Implementation of an Individual English Oral Training Robot System

Chen-Yu Lin[1], Wei-Wei Shen[2], Ming-Hsiu Michelle Tsai[2], Jim-Min Lin[1](✉), and Wai Khuen Cheng[3]

[1] Department of Information Engineering and Computer Science, Feng Chia University, Taichung City 40724, Taiwan
qwell1845@gmail.com, jimmy@fcu.edu.tw

[2] Department of Foreign Languages and Literature, Feng Chia University, Taichung City 40724, Taiwan
{wwshen,mhtsai}@fcu.edu.tw

[3] Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Malaysia
chengwk@utar.edu.my

Abstract. To improve oral English ability, in addition to learners' willingness to practice more, the learning effect will be more obvious if tutor assistance is provided and can involve one-on-one individual tutoring. However, due to the scarcity of English teacher manpower, teachers cannot take care of every student in class, nor can they teach students one by one after class. Robot-Assisted English Speaking may provide a feasible solution. Therefore, this research has developed an educational robot system that can support individual tutoring of English speaking after class. It is called "English Oral Training Robot Tutor System (EOTRTS)", which can actively lead students to learn through robots, and help students improve their oral English practice through individual tutoring and interactive methods. The implementation of this system is to use the social robot NAO's voice recognition, QR code scanning, humanized limbs, and various sensor functions as well as the ability to interact with people and other features to help students learn oral English after class. The experimental results show that, in addition to a slightly lower satisfaction with the robot's gesture performance, the students' acceptance of the EOTRTS system is promising.

Keywords: Educational robot · Robot assisted language learning · English Oral Training Robot Tutor System (EOTRTS) · Google Cloud Automatic Speech Recognizer (ASR) · NAO

1 Introduction

With the development of technology, many robots have been used in real life today, and the existence of social robots allows us to interact with them through dialogue or touch. The interaction with the robot has a positive impact on the body and mind [1]. Mubin et al. [2] also enumerate the various advantages of using robots in a teaching environment, such as dull and boring tasks without fatigue and remote teaching. Kang et al. [3] assumes that the robot is humanized and has the proper appearance to interact directly with the person. Extending these ideas, robots are an attractive tool that is brought into the second language learning arena and used to meet different teaching needs. Kanda et al. [4] and others placed robots in the primary school classroom for two weeks and compared the frequency of interactions between students and their English test scores. Although the two-week robot-assisted learning did not have any significant impact on students' oral English and listening test scores, students who showed great interest at the initial stage had significantly improved English scores.

In the previous research, most of the robots were used in the classroom. The purpose of using robots in the classroom was to attract students' attention and enhance the interest of learning. An ETAR system [4] was implemented to be a teaching assistant in class. It can help to improve the motivation of English learning in class and bring convenience to classroom teaching for English teachers. Research [5] was also proved to be effective. However, the robot cannot take care of every student in the classroom; likewise, after the class, the teacher cannot teach each student individually. Therefore, a robot system for after-school tutoring, called English Oral Training Robot Tutor System (EOTRTS) [6] is proposed. This paper presents how the system is implemented. In this system, the educational robot leads the students to learn in a one-to-one manner, using a social robot NAO with interactive functions such as voice function and touch as a personal lecturer for the learner to conduct after-school tutoring for oral English training. Different from the past research, the robot instructor can actively guide the learner to learn. It does not need to be triggered by the helper to guide the coaching process. Instead, the robot instructor NAO actively guides the students according to the progress of the student's curriculum to practice single words, recitation of texts, recording exercises, and evaluate the student's speech rate and correct reading rate through Google Cloud Voice to Text and Word Error Rate Calculation Formula.

The structure of this article is as follows. Section 2 explains how the system is designed, and describes the design of the student's interaction with the robot. The details of the experimental procedures is described in Sect. 3. Next, the analysis and discussion of the result are then explained in Sect. 4. Finally, a brief conclusion is made and the future research direction is explored in Sect. 5.

