, and it is unclear whether the same results can be produced for other speech recognition functions. Also, experiments have yet to be conducted using languages other than Japanese.

Therefore, this study focuses on learning English, often studied as a second foreign language in non-English speaking countries. Then, when practicing English pronunciation by self-study, the authors propose to use the speech recognition function to check whether the pronunciation is correct. Furthermore, since several types of speech recognition functions are currently provided by various companies, this experiment aims to confirm which is suitable for self-studying English pronunciation.

II. Previous work

A. Second language learning and automatic speech recognition

In second language learning, vocabulary knowledge is considered closely tied to oral ability, particularly fluency [2]. Pronunciation holds an equal importance to vocabulary. It has been highlighted that the ability to effectively communicate in a second language is strongly linked to the speaker's level of pronunciation proficiency [3]. Additionally, in the context of pronunciation learning, immediate feedback following pronunciation is deemed crucial [4]. Nevertheless, providing individualized feedback to every learner is impractical for teachers due to time and cost constrains. To address these challenges, many studies have explored the utilization of automatic speech recognition technology for second language learning [5]. Furthermore, there has been recent research on various automatic speech recognition technologies, such as free speech text processing (Windows Speech Recognition, Google Speech Recognition, Apple Siri, etc.). For instance, a previous study [6] conducted an accuracy assessment comparing two programs, Windows Speech Recognition and Google Speech Recognition. Additionally, another study [7] investigated the transcription accuracy of second language learners' speech using two engines, Apple's Siri and Google Speech Recognition. The findings concluded that Google Speech Recognition exhibited higher accuracy in voice transcription and was easier to implement. However, these prior studies primarily focused on analyzing speech recognition accuracy and did not assess proficiency in pronunciation. In this study, the authors specifically evaluate pronunciation proficiency.

B. Pronunciation checklist proposal

In previous research [8], researcher proposed a pronunciation checklist using self-monitoring. Self-monitoring means having the learner read the pronunciation checklist, record it, and check the recorded speech by himself/herself using the checklist. In this study, researcher created a checklist based on textbooks for learning Japanese speech for Korean speakers. This checklist includes minimal pairs of difficult-to-pronounce sounds for Korean speakers, such as "tsu" and "chu," as well as the Japanese "a" column and "ha" column. A pronunciation checklist was employed and the word was relearned if a mistake was made. The method of checking was to record the student's voice, which was then listened to and rated by the evaluator and the learner himself/herself. Thus, students who consistently made mistakes tended to regard their incorrect pronunciation as appropriate. On the other hand, after hearing incorrect pronunciations, even students who made a lot of mistakes could determine that some words were improper. This study contends that it is possible to learn step by step by practicing with such words.

C. Revision of pronunciation checklist using speech recognition function

The pronunciation checklist proposed in the prior study [8] was revised in the previous study [1] to allow the speech recognition function to use the checklist. The speech recognition function used at that time was Siri (iOS12.0). In this study, researcher used a speech recognition function (Siri) in addition to the conventional evaluator's evaluation when confirming whether the pronounced words and sentences were correctly recognized. The experiment was conducted in a class of 9 foreigners with an intermediate level of Japanese (including 5 Korean-speaking students). Comparing the evaluation by the evaluator and the evaluation by the speech recognition function, some of the evaluation results were consistent, but there were also differences. However, when the same misuse tendency category was considered, many coincidences were discovered. Therefore, the speech recognition function is also regarded to be accurately recognized, and it is decided that using the speech recognition function for learning pronunciation is worthwhile.

D. About English pronunciation

Previous study [9] investigates the reasons for several aspects of Japanese English pronunciation (hereinafter referred to as "Japanese English") and recommends countermeasures. Causes of poor pronunciation according to phonological level were clarified. Concretely, the authors enumerate the characteristics of Japanese English at the levels of segmental sounds, syllables, words, phrases, and sentences. Several pronunciation issues, particularly at the segmental sound level were pointed out. It was argued that this is due to a lack of correct vowel and consonant pronunciation practice at the beginning of English learning. This study concludes that pronunciation instruction is essential for English learning and is most effective for teachers to teach pronunciation firmly.

