Adopting Speech Recognition in EFL/ESL Contexts: Are We There Yet?

Lan Vu (vulan@siu.edu)
EPAS Learning Service, USA
https://orcid.org/0000-0002-5147-9125
Thom Thibeault (tthibeau@samford.edu)
Samford University, USA
https://orcid.org/0000-0002-6434-8896
Phu Vu (vuph@unk.edu)
University of Nebraska at Kearney, USA
https://orcid.org/0000-0003-2862-9110

Received: December 21, 2020
Accepted: January 6, 2021
Published: January 12, 2021

Abstract: This paper reviews the advancement of using speech recognition (SR) technology in EFL/ESL classrooms in the last few decades, addresses researchers' and educators' concerns about the limitation of this technology and examines how far SR technology has been evolving in its own field. Finally, potential pedagogical implications of SR technology for EFL/ESL, its limitations and suggestions for further studies are discussed.

Keywords: speech recognition, SR, technology, ESL, EFL, FL

Introduction

In 1918, Yale professor Charles C. Clarke (1918) discussed the potential of using a "talking machine" in teaching foreign languages (FL). He elaborated on the advantages of utilizing this technology for teaching pronunciation to FL students but also acknowledged the hurdles that prevented this technology from being effectively implemented in FL classrooms, especially the steep learning curve for some teachers and students as well as the cost of obtaining speech samples appropriate for FL instruction. Clarke (1918) also noticed "The silent verdict brought in by its general abandonment is that it is not worth the trouble it involves" (p. 116). Exactly 100 years later in 2018, inspired by Clarke's work, we attempted to revisit this topic by examining a new version of the talking machine, speech recognition (SR), in FL instruction.

In plain language, SR is the capability of an electronic device to recognize spoken words and respond to them accordingly. There are two types of SR. One common type of SR is "speech-to-text" or "dictation" software that allows users to record text as well as perform basic system commands. The other type of SR is interactive speech that allows users to speak to a device and receive responses from the device. The latter type of SR has two modes. The basic mode receives oral or verbal commands from users and makes responses to them with pre-programmed or pre-set knowledge that the designer of the SR wants the devices to respond to. In other words, the basic mode in SR technology can only respond to the users' commands, based on however they are programmed. The advanced mode uses artificial intelligence and/or machine learning technology that gives devices the ability to learn without being explicitly programmed to respond to the users' commands. Within the scope of this paper, we focus only on the basic mode of SR technology in FL instruction to 1) address two main concerns that Clarke mentioned: the steep learning curve for some FL teachers and students before operating it and the cost of obtaining speech samples appropriate for FL instruction, 2) examine how far SR technology has been evolving, and 3) provide several pedagogical implications, limitations and suggestion for further research about SR technology.

Also, in this paper, when we refer to the term "speech recognition", we loosely include voice recognition technology although, technically speaking, they are quite different. According to Coniam (1999), voice recognition requires machine training and is speaker-dependent while speech recognition is speaker-independent. In his research project, Coniam specifically named the software Dragon Naturally Speaking as "voice recognition" but Dragon Naturally Speaking itself was labelled as a speech recognition program. In our paper, it is not our intent to make any clear distinction between those two terms. It is argued that voice recognition is actually the early stage of the current speech recognition technology.

Literature Review

As acknowledged by Clarke (1918), "talking machine" technology is by no means new, and neither is SR technology. Computer scientists and engineers have been exploring SR as a mode of input to computers instead of using the keyboard since International Business Machines Corporation (IBM)'s initial experiments with its product called VoiceType in the 1950s. Nonetheless, its unreliability, high cost and lack of sophistication are often the reasons that critics cited that it was not worth the trouble it involved. For instance, previous generations of commercially available SR products such as Kurzweil's Voice or IBM's ViaVoice, a replacement of VoiceType, required discrete speech as the form of input in which each word needed to be uttered basically in its citation form (Coniam, 1999). In addition, Neri, Cucchiarini and Strik (2003) pointed out that SR technology had limitations in two main areas: the ability to recognize accented or mispronounced speech, and the ability to provide meaningful evaluation of pronunciation quality. Benzeghiba et al. (2007) added several other concerns about SR technology that affected the quality of speech recognition, including regional and sociolinguistic factors, sensitivity to the environment (background noise), and/or the weak representation of grammatical and semantic knowledge.

