Investigating the Effect of Chinese Pronunciation Teaching Materials Using Speech Recognition and Synthesis Functions (利用语音识别及语音合成功能的汉语语音教材的教学效果研究)

Watanabe, Yukiko
(渡邉ゆきこ)
Okinawa University
(沖縄大学)
watanabe@okinawa-u.ac.jp

Omae, Tomomi
(大前智美)
Osaka University
(大阪大学)
omae@cmc.osaka-u.ac.jp

Odo, Satoru
(小渡悟)
Okinawa International University
(沖縄国際大学)
sodo@okiu.ac.jp

Abstract: Research on teaching pronunciation using computers has finally become prominent in recent Chinese language education in Japan. A more focused method of teaching pronunciation, computer assisted pronunciation teaching (CAPT), which uses speech recognition, has become a subject of research because of its effectiveness. This paper gives an overview of the current state of CAPT in Chinese pronunciation education in Japan and China. It introduces the conception and implementation of the ST Lab (Speech Training Laboratory), an original CAPT teaching platform that uses speech recognition and speech synthesis functions. ST Lab features a teaching material creation interface and thus can be used in conjunction with any textbook or existing learning material. The paper also includes a section that provides preliminary evidence for the ST Lab system’s effect on learning in Chinese language class after one and a half years.

摘要：近年来在汉语教育中，一些使用电脑教学的语音教学研究逐渐增加，被称为“CAPT”（电脑辅助语音教学）。其中，利用语音识别功能的语音教学研究尤以其显著效果备受瞩目。本文概述 CAPT 在日本和中国的汉语语音教学之现状，并讨论利用语音识别和语音合成功能开发具有制作语言教材功能的教学系统 ST Lab 之功能。最后以一年半的实际教学状况证实其教学的初步效果。

Keywords: Chinese pronunciation, CAPT, speech recognition, speech synthesis

关键词: 汉语发音、电脑辅助语音教学、语音识别、语音合成

1. Introduction

Modern Chinese language education in Japan has lagged behind English in educational theory, resources, development, and implementation. It was only in the late 1990s that the movement to use computers grew in Chinese language education communities in Japan. According to Tanabe's (2004) work on eLearning, the demand for a transition in Chinese language education "from a structure-centric grammatical syllabus to emphasis on content transfer, and the introduction to meaningful transfer activities through a formalistic focus" gained momentum. As a part of this trend to introduce meaningful transfer activities, the development of autonomous learning systems related to the digitization of textbooks, grammar, and vocabulary has begun. However, until recently pronunciation, which is the first major challenge that learners encounter when learning Chinese, was not a subject of study in eLearning systems.

Most Chinese classes at Japanese universities are second language classes for students who are learning Chinese for the first time. The average class time is two classes a week (3 hours), with less than 100 contact hours per year. About ten of these sessions (30 hours) are usually dedicated to the explanation and practice of tones, phonemes, and rhyme; most of them end after an explanation of the four tones and the introduction of about 400 syllables. This is a considerable amount of material to cover in such a period of time. At present, it is difficult to spend enough time for pronunciation practice of two syllable words, which account for 70% of vocabulary frequently used in Chinese. The size of the class is from 30 to 40 students, which is not small. Also, it is hard to say if students have much motivation to study. In such a situation, it is impossible for teachers to repeatedly check each student's pronunciation and to help them correct their pronunciation by asking them to listen until they have the correct grasp of the word sounds, and then repeat practicing until they can pronounce the words correctly. Computers, however, can solve this problem.

One method of teaching using the CALL system is to have students record their own voices. But as Iwai (2015) pointed out, it is difficult for beginners to listen to a model voice and notice their mistakes by themselves. Teachers can also collect recorded audio files, score them, and return them to the learners, giving them appropriate suggestions. However, these classes are usually only twice a week, and feedback a few days after practice is unlikely to be very effective.

As natural language processing and AI technology have developed in recent years, computer technology has finally become a helpful tool in learning Chinese pronunciation. According to Li (2016), the number of studies on the use of speech recognition functions of smartphones and tablets for foreign language education has increased since 2000. There are also other studies that have used methods for visualizing and analyzing speech for pronunciation education, which are called CAPT (computer assisted pronunciation teaching).

In this paper, we discuss previous research of CAPT applications, including cases in which speech recognition was used and another situation in which speech recognition was done using a large-scale Chinese interlanguage speech corpus and deep neural networks (DNN). After that, we introduce ST Lab, a CAPT system, which is the subject of this current study. After clarifying the difference between this study and previous research, the development concept and functions of the developed pronunciation practice system ST Lab and its initial learning effects are shared.

2. Literature Review

Iwai (2015) mentioned the following advantages of speech recognition software in foreign language acquisition:

1. Immediate feedback on pronunciation in a form that learners can recognize.
2. Immediate feedback helps improve and maintain student's motivation to practice.
3. The learners get motivation from the feedback that occurs right after they pronounce a certain word or phrase. Since they can see the result immediately, even without a teacher's instructions they can keep practicing on a device until it recognizes their pronunciation as correct.

It is important to note that speech recognition did not initially serve educational purposes, but aimed to simplify character input for native speakers. Furui (2000) stated that a computer recognizes not only sounds but other information, such as grammar, vocabulary, and context. There are cases when the latter play an even more significant role than sounds. In other words, even if the pronunciation is not perfect, the speech recognition system can also guess the correct words from context. So, there are cases in which incorrect pronunciation may not cause input failure.

The tendency of computer speech recognition systems to guess words correctly is greater when analyzing them in a sentence using its context rather than recognizing single words with no contextual information provided by a sentence. When a computer speech recognition system is given a single word there is a greater likelihood that the computer will not recognize the word correctly. Also, factors such as microphone performance, voice volume and clarity, enunciation, and surrounding noise may affect the accuracy of the input. When using the speech recognition function, it is necessary to bear in mind that "successful input ≠ correct pronunciation" and "input failure ≠ incorrect pronunciation."

There are several studies on language acquisition using speech recognition. Schwartz et al. (2009) researched English education and mistakes in English pronunciation. Hayashi, Mizuochi, Kiryu, and Kanzaki (2012) investigated the reading out loud of translated content by using a voice recognition system. Iwai (2013) studied the automatic conversion of speech into text by using speech recognition. Finally, Zhao, Tomita, Konno, Ohkawa, and Mitsuishi (2019) developed an application, KoToToMo, for Chinese language acquisition.

KoToToMo is a comprehensive review application based on the beginners textbook *KOTOTOMO: Words as Friends* written by Zhao et al. (2019), and includes writing tasks by rearranging words.[1] In terms of pronunciation, each section offers practice exercises using the voice synthesis and word learning functions of the online learning tool Quizlet, followed by repeating and shadowing using the recording function. Finally, there is a "trial" using the voice recognition function. The speech recognition task screen displays a button with a pronunciation example, also with pinyin (the Romanization of Chinese characters). That way, the learner can listen to the model voice and repeat it. To increase learner motivation, a character, "Teacher Panda," appears on the screen, cheering the learner if the input is correct. After this, the screen switches to the next task. In case of incorrect input, the "failure" screen appears, Teacher Panda expresses regret, and the screen returns to the same task again, offering learners the opportunity to redo it. It is reported that the learning effectiveness is remarkable.

A study of language education using a large-scale Chinese inter-language speech corpus and DNN speech recognition system looked at 尔雅中文 [Ěryǎ Zhōngwén], an application created by Wei and Zhang in 2018. The application software is based on the textbook series *Ěryǎ* [尔雅]. It provides reading practice for each new word, example sentences, and conversation reading practice. It visualizes pronunciation, dividing it into 3 parts: initial [声母 shēngmǔ], final [韵母 yùnmǔ], and tone [声调 shēngdiào]. It then compares this with the voice of a native speaker and calculates the difference between each part. If the difference is substantial, it indicates that the pronunciation is incorrect, changes the color of the corresponding part of the pinyin displayed on the screen, and shows the correct pronunciation as a percentage.