2 System Design

2.1 System Structure

Figure 1 shows the system structure of the educational robot (called EOTRTS) proposed in this study. This article uses NAO robots to conduct after-school tutoring and lead students to learn text materials. EOTRTS is mainly divided into three parts: cloud services, servers, and robots. The system software functions are implemented in Python language, and Socket is used as a network communication mechanism to connect various system components. Cloud services use Google Cloud services.

Server. The server is responsible for accessing the database and the cloud, obtaining course data to send to the robot, recording the session test data returned by the robot, etc., and respectively storing each student's recording files after reading the text. After

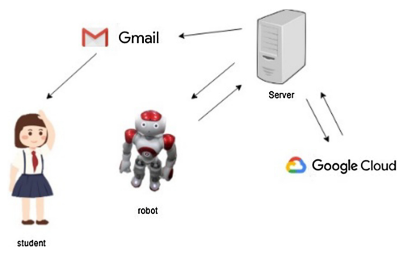


Fig. 1. System structure of EOTRTS.

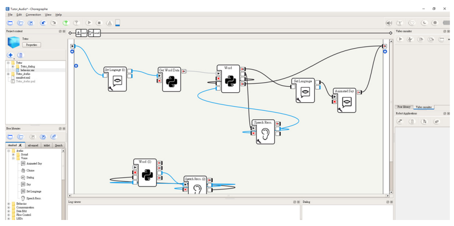


Fig. 2. Choregraphe operational interface.

the server uploads the data and the data reaches the cloud, it uses Google Cloud Speech to Text (STT) and Word Error Rate (WER) formulas to calculate the accuracy of reading aloud and speaking rate of the text recording, and sends the recognition results of the text and the recognition results of the conversation test through SMTP (Simple Mail Transfer Protocol) after the session test of the unit is over.

Robot. The robot end mainly designs the functional process of the robot, including scanning the QR code to log in to the account, requesting the server to confirm the account, and returning the course progress and course data. The TTS technology of the Nao robot itself is used for the pronunciation of the words, and the storage of audio files, playback, and recording are implemented using the NAOqi SDK. Google Cloud STT is used for the real-time recognition function of the conversation test. After the result is recognized, the recording file is read to calculate the reading speed, and the result is finally returned to the server.

- Hardware – NAO robot

The NAO robot has enough sensors, so it is quite helpful for this research. This research uses sensors such as head and hand to enhance the interaction of tactile sense. It is also possible to use touch to continue the course to enhance the motivation of students to learn. There are 6 touch points (Head-front, Head-middle, Head-rear, Left hand, Right-hand, Right foot) provided for users to interact with EOTRTS.

- Software development – Choregraphe

EOTRTS uses the Choregraphe development platform because Choregraphe can directly use local calls to create modules for the use of various methods. The program in Choregraphe is coded in a way that each functional module has a block, and then it controls the process by connecting various modules. Therefore, it is relatively easy to know how each function is performed in terms of the program structure, and the process of after-school tutoring can also be simply modified according to the needs of the tutor. Choregraphe can visualize each function using the Python programming language. Figure 2 shows the operational interface of Choregraphe platform.

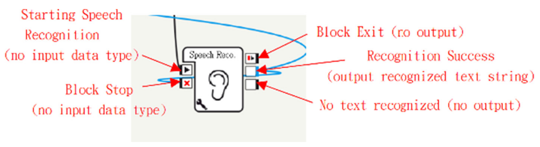
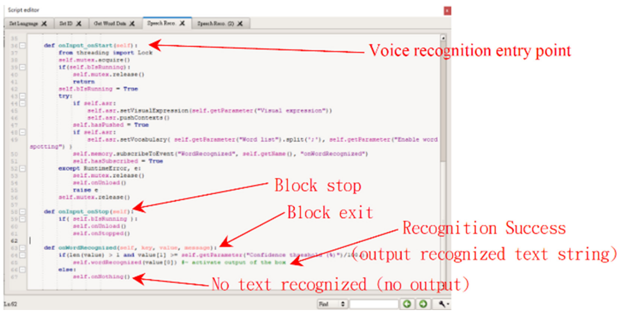

