Development, Evaluation, and Further Research of Voice-enabled Chatbot for English as a Foreign Language

Julio Christian Young[1] Makoto Shishido[2]
Graduate School of Advanced Science and Technology, Tokyo Denki University
[1]{julio.christian.young}@gmail.com [2]{shishido}@mail.dendai.ac.jp

Abstract

Studies show that chatbots can help new EFL students practice their communication skills. However, several problems in the EFL learning chatbot still have not been fully addressed. Two of the problems are the appropriateness of TTS to produce audio materials within the chatbot and a standalone chatbot that could work without an internet connection. This study compared quantitative and qualitative aspects of TTS- and native speakers-produced to address the first problem. Besides that, we created and evaluated a standalone chatbot using lightweight speech recognition to address the second. Furthermore, we discuss other topics worth investigating based on these two results.

1 Introduction

Learning a new language can be daunting, especially for English as a Foreign Language (EFL) students. As EFL students often lack time and opportunity to practice, they will feel awkward and fear being judged when they get a chance. In the long term, this situation can lead students to doubt their abilities and lose motivation to learn [1, 2].

Several studies in the past [3, 4, 5, 6, 7, 8] showed that chatbots could be ideal learning partners, especially for EFL students with low language proficiency. Research in [3, 5, 6] argued that students tend to feel less anxious when practicing with chatbots. Based on their input type, two kinds of chatbots that often used in the EFL learning context: text-based and voice-enabled chatbots.

While text-based implementations often interact with their students in free-typed text format, voice-enabled chatbots are usually implemented via multi-choice text inputs that students can choose by read-aloud one of the choices available. This mechanism was preferred as chatbots that process direct voice input often led to communication breakdown. A communication breakdown might occur as the chatbot tries to reply based on an incorrect transcription by the speech recognition module in the chatbot.

Even though implementing multi-choice text inputs limits students' interactions, researchers believe it could still be helpful to support students with low language proficiencies [5]. As they still have limited vocabulary to produce their sentences, predefined choices could help them to interact with the bot. Moreover, by employing a mechanism that compared the transcription result with the reference text, the application could provide meaningful feedback regarding students' mispronunciations.

Besides the SR module, the quality of the voice response from the chatbot is another factor that influences the success of EFL learning chatbots. Ideally, pre-recorded responses produced by native speakers would be best to mimic an actual conversation experience with a human partner. However, involving native speakers will increase the cost and time of application development.

In [5], the researcher demonstrated how TTS technology could substitute the involvement of native speakers. Even though TTS produced all audio used in [5] almost 85% of participants in the experiment felt the audio sounded natural. The overall evaluation of the chatbot also revealed that most participants enjoyed their learning experience using the chatbot.

In our study, to support EFL students' English learning journey, we developed a voice-enabled chatbot to help students practice their speaking skills. Although similar studies have been conducted, our study will address the remaining challenges that have not been adequately addressed.

2 Problem Statement

Despite many previous studies on EFL learning with chatbots, several issues still have not been entirely

covered. In our study, there will be two issues that we want to address.

Lack of comparison of audio materials produced by native speakers and TTS technology - None of the previous studies explicitly compared the use of TTS- and native speakers-produced materials for EFL learning chatbots. In our study, we compared the qualities between TTS- and native speakers-produced materials for sentences in dialogue format. Since chatbots process sentences in dialog format, this comparison could reflect the suitability of TTS-produced materials in them.

Implementation of the SR in the chatbot relies on the Internet - To the best of our knowledge, all previous studies that explored the usage of voice-enabled chatbots for EFL learning relied on cloud-based SR services [3, 4, 5, 6], which rely on the internet. As they rely on the internet, students' interaction would be disrupted with the app when the internet connection is unstable. In the worst scenario, students with a bad Internet connection would be unable to use the app.