Wei and Zhang (2018) used the application for a six-week experiment in two classes with a total of 36 international students. Consequently, the learners practiced pronunciation 28,101 times in total and an average of 780.6 times per person. The application was used 8.1% of the time during class hours and 91.1% of the time during extracurricular hours. The number of exercises on the day the teacher asked students to practice pronunciation as homework has increased significantly. Regarding the effectiveness of the application's pronunciation correction, 83.2% of the corrections were improved, showing a remarkable effect.

Both CAPT methods mentioned above have tried to improve pronunciation by giving feedback quickly, and positive results have been achieved. It has been confirmed that many learners practice pronunciation outside of class. This makes such exercises excellent teaching materials for pronunciation education when there is a shortage of study time and the need to undertake extracurricular study.

In August 2012, a report of the Central Education Council, published by the Japanese Ministry of Education, Culture, Sports, Science, and Technology, proposed the following new educational reforms to be pursued:

> It is necessary to shift from conventional classes centered on the

[1] Zhao, Zhang, Ueno, Konno, and Mitsuishi (2018) also described the design principles of KoToToMo, namely how to conduct a blended learning class using this teaching material.

> transfer and injection of knowledge to autonomous learning (active learning) where teachers and students communicate, work hard together, and develop intellectually while stimulating each other, and students proactively discover problems and find solutions.[2]

The authors of this paper believe that "active learning" is the direction that CAPTs using Information and Communication Technology (ICT) should take, and that blended learning, which combines CAPTs and face-to-face classes, is a form of education that should be pursued now.

Blended learning is a new form of education in which eLearning and face-to-face teaching are designed to overcome the shortcomings of both. Not only by simply providing students with Internet-enabled educational materials, computers, tablets, and other equipment and the Internet environment. Teachers can understand and use ICT equipment to motivate and keep students engaged as well as to facilitate communication between students and teachers.

The simplicity of sharing and editing teaching materials is also a major advantage of ICT utilization. Even if the same teaching materials are used, there are many cases in which the practice questions prepared by other teachers cannot be used as they are. However, ICT technology makes it much easier to create, edit, and share teaching materials. Therefore, in this research we developed a system, Speech Training Laboratory (ST Lab)[3] that combines a teaching material creation function and a pronunciation practice function by using speech recognition technology and speech synthesis technology. Subsequently, this study uses ST Lab as blended learning material in the classroom to verify its initial effects.

3. The Development of ST Lab

3.1. The Goal of Pronunciation Education and the Definition of a Correct Answer

The ST Lab developed in this research is a pronunciation training system that specializes in teaching "speaking" and "listening" using the speech recognition and speech synthesis functions provided by the Web Speech API (application programming interface). In addition to teaching pronunciation, it also can create, edit, and manage teaching materials, which makes it significantly different from other CAPT applications.

[2] Ministry of Education, Culture, Sports, Science and Technology, Central Council for Education, (2012). "Toward a qualitative transformation of university education to build a new future: To be a university that empowers students to pursue life-long learning and the ability to think independently" (Report) p. 9. [文部科学省，中央教育審議会, (2012). 新たな未来を築くための大学教育の質的転換に向けて～生涯学び続け、主体的に考える力を育成する大学へ～（答申），本文，p. 9.]

[3] https://stlab-elearning.jp/#/login/

The Web Speech API that is used in this study was developed in 2012 by the Speech API Community Group under the W3C (FSA) Final Specification (World Wide Web Consortium).[4] The API has two interfaces: a speech synthesis interface and a speech recognition interface. Among other advantages over other software and API, it is free of charge, easy to learn even for beginners in programming, and can be implemented through JavaScript.[5]

ST Lab does not aim for the "correct pronunciation," but for the mastery of "communicational pronunciation." The speech recognition function is developed on the assumption of native speaker input. The correct speech input does not necessarily mean that the pronunciation is "correct," as described above. Still, pronunciation is considered "correct" when it reaches a level where the speaker can communicate with native speakers. In other words, it means that it was sufficiently practical and fulfilled the conditions to set it as a learning goal.

There are three types of training modes that use speech recognition technology: "Reading Aloud Practice," which reads simplified Chinese words and short sentences aloud; "Pinyin Reading Aloud Practice," which reads pinyin aloud, and; "Simulation Interpretation Practice," which inputs Japanese instructions in Chinese.[6] These voice-recognition trainers also have "help" buttons, which allow learners to hear the voice of Chinese texts throught text-to-speech technology. ST Lab is also equipped with a four-choice listening task that uses the speech synthesis function to learn the four tones [四声 (sìshēng], an aspect of learning to speak Chinese that is challenging to master.

It is preferable to listen to a native speaker's correct example repeatedly to master the correct pronunciation. However, outside of Chinese-speaking regions, recording voices by native speakers and making audio teaching materials through editing is a task that requires enormous effort, time, and money. That audio materials could not be created without this hard work may have been a factor preventing the spread of CAPT. Therefore, text-to-speech technology is used for all speech in ST Lab. Hence, it is able to create and edit audio teaching materials as efficiently as editing texts. Consequently, an environment where anyone can create audio teaching materials easily and quickly has been created.

Some may question the accuracy and acceptability of using synthetic speech for language education rather than an actual native speaker. But speech synthesis technology has already become widespread enough to be used in many everyday situations, and is widely used as a highly efficient technology. Moreover, Sunaoka and Iwami (2009), who

[4] MDN web documents: "Web Speech API". https:// developer.mozilla.org/en-US/docs/Web/ API/ Web Speech API.

[5] ST Lab currently can run speech recognition and speech synthesis function in 16 languages: Chinese (simplified and traditional), Cantonese (Hong Kong), German, Spanish, French, Italian, Korean, Dutch, Polish, Portuguese (Brazil), Indonesian, Hindi, Japan, British English, American English. It is also available to create teaching materials for these languages and to practice pronunciation.

[6] Japanese is used because it is a class for Japanese students. It can be changed to any language that allows text input.

analyzed the degree of acceptance of learners to speech synthesis, argued that learners do not feel uncomfortable with synthesized speech. However, some learners complained that the pitch of synthesized speech was too high or too fast. To resolve user dissatisfaction, a new function, the "Synthetic Voice Adjustment Menu," has been included in all training interfaces for learners. Therefore, users are now able to adjust freely "volume," "pitch," and "rate" of synthesized speech. Users can also choose between male and female voices, depending on their device's operating system.

This system also considers the popularization of pronunciation education using speech recognition. Therefore, ST Lab is designed for compatibility across different device operating systems. It can run on MAC OS and Windows, and on mobile devices using Android, iPad OS, or iOS systems so that it can be used on any PC, tablet, or smartphone.

A prototype of ST Lab was completed in March 2018, and in April it became included in Chinese language study material for first-year students at Okinawa University. It has been used every week for a year and a half. The service, which was released to the public in August, 2019, is already used by 18 universities and high schools in Japan, totaling more than 300 learners.[7]

