Using ASR on iOS for pronunciation training in EFL

– A promising future in the backdrop of advancing technology –

（英語教育講座） David R. BOGDAN

(Accepted 9/1/2021)

This study constitutes a followup to a mini experiment the author (DRB) conducted in 2019 in which he looked at how the "dictation" function on Mac OS might differ when compared to humans with regard to perception/comprehension. Here, the main focus was on the use of the automatic speech recognition (ASR) function in Apple's iOS (including iPadOS) and Apple Watch. As before, the question was how might dictation activities through ASR be used to aid pronunciation training for ELLs (English Language Learners), but, while the earlier study concentrated on the receiving end of things–i.e., how the computer and participants were able to perceive, process, and recognize speech–here, because of the improvements in ASR with the app used, the focus shifted to more on production. The original test itself was run with only four participants (including the author) and, therefore, did not make any pretense of being a statistically-based study. It did, however, provide some insights into the differences in how ASR works when compared to a human interlocutor. While the previous study dealt primarily with the receiving end and how the problems with ASR might reduce motivation in language learners and hinder progress in pronunciation, here it can be seen that advances in the technology allow it to show much more promise in advancing pronunciation training.

BACKGROUND

In August of 2019, the author, DRB, decided to run a quick test to compare (or perhaps more accurately, contrast) the way a Mac and human listeners processed dictated text. He had for several years been interested in how the native automatic speech recognition (ASR) capability of MacOS and iOS (including the iPhone) might help in pronunciation training for foreign language learning and instruction. He himself had been using the built-in speech-to-text function in iOS for his own foreign language learning on Duolingo (a very popular language-learning platform).

Based on his own experience, the premise was that the computer, with its more limited database of real-world knowledge, would lack some of the context that a human had in making inferences and judgements helpful in recognizing and processing a stream of speech—especially in isolation—thus making it a harder taskmaster in training a language learner to pronounce in a way acceptable to it.

In a similar vein, when a grad student advisee of the author conducted a study in which he created specific tasks for students to do in order to help them to improve their English pronunciation and to gain confidence in their speaking ability, there were many instances where the ASR was often not able to accurately process what the students were trying to communicate.

DRB was concerned that this might have a detrimental educational effect on the students by causing them to lose confidence in their pronunciation ability. There has always been concern about how correcting, and especially over-correcting, a language learner could hinder them in their path to mastering a foreign language, and pronunciation poses a special challenge because it tends to be very hard to teach and for the student to gain confidence in. Ahn (2016) goes as far as to state that "speaking is the most challenging language skill for English as a foreign language (EFL) students, and the EFL classroom is unable to offer enough opportunities for speaking practice."

The concern regarding confidence led the author to suggest to students who might be interested in using this dictation function to improve their English pronunciation that ASR would probably have more problems than a human in handling the variation in pronunciation that we observe in actual speech, not only between individuals but also with the same speaker during different utterances of the same material.

We are definitely not perfect machines, and every time we attempt to produce the same linguistic oral unit, even though the intent may be to faithfully reproduce the previous utterance, there is going to be some variance. Human listeners, depending on their competence in the language, are able to make allowances for such variation, mainly due to their vast database of world knowledge which lets them make inferences based on contextual probabilities.

Because of this disparity, the author emphasized that students should not lose heart whenever there are problems with their dictation, but should rather be cognizant of the limitations in ASR and try to work on carefully enunciating items in order to overcome the limitations. This would, in the long run, make them better speakers. Working on accuracy first would give them the basis and ability to work on speed later on as they gain more confidence.

The hope that the author had, however, was that ASR would at some time improve to the extent to where it would be more useful in pronunciation training. This study suggests that this point may have arrived.

THE ORIGINAL TEST

The experiment was carried out in August, 2019, and prior to running it, the author explained to the participants that they would be recorded and that these recordings might be used in further research. They gave their consent both prior to and following the exercise. These recordings were used in this study. Both in the original report and here, references to the participants are anonymized to protect their privacy.

The original report (Bogdan, 2019) describes the experiment in detail, but a short description follows here.

The original participants, conditions, and procedure

Three of the participants were students in the EFL Department of the Faculty of Education at Ehime University, and they were all native speakers of Japanese. One was a sophomore, another a junior, and the third, a senior. The labels *LL1, LL2,* and *LL3* (The "LL" standing for Language Learner) were randomly assigned to the students to keep each student anonymous. The final participant, *NS*, was the author himself, a native speaker of North American English.