III. Proposal

A. Overview

Previous research [1] found that learning using the speech recognition function is worth using. In this research, the authors will find out which speech recognition function has a greater learning effect by comparing it with speech recognition functions other than those used in previous studies [1]. In addition, Japanese was used in previous research [1], but this research targets English. It will be also investigated whether this learning method is effective even for self-study without an evaluator and the difference in the effect of learning using and not using the speech recognition function. Specifically, 28 participants were randomly assigned to four groups and engaged in a seven-day randomized controlled trial for English pronunciation practice. Subsequently, an analysis was conducted to compare the average number of correct pronunciations on the first and the last days of the seven-day period across each group. Additionally, the number of accurate response in the daily pronunciation checks was analyzed.

B. Hypothesis

Previous research [1] has established the value of utilizing speech recognition functions for pronunciation practice. Consequently, the speech recognition function is regarded as a useful tool for self-study pronunciation practice. There exists a notable distinction between learning with and without the speech recognition function, as visual feedback on the accuracy of recognition is readily available when employing the function. This visual feedback serves to enhance motivation during the learning process and facilitates the acquisition of correct pronunciation, even through self-study. Learning with the speech recognition function is more effective for acquiring accurate pronunciation compared to learning without it.

IV. Experimental overview

A. Words and sentences used in the experiment

Figure 1 shows the list of words and sentences used in the experiment. This list was created with reference to previous research [8] [1], focusing on minimal pairs and those with "l" and "r" that are difficult for Japanese to pronounce.

1	athlete	5	light	9	match
2	itinerary	6	right	10	usually
3	refrigerator	7	jewelry	11	I am reading a book in the library.
4	woman	8	march	12	I know you like it.

Fig. 1. Words and sentences used in the experiment

B. Experiment flow

First, the experiment participants confirmed the pronunciation of the 12 words and sentences on the checklist using the speech recognition function determined for each group. Then they counted how many words and sentences the speech recognizer correctly detected. Following a pronunciation check, they spent 1-2 hours practicing their pronunciation (when the authors surveyed the actual situation after the experiment, the authors found that many participants actually practiced for about 30 minutes. Some participants practiced for as little as five minutes). Participants can utilize Google's read-out function to check if they are pronouncing a word or sentence correctly if they are unsure. Pronunciation confirmation is standardized for all experiment participants using Google's read-out function. The Google read-out function described here differs from the speech recognition functions that assess pronunciation accuracy, as explained later. It is solely utilized to determine the correct pronunciation. This process is performed for seven days. Participants recorded their pronunciation on the first and last days of the experiment and asked a native English speaker (hereafter referred to as the evaluator) to judge whether the pronunciation was correct. Regarding the recording method, each participant was instructed to use the recording function on their smartphone.

At the beginning of each day's study, group participants using the speech recognition function (Groups A, B, and C) must use the function to check all the checklists. Pronunciation practice with the speech recognition function entails practicing pronunciation until the speech recognition function identifies the words and sentences on the checklist correctly. The experiment consisted of 28 participants, randomly divided into groups of seven individuals each. The participants in each group conducted the experiment using different speech recognition functions. Seven participants (A_1 to A_7) practiced pronunciation using Google's speech recognition function, seven participants (B_1 to B_7) practiced pronunciation using iOS's function, and seven participants (C_1 to C_7) practiced pronunciation using Microsoft Edge's function. In addition, the seventh participants (D_1 to D_7) practice pronunciation without using the speech recognition function.