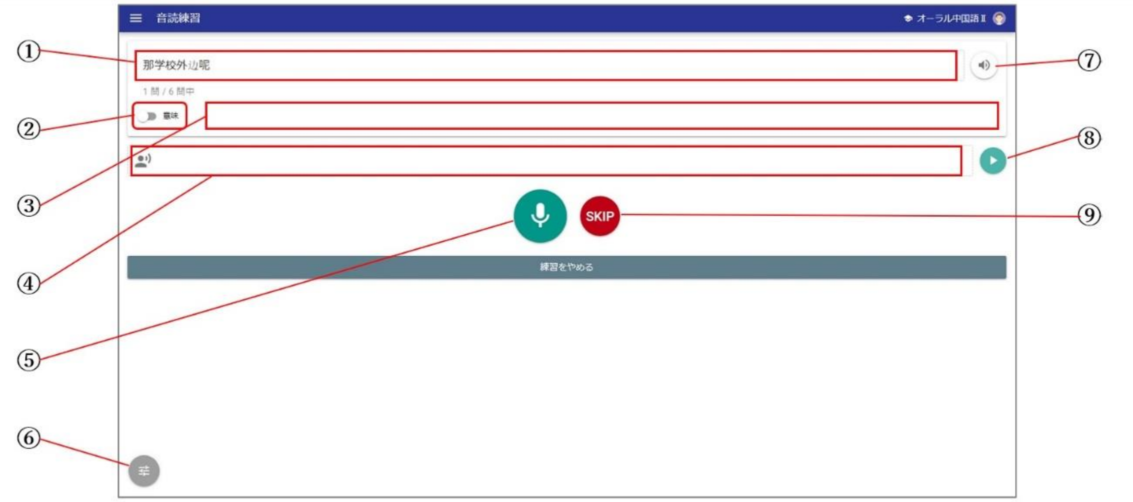
3.2 Training Menu of ST Lab

3.2.1 Reading Aloud Practice

"Reading Aloud Practice" is an exercise for reading questions written in simplified Chinese. There is also a function to read aloud the question using synthetic speech (see Figure 1, #5). The meaning can be displayed as a supplementary function (see Figure 1, #3). There is no time limit for responses, and the learner can repeat the question as many times as desired until the correct answer is given. If providing the correct answer is too difficult, a skip button (see Figure 1, #9) can be used to skip over the problem. When the synthesized voice adjustment button is pressed the adjustment menu appears (see Figure 2), and the

[7] The ST Lab currently in operation previously had a beta version and a prototype. The prototype was upgraded to a beta version in October 2018, which became the current version in August 2019. The current four practice modes have been in ST Lab from the beginning. But there were also three other menus. For details on previous versions, see Watanabe et al. (2019) and Watanabe et al. (2018).

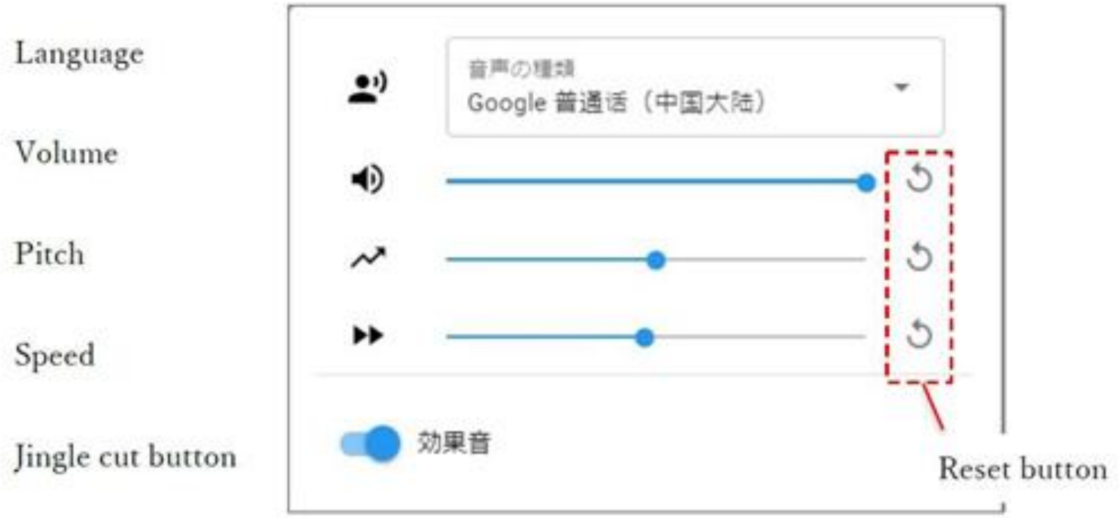
student can adjust the volume, pitch, and speed of the synthesized voice in any exercise.



那学校外边呢

Figure 1 Reading Aloud Practice screen

① Question
② Switch to display meaning
③ Meaning
④ Speech recognition result
⑤ Speech recognition button
⑥ Synthetic voice adjustment menu display button
⑦ Synthetic voice of question
⑧ Synthetic voice of incorrect answer
⑨ Button to skip the question



音声の種類
Google 普通话（中国大陆）

Figure 2 Synthesized Voice Adjustment screen

The training menus used in the classroom are “Reading Aloud Practice” and “Simulation Interpretation Practice.” The questions are created by the teacher using the ST

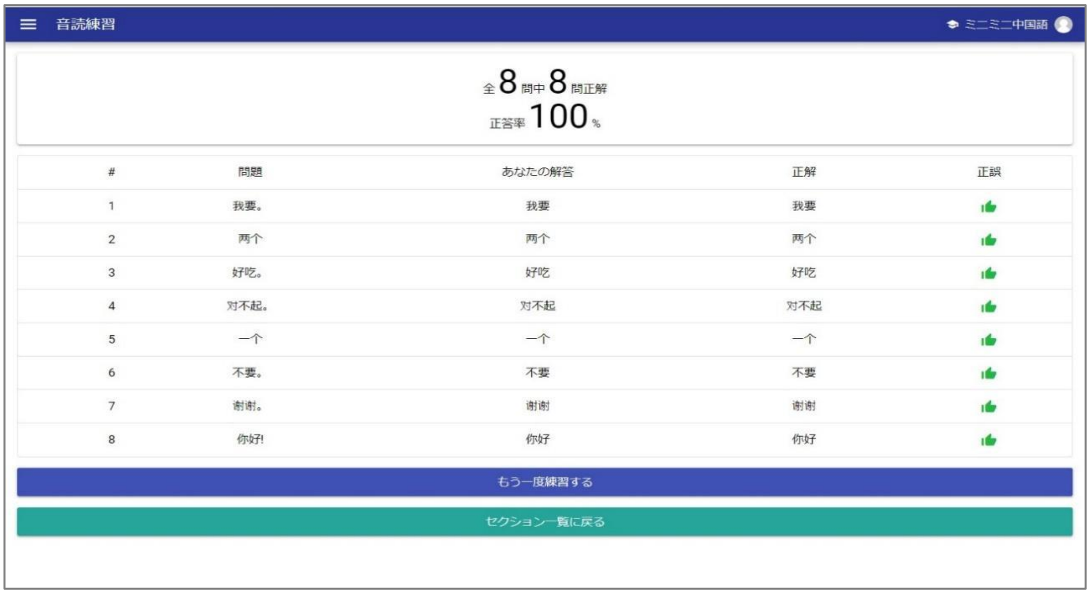
Lab material creation function based on the textbook used for the class. The addition of a material creation function is unique to ST Lab and was created by the author as this function is not available in any other CAPT system. "Reading Aloud Practice" includes "applied practice," in which each lesson's text and the grammatical points of that lesson are learned using previously learned words. When using this mode, students first learn sentence patterns in a standard lecture format, then practice pronunciation by chorus reading, and finally use ST Lab to read aloud the lesson ("Reading Aloud Practice"). Students who have reached 100% in providing the correct answer can go on to applied practice and follow up with "Simulation Interpretation Practice" to measure their understanding. The practice items are prepared so that students can practice according to their progress. Additionally, to make it easier to overcome students' weaknesses through focusing practice on the pronunciations that are found challenging during the learning process, a "weakness to overcome exercise" option is also included in the "Pinyin Reading Aloud Practice."

All questions are given randomly, and one problem is a set of 6-8 items. It is designed so that a student can repeatedly attempt a wrongly-answered question over and over again in a consecutive interval of time. Also, as mentioned above there is no time limit, and even if the wrong answer is repeated many times as long as the correct answer is entered in the end the item will be considered as the "correct answer." This is a form of repeated practice until the learner can answer correctly without fear of mistakes.

The speech recognition result is displayed in the answer text box (see Figure 1, #4). If it matches the question sentence, the "correct answer" message appears along with a chime and moves to the next question. If the answer is incorrect, an "incorrect" message appears, and the recognition result will continue to be displayed. The learner can listen to the wrong answer using the speech synthesis function by pressing button #8 seen in Figure 1. This function makes it easier to quickly realize what is incorrect in their pronunciation by comparing the difference between the correct answer and the student's incorrect answer.