Starting Speech Recognition (no input data type)
Block Stop (no input data type)
Speech Reco.
Block Exit (no output)
Recognition Success (output recognized text string)
No text recognized (no output)

Fig. 3. Speech recognition block in choregraphe.



```
    def onInput_onStart(self):
        from threading import Lock
        self.mutex.acquire()
        if(self.bIsRunning):
            self.mutex.release()
            return
        self.bIsRunning = True
        try:
            if self.asr:
                self.asr.setVisualExpression(self.getParameter("Visual expression"))
                self.asr.pushContexts()
            self.hasPushed = True
            if self.asr:
                self.asr.setVocabulary( self.getParameter("Word list").split(';'), self.getParameter("Enable word
spotting") )
            self.memory.subscribeToEvent("WordRecognized", self.getName(), "onWordRecognized")
            self.hasSubscribed = True
        except RuntimeError, e:
            self.mutex.release()
            self.onUnload()
            raise e
        self.mutex.release()

    def onInput_onStop(self):
        if( self.bIsRunning ):
            self.onUnload()
            self.onStopped()

    def onWordRecognized(self, key, value, message):
        if(len(value) > 1 and value[1] >= self.getParameter("Confidence threshold (%)")/100.):
            self.wordRecognized(value[0]) #~ activate output of the box
        else:
            self.onNothing()
```

Fig. 4. Corresponding program python codes of Fig. 3

As shown in Fig. 2, there are many boxes, each of which is a module with the specified functions. By connecting each box to integrate the system, the box is normally the code, but it can also be Block flow chart as in the middle of Fig. 3. The advantage of developing on Choregraphe is that you can see which module is currently actively running, which is convenient for programmers to control the process of the system. However, the design of each module is not as simple as in writing usual programs. It is necessary to set the input data type and output format with a specific code-writing method for each module. Figure 3 and Fig. 4 illustrate the functional design examples of the block. Figure 3 is the speech recognition block, and Fig. 4 is the compiled version of the corresponding Python codes.

The entry point and exit point of each recognition box have their own input and output to be set, which have different colors according to different output data types. For example, the output string is blue; the type and length of the output data needs to be properly set in advance, otherwise the program will have an error occurred or fail to receive the data that should be obtained. For example, the voice recognition entry point in Fig. 4 will go to *def onInput_onStart(self)*: The method in this line will start to operate. If one want to get data, he must set the data input in the box and set the data type of the team, and then change the Python code part to *def onInput_onStart(self, data)*: to get the transferred data.

The above is briefly the preliminary process of designing each block. There are requirements to use the microphone and so on. The *ALSpeechRecognition* and the *ALMemory* modules can be used through the above mentioned NAOqi methods. Then the voice recognition function can be turned on through the methods inside.

Cloud Services. An important function of this system–"voice recognition"–is provided through cloud services. The robot saves the recording file of each student on the server after the text is read aloud, and then uploads it to the cloud service from the server. We use Google Cloud Speech-to-Text (STT) service to recognize the text recording and then convert it into a corresponding text file. Finally, we use the Word Error Rate (WER) formula to calculate the accuracy and speed of the student's reading aloud. After the end of the session test of a unit, the recognition result of the text read aloud and the session test will be sent to the student via SMTP (Simple Mail Transfer Protocol) by email.

- Google Cloud Platform Speech-To-Text (STT) Service

Google Cloud Platform has a forward-looking infrastructure, powerful data analysis services. EOTRTS uses Google Cloud Automatic Speech Recognizer (ASR) for speech recognition and its STT API for speech to text translation, in which a powerful neural network is added to the Cloud STT API. This model is convenient for developers to convert their audio messages into text. Additionally, because it is a voice recognition model, to have a higher recognition rate, it requires a complete and clear pronunciation for single words to be read. Therefore, it is very suitable for students to practice reading aloud. As a result, students would be able to read English clearly and completely, thereby improving their reading ability. This API mainly has two functions, namely:

- Synchronous voice recognition: For short voice message (less than 1 min), the recognized text will be recognized by a built-in STT function in NAO, and returned in the response immediately. EOTRTS uses this part in the robot conversation test. This function is particularly suitable for the conversation test as a conversation has fewer sentences and needs to be quickly identified and then sent back.
- Asynchronous voice recognition: Performing voice recognition on batches of long audio files will start long-running audio processing operations. Asynchronous voice recognition is used to recognize an audio message that is longer than 1 min and stored in Google Cloud Storage. The system then uses this API to identify the part of the text to be practiced in the system.

