Speech-to-text applications’ accuracy in English language learners’ speech transcription

Akiyo Hirai, University of Tsukuba
Angelina Kovalyova, Sophia University

Abstract

Speech-to-text applications have great potential for helping students with English language comprehension and pronunciation practice. This study explores the functionality of five speech-to-text (STT) applications (Google Docs voice typing tool, Apple Dictation, Windows 10 Dictation, Dictation.io [a website service], and “Transcribe” [an app on iOS]) to measure their speech transcription accuracy of American English. The experiment involved 30 nonnative speakers, who were asked to perform four speaking tasks and whose speeches were recorded and transcribed with these applications. The transcriptions produced by the applications were then compared with human-made transcriptions to evaluate the accuracy rate of each application’s speech transcription ability. The results revealed that the accuracy rate of speech transcriptions depends not only on the applications’ automatic speech recognition ability but also on the types of speech produced, as well as each speaker’s L1 influence on L2 (English). The study also offers examples of Japanese speakers’ pronunciation errors attained through STT transcription, demonstrating great pedagogical potential for pronunciation practice and assessment in English classrooms.

Keywords: Automatic Speech Recognition, Speech-to-text Applications, Pronunciation, Loanwords

Language(s) Learned in This Study: English

APA Citation: Hirai, A., & Kovalyova, A. (2024). Speech-to-text applications’ accuracy in English language learners’ speech transcription. Language Learning & Technology, 28(1), 1–21. https://hdl.handle.net/10125/73555

Introduction

In this digital age, automatic speech recognition (ASR) technology has become a ubiquitous tool available in many digital devices. It helps us communicate with a computer by converting speech to text (Microsoft, 2004) and is especially useful when human support is unavailable. Currently, ASR software is commonly represented by digital assistants, such as Siri and Alexa, which can execute small, pre-programmed requests on our digital devices. There is also a myriad of speech-to-text (STT) transcription services, compiling transcriptions of speaking events (conversations, meetings, presentations) to provide subtitles or event scripts.

More recently, speech recognition software has attracted attention as an educational tool for English as a foreign language (EFL) learners to practice pronunciation (Evers & Chen, 2020; McCrocklin, 2019; Tejedor-Garcia et al., 2020; Tejedor-García et al., 2021; Vaughn et al., 2019). However, our interactions with the digital world have progressed only to the extent to which technology has moved forward, and despite these efforts, speech recognition software cannot fully cope with the spontaneity and unpredictability of human speech. There are still noticeable gaps in ASR technology in comprehending the prosody of nonnative speakers (NNS; O’Brien et al., 2018) and assessment of NNS’s segmental properties (Tejedor-García et al., 2021). Thus, this study aims to examine the intelligibility of nonnative speech of adult English language learners (American English variety) using current ASR software in English language classrooms.
Speech Recognition Software

The use of speech recognition software in English language classrooms has been of interest to language researchers since the 1990s when speech recognition software became available to consumers (Coniam, 1998; Golas, 1995; Kincaid, 2018; Manes, 1997; Noyes & Starr, 1996). The software uses technology that allows a computer to convert audio into text (Microsoft, 2004), which allows humans to not only interact with a computer but also to record their speech for various purposes. Following the progressive development of computing power and the Internet, speech recognition technology has improved considerably and has been extensively used for pedagogical purposes (Hwang et al., 2012; Kuo et al., 2012; McCrocklin, 2019; Tejedor-García et al., 2020).

The process of speech recognition involves several steps, the goal of which is to break apart the spoken text into phonemes and find the most probable combinations of these phonemes in a particular language using a statistical model (Zajechowski, n.d.). This process relies on an acoustic model (a reference to examples of phonemes and their frequency), a language model/dictionary (examples of words and their sequences), a statistical model (a tool that considers the probability of an utterance), a cloud server (a place that stores all the data), and a computer that can implement this process (Kang & Ginther, 2017, p. 139).

Today, the digital marketplace offers a myriad of speech recognition tools to choose from. However, these platforms differ in their purpose, functionality, and accuracy rate, and it becomes very challenging for educators to determine how reliably an application can be utilized for language practice.