However, pioneers or early adopters of SR technology demonstrated its potential to help improve FL learners' pronunciation and speaking skills, but they also called attention to possible challenges. One of the early adoptions of SR technology in FL classrooms was a project by Coniam (1999), which examined the potential of using voice recognition technology with second language speakers of English. The researcher used SR software Dragon Systems 'Dragon Naturally Speaking', a commercial program, to test a small group of 10 English native speakers and another small group of 10 competent ESL learners/speakers. Each of the participants of both groups read a training text of approximately 3800 words, and then read a test text consisting of approximately 1050 words to the computer that ran the Dragon Naturally Speaking analysis. It was concluded that voice recognition technology was still at an early stage of development in terms of accuracy and single-speaker dependency. However, the analysis results indicated that the development of a voice recognition technology-based assessment might have some potential. Sharing some of Coniam's views on the potential of SR technology, Derwing, Munro, & Carbonaro (2000) contended that the possibilities for integrating SR technology in FL teaching settings were promising because it could offer useful negative feedback to learners in a nonthreatening context. However, those researchers also argued that until SR software satisfied the two criteria, namely reasonable accuracy levels to avoid frustration along with humanlike recognition patterns, it cannot be considered to be of benefit to FL learners either in classrooms or in business and personal contexts.

Not discouraged by the previous skepticism toward SR technology, Hincks (2002) took a shot at SR technology by examining whether unlimited access to a speech-recognition-based language learning program would help increase ESL learners' pronunciation competency. Eleven ESL learners were provided with a copy of the commercial SR-based program "Talk to Me" by Auralog as a supplement to a 200-hour course in Technical English. The researcher noticed that the substantial limitation of this commercial SR technology system was its inability to diagnose specific articulatory problems. Except for that drawback, learners in that research project reported high satisfaction with the software. It was also indicated that practice with the program was beneficial to those learners who started the course with a strong foreign accent but that learners who started the course with a moderate foreign accent did not show the same progress. Taking a different approach to Hick's research project of using a commercial SR product, Chen (2011) explored the potential of free SR technology in EFL by using the Microsoft Speech Application Software Development Kit (SASDK) to develop an oral skills training website for EFL learners. The SR-integrated website provided six different types of online exercises allowing EFL learners to practice their speaking skills and receive immediate feedback on their performance. Research participants consisted of a group of 25 college students and 35 pre-service EF teachers. Two research surveys were conducted to examine the participants' perceptions of using the SR-integrated website to practice their speaking skills. The findings showed that most participants enjoyed using the website, expressing that it could help improve their English speaking skills. The participants pointed out one of the main benefits of the SR-integrated practice: it provided many different types of exercises that could motivate learners to practice more in a stress-free environment. In addition, the participants mentioned some of the limitations of SR technology, including the insufficient feedback and the high benchmark they were supposed to meet in order to pass the tests. According to Chen (2011), the study's findings could be beneficial for teachers who are interested in implementing SR technology in teaching and for CALL researchers who want to develop better ASR-based systems for foreign language learning and teaching.

Although ASR has been criticized for its lack of reliability in language learning contexts, one study made use of this factor to identify words in speech samples that may be problematic for L2 learners. The study by Mirzaei, Meshgi, Akita, & Kawahara (2015) used ASR along with a Partial and Synchronized Caption system (PSC) designed to foster L2 listening skills. The algorithm in PSC is designed to generate captions on videos showing only the words that may be more problematic for L2 learners. For example, in a TED Talk video containing the phrase, "We are evolving to be a more collaborative and hearty species", the words "evolving", "collaborative", and "hearty" are isolated and only these words appear in the caption. In the study, transcription errors using ASR and problematic words identified by PSC were compared and overlapping words were used to improve the PSC algorithm for identifying which words should be marked as difficult for L2 learners. This was believed to improve the ability of instructors to customize learning activities based on the most problematic words in a given speech sample.

While previous research projects about SR integration in EFL had participants who were adults, Liaw (2014) examined the use of SR technology to support EFL learners' reading skills at the elementary level. The researcher tested SR technology from IBM called IBM RC with a group of eleven students in an elementary school in Taiwan. A tracking data system was used to monitor and analyze participants' performance. In addition, the researcher conducted field observations and interviews with both students and the teacher to understand how they responded to the feedback provided by the program and their perception of SR technology for learning. The findings revealed that participants thought the SR-integrated program helped them be more aware of their English pronunciation problems so that they could make corrections. The participants felt that they became more motivated and confident in reading and learning English than before. However, they also raised concerns about many limitations of this technology, the most concerned of which were over-correction, ineffective feedback, phonetic recognition restrictions, system breakdowns, slow reading speed, all-English interfaces, etc. In terms of performance, the tracking data indicated that the SR-integrated program did not assist to significantly raise the participants' reading accuracy level. Similar to Liaw's research project, also in the same year, Elimat and AbuSeileek (2014) conducted a study using an SR-based program called "Tell Me More Performance English'' with a group of 64 third grade students at an elementary school in Jordan. The participants were randomly assigned into four groups, three experimental and one control. The three experimental groups were taught using "Tell Me More Performance English'' program to practice pronunciation while the control group was taught using regular instruction. All of the groups were instructed by the same teacher. The results of the study indicated that the SR-based program provided a great opportunity in teaching and learning pronunciation. Participants in the experimental group gained better results on the pronunciation test than their peers in the control group.