The recognition rate can be obtained by comparing the original text with its recording of student's read aloud voice. The ASR recognition rate is more than 80%, even up to as high as 92%, which is an acceptable recognition rate.

When using asynchronous voice recognition, only the audio files stored in Google Cloud Storage. The system creates a *bucket* named *speech_to_text_class* on the cloud platform to store the student audio files to be processed. Then we can upload audio files to this bucket from the server. In order for the data to be reusable, we use the *Regional* level storage space and grant permissions through Cloud Storage ID, so that project members can access Cloud Storage data based on their project roles. Server-end account access rights are also granted through Cloud Storage ID.

- Correct rate of reading aloud and calculation of speaking speed

WER is a common measurement of speech recognition or machine translation system performance. It was proposed by Levenshtein [7]. The calculation formula is as Eq. (1). The calculation method is as follows:

S is the number of different words in the original sentence, *D* is the number of missing words, *I* is the number of inserted words, and N is the number of correct words. The spaces, commas, and other symbols are removed in the calculation of correct rate.

$$WER = 100 * \frac{S + D + I}{N}\% \tag{1}$$

The formula for calculating the speed of reading aloud is as Eq. (2), and Eq. (3).

$$Speaking\ Speed = \frac{text\ length\ (words)}{recording\ time(min)},\ if\ \geq\ 1\ minute \tag{2}$$

$$Speaking\ Speed = 60 * \frac{text\ length(words)}{recording\ time(min)},\ if\ <\ 1\ minute \tag{3}$$

2.2 English Oral Educational Functions in EOTRTS

EOTRTS has 4 main oral English education functions: single-word pronunciation exercises, text reading exercises, text recordings, and conversation tests. The robots in the process have operation tips to prevent students from forgetting how to operate them.

- Single-word pronunciation exercises

After entering the course practice, students begin to learn the vocabulary in the new course. The robot reads the words first, and the students follow along. Students learn the meaning of the words, so that the students will not have difficulty to read the new words in the text reading practice afterwards. After reading a word, there is a waiting interval of 3 s for students to practice the pronunciation of the word, and then NAO will read the next word. Vocabulary practice is about two and a half minutes, and students can follow their own abilities to choose whether to practice again.

- Text read-aloud exercises

The student touches NAO's head to enter the text reading exercise. The NAO robot will play the text audio files (real-person English pronunciation) of the practice unit, and can make some actions to attract the students' attention. It will stop at the appropriate semantic paragraph. After students listen to the audio file and practice reading the passage, they can listen to the audio file of the next passage by touching NAO's right hand or right foot. If they don't understand this passage, they can repeat this audio file by touching NAO's head (front). Then NAO will play the text audio file of the whole unit. If students want to practice again, they can touch NAO's head (middle) to listen to the entire audio file again. The estimated practice time for this process is around 10 min.

- Text read-aloud recordings

When the text reading exercise is over, the NAO robot prompts, "Touch my front head to start recording." After that, NAO will not take any action and wait. When a student thinks that he/she can start the next process, he/she will touch NAO's front head to start recording. After recording, the student can touch NAO's middle head to stop recording. Students choose whether to re-record the read-aloud again orally by saying "YES" or "NO". If students do not choose to re-record, the robot will upload the audio recording file to the server.

- Conversation tests

EOTRTS will firstly read the questions of the conversation test in the textbook. In order to train the student's listening ability, the test paper we give will only have the answer part, and the question will be read by the robot. The robot reads the question twice with 2 s apart each time. Student will first answer which choice (A-D) it is, and the robot will ask the student to read the complete sentence of that answer choice next. For example, if the student says, "A", the robot will ask the student to read the sentence of his answer choice. Students touch NAO's forehead to activate its short speech recognition function, and touch the middle of NAO's head to end the function after finishing reading. After that, an email with the text recognition content of the unit and the recognition content of the conversation test will be sent immediately to the student.