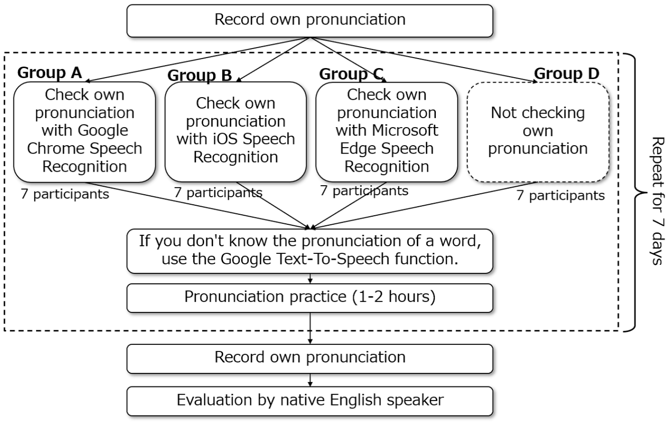
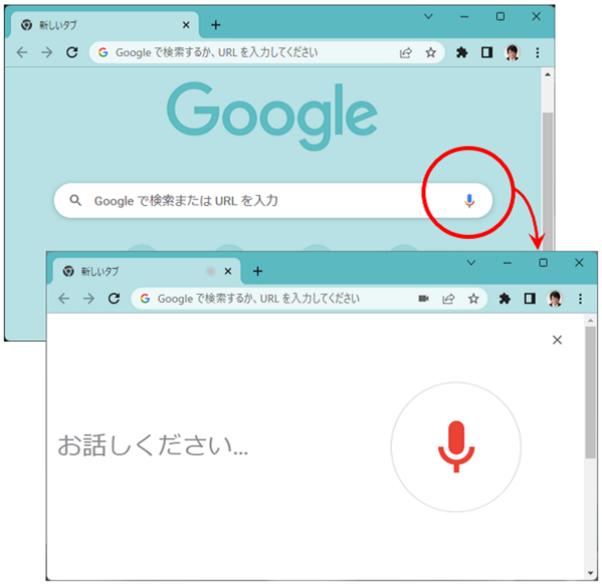

Record own pronunciation
Group A
Group B
Group C
Group D
Check own pronunciation with Google Chrome Speech Recognition
Check own pronunciation with iOS Speech Recognition
Check own pronunciation with Microsoft Edge Speech Recognition
Not checking own pronunciation
7 participants
7 participants
7 participants
7 participants
Repeat for 7 days
If you don't know the pronunciation of a word, use the Google Text-To-Speech function.
Pronunciation practice (1-2 hours)
Record own pronunciation
Evaluation by native English speaker

Fig. 2. Experiment flow

C. About the speech recognition function used in the experiment

As shown in Figures 3 to 5, the speech recognition function used in the experiment is the standard voice search function of three types of Web browsers: Google Chrome (Version 107.0.5304.63) running on Windows 11, Safari running on iPhone (iOS 15.0), and Microsoft Edge (Version 107.0.1418.24) running on Windows 11. The language setting of all web browsers was set to English.


新しいタブ
Google で検索するか、URL を入力してください
Google
Google で検索または URL を入力
新しいタブ
Google で検索するか、URL を入力してください

お話しください...

Fig. 3. Speech recognition function of Google Chrome

V. Experimental results and evaluation

A. Evaluation of final pronunciation practice results

As shown in Figure 2, participants were recorded on the first and last days of the experiment. This section evaluates whether participants can finally pronounce words and sentences. The evaluation method is to judge whether the recorded material is