The experiment took place in the ICT Lab in the Faculty of Education, equipped with relatively new 2017 iMacs. The participants took turns dictating the text that they had chosen into one of the iMacs that had been equipped with an external microphone because the acoustic conditions in the room were far from optimal.

The "speaker" read each of their two sentences twice. These sentences were dictated into a TextEdit document using the native Mac ASR. They were simultaneously recorded by QuickTime. Both TextEdit and QuickTime come installed on any Mac. While this was happening, the three "listeners" attempted to transcribe the sentences as they were being read out loud. Again, because this was a dictation exercise, no "listener" had prior access to the sentences before they were dictated. They were also separated from the "speaker" by the iMac itself and therefore could not observe the "speaker's" face as the "speaker" was speaking.

The language data

The language data used for dictation consisted of just eight sentences. Each of the four participants—including the author—dictated a pair of sentences they chose both into the iMac and to the three other participants. They chose these sentences from three different EFL/ESL textbooks: *Power On English Communication I* (Asami, 2019), *All Aboard! English Communication I* (Kiyota, 2019), and *Prominence English Communication I* (Tanabe, 2019). All three textbooks were published in Japan and used for teaching English to Japanese high school sophomores.

Each of the three students who participated was asked to find a pair of sentences from one of the texts—the first, a shorter, simpler one and the other, a longer more complex one. They avoided sentences that contained foreign personal names and place names in order to give the ASR a better chance of recognizing all of the vocabulary. The author also chose two sentences from one of the textbooks used by the students, making sure that he chose from different pages than the students had used.

Up until the point that the sentences were being dictated, none of the participants, including the author, had any advance knowledge of what the sentences of any of the others were.

In (1), we see the original target sentences used for the dictation exercise in the order in which they were dictated.

(1) S1(LL1) Some animals sleep longer than others do.
S2(LL1) Some ocean animals, such as whales, have to swim to the surface for air.

S3(LL2) What course do you take at your school?
S4(LL2) During their school life, the students often face the lives and deaths of their animals.

S5(LL3) I had no interest in trying the phone in our new house.
S6(LL3) When I was very young, my family had one of the first telephones in our neighborhood.

S7(NS) By deep frying noodles, they become hard and dry.
S8(NS) The angels celebrate the birth of [Jesus] Christ by singing and playing musical instruments.

The target sentences are numbered S1 - S8. The *LL#* represents each of the three language-learning students, while *NS* is the native speaker of English. The first and second sentences for each informant differ in length and complexity, with the second one of each pair (the even-numbered sentences) being longer and more complicated. Each sentence was repeated twice to aid in the dictation.

Some observations from the earlier study

The extremely limited database, both in terms of participants and linguistic tokens, made it impossible to make any statistically-backed conclusions, but, it appeared that the native speaker had, in nearly every instance, better results than the iMac with his speech recognition, especially after hearing each sentence a second time. On the other hand, the computer on the whole produced somewhat more accurate transcriptions than the non-native ELLs when processing the author's utterances.

CHANGES IN THE INTERIM

As noted earlier, the original test was run in August, 2019, and the author had recorded the participants as they were dictating the sentences. There were two major reasons for doing so. One was to to make it possible to provide feedback by letting the speakers listen to themselves. Because, however, the study was concentrating on perception rather than production, this feedback only took place in a very minimal fashion; at the end of the experiment, the participants all listened to the recordings just once. There was no individual follow-up in order to provide pronunciation training.

Another motivation for making and keeping the recordings was to have the data available in case ASR improved in a way that would allow a more accurate rendering of the spoken text. As we will see in this report, such improvement has now taken place.