The role that teachers play in this pronunciation practice is also essential. For example, they could help a student who cannot hear the sounds from the headset properly or who has problems with answering correctly by using the patrol function of the CALL system. Teachers can check answers and give feedback in the case that it is incorrect. They can also make suggestions about pronunciation if a wrong answer is given. For example, if a learner has entered 徒弟 [túdì; apprenticeship] for a question that is pronounced 土地 [tǔdì; land], the teacher can point out that the tone of the first syllable is incorrect and give the correct answer. This way, there is continuous communication between the teacher and student, which motivates the student to set clear goals and encourages him or her to keep practicing until the input is correct, even after several failures. This way of interaction makes the process of learning a fun experience.

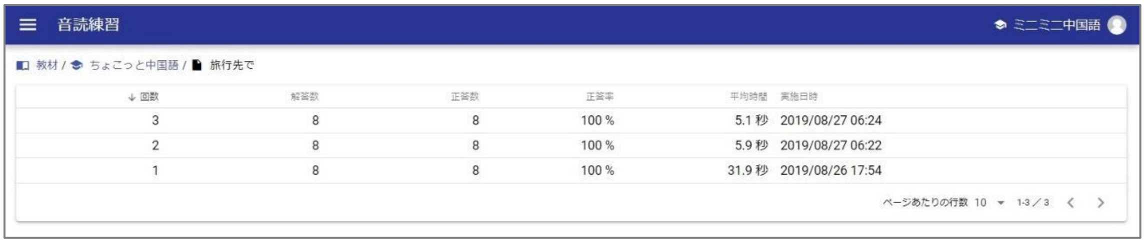
After completing a set of questions, students see their score (see Figure 3). All of these grades are recorded for each question, and students can quickly check past grade changes from the questions interface (see Figure 4). Also, the average time that was necessary to submit the correct answer is also displayed on this screen, and it is possible to check the progress of pronunciation learning that cannot be known only by the correct answer rate.



全8問中8問正解
正答率100%

#	問題	あなたの解答	正解	正誤
1	我要。	我要	我要	👍
2	两个	两个	两个	👍
3	好吃。	好吃	好吃	👍
4	对不起。	对不起	对不起	👍
5	一个	一个	一个	👍
6	不要。	不要	不要	👍
7	谢谢。	谢谢	谢谢	👍
8	你好!	你好	你好	👍

Figure 3 Accuracy Rate Display screen



教材 / ちょこっと中国語 / 旅行先で

↓回数	解答数	正答数	正答率	平均時間	実施日時
3	8	8	100 %	5.1 秒	2019/08/27 06:24
2	8	8	100 %	5.9 秒	2019/08/27 06:22
1	8	8	100 %	31.9 秒	2019/08/26 17:54

Figure 4 Learning History screen

3.2.2 Pinyin Reading Aloud Practice

"Pinyin Reading Aloud Practice" was created using the same template as "Reading Aloud Practice." The difference, however, is that pinyin is displayed in the text box that displays the Chinese sentence that forms the question in "Reading Aloud Practice." Other features, such as providing correct answers, reading incorrect answers, adjusting synthesized speech, skipping questions, etc. are the same as "Reading Aloud Practice."

Pinyin is not only a Chinese notation system.[8] It is also used for computer input. Therefore, it is important that Chinese learners master pinyin in the introductory stage. However, because Japanese learners can understand kanji, it is rather easy for them to understand the approximate meaning. So they frequently neglect practicing pronunciation. Therefore, "Reading Aloud Practice" includes exercises using simplified Chinese characters and "Pinyin Reading Aloud Practice" using pinyin, both of which are set as

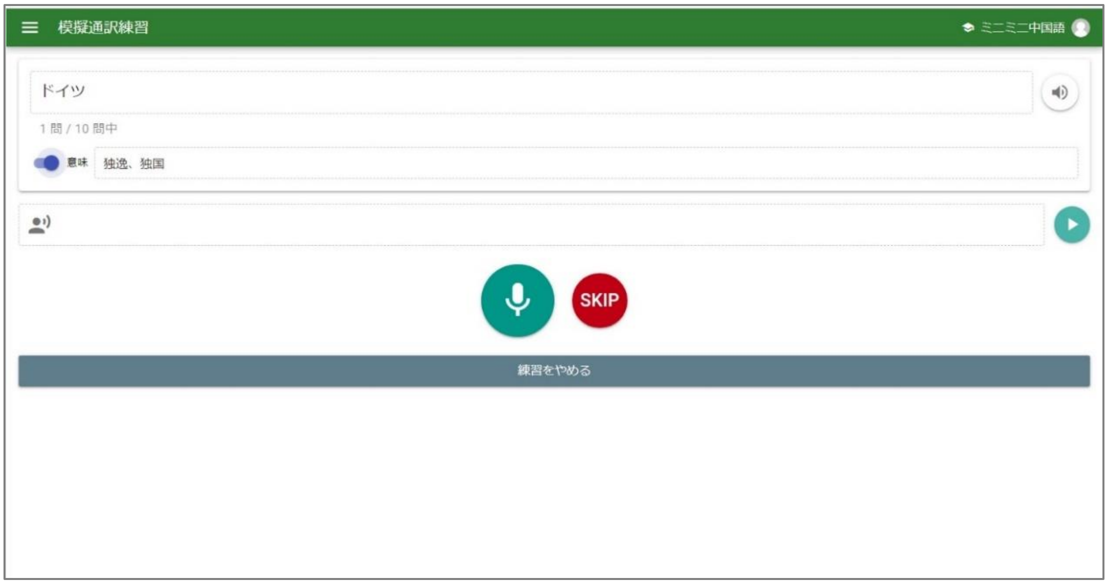
[8] There are other Romanization systems for Chinese, such as Wade-Giles and Yale, but they are not currently used in Chinese language education in Japan. In Taiwan, Zhuyin phonetic characters {注音符號 zhùyīn fúhào] are used instead of Romanized letters to transcribe Chinese sounds. This system is not taught in Chinese language education in Japan either.

separate practice items.

3.2.3 Simulation Interpretation Practice

Even if a learner can reproduce a sound, it does not mean that the learner has mastered pronunciation unless one pronounces correctly and understands the meaning. For this reason, a mode of practice, “Simulation Interpretation Practice,” is also included in ST Lab. It displays Japanese, asks the student translate the text into Chinese, and record it as speech.

“Simulation Interpretation Practice” uses the same template as “Reading Aloud Practice.” But Japanese text is displayed instead of simplified characters in the text box where the question is posted. There are several ways of interpreting. However, since only one correct answer can be set in ST Lab, the exact answer provided to the learner is also a hint for learners to answer correctly. Other functions are the same as in “Reading Aloud Practice.”



ドイツ

1 問 / 10 問中

意味 独逸、独国

Figure 5 Simulation Interpretation Practice screen

3.2.4 Four Tones Listening Practice

The four Chinese tones are fundamental to Chinese pronunciation and are the first and most significant difficulty that learners encounter when learning Chinese sounds. Also, many students cannot speak Chinese well because the tones are challenging to master. Even if they have mastered grammar items to some extent they cannot complete their tone learning. The authors believe that tone learning practice is imperative for learning adequately. Hence, a “Four Tones Listening Practice” mode is included in ST Lab.