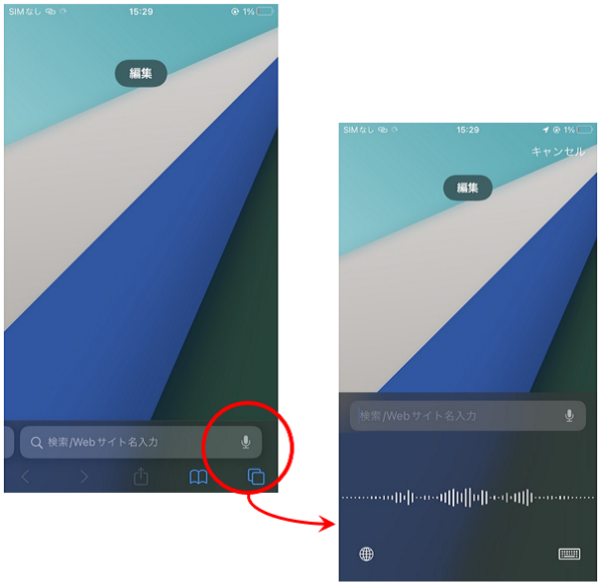
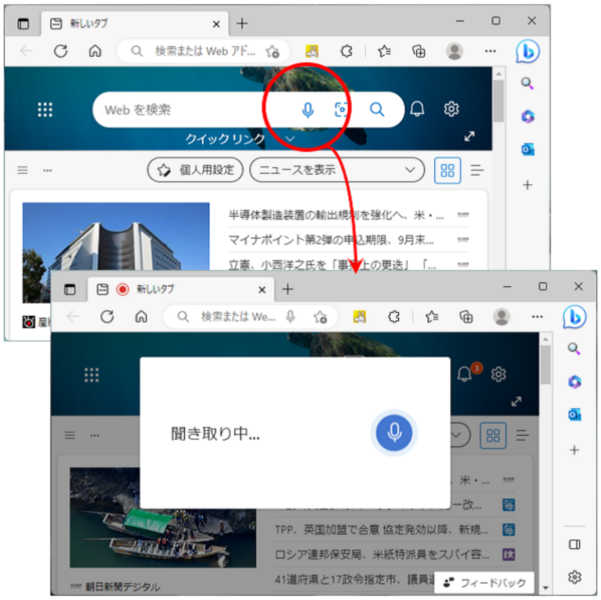


Fig. 4. Speech recognition function of iOS



半導体製造装置の輸出規制を強化へ、米・...
マイナポイント第2弾の申込期限、9月末...
立憲、小西洋之氏を「事実上の更迭」「...
TPP、英国加盟で合意 協定発効以降、新規...
ロシア連邦保安局、米紙特派員をスパイ容...
41道府県と17政令指定市、議員選...

Fig. 5. Speech recognition function of Microsoft Edge

TABLE I
RESULTS OF CHECKING BY NATIVE SPEAKERS ON THE FIRST AND LAST DAY FOR EACH GROUP (NUMBER OF CORRECT ANSWERS)

Group A (Learned on Chrome)		
partici-pants	first day	last day
A_1	6	7
A_2	6	7
A_3	6	8
A_4	6	7
A_5	7	9
A_6	6	7
A_7	8	9
Ave.	6.4	7.7

Group B (Learned on iOS)		
partici-pants	first day	last day
B_1	7	8
B_2	9	9
B_3	6	7
B_4	6	7
B_5	7	7
B_6	6	7
B_7	5	6
Ave.	6.6	7.3

Group C (Learned on Edge)		
partici-pants	first day	last day
C_1	5	6
C_2	6	7
C_3	7	8
C_4	4	6
C_5	6	8
C_6	6	7
C_7	9	10
Ave.	6.1	7.4

Group D (No speech recog. func.)		
partici-pants	first day	last day
D_1	6	8
D_2	5	7
D_3	7	8
D_4	6	7
D_5	7	8
D_6	8	8
D_7	5	6
Ave.	6.3	7.4

pronounced correctly by a native English speaker. In addition, the criteria for this evaluation are similar to the conventional study [9], and judge whether it sounds like native English rather than Japanese English.

Table I shows the check results (number of correct answers) by native speakers of the group that learned without using the speech recognition function and the group that learned using each speech recognition function.

The authors tested whether there was a significant difference between the number of correct answers on the first day and on the last day in each group. First, when the authors performed an F-test to check whether the variances were equal or not, it was found that all groups had "$p > 0.1$" and that the variances were equal. Therefore, the authors performed t-test with two samples assuming equal variances for all groups. p-values are shown in table II and Figure 6. A one-sided t-test was used to check whether the number of correct answers was higher on the final day.