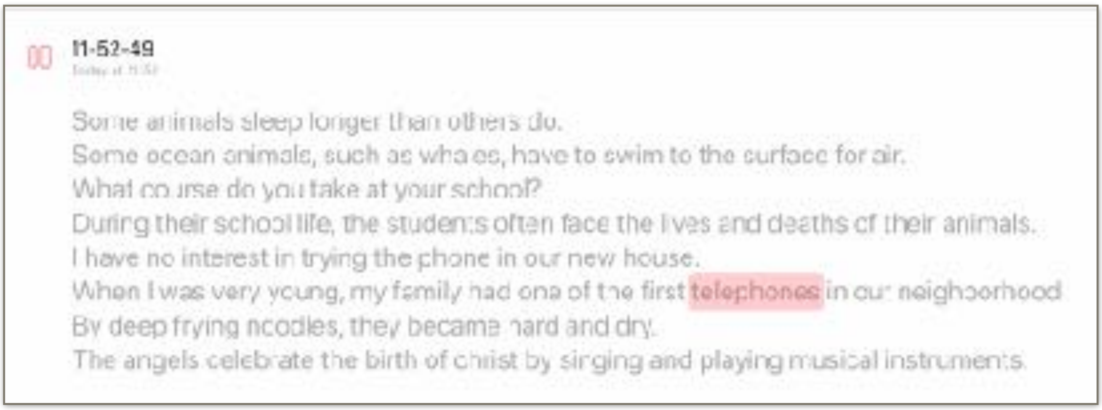
Apple has reportedly improved on the accuracy of the built-in dictation function somewhat, but there are still some inconveniences involved. Much to the disappointment of many, including the author, Apple removed the separate Enhanced Dictation function in the two most recent iterations of its OS: Catalina and Big Sur. They say the same enhanced dictation functionality can be achieved if the user opts in on Voice Control, which allows the user to navigate and interact with their device using their voice. The author attempted this on a Mac running Big Sur, but ran into problems with the machine misinterpreting some of his normal speech as commands, leading to false starts and other unintended actions occurring on the Mac. He eventually had to turn the feature off and is now limited to doing his dictation on an older Mac still running Mojave. Apple is reportedly reintroducing a stand-alone enhanced dictation feature, which will supposedly only run on the the newest processors, M1 Macs and iDevices with the A12 chip.

A third-party solution

While the move away from enhanced dictation has been, to say the least, discouraging, recently the author became aware of a dictation application called JustPressRecord (Open Planet Software)–available both on the Mac and iOS App Store–and he decided to give it a try, especially because it was quite reasonably priced, and the iOS version would also work on the Apple Watch.

Although this report is not intended to be an app review, and there are no doubt other applications and platforms that may have similar functionality, the author has been quite impressed with the accuracy of the transcriptions, but even more so by the fact that it keeps a copy of the audio input that can be listened to while viewing the transcription. Not only this, but you can play the audio file and simultaneously view the text being spoken as highlighted transcription, with the highlighting progressing together with the recording along the timeline. A video of this here would, of course, provide the reader a much better idea of how this works, but Figures 1 and 2 do provide screenshot examples visually illustrating this.

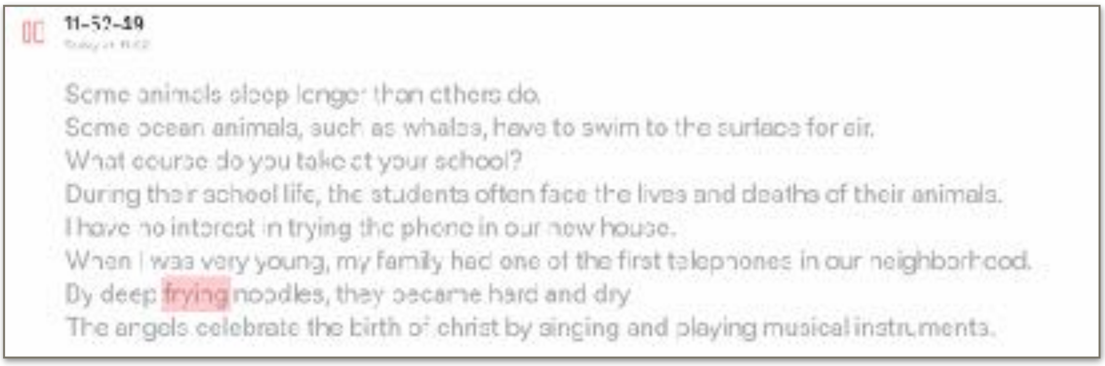
Another very useful feature, especially from a pedagogical standpoint, is that you can click on a word or phrase in the transcription and the playback will back up or fast forward to that point and start from there. This could be a very useful tool in going over problem areas with students, providing them with both audio and visual feedback simultaneously. We shall see some of this feedback in action in the section with the follow-up interviews with two participants from the earlier study.



11-52-49
Today at 11:52

Some animals sleep longer than others do.
Some ocean animals, such as whales, have to swim to the surface for air.
What course do you take at your school?
During their school life, the students often face the lives and deaths of their animals.
I have no interest in trying the phone in our new house.
When I was very young, my family had one of the first telephones in our neighborhood
By deep frying noodles, they became hard and dry.
The angels celebrate the birth of christ by singing and playing musical instruments.