3 EOTRTS Experiment

3.1 Participants

The experiment is aimed at students in the science-related departments of a university in central Taiwan. The total number of students is 19, including 17 boys and 2 girls, with an average age of 21.65 years and a standard deviation of 1.06. A total of 19 people in the experimental group and the control group participated in the experiment.

3.2 Teaching Material

The textbook used in this study is 4000 Essential English Words 1. It's an English textbook selected for freshmen. It has a certain correlation with the participating students. Students have certain familiarity with the teaching materials, and the difficulty of the teaching materials is relatively low. It is suitable for students of science and engineering background for the purpose of improving their insufficient oral English ability.

3.3 Procedure

The experiment lasted a total of 6 weeks. In the first week, the overall experimental process and the pre-test group were introduced and the English learning motivation and English learning anxiety questionnaire were filled out. In the 2nd to 5th weeks, the experimental group and the control group were arranged in a quiet and undisturbed room space. The experimental group used EOTRTS for learning, while the control group learned English in a self-study manner. Participants spent half an hour per unit, and 1 h for two units per week. There were four main learning steps for each unit: single-word pronunciation practice, text reading practice, text reading-aloud practice and conversation test. In addition, the two groups of students recorded their text reading-aloud practices, and the session tests were recorded afterwards as well. The experimental group used the robot to record, and the control group used the mobile phone to record. In the sixth week, the two groups of students engaged in post-testing, and filled out the English learning motivation and English learning anxiety questionnaires, while the experimental group also filled out the technology acceptance questionnaire, and the control group filled out the learning satisfaction questionnaire.

4 Results: Analysis of Technology Acceptance Questionnaire

Limited by the paper length, this paper will only report the analysis of the Technology Acceptance Questionnaire. More results will be presented in our future papers. This questionnaire is designed based on the five-point Likert scale. All questionnaires have been evaluated by two English teachers to ensure that the questions are clear and structured. The technology acceptance questionnaire was conducted in the experimental group to understand students' satisfaction and acceptance of EOTRTS. The average of the item classification is shown in Table 1. Table 2 shows the questions with answers of lower mean.

Table 1. Technology acceptance questionnaire topic classification.

Evaluation items	Ease of use	Usefulness
Sound	4.3	4.2
Action	3.4	3.9
Procedure	4.4	4.3
Interaction	4.1	4.1

In terms of ease of use, except for the part of the gesture "action", the rest of the evaluation items receive all 4.0 or more, which means that the usability in sound, process, interaction, results and other items is acceptable. However, the action items got the lowest point at 3.4. After asking most students in the experimental group, the reason is that they think the robot had too many actions and big moves in its gestures. As shown in Table 2, it is indeed as low as 3.6 and 3.3 points.

In terms of usefulness, except for the part of gesture action, the other evaluation items all receive points above 4.0, which means that the usefulness of items such as sound, procedure, interaction, and results also has a good effect. Although the robot's action can attract the attention of students, its excessive movements can also distract students. The lowest-rated item in the questionnaire is the STT ability to recognize speech. This is partly because EOTRTS is recognized through the STT function of

Table 2. Technology acceptance questionnaire with answers of lower mean.

No	Questions	1	2	3	4	5	Mean
3	I think the speed of the Educational robot is appropriate	11.1%	11.1%	22.2%	22.2%	33.3%	3.6
4	I think that the display of Educational robot movements with spoken content is appropriate	11.1%	11.1%	22.2%	44.4%	11.1%	3.3
14	I am satisfied with the system ability to identify the voice recognition function of the server (the identification result sent back by email)	0%	22.2%	33.3%	11.1%	33.3%	3.6
15	I am satisfied with the Educational robot's ability to read English in real time	0%	22.2%	44.4%	22.2%	11.1%	3.2

strongly disagree (1) → strongly agree (5)

Google Cloud. It would have a great impact on the recognition of short speech, including the speech sound level, environmental noise, and even the name of the person, etc. Because the conversation test answer is a sentence, up to a dozen words, as long as one or two words are not recognized, it would have a very big impact on the scoring. Therefore, it would also affect the real-time STT recognition greatly, as shown in Table 2 in which the average score is only 3.2. On the contrary, the impact of that on the server's text reading-aloud recording recognition ability is relatively lowered. Because each text is more than 300 words, the impact on names or small number of errors is not very large, and the average recognition rate is 0.4 higher than that of real-time recognition.