Unlike previous research that relied on cloud-based SR services, our study tried to explore the possibility of using a small SR model. As a small SR model only require low computational power, it could be run on students' own devices, thus making it works without the internet. A small SR model developed in [9] shows good transcription accuracy despite using less than 400Mb of memory. While big models usually need about 16Gb of memory to achieve a word error rate (WER) of 5.6%, this model achieves a WER of 9.85% with far less memory.

Despite its promising performance, no previous studies have attempted to implement small SR. Even though the small SR model has relatively lesser performance than the big one, its implementation potential for achieving standalone chatbot EFL learning is worth investigating. Therefore, our study tries to measure the suitability and performance of a small SR model in EFL learning chatbots.

3 Research Settings

3.1 The Comparison of Two Audio Types

We designed a blind comparative test scenario involving native speakers and TTS audio materials to address the first problem. Sixty undergraduate EFL students participated in the test. All students had studied English in formal learning settings for about 13 to 14 years. During the test, participants were asked to listen to each audio material in random order without knowing whether native speakers or TTS produced it. After that, they were asked to judge the audio material and transcript its content.

Moreover, ten audio transcriptions used in the test were taken from an English learning textbook in [10]. As there were two audio materials for each audio transcription, each participant needed to listen to twenty audio materials to finish the test. While native speakers' audio materials were taken from supplementary materials with the book, TTS-produced materials were produced using WaveNet TTS with North American English. WaveNet TTS was chosen since it produced significantly better audio quality than other TTS methods [11].

In the test, we compared native speakers- and TTS-produced audio materials qualitatively and quantitatively. The qualitative aspects measured in the experiment are pronunciation accuracy (PA), naturalness (N), comprehensibility (C), and intelligibility (I). Comprehensibility measures how easy to hear given audio. Intelligibility measure how well they can understand it. Data related to these criteria will be collected using a 6-point Likert scale question.

Other than that, word error rate (WER) was chosen as a metric to compare two audio material groups quantitatively. The WER on each audio material was calculated by comparing the reference transcription from a given audio and each participant. The WER score can represent how effortlessly students understand specific audio materials.

3.2 Usability of a Small Speech Recognition Model for EFL Learning Chatbots

We developed an android-based EFL learning chatbot for practicing English speaking skills to measure the suitability of a small SR model. The developed chatbot is a voice-enabled chatbot with multi-choice text inputs with a small SR module from [9]. Fifteen undergraduate students with 13 to 14 years of English learning experience participated in this experiment. Participants were asked to use the app for a week and finish at least two topics within the chatbot. All participants' resulting transcriptions were recorded to be analyzed after.

In the app, the bot will always initiate the conversation, and users can choose one from three or four available responses as a reply. Based on a user reply, the bot will continue the conversation until there is no other response that the bot cannot return. Whenever users choose a reply, they can click a new bubble chat containing their response to practice their speaking skills. The app will mark parts of the sentence that do not appear in the resulting transcription by the SR module in red. Otherwise, in green. Figure 1 depicts the chatbot and user interaction in the app.

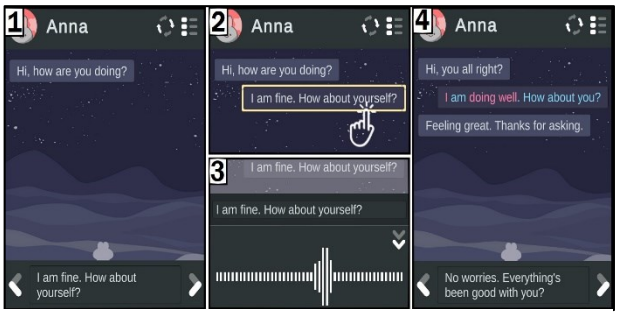

Figure 1. The bot and user interaction in the app

After a week, they were required to fill out a post activity questionnaire to capture their perceived usefulness (PU), perceived ease of use (PE), and attitude toward using (TA) the EFL learning chatbot. The questions asked to the participants were 6 Likert-type questions given in Table 1.