Also interested in the potential of SR technology for the improvement of EFL learners' speaking proficiency, but from a different perspective, Lee (2016) examined their experience of a mobile-based learning system enhanced by SR technology, using Google Voice. Three hundred and two students from five middle schools were provided with the SR-based application to use in and out of the classroom for two weeks. After using the application, the researcher sent out a survey to the participants for their responses. The results revealed that the participants were generally satisfied with the SR-integrated application because it made speaking practice more interactive, enjoyable, and motivating. Recent studies about the use of SR technology in classrooms (Chen, 2016; Tatman & Kasten, 2017) also yielded the same results indicating that students enjoyed learning with SR technology, and they perceived that this technology could help improve their English oral skills.

In summary, while expressing some concerns about SR limitations in its early versions, researchers and educators in the field of EFL/ESL seemed to agree that SR technology has potential in EFL/ESL. It is also noticed in the literature that over the years reports on SR adoption with positive impacts were more dominant.

Current Stages of SR Technology

A quick search of SR programs or software on dominant search engines such as Google or Bing can easily yield hundreds of results related to SR, ranging from common commercial programs such as Dragon Naturally Speaking to free ones such as Google's web speech recognition API. To address two main concerns by Clarke: the cost of obtaining speech samples appropriate for FL instruction and the steep learning curve for some FL teachers and students before implementing it, we selected the Google web speech recognition API, a speech recognition program available on the Chrome web browser and/or Android-operated mobile devices, to demonstrate how this technology has developed and how we can use it in EFL/ESL contexts.

The Google web speech recognition API is totally free and accessible to any users/learners who have Internet-connected devices. According to Statcounter Global (2018), a web-based tool specializing in measuring Internet usage trends and browser traffic tracking, a Chrome browser, in which Google web speech recognition API is integrated, accounts for 56.31% web browser usage worldwide. In terms of its usability, which refers to the ease of use of the program, it is argued that the Google web speech recognition API is one of the easiest SR technologies available. There is no learning curve for the user. They simply click on the microphone icon on the screen, speak, and receive the responses, feedback or commands, depending on what specific program the users are using. As its name suggests, the Google web speech recognition API was originally developed and deployed for computing devices, mainly computers that run Google Chrome web browser. To be able to use this SR technology, users need to have access to Internet-connected computers that have the Chrome browser. The simple and intuitive design of the Google interface has made the adoption of SR into EFL/ESL arguably effortless. However, one of the limitations is the feeling of a formal learning process in which users need to turn on their computers whenever they need to learn or practice, which is not the goal of "learn anytime and anywhere".

The popularity of mobile devices in the last decade has pushed SR technology to another level where it has been integrated into virtual assistants that can talk to users. EFL/ESL learners can now learn and practice speaking anytime and anywhere at their fingertips. Many technology enthusiasts even claimed that SR-powered apps would soon disrupt conventional EFL/ESL educational approaches, especially in the field of speaking and pronunciation. While the optimism is there, holding a device in front of the users' mouth to speak and practice pronunciation is still not a natural learning process.

The third generation of SR technology, smart home devices such as Google Home or Amazon's Echo or Alexa, probably has more meaningful and significant impacts on

EFL/ESL speaking and/or pronunciation teaching. Serving as an artificial intelligent chatbot, those smart home devices like Google Home can be anywhere in the room, listen to the user's utterance and respond to it naturally. In our professional development project to help improve English pronunciation-teaching skills for 155 EFL teachers in Vietnam, we created an SR-based pretest and post-test evaluation and had all of the participants take the tests. Since all of those teachers were from remote and/or mountainous regions in Vietnam, taking those SR-based tests was their first exposure to this technology. They talked to the machine without any major issues except for a few minor Internet problems. The experience of using SR technology in our professional development sessions, even for the very first time, was so easy and smooth that all the participants expressed their excitement and desire to further explore the use of this technology in their classrooms.