5 Conclusion

This article proposes an educational robot system EOTRTS to provide individual tutoring to students to improve their oral English ability. It also proposes to use Google Cloud-STT and WER calculation formulas to evaluate the accuracy of students' reading aloud, so as to improve the accuracy of their English reading and reading rate. After six weeks of experiments, EOTRTS can indeed play the role of a tutor, helping students learn English in a one-on-one manner. The results of questionnaires show that, regardless of ease-of-use and usefulness, this system has good system acceptance for users, except for the robot action factor. It signifies that using robots as a language learning tool is a new way of oral English learning that can be accepted by students.

Future Works. EOTRTS can incorporate more robots with different functions, appearances, and operating types in the future to support language learning on different robots. In addition, EOTRTS can also be used for different foreign languages in the future, such as Japanese and Korean, etc. because NAO robots can support multiple languages. Furthermore, in the learning process, many people are attracted by the vivid and interesting appearance and actions of the robot and the lively and delicate nature of the interaction between the robot and the human. Therefore, EOTRTS can also be combined with the previous research results of [8–10] in the future to design more lively and vivid robot programs.

Acknowledgement. This study is partly supported in finance by the Ministry of Science and Technology, TAIWAN under the contract numbers MOST106-2511-S-035-003-MY2, MOST107-2511-H-035-004-MY2, and MOST108-2511-H-035-002-MY2.

References

1. Grewen, K.M., Anderson, B.J., Girdler, S.S., Light, K.C.: Warm partner contact is related to lower cardiovascular reactivity. Behav. Med. **29**(3), 123–130 (2003)
2. Mubin, O., Shahid, S., Bartneck, C.: Robot assisted language learning through games: a comparison of two case studies. Aust. J. Intell. Inform. Process. Syst. **13**, 9–14 (2013)
3. Kang, S.C., Chang, W.T., Gu, K.Y., Chi., H.L.: Robot Development Using Microsoft Robotics Developer Studio. CRC Press, Boca Raton (2016)
4. Kanda, T., Hirano, T., Eaton, D., Ishiguro, H.: Interactive robots as social partners and peer tutors for children: a field trial. Hum.-Comput. Interact. **19**, 61–84 (2004)
5. Shen, W.W., Tsai, M.M., Wei, G.C., Lin, C.Y., Lin, J.M.: ETAR: an english assistant robot and its effects on college freshmen’s in-class learning motivation. In: L. Rønningsbakk et al. (eds.): ICITL 2019, LNCS 11937, pp. 77–86 (2019)
6. Lin, C.Y.: EOTRTS: Research into an English Oral Training Robot Tutor System to Support After-Class Individual Learning. Master Thesis, Feng Chia University, Taiwan (2019)
7. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Sov. Phys. Dokl. **10**(8), 707–710 (1966)
8. Li, Y.H., Lin, J.M., Lee, K.Y., Chiou, C.W.: A robot motion control mechanism using robot drama markup language. J. Comput. Appl. Sci. Educ. **5**(2), 17–35 (2018)
9. Luo, B., Shen, W.W.: Repeated reading: a practical approach when using an ESL textbook? Feng Chia J. Hum. Soc. Sci. **30**, 77–103 (2015)
10. Wei, G.C., Shen, W.W., Tsai, M.H.M., Lin, J.M.: An education robot supporting adaptive learning for english as a foreign language. In: Lin, J.M., Chang, C.-C, Hwang, Y.M. (eds.) The 9th Conference on Engineering, Technological and Technology Education CETTE 2020, pp. 44–55 (2020)