TABLE II
p-VALUE (ONE-SIDED) OF t-TEST OF THE DIFFERENCE BETWEEN THE MEANS OF THE FIRST DAY AND THE LAST DAY

Group	p-value	Remarks
Group A	0.0087 (< 0.05)	Significantly higher
Group B	0.1286 (> 0.05)	Not significantly higher
Group C	0.0660 (> 0.05)	Not significantly higher
Group D	0.0233 (< 0.05)	Significantly higher

From Table II, it was found that the group that learned using Chrome (Group A) and the group that did not use the speech recognition function (Group D) had a higher learning effect on pronunciation. The group that learned using Edge (Group C) did not show a significant difference, but the p-value was close to 0.05. On the other hand, the group that learned using iOS (Group B) clearly had a "$p > 0.05$", and it cannot be said that the learning effect increased. It is fascinating to note that learning on iOS is less effective than learning without using the speech recognition function. Then, another analysis is undertaken in the following section to determine why such a result was obtained.

B. Daily evaluation using speech recognition function

As shown in Figure 2, the participants repeatedly practiced daily and checked their pronunciation. During the 7-day experiment, the authors evaluated how many words and sentences

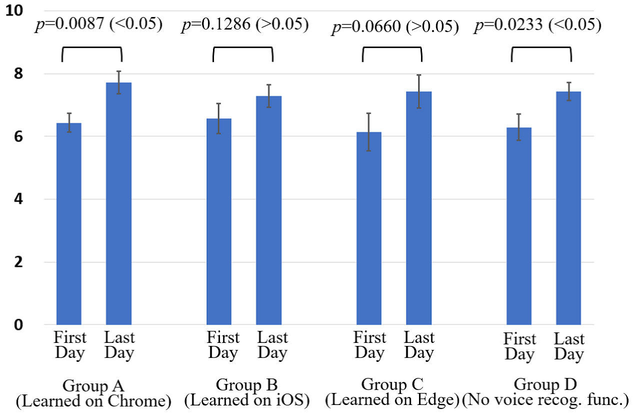


Fig. 6. Comparison of the average number of correct answers on the first and last day (Error bars are standard errors)

were correctly recognized using each speech recognition function. This clarifies whether the speech recognition function is suitable/unsuitable for self-study of English.

Table III to table V show the check results (number of correct answers) using the speech recognition function for each day of the experiment period (7 days). D1 to D7 in the table represent Day 1 to Day 7.

TABLE III
RESULTS OF CHECKS BY CHROME (NUMBER OF CORRECT ANSWERS)

Participants	D1	D2	D3	D4	D5	D6	D7	Ave.
A_1	5	6	5	6	4	5	5	5.1
A_2	6	8	7	8	7	9	7	7.4
A_3	8	8	9	8	10	8	9	8.6
A_4	7	7	6	6	7	6	7	6.6
A_5	7	8	6	6	7	8	8	7.1
A_6	6	7	6	7	7	6	7	6.6
A_7	8	8	9	8	9	9	9	8.6
Ave.	6.7	7.4	6.9	7.0	7.3	7.3	7.4	7.1

TABLE IV
RESULTS OF CHECKS BY IOS (NUMBER OF CORRECT ANSWERS)

Participants	D1	D2	D3	D4	D5	D6	D7	Ave.
B_1	8	10	9	9	10	10	11	9.6
B_2	10	10	11	10	9	11	11	10.3
B_3	10	9	9	9	10	11	9	9.6
B_4	9	10	11	9	10	9	10	9.7
B_5	8	9	10	9	10	9	10	9.3
B_6	9	10	9	10	10	12	11	10.1
B_7	11	9	11	10	10	11	10	10.3
Ave.	9.3	9.6	10.0	9.4	9.9	10.4	10.3	9.8