Figure1: Example1 of Highlight Movement



11-52-49
Today at 11:52

Some animals sleep longer than others do.
Some ocean animals, such as whales, have to swim to the surface for air.
What course do you take at your school?
During their school life, the students often face the lives and deaths of their animals.
I have no interest in trying the phone in our new house.
When I was very young, my family had one of the first telephones in our neighborhood.
By deep frying noodles, they became hard and dry
The angels celebrate the birth of christ by singing and playing musical instruments.

Figure2: Example2 of Highlight Movement

Accuracy

As mentioned earlier, the author had been keeping the recordings (in the form of m4v audio files) of the dictations from the 2019 test just in case they could come in handy for giving feedback or as possible input for improved ASR whenever that came along. Perhaps the biggest feature of the JustPressRecord application that makes it stand out from Apple's native dictation function is that it will accept and transcribe audio files. The author ran the 16 audio files (two for each sentence) through the app, and it processed the material quite well, in spite of the fact that the recording had been made in less than optimal acoustic conditions.

Appendix 1 provides a comparison of how the eight sentences were recognized in the form of transcriptions by: (1), the Mac OS ASR from the first experiment (*C*); (2), by the revised transcriptions made from the audio files by JustPressRecord (*JPR*); and (3), by the author (*NS*) during the first experiment. The data are divided into eight sections, one for each sentence, which are labeled with the sentence number and with the "speaker" indicated in parentheses. In each section, we then see the target sentence followed by the transcriptions rendered by *C*, *JPR*, and the

NS, respectively, for both the first and second readings of each sentence. The two minor exceptions to this are the final two sentences. They were read by the author; therefore, there is no transcription by *NS*.

FOLLOW-UP INTERVIEW WITH TWO PARTICIPANTS

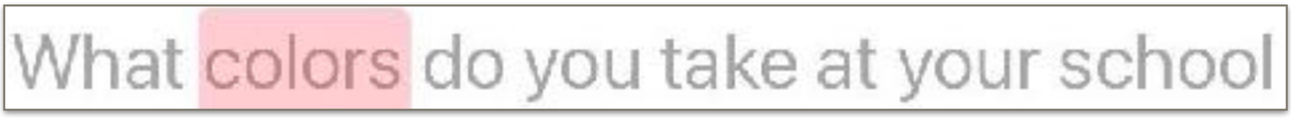
Once the author had seen how well the application processed the spoken text from the original test, he decided that it would be beneficial to show the results and the application to the participants and to get feedback from them. Two of the participants were still students at the university, and he was able to set up a Zoom interview to consult with them. The interview lasted over an hour.

Recap and software presentation

The first thing that the author did was give an oral recap of what had been done during the experiment of two years prior. The students had forgotten a good bit of it, so it was a good chance to refresh their memories. Prior to this interview, the author made a short slide-show presentation based on Appendix 1 to show the interviewees how the iMac and the *NS*, both during the actual dictation, compared with the JPR application in terms of recognition accuracy. It was evident from the presentation that the JPR handily beat out the native ASR of the iMac, but another interesting thing was that it also did better than the author in places. Following the presentation and discussion, DRB then demonstrated to them how the application worked by using a screen-sharing session of it showing it in action on his iPad. The students appeared to be impressed with the combined audio and visual effects seen with the moving transcription display feature. This feature was used to go over the sentences that they themselves had dictated, providing them feedback as to why portions of the transcriptions might have gone awry.

Feedback activity

For example, when reviewing the transcription and audio for S3 (See Figure 3), *LL2*, who had originally dictated the sentence, commented on the difference in the pronunciation of "course" between the first and second readings, noting how carefully it had been pronounced the second time around. The author and the other student also noticed the clear audible difference.