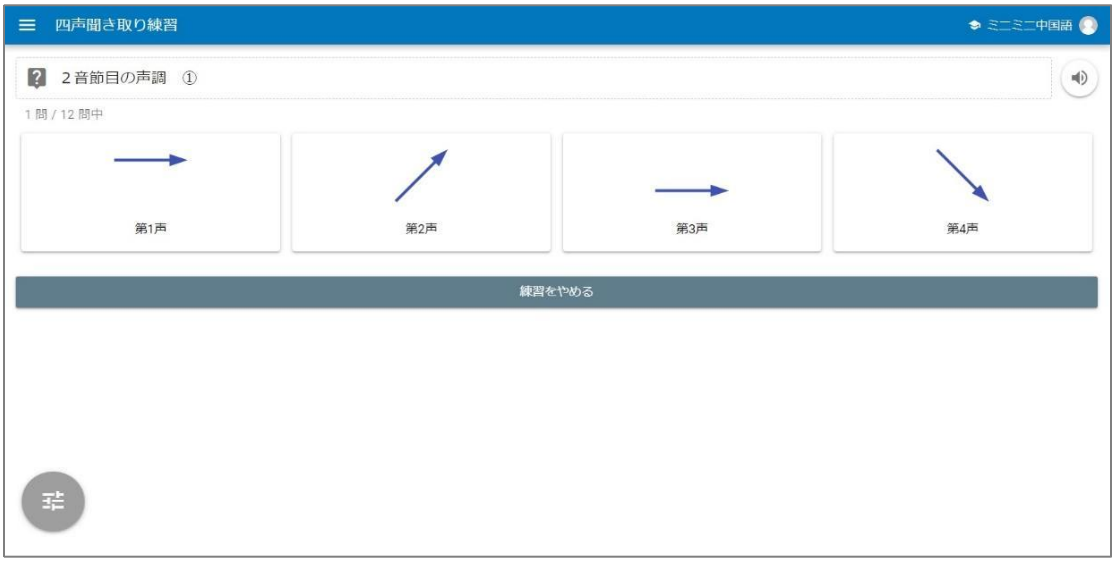


Figure 6 Four tones listening practice

In this mode, no letters are given, and the student learns to distinguish between the four tones, relying solely on the synthesized speech. In general, vowels and “ma” sounds are used as teaching materials to distinguish the four tones. In addition, ST Lab has prepared exercises using another voice to identify four different tones. Having more types of exercises available makes it easier for learners to discern the four tones of more types of sounds.

Learners listen to synthesized speech and answer a four-choice question. The synthesized speech can be heard as many times as necessary until it is answered. Once an answer is selected, a “correct answer” or “incorrect answer” message will appear along with the chime, and the next question will proceed regardless of the answer.

Each problem is 12 questions per set. After completing one set, the correct answer rate is shown as in the other exercises, and the learner can check what tones were missed to know where their own deficiencies in listening to the four tones are. It is also possible to check past learning history, the same as in other exercises.

The four tones form the basis of Chinese pronunciation and should be practiced over and over in the learning process. However, Japanese students often stop exercising after they mastered the differences between the four tones of a single vowel or one syllable. Such insufficient pronunciation practice is also a factor that strengthens the learner’s notion that “Chinese pronunciation is too difficult to master.” Thus, at Okinawa University, practice of the four tones is done occasionally and repeatedly throughout the whole year.

3.3 Interface for Management of Course and Materials by the Teacher

3.3.1 Organization, Subject, and Course Management Interface

The teacher’s screen has two interfaces: subject and course management and creation of teaching materials. In either case, anyone who has administrator privileges can access them. At present, students are not allowed to make their own practice questions.

“Organization,” “Subject,” and “Course” have a hierarchical structure. Multiple “Subjects” can be created under one “Organization,” and numerous “Courses” can be created in one “Subject.” There are no restrictions on the number of sections. Only administrators can add and manage “Organizations.”

Only the administrator can manage the user list of the entire system, and users with teacher status can browse the student list of the course they are in charge of. Teachers can also delete or add students if necessary. However, students who can be added are limited to those who belong to the same organization and have already registered. The student list is automatically created by adding names when a student registers. If the teacher has privileges, a student can be deleted from or added to the course attendee list.

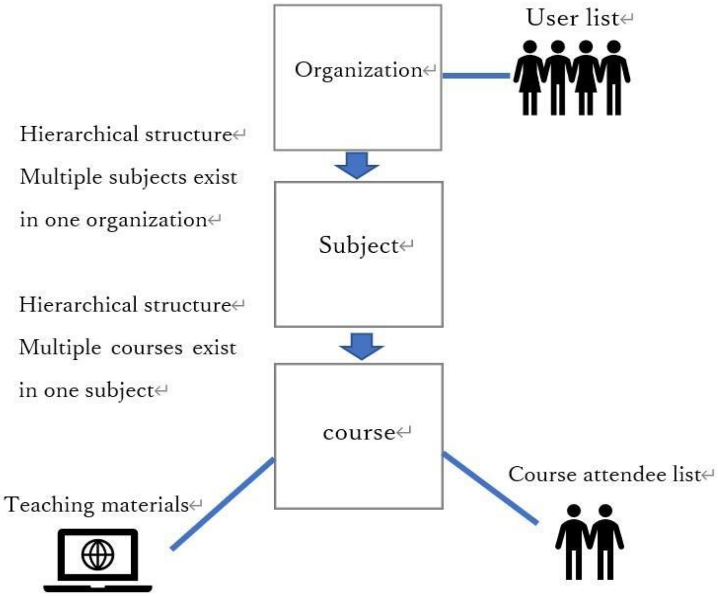


Figure 7 ST Lab object conceptual diagram

Teaching materials are linked to the course (see Figure 7). Instructors with teacher privileges can select teaching materials from a drop-down menu. These materials can be set to “Open to all” or made available only in their own course.

3.3.2 The Teaching Materials Creation Interface

The ability to create teaching materials is a unique feature not found in other pronunciation training software. With this feature, ST Lab allows any teacher to develop exercises based on any textbook or material. Only teachers and administrators have the authority to create teaching materials, and only the creator has the power to edit generated content.

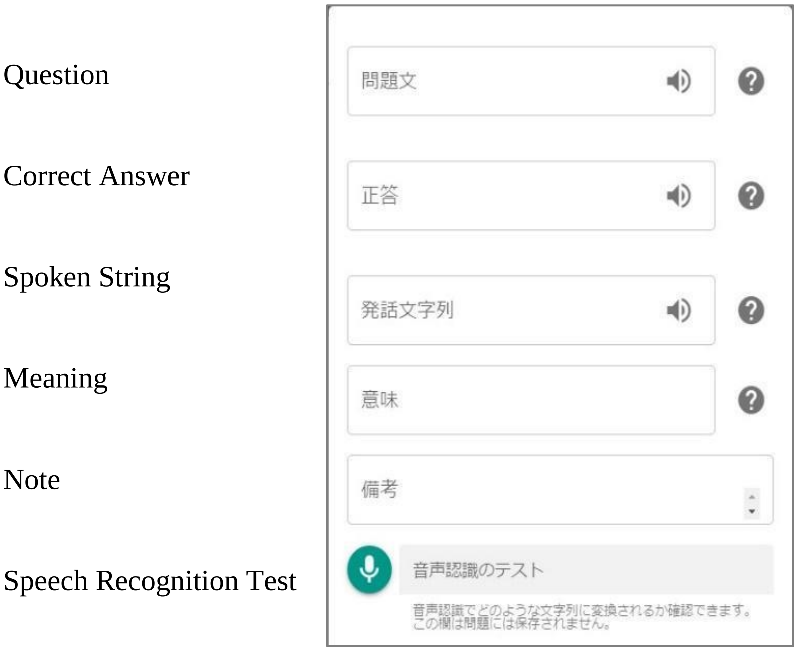
To create teaching material, users first select a language to use for the speech recognition and speech synthesis from the drop-down menu. They then choose the scope of the teaching material to be made available from one of the following: “All courses,” “Only the courses I am responsible for,” or “Only myself.”

Although it is possible to create hierarchical teaching materials (there is no limit to the number of levels either), it is still recommended that teachers develop only a few levels. This will make it easier to share materials with other teachers.

The teaching materials creation interface is composed of “Question,” “Correct Answer,” “Spoken String,” and “Meaning” (see Figure 8). Among them, “Question” and “Correct Answer” are indispensable items for “Reading Aloud Practice.” In “Pinyin Reading Aloud Practice” and “Simulation Interpretation Practice,” “Question” and “Correct Answer” are different, so three items up to “Spoken String” are required. Because of this, the priority of speech synthesis in this system is set as “Question” < “Spoken string.”

“Meaning” is not an essential item, but learners should understand meaning while mastering pronunciation. Therefore, it is recommended that teachers include “Meaning” in “Reading Aloud Practice” and “Pinyin Reading Aloud Practice.” Also, note that to ensure that ST Lab can be operated on any OS, punctuation is removed from the “Correct Answer” string and the string captured by speech recognition before the two are compared.