Using the mean values for each participant in Tables III to V (the rightmost column of the tables), the authors test whether there is a difference between the mean values of Groups A and B, Groups B and C, and Groups A and C. First, when the authors performed an F-test to check if the variances were equal, the authors discovered that all groups had ”$p > 0.1$” and that the variances were equal. Therefore, the authors performed t-test with two samples assuming equal variances for all groups. The p-values are shown in Table VI. A two-sided t-test is also used to determine whether there is a difference in the mean values of the two groups. Note that this t-test is repeated three times, so there is a problem of multiplicity. Accordingly, the Bonferroni method is used to avoid multiplicity. Precisely, it is adjusted by dividing the p-value threshold of 0.05 by the multiplicity (3 in this case).

TABLE V
RESULTS OF CHECKS BY EDGE (NUMBER OF CORRECT ANSWERS)

Participants	D1	D2	D3	D4	D5	D6	D7	Ave.
C_1	5	5	7	5	6	7	6	5.9
C_2	6	8	7	6	8	7	9	7.3
C_3	6	7	6	6	8	7	6	6.6
C_4	7	8	8	9	8	9	10	8.4
C_5	5	7	8	6	7	6	8	6.6
C_6	7	8	10	9	10	11	10	9.3
C_7	7	8	6	5	7	8	7	6.9
Ave.	6.1	7.3	7.4	6.6	7.7	7.9	8.0	7.3

TABLE VI
p-VALUE (TWO-SIDED) OF t-TEST OF THE DIFFERENCE BETWEEN THE MEANS BETWEEN GROUPS

Group	p-value	Remarks
Group A and B	0.000129 ($< 0.05/3$)	Significant difference
Group B and C	0.000146 ($< 0.05/3$)	Significant difference
Group A and C	0.828017 ($> 0.05/3$)	Not significant difference

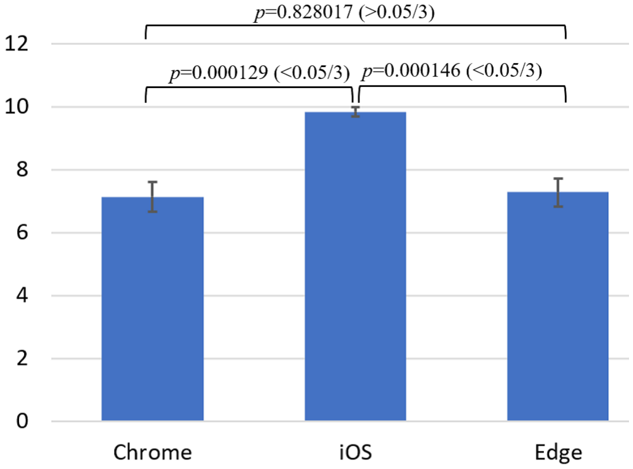


Fig. 7. Comparison of averages for different speech recognition functions (Error bars are standard errors)

Table VI and Figure 7 show the two-sided p-values. As a result, it can be said that the speech recognition function on iOS has a significantly higher recognition rate than the other speech recognition functions of Chrome and Edge.

VI. CONSIDERATION

From Table II and Figure 6, the correct response rate did not increase considerably on the first and last days of iOS pronunciation practice. Specifically, the average score on the first day was 6.57, and on the last day, it increased to 7.29. The p-value obtained from the t-test was 0.1286, indicating no significant difference. In contrast, looking at Table VI and Figure 7, the average number of correct answers from daily checks on iOS is much higher than the average number of correct answers from checks on other Chrome and Edge. Moreover, the average score for iOS was 9.81, while Chrome scored 7.13, and Edge scored 7.24. Both Chrome and Edge showed significant difference compared to iOS. Considering that the number of correct answers did not significantly increase in the final native check, despite the higher number of correct answers in the daily checks, it is possible that the iOS speech recognition function accurately recognize pronunciation even if it deviates from native pronunciation. Figure 7 shows that iOS demonstrated superior speech recognition performance compared to other functions and tends to accurately recognize pronunciations that may sound unusual to native speakers. In other words, it is inferred that pronunciation was not improved through daily practice, resulting in no increase in the number of correct answers in the final native check.