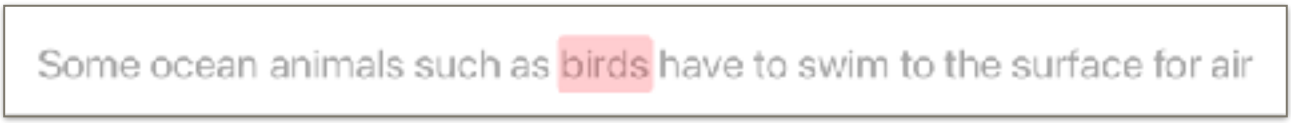


What colors do you take at your school

Figure3: "colors" vs. "course" pronunciation in S3

Figure 4 applied to the other student (*LL1* in the dictation exercise). The word "whales" posed problems in both the first and second readings: coming out as "bears" for both the iMac and

JPR for the first and as "bears" and "bird", respectively, in the second. The *NS*, on the other hand, had been able to correctly figure out "whales" during the second reading, probably due to a real-world knowledge of what sort of "sea" animals surface for air. *LL1* noted that they needed to work on the pronunciation for that item.



Some ocean animals such as birds have to swim to the surface for air

Figure4: "birds" vs. "whales" pronunciation in S2

Mini-test in distance-learning pronunciation training

Following the somewhat belated feedback activity, the students volunteered to take part in a small distance-learning dictation exercise in which each of them read out a piece of text that they had chosen into their microphones they were using for the Zoom meeting while the author had JustPressRecord on his Apple Watch recording from the speakers on his end. The recording/ transcription of one of the students did not go so well at first, but it turned out that merely turning up the speakers on the author's iMac solved the issue, and the application then did a fairly accurate job of transcribing the input from both students. Each of the two transcriptions had only a slight error at the beginning, and that probably stemmed from the author fiddling with his watch to start the recording and get it pointed at the speakers.

Naturally, it would be much more efficient and no doubt more accurate to have the recording done by the Zoom application itself rather than going through the convoluted process of having the watch record the sound coming out of the speaker. The JPR application could then just process the resulting audio file. For an on-the-fly pronunciation training exercise, however, it did not go too badly.

SOME OBSERVATIONS

Following the mini pronunciation training exercise, the interview was rounded off with a general discussion about impressions and observations. In addition, the students promised to send emails with their observations, and they did so quite promptly. A rough translation of their observations and comments can be seen in Appendix 2.

The comments make it abundantly clear that the students had a very positive impression of the software and could easily foresee its possible uses for them both personally as language learners and as educators providing help to students in achieving pronunciation competency.

CONCLUSION

The application shows great promise as an assistive language-teaching tool for giving learners feedback on their pronunciation, and, as such, can really help reduce some of a teacher's workload by taking over some of the burden in an extremely time-consuming activity. It can also help teachers who need more confidence in their own pronunciation.

While it would be ideal if it were completely cross-platform and not limited to the Apple ecosystem, it is making use of Apple's Speech-to-Text framework to perform the transcriptions, which means duplicating the process exactly on another platform might prove problematic. There are also privacy protection concerns. Apple has been positioning itself–especially in recent times–as being a champion of the personal privacy of its consumers, and this may not be as great of a priority for other tech companies.

ACKNOWLEDGEMENTS

I would like to again thank the three undergraduate students who participated in the earlier experiment, and especially to thank the two who were able to take part in the interviews and who provided such valuable feedback (in Appendix 2). Thanks also to T. Ando, a grad student and English teacher at a local high school, for agreeing to meet with me to see how the application works and who also gave some very insightful and helpful comments on how it might be useful in an actual teaching situation.

As I have done many times in the past, I would to thank my colleague at the university, R. Vergin, who again agreed to proofread my English and make very useful comments on the content. Again, however, the responsibility for any errors lies totally with me.

REFERENCES

Ahn, T. Y., & Lee, S.-M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. British Journal of Educational Technology, 47(4), 778-786.

Asami, Michiaki. (2019). *Power On English Communication I*. Tokyo: Tokyo Shoseki.

Bogdan, David R. (2019). "Differences between Computer and Human Speech Recognition: A Brief Look at Dictation on MacOS/iOS". http://www.ed.ehime-u.ac.jp/~kiyou/2019/pdf/15.pdf.

Kiyota, Yoichi. (2019). *All Aboard! English Communication I*. Tokyo: Tokyo Shoseki.

Open Planet Software. *Just Press Record*. https://www.openplanetsoftware.com/just-press-record/.

Tanabe, Masami. (2019). *Prominence English Communication I*. Tokyo: Tokyo Shoseki.

Appendix 1: A Comparison of MacOS and JustPressRecord ASR Results

S1(LL1)
Some animals sleep longer than others do.