When the speaker button is pressed in the question creation window, the synthesized voice can be heard. Voice recognition can be run by pressing the microphone button below it. This is a function to check whether synthesized speech and speech recognition are operating correctly. This feature was added because the need for it was realized through actual practice. For example, “1:45” can be written as “一点三刻” in Chinese, but in real speech recognition it will be displayed as “1 点三刻.” “一点三刻” is correct in Chinese, but the system determines that the answer is incorrect, so the correct answer must be reset to “1 点三刻.” This feature was added to allow a check in advance to avoid problems with speech recognition and speech synthesis that may occur when creating materials.



音声認識でどのような文字列に変換されるか確認できます。
この欄は問題には保存されません。

Figure 8 Question creation interface

Since "Four Tones Listening Practice" does not display "Question," the required items are "Correct Answer" and "Question" or "Spoken String," and "Correct Answer" is a half-width number. Each problem has 12 questions so that learners can listen to each tone three times in a four-choice listening problem. If four questions that differ only in tone are repeated, just set a group of queries made of four different tones and automatically replicate the issue three times at random. All items could also be arranged in different speech sounds.

4. Initial Outcomes of ST Lab on Learning Pronunciation

Since April 2018, ST Lab has been used for a year and a half in "Oral Chinese I" (first semester) and "Oral Chinese II" (second semester), which are common subjects for first-year students at Okinawa University. During the first semester of 2018 and the first semester of 2019, there were two class sessions a week with 25 students. None of the students had any previous Chinese learning experience. In the first semester of 2018, the first six weeks were used to learn pronunciation starting with "four tones" followed by practice of conversational sentences.

To make the practice of meaningless syllable pronunciation as short as possible, the speech-only lesson ended in half a month, a month earlier than usual. After that students started using "Reading Aloud Practice." The set of questions practiced was 16 combinations of 4 tones, excluding the neutral tone [轻声 qīngshēng], each with two words, totaling 32 words. Including explanations of pronunciation and meaning as well as demonstrations, this took two class periods (3 hours). In these three hours, 25 students practiced 32 words 11,198 times (an average of 448 times per person), and this demonstrates that ST Lab promotes autonomous learning.

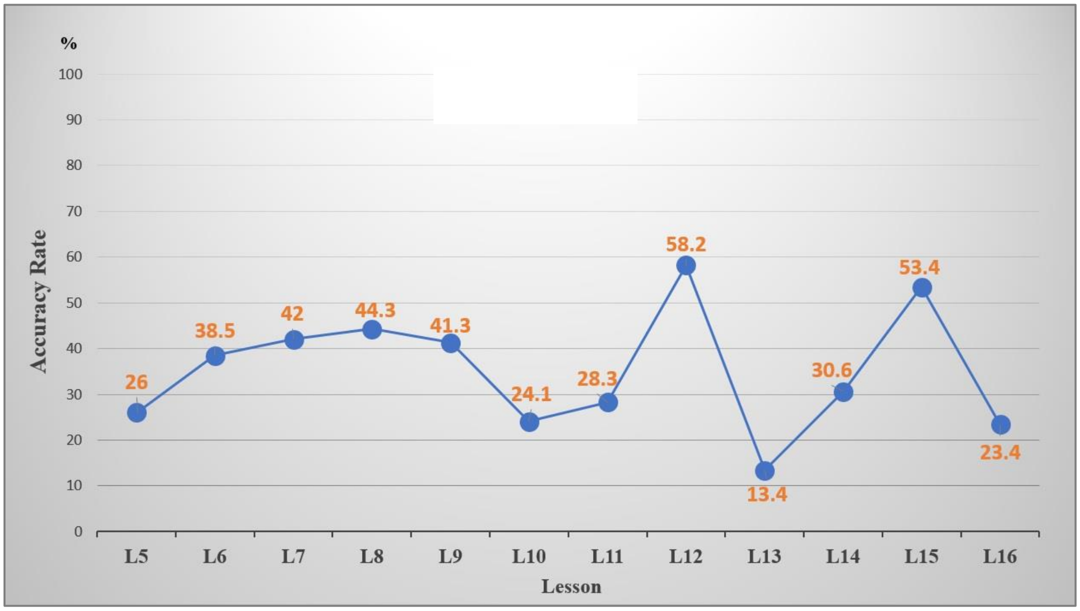


Figure 9 Accuracy rate for “Reading Aloud Practice” for each lesson in the first semester of 2018

Also, in the first semester of 2019, the correct answer rate did not increase as quickly as in the graph (see Figure 9). Although the correct answer rate is low in the practice of reading aloud in “Oral Chinese I,” it can still be seen that learners continue to maintain high motivation to succeed in the following practice sessions despite drops in correct answer rates during some exercises (see Figure 9).

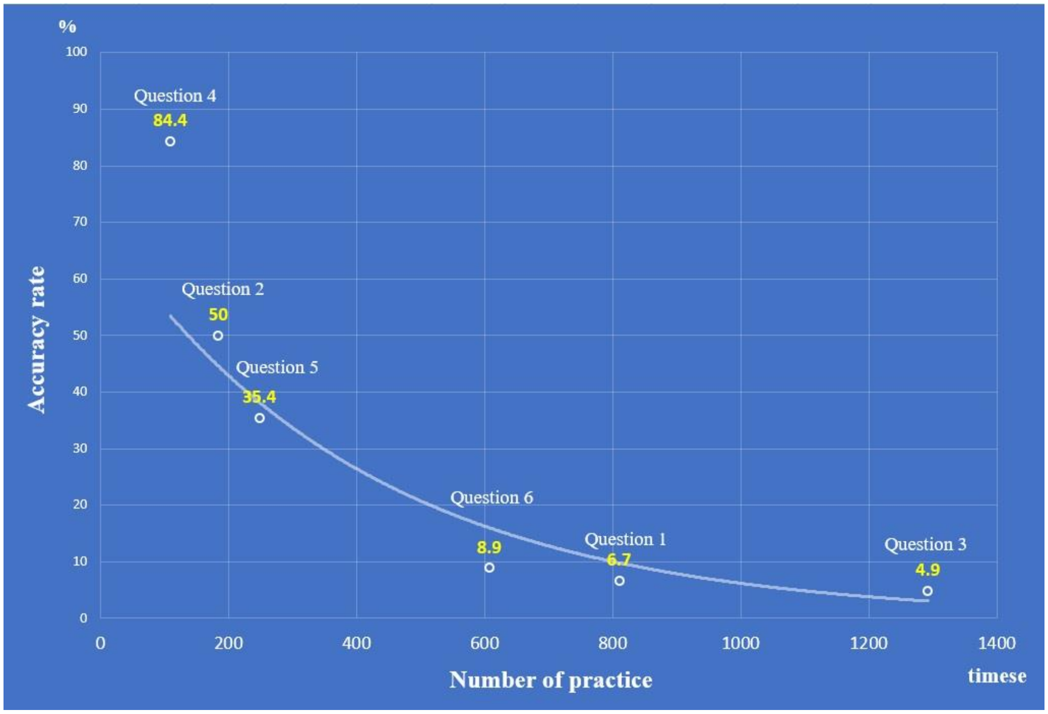


Figure 10 **Correlation between the accuracy rate and the number of practice sessions of the 6 questions of Lesson 13 in the first semester of 2019**

Figure 10 shows the correlation between the average accuracy rate of the six exercises in Lesson 13 and the total number of practice attempts performed by 25 students in the first semester of 2019. These correlation coefficients correlated strongly with r = -5.6. This means that the lower the correct answer rate, the more learners increased the number of times they practiced. If the learner did not get the right answer, the learner did more practice voluntarily. It is an ideal form of autonomous learning.

Also, in the last class of "Oral Chinese I" in the first half of the 2019 semester, a class evaluation questionnaire was conducted for all 25 students using Google Forms. In response to the question "do you think that pronunciation practice using the voice recognition function is effective?" nineteen students answered, "I strongly agree," and two students answered, "I think so." In total, 96% of students in the class thought that "practice using the speech recognition function is useful for learning pronunciation" (see Figure 11).