In addition, from Table II and Figure 6, the group that learnt without using the speech recognition function (group D) had a much higher number of accurate responses on the native check on the last day. The experiment revealed that pronunciation is enhanced by not employing the speech recognition function rather than applying the speech recognition function with a high recognition rate to self-study pronunciation. From the experimental results, preferable to use the speech recognition function for pronunciation practice, which does not recognize correctly until the pronunciation is firm.

VII. Summary and future work

In this study, the authors designed and tested a method for checking the daily learning outcomes of self-study pronunciation practice using three different speech recognition functions and evaluated its effectiveness through experiments. Specifically, the pronunciation on the first day of self-study and the last day (seven days later) was assessed by a native English speaker to see whether the pronunciation was correctly pronounced. Moreover, during the seven-day experiment, the authors used the speech recognition function daily to check whether the pronunciation was recognized. The study concluded that the iOS speech recognition function is more powerful than others and recognizes pronunciation even when not pronounced correctly, making it inappropriate for self-learning pronunciation. It should be noted that the results of this experiment are likely to be dependent on the version of the speech recognition function employed; therefore the results may alter in the future if a new version of the speech recognition function is used.

The authors believe that future experiments involving a bigger sample size are required and that more target phrases and terms should be used. It is necessary to consider potential limitations in future research, such as evaluating the effectiveness of the speech recognition function for individuals with specific speech impediments or dialects.

Research ethics

All experiments were approved by the Research Ethics Committee of Shonan Institute of Technology. The authors also received written informed consent from the participants and parents or guardians.

Acknowledgments

Part of the work reported here was carried out as a part of the research project “Research on e-learning for next-generation” of Waseda Research Institute for Science and Engineering, Waseda University. Part of this work was supported by JSPS KAKENHI Grant Numbers JP22H01055, JP21K18535, and JP20K03082. Research leading to this paper was partially supported by the grant as a research working group “ICT and Education” of JASMIN.

References

[1] Toshiyuki Kawano. The pronunciation checklist using speech recognition and its effect. *Japanese Language Education Methods*, Vol. 25, No. 2, pp. 28–29, 2019.

[2] Takumi Uchihara and Kazuya Saito. Exploring the relationship between productive vocabulary knowledge and second language oral ability. *Language Learning Journal*, Vol. 47, No. 1, pp. 64–75, 2019.

[3] Christine C. M. Goh and Anne Burns. Teaching speaking: A holistic approach. *Cambridge University Press*, pp. 1–11.

[4] Catia Cucchiarini, Warda Nejjari, and Helmer Strik. My pronunciation coach: Improving english pronunciation with an automatic coach that listens. *Language Learning in Higher Education*, Vol. 1, No. 2, pp. 365–376, 2012.

[5] Ewa M. Golonka, Anita R. Bowles, Victor M. Frank, Dorna L. Richardson, and Suzanne Freynik. Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, Vol. 27, No. 1, pp. 70–105, 2014.

[6] Shannon McCrocklin, Abdulsamad Humaidan, and Idée Edalatishams. Asr dictation program accuracy: Have current programs improved. *In Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, pp. 191–200, 2019.

[7] Paul Daniels and Koji Iwago. The suitability of cloud-based speech recognition engines for language learning. *JALT CALL Journal*, Vol. 13, No. 3, pp. 229–239, 2017.

[8] Toshiyuki Kawano and Yoshiro Ogawara. Making the pronunciation checklist in the textbook and its use. *Japanese Language Education Methods*, Vol. 14, No. 2, pp. 10–11, 2007.

[9] Koji Ono. English pronunciation by thejapaneseand theways to correct it. *Faculty of Culture and Education, Saga University, Departmental Bulletin Paper*, Vol. 17, No. 1, pp. 57–78, 2012.