Pedagogical Implications of Speech Recognition Technology for ESL/ESL

SR technology has a number of applications for EFL/ESL. It can be used as conversational chatbots to help EFL/ESL learners practice daily conversations in nonthreatening judgement-free contexts. Artificial intelligence (AI)- powered home devices with SR integration such as Google Home, Alexa from Amazon or Microsoft’s Cortana are typical examples of conversational chatbots that EFL/ESL learners can use to practice English 24/7 by simply talking to the device and it will respond in a human voice, answer questions, or even crack jokes. SR can also help learners to engage in conversation, especially those who do not have access to native speakers to practice pronunciation. Even without a partner being available, learners may work alone to dictate a text or act out a dialogue so that they are less reliant on their teacher or native speakers for constant feedback. Moreover, teachers may use chatbots to offer personalized learning experiences to learners with different learning levels. For instance, SR can be served as a remedial function for learners with a beginning level of pronunciation and/or speaking by allowing them to see the words on screen as they dictate or "speak" so that they may gain insight into key elements of phonemic awareness. SR may also be deployed to personalize learning paths for ESL gifted students or high ability learners whose learning needs are different from their peers. Given the advance of increasing technological innovation, the potential of SR technology in ESL/EFL is promising. If appropriately implemented, SR can have significant impacts on students' EFL/ESL performance.

Limitations of SR Technology in ESL/EFL

Like any other technologies, educators should proceed with caution. Learners' privacy and security may be jeopardized because SR-powered devices can listen to the users all the time without their awareness (Vu, Fredrickson, & Gaskill, 2019). One typical example is the Google speech recognition API. There may be concerns about privacy, especially in a learning environment where minors are involved. Google has recently come under scrutiny for the manner in which it collects and distributes data on those who use their products. This scrutiny focuses on Google products such as Google Search, Youtube, and Chrome. As of this writing, Google does not collect data on recordings or text generated through its SR software. There is a logging option but, by default, no data is logged. More details can be found at cloud.google.com/speech-to-text/docs/data-logging.

Suggestions for Further Research

As discussed in the “Limitations” section, future research may examine the effect of solely using SR technology on the development of learners' learning autonomy. In addition, more rigorous experimental studies to compare groups of students who use SR in their learning and those who do not are needed to measure the effectiveness of this promising SR technology. Finally, aspects of users' data and learning ethics should also be investigated to address privacy concerns.

References

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... & Rose, R. (2007). Automatic speech recognition and speech variability: A review. *Speech communication*, *49*(10), 763-786.

Clarke, C. C. (1918). The phonograph in modern language teaching. *The Modern Language Journal*, *3*(3), 116-122.

Chen, H. H. (2011). Developing and evaluating an oral skills training website supported by automatic speech recognition technology. *ReCALL : The Journal of EUROCALL, 23*(1), 59-78. http://dx.doi.org/10.1017/S0958344010000285

Chen, H. H. J. (2016, October). Developing a Speaking Practice Website by Using Automatic Speech Recognition Technology. In *International Symposium on Emerging Technologies for Education* (pp. 671-676). Springer, Cham.

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. *System*, *27*(1), 49-64.

Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech?. *TESOL quarterly*, *34*(3), 592-603.

Elimat, A. K., & AbuSeileek, A. F. (2014). Automatic Speech Recognition Technology as an Effective Means for Teaching Pronunciation. *JALT CALL Journal*, *10*(1), 21-47.

Hincks, R. (2002). Speech recognition for language teaching and evaluating: a study of existing commercial products. In *INTERSPEECH*.

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.

Liaw, M. L. (2014). The affordance of speech recognition technology for EFL learning in an elementary school setting. *Innovation in Language Learning and Teaching*, *8*(1), 79-93.

Mirzaei. M. S., Meshgi, K., Akita, Y., & Kawahara, T. (2015). Errors in automatic speech recognition versus difficulties in second language listening. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), Critical CALL – Proceedings of the 2015

EUROCALL Conference, Padova, Italy (pp. 410-415). Dublin: Researchpublishing.net. http://dx.doi.org/10.14705/rpnet.2015.000367

Neri, A., Cucchiarini, C., & Strik, W. (2003, August). Automatic speech recognition for second language learning: how and why it actually works. In *Proc. ICPhS* (pp. 1157-1160).

Tatman, R., & Kasten, C. (2017). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. *Proc. Interspeech 2017*, 934-938.

Vu, P., Fredrickson, S. & Gaskill, M. (2019). One-To-One initiative implementation from insiders' perspectives. Tech Trends, 63(1). https://doi.org/10.1007/s11528-018-0359-5