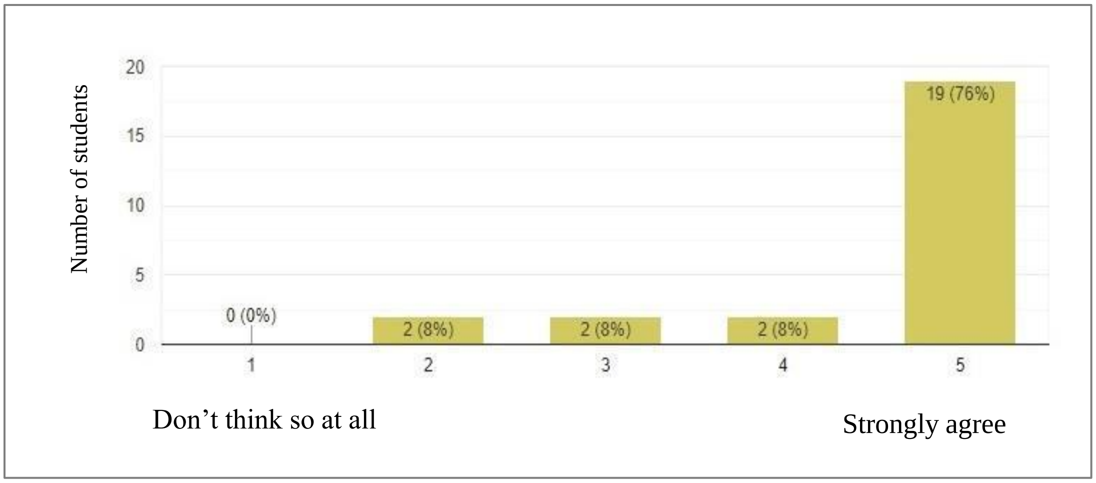


Figure 11 Questionnaire results on the effectiveness of speech recognition for learning pronunciation

The following are the results of asking “why it was useful” for students who replied that ST Lab speech recognition is useful for learning pronunciation (multiple answers allowed).

- Because users can know if their pronunciation is correct or not right away (21 people; 87.5%)
- Because users can see where the problem is in their pronunciation (17 people; 70.8%)
- Because the results are immediately known, so practice can proceed effectively (12 people; 50%)
- Because users can hear the sound of the correct answer repeatedly (19 people; 79.2%)
- Because users can concentrate without worrying about other people seeing them (4 people; 16.7%)
- Because users do not feel it is hard to repeatedly practice (7 people; 16.7%)
- Because users can feel that their pronunciation is improving (13 people; 54.2%)

ST Lab has been used in the class for only one and a half years, and the number of samples is not large enough. Nevertheless, pronunciation learning using a system (such as ST Lab) that provides immediate feedback on pronunciation shows in these preliminary results that it keeps the learner motivated to practice pronunciation and makes it easier to achieve a sense of accomplishment with the practice results. The above data and questionnaire results are considered to provide initial evidence to demonstrate these connections.

There were no students who complained about synthesized speech. To the question, “do you think that the speech synthesis function was useful for pronunciation practice?” twenty students (80%) replied, “I strongly agree,” and four students (16%) answered, “I think so.” This shows that the learners feel not only comfortable but also see synthesized speech actively and positively (see Figure 12).

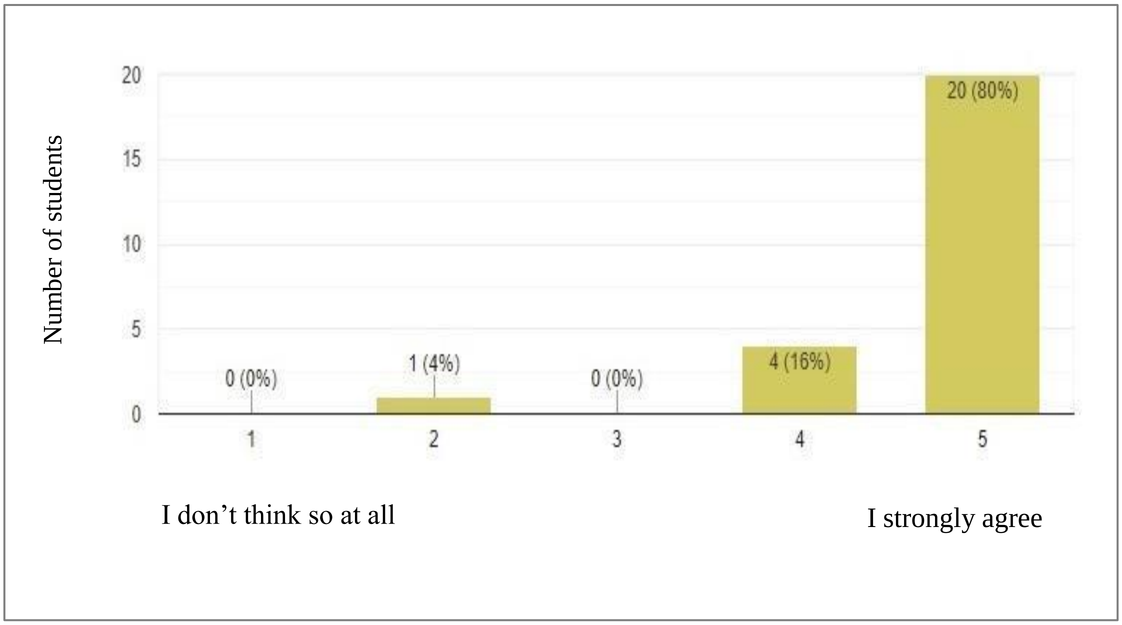


Figure 12 Questionnaire results on effectiveness of synthetic speech for pronunciation learning

In class, students were asked to practice listening using a different eLearning system as an extracurricular activity, and pronunciation practice was not required as an extracurricular activity. Despite this, two students (8%) voluntarily "practiced frequently outside school," and 17 students (68%) answered, "sometimes practiced." In total, nearly 80% of students practiced pronunciation independently.

Furthermore, when the students who practiced pronunciation outside school were asked, "did you use ST Lab during extracurricular practice?" it was found that students were also practicing pronunciation using the ST Lab outside school. This data also provides preliminary evidence that pronunciation learning using ST Lab can help students undertake "autonomous learning."

5. Conclusion

In this paper, we examined whether teaching Chinese pronunciation using speech recognition software is useful for learning pronunciation.

In order to answer this question, a new CAPT platform has been created, ST Lab, which provides new features not found previously in other platforms. A new feature that allows teachers to create and edit teaching material enables the class to work from any textbook and from any other language because the teacher can create exercises in ST Lab. Another addition is the inclusion of the "Simulated Interpretation Practice" feature, which allows ST Lab to help students practice the meaning of Chinese words and phrases where other platforms cannot. Even if students do not understand the meaning of a Chinese character, they can still learn the character through pinyin and listening to speech synthesis.

ST Lab can truly help students of different language backgrounds to learn Chinese because menus, prompts, and "Simulated Interpretation Practice" questions can be configured in any language to display text and synthesize speech while the speech recognition remains only in Chinese. Last, ST Lab is designed to work with speech recognition on any OS because it strips punctuation from both the teachers' "correct answer" and the result of speech recognition before the two are compared.

Utilizing ST Lab in the classroom, data was collected so as to ascertain the effectiveness of speech recognition and ST Lab's features to help Chinese-language learners. In the combination practice of two-syllable words in May 2018, 25 students practiced 11,198 times over 38 questions in less than 3 hours of class time. In the lessons from the first semester of 2019, it was found that the number of repeat attempts for exercises did not decrease, even for problems with a low average correct answer rate in the class. The students voluntarily repeated the exercise until the correct answer was given. Furthermore, there was a correlation between practice time becoming longer as speech input requested from students was becoming more difficult. Pronunciation practice is similar to sports, and it is traditionally believed that many repetitions of listening and speaking exercises will lead to progress.