Besides that, the WER calculation was also done with the collected transcriptions. The purpose is to measure the WER of a small SR model, specifically for EFL learning chatbots. As the SR model implemented in the chatbot also has dynamic vocabularies features to target only specific vocabularies, the resulting WER might be lower than the previously mentioned WER in [9].

4 Results

4.1 The Comparison of Two Audio Types

The result on qualitative aspects of native speakers- and TTS-produced audio materials can be seen in Table 1. The slight difference in another three criteria (except N) between the two audio types shows that TTS-produced audio can be appropriate for producing listening materials in dialogue format, thus making it suitable for a chatbot. On top of that, a significant standard deviation from naturalness on TTS-produced materials indicates that several students might believe that sounds natural while others do not.

Table 1. Participants perceived audio quality

Questions
Has ASR correctly recognized and evaluated what you said? (PU-1)
Does practicing speaking skills with ASR give you a more controllable and personalized learning environment? (PU-2)
Do you feel comfortable practicing your speaking skill with the help of ASR? (PE-1)
Do you feel convenient as you need to emphasize clarity (speaking slowly) when you practice using ASR? (PE-2)
How interested are you in a learning system with ASR that can help practice your English speaking skill?
Do you want to adopt the learning system with ASR to practice your English speaking skill?

Table 2. Perceived Quality of Both Audio Groups

Audio Type	Indicator	Criteria			
		PA	N	I	C
Human	Mean	5.60	5.33	5.71	5.69
	Std Dev.	0.75	0.92	0.62	0.67
TTS	Mean	5.56	4.79	5.65	5.65
	Std Dev.	0.75	1.34	0.68	0.67

Furthermore, the WER on each audio materials group was calculated using the participants' resulting transcriptions. While transcriptions using native speakers-produced materials got a WER of 0.068, the TTS-produced ones got 0.062. In contrast to the result in the comprehensibility criterion that favors native speakers-produced materials, low WERs on both audio types indicated that participants could easily understand both.

Following such results, we investigated what kind of sentences resulted in more errors in the native speakers-produced materials. After looking further, we notice that transcriptions based on native speakers-produced materials miss words more often than TTS ones. As native speakers often reduce, contract, and mash some combination of words in their spoken form, low- to intermediate-level students might miss one of the words while listening to them. However, further investigation is needed to support this claim.

4.2 Usability of a Small Speech Recognition Model for EFL Learning Chatbots

The summary of results from the post activity questionnaire towards the EFL Learning chatbot with a

small SR model is given in Table 3.

Table 3. Post-Activity Questionnaire Result

Question	PU-1	PU-2	PE-1	PE-2	TA-1	TA-2
Mean	5.06	5.13	5.33	4.13	5.20	5.26
Final Mean	5.1		4.73		5.23	

The result concluded that participants agreed that the chatbot could correctly recognize and evaluate their speech, thus enabling them to practice their speaking skills. Moreover, the PE-1 value showed that participants agreed to feel comfortable speaking with the chatbot. However, the low PE-2 value indicates that they might feel awkward as they need to speak slowly while using the chatbot. Finally, participants seemed interested in practicing English using the chatbot based on the final average from TA-1 and TA-2.

Next, the WER was calculated to determine the significant performance gain of dynamic vocabulary features in the SR model for a specific application. While the SR model used achieved a WER of 9.85% in a more general dataset, it achieved a WER of 7.42% in the experiment. The lower WER score indicated that implementing the dynamic vocabulary feature allows the model to perform better.

Following such findings, we also did an open-ended group discussion with participants to gather opinions about their learning experiences while using the app. A participant in the discussion mentioned feeling less enthusiastic about using the app after using it for some time. The participant also added that it would be interesting if there were more features besides speech evaluation in the application. Following this statement, other participants also agree that different types of exercise might make the application more challenging, thus making it more fun.