C-1a: I must say wrong about us do Sam I'm saying
JPR-1a: Some animals three longer than others do
NS-1a: Some animals _______ longer than others do

C-1b: Somebody Mars Street on the other of us do
JPR-1b: Some animals sleep longer than others do
NS-1b: Some animals live longer than others do

S2(LL1)
Some ocean animals, such as whales, have to swim to the surface for air.

C-2a: Some astronomers such as bears have to seem to have faith for air
JPR-2a: Some other animals such as bears have to swim to the surface for air
NS-2a: Some _______ animals ____________________________________

C-2b:: _______ Ocean such as bears have to seem to have phase for air
JPR-2b: Some ocean animals such as birds have to swim to the surface for air
NS-2b: Some sea animals, such as whales, have to swim to the surface to breathe.

S3(LL2)
What course do you take at your school?

C-3a: What colors do you take at your school
JPR-3a: What colors do you take at your school
NS-3a: What course do you take at your school [courses?]

C-3b: What course do you take at your school
JPR-3b: What course do you take at your school
NS-3b: What course do you take at your school

S4(LL2)
During their school life, the students often face the lives and deaths of their animals.

C-4a: During their school life the students often freeze dog lives and death of their I'm else
JPR-4a: During the school life the students often FaceTime lives and deaths of their animals
NS-4a: During their school life, the children often _______ the life and death of their animals

C-4b: During their school life the students often face lives and this is of their animals
JPR-4b: During their school life the students often face the lives and deaths of their animals
NS-4b: During their school life, the students often face the life and death of their animals

S5(LL3)
I had no interest in trying the phone in our new house.

C-5a: I have no interest in trying the phone eat our new house
JPR-5a: I have no interest in trying the phone in our new house
NS-5a: I have no interest in _______ into our new house

C-5b: I have no interest in trying to cell phone eight our new house
JPR-5b: I have no interest in trying the phone in our new house

NS-5b: I have no interest in trying _______ in our new house

S6(LL3)
When I was very young, my family had one of the first telephones in our neighborhood.

C-6a: Bananas very young I find a hot while I'm at the foster funds eat our neighborhood
JPR-6a: When I was very young my family had one of the foster phones in our neighborhood
NS-6a: When I was very young, our family had one of the very first telephones in our neighborhood

C-6b: When I last Verianc my family hot while I was in foster phones in our neighborhoods
JPR-6b: When I was very young my family had while I was fast telephones in our neighborhood
NS-b: When I was very young, my family had one of the very first telephones in our neighborhood

S7(NS)
By deep frying noodles, they become hard and dry.

C-7a: Jo buy the frying noodles they became hard and dry
JPR-7a: Buy deep frying noodles they became hard and dry

C-7b: Buy deep frying noodles they became hard and dry
JPR-7b: Buy deep frying noodles they became hard and dry

S8(NS)
The angels celebrate the birth of (Jesus) Christ by singing and playing musical instruments.

C-8a: The Angels celebrate the birth of Jesus Christ by singing and playing musical Yzerman
JPR-8a: The angels celebrate the birth of Jesus Christ by singing and playing musical instruments

C-8b: The Angels celebrate the birth of Christ by singing and playing musical instruments
JPR-8b: The angels celebrate the birth of Christ by singing and playing musical instruments

Appendix 2: Interview Comments

LL1:
I found JustPressRecord (JPR) to be a great way for students to practice pronunciation on their own. If you have JPR, you can practice pronunciation at school or at home anytime, so when I become a teacher, I would like to introduce JPR to my students. Also, I would like to practice pronunciation using JPR for my English learning.

LL2:
I think there are quite a few teachers who have enough concerns about their own pronunciation that they might not feel confident in teaching it. For me, if I were able to use this app, I would feel more at ease about teaching pronunciation, and I feel that it would be good for both teachers and students alike. I definitely want to use it when I become a teacher!

Up to this point, I have tried to practice pronunciation by reading English textbooks aloud, but I couldn't objectively identify which parts were difficult to pronounce and might be difficult for listeners to clearly comprehend. So this app would be quite good for me as an English learner to objectively identify my own weak points. (such as, for example, the distinction between "th" and "d").

When I was a junior high school student, my English teacher taught us phonetic symbols and phonics, but we didn't really have all that much opportunity to actually pronounce English. So I now wish we had had the opportunity to use this app to practice pronunciation then. Also, this app gives the students a more natural experience of practicing pronunciation. Therefore, they can remember the learning experience in a more positive light, and this will make it easier for them to master English pronunciation.