Based on the questionnaire results, students believe that it is useful and practical to practice using a speech recognition function that can instantly determine whether the answer is correct or not. Additionally, the result of the survey demonstrates more than half of the students felt the system was useful and they used the system voluntarily during extracurricular hours. Both of these responses from students are desirable because a belief in the effectiveness of practice, together with a willingness to practice during free time, indicates that students will continue pronunciation practice independently.

From the above information, the authors found that teaching Chinese pronunciation using speech recognition is useful; it improves the learners' motivation and helps them to undertake "autonomous learning."

6. Future work

The first tasks are to increase the number of teachers using the system, to continue teaching with the system, to collect more data, and to prove the learning effectiveness of the system more scientifically. As mentioned earlier, this system has been available to the public since August 2019 and already has more than 300 users. In the future, the accumulation of data is very likely to provide more empirical evidence.

Still, there is room for further improvement. First, to make the teachers' responses more active, the cause of the incorrect answer derived from the wrong answer needs to be more scientifically analyzed. There is also a need to develop a system manual. For example, the learner's incorrect answer is not just a mistake of misunderstanding "徒弟" (apprenticeship) as "土地" (land). Because it occurs in multiple syllable tones, initials, and finals, the "incorrect answer" shows much different information. One author has used it for over a year and a half, so it is known from experience what advice to use based on the content of the "incorrect answer" and what priority to use. However, it is difficult to share experiences because it is not organized intuitively. More effective feedback requires theoretical and systematic error analysis.

Currently, only students have access to their learning history. Teachers cannot access it, which creates a decrease in teacher-student interactions and understanding of student's progress. Teachers need to keep track of each student's learning progress to give the right guidance to every student and to get a better idea of their learning abilities. By using data from a server, what difficulties every student encounters while learning would be revealed, as well as how frequently they practice. It will be helpful in the future to have this data displayed graphically. That way, teachers will be able to follow the students' progress and the correct answer rate. It will also contribute to a more accurate understanding of learning progress. This feature is planned to be available in the next phase.

Future work also includes the implementation of a materials sharing function. As mentioned above, ST Lab can make the teaching materials settings "open to the public" and share them with other teachers. It is relatively easy to create teaching materials using the ST Lab's teaching material creation function. Still, it is unproductive for many teachers to develop similar materials from scratch. It is particularly easy to share such materials as "Four Tones Listening Practice." It is aimed to implement a material sharing community that will allow every teacher to add documents to existing ones.

Acknowledgements: This work was supported by JSPS KAKENHI Grant Number JP19K00879.

Reference

Furui, S. (2000). Perspectives on speech recognition technology. *Journal of the Phonetic Society of Japan*, *4*(3*)*, 60-63. [古井貞煕. (2000), 音声認識技術の現状と課題. *音声研究, 4*(3), 60-63.]

Hayashi, T., Mizuochi, Y., Kiryu, T., & Kanzaki, H. (2012). A case study on translation activities by the speech recognition function of a tablet type terminal in elementary schools foreign language activities. *Japan Journal of Educational Technology*, 36 (Suppl.), 45-48. [林俊行, 水落芳明, 桐生徹, 神崎弘範. (2012). 小学校外国語活動におけるタブレット型端末の音声認識機能による翻訳活動に関する事例的研究. *日本教育工学会論文誌*, 36(Suppl.), 45-48.]

Iwai, H. (2013). Practice and research on German voice training with a speech recognition application. *Journal of Osaka University Graduate School of Human Science*, 2, 11-18. [岩居弘樹. (2013). 音声認識アプリを用いたドイツ語発音学習の実践と検証. *大阪大学高等教育研究*, 2, 11-18.]

Iwai, H. (2015). German voice training with a speech recognition application.

Osaka University Higher Education Studies, *3*, 1-15. [岩居弘樹. (2015). 音声認識アプリを活用したドイツ語発音トレーニング. *大阪大学高等教育研究*, *3*, 1-15.]

Li, Z. (2016). Trend of ICT utilization in foreign language education in Japan. *Journal of Osaka University Graduate School of Human Science*, *42*, 329-341. [李哲. (2016). 日本の外国語教育における ICT 活用の研究動向. *大阪大学大学院人間科学研究科紀要*, *42*, 329-341.]

Schwartz, A., Huang, C., Tao, J., Van T., Wolf, P., & Harada, Y. (2009). Evaluation the use of speech recognition in CALL systems. *The Technical Report of The Proceeding of the Institute of Electronics*, *109*(297), 29-34. [Schwartz, Alan, Huang, Caroline, Tao, Jidong, Van Thong, J.M., & Wolf, Peter, 原田康也. (2009). ビデオ配信と音声認識を活用した英語学習システム：ＣＡＬＬシステムにおける音声認識利用の評価. *電子情報通信学会技術研究報告*. *TL, 思考と言語*, *109*(297), 29-34.]

Sunaoka, K., & Iwamida, H. (2009). Listening education and its effect using Chinese Text-to-speech —Changing language learning tools and learner's acceptability—. *Computer & Education*, *27*, 28-32. [砂岡和子, 岩見田均. (2009). 中国語合成音利用の聴取教育とその効果 —変わる語学ツールと学習者の受容能力—. *コンピュータ＆エデュケーション*, *27*, 28-32.]

Tanabe, T. (2004). A study on the design of computer teaching materials in Chinese language education: Development of tone and autonomous learning materials in the introduction period. *Doshisha Studies in Language and Culture*, *7*, 49-64. [田邉鉄. (2004). 中国語教育におけるコンピュータ教材デザインに関する考察—導入期の声調自律学習教材の開発—. *言語文化*, *7*, 49-64.]

Watanabe, Y., & Omae, T. (2018). Developing Multilingual Learning Materials to Encourage Speech – The Potential of Speech Recognition and Synthesis APIs in Foreign Language Education. *2018 PC conference*, 56-59. [渡邉ゆきこ, 大前智美 (2018). 発話を促す多言語教材の開発–外国語教育における音声認識・合成 API の可能性. *2018 PC Conference 論文集*, 56-59.]

Watanabe, Y., & Omae, T. (2019). Development of Chinese phoneme retrieval system and trial of efficient pronunciation learning using speech recognition function. *2019 PC Conference*, 35-38. [渡邉ゆきこ, 大前智美. (2019). 中国語音韻検索システムの開発と音声認識機能を使った効率的発音学習の試み. *2019 PC Conference 論文集*, 34-38.]

Wei, W., & Zhang, J. (2018). An intelligent Chinese pronunciation teaching App and the preliminary result of a teaching experiment. *Journal of Technology and Chinese Language Teaching*, *9*(2), 83-97. [魏巍, 张劲松. (2018). 一款汉语智能语音教学 App 及教学实验初步结果. *Journal of Technology and Chinese Language Teaching*, *9*(2), 83-97.]

Zhao, X., Zhang, L., Ueno, T., Konno, F., & Mitsuishi, T. (2018). The practice of blended learning for beginning learners of Chinese at Tohoku University: Evaluation of the development of the textbook "KOTOTOMO". *Bulletin of the Institute for Excellence in Higher Education, Tohoku University*, *4*, 149-163. [趙

秀敏, 張立波, 上野稔弘, 今野文子, 三石大. (2018). 東北大学初修中国語におけるブレンディッドラーニングの実践: 開発した教科書『KOTOTOMO』の検証を中心として. *東北大学高度教養教育・学生支援機構紀要*, *4*, 149-163.]

Zhao, X., Tomita, N., Konno, F., Ohkawa, Y., & Mitsuishi, T. (2019). Development and practice of review material KoToToMo for use on smartphones in blended learning by beginning learners of Chinese in university. *Transactions of Japanese Society for Information and Systems in Education*, *36*(2), 131-142. [趙秀敏, 冨田昇, 今野文子, 大河雄一, 三石大. (2019). 大学初修中国語ブレンディッドラーニングのためのスマートフォン利用復習教材「KoToToMo」の開発と実践. *教育システム情報学会誌*, *36*(2), 131-142.]