5 Discussions

5.1 Further Evaluation on TTS and Native-speakers Audio for EFL Learning

Our evaluation results indicate that students made fewer errors when transcribing TTS materials than ones produced by native speakers. Native speakers sometimes link words together to make them sound natural, yet make them sound different and challenging to be understood by students. Alternatively, TTS materials might pronounce each word in the sentence individually, thus making it easier for students to grasp but sounding unnatural. However, as only ten audio pairs were compared in our study, they would be enough to capture different types of linking in connected speech. Therefore, further studies are needed to analyze the differences between TTS- and native-speaker-produced materials to prove such a claim.

Furthermore, we suggest that future research involves using speech recognition technology rather than involving students in the audio transcription process. SR technology enables the comparison process for large amounts of data, thus making the research result more reliable. On top of that, various SR models with different transcription performances could capture the difference between two audio groups across different levels of competency

5.2 Improving Chatbot Interactivity

There are a few ways we can try to improve interactivity within the chatbot. One approach could be to make the chatbot more engaging by incorporating a wider range of content and activities. Presently, no active learning activities are done while students listen to audio responses by the chatbot. Therefore, in the future update, we plan to implement the fill-in-the-blank and sentence scramble features in the app.

To implement the listening fill-in-the-blank feature, we can modify the text response returned by the bot by replacing a random word with a blank line. Thus, instead of passively listening to the bot's response, students can actively interact with the app by filling in the blank line after each listening session. By listening to a spoken language recording and filling in the missing words, students can practice using the vocabulary in context and become more familiar with it, thus reinforcing their understanding of such words and how to use them correctly.

Similarly, the sentence scramble feature can be employed by modifying the text response by the bot. In this activity, students will get a jumbled sentence and its spoken form in the correct order for each response from the bot. Then, based on the jumbled sentence, they need to rearrange the words to form a grammatically correct and coherent sentence. By completing sentence scramble activities, students can improve their ability to understand and produce grammatically correct sentences.

References

[1] Aiello, J., & Mongibello, A. (2019). Supporting EFL learners with a virtual environment: A focus on L2 pronunciation. Journal of e-Learning and Knowledge Society, 15, 95– 108. https://doi.org/10.20368/1971-8829/1444.

[2] Huang, X., & Jia, X. (2016). Corrective feedback on pronunciation: Students' and teachers' perceptions. International Journal of English Linguistics, 6(6), 245–254. https://doi.org/10.5539/ijel.v6n6p245.

[3] Han, D. E. (2020). The Effects of Voice-based AI Chatbots on Korean EFL Middle School Students' Speaking Competence and Affective Domains.

[4] Shishido, M. (2018, June). Developing e-learning system for English conversation practice using speech recognition and artificial intelligence. In EdMedia+ Innovate Learning (pp. 226-231). Association for the Advancement of Computing in Education.

[5] Shishido M. (2021). Developing and Evaluating the e-learning Material for Speaking Practice with the Latest AI Technology ISSN: 2189-1036 – The IAFOR International Conference on Education – Hawaii 2021 Official Conference Proceedings.

[6] Shishido, M. (2019, June). Evaluating e-learning system for English conversation practice with speech recognition and future development using AI. In EdMedia+ Innovate Learning (pp. 213-218). Association for the Advancement of Computing in Education.

[7] Kim, N. Y. (2018). A study on chatbots for developing Korean college students' English listening and reading skills. Journal of Digital Convergence, 16(8), 19-26.

[8] Shin, D., Kim, H., Lee, J. H., & Yang, H. (2021). Exploring the Use of An Artificial Intelligence Chatbot as Second Language Conversation Partners. Korean Journal of English Language and Linguistics, 21, 375-391.

[9] Alpha Cephei. (2020). Vosk Offline Speech Recognition Performance. [Online]. Available: https://alphacephei.com/vosk/models

[10] Shishido, M., "Checking in at the airport - Listening Practice," in Virtual Travel around the World, バーチャル世界旅行:チャットボットで学ぶ旅行英会話, 1st ed. Tokyo, Japan: Seibido, 2022

[11] Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.