Problems with Accuracy

Neri et al. (2003, p. 1158) expressed a concern that some speech recognition technologies designed mainly for dictation were built to recognize acoustic properties of native speakers (NS), and current research shows that favoring NS might still be common among ASR applications (e.g., McCrocklin et al., 2019). In 2017, Google claimed to have achieved a 95% accuracy rate in English (Worthy, 2019). At the same time, Microsoft produced an accuracy rate of 93.1% in the switchboard test, with a possibility of 99% accuracy with additional training (Hachman, 2017). Both assessments were conducted using the NS (English) dataset from the National Switchboard Corpus prepared by Godfrey and Holliman (1993).

Despite continuous improvements in speech recognition technology, more recently, Koenecke et al. (2020) assessed the speech recognition accuracy of NS among five ASR systems (Amazon, Apple, Google, IBM, and Microsoft), comparing speech recognition of White and Black English language speakers. The study reported an average accuracy rate of 81% (85% for Microsoft ASR) for White speakers of English. Interestingly, the results also showed a 65% accuracy rate for Black speakers of English. Researchers associate this significant drop in accuracy with African American Vernacular English (AAVE), a variety of English used by African American speakers. Researchers believe that the acoustic models of the studied ASR systems do not account for the disparities in pronunciation and prosody of different language varieties because they lack AAVE data samples (Koenecke et al., 2020, p. 7687).

In addition, one of the biggest concerns remains the accuracy of recording NNS by ASR applications (McCroclkin et al., 2019, p. 193). McCrocklin et al. (2019) tested the performance of Windows Speech Recognition and Google Voice Typing, comparing the accuracy of English NS and NNS recognition. When analyzing the accuracy for rate of speech recognition of English NNS, the results showed an accuracy rate of 53.50% (free speech) and 74.44% (reading sentences) for Windows Speech Recognition and 93.47% (free speech) and 88.61% (reading sentences) for Google Voice Typing. Although the reason for such disparity in accuracy scores was unclear, a further drop in accuracy rates of NNS recognition was observed. Considering that ASR technology cannot recognize the meanings of words (relying on acoustic model and statistical analysis instead), the ASR’s ability to make sense of segmental features (such as training the pronunciation of minimal pairs) should be a priority to increase intelligibility, that is, the accuracy of speech recognition (Thomson, 2018).
Major issues remain unclear with voice recognition of (a) spontaneous speech and (b) unfamiliar vocabulary and local varieties of a language (Worthy, 2019). When talking about (a) automatic recognition of spontaneous speech, experts in the field note that the current language model of speech recognition software is not yet suitable for analyzing disorganized spontaneous speech well, but that the task is suitable for artificial intelligence (Jarnow, 2016). Regular human speech is full of false starts, repetitions, hesitations, corrections, and mispronunciations, which poses a challenge for the algorithm. However, we are not sure to what extent the transcription accuracy declines from prepared speech to spontaneous speech. Regarding (b) treatment of unfamiliar words by STT applications, for some cultures, such as the Japanese language, pronunciation features in the native language noticeably differ from English pronunciation. Specifically, the use of katakana, a subset of Japanese characters that helps transliterate English words, strongly affects the common pronunciation of some English words, making it difficult for native speakers to understand (Koon, 2018, p. 84). In fact, by using katakana, Japanese speakers fossilize pronunciation features that are uncommon for English (Martin, 2004), which inevitably results in a lower accuracy rate when it comes to transcribing the English speech of Japanese speakers. However, how STT applications treat these foreign words when embedded in English speech has not been investigated. This point is important, as speakers sometimes include their names and loanwords when speaking English. Thus, it is worth investigating how STT applications are affected by these words in both prepared and spontaneous speech conditions.

As a way to battle ASR’s low accuracy scores, Computer Assisted Pronunciation Training (CAPT) programs started being developed to help NNS target their pronunciation issues. However, with the limited functionality of CAPT (McCrocklin et al., 2019, p. 192) and focusing mainly on minimal pairs (Tejedor-García et al., 2021), interest in STT applications remains. Most recent research on nonnative speech assessment has shifted its focus to advancing the tools that can recognize comprehensible and intelligible speech rather than native-like speech (O’Brien et al., 2018). Thus, it is crucial to continue to assess the advancement of ASR technology in this new direction to ensure that educators can use speech recognition tools for language practice in classrooms full of NNS who hesitate, rephrase, and mispronounce words while learning a new language.

**Five STT Applications**

Speech recognition tools available online for free or for a fee vary in quality and purpose, as well as in terms that describe their role, such as ASR system, speech recognition software, dictation software, voice-to-text software, speech-to-text software, and transcription software. For the purposes of this study, the term STT software was adopted to refer to the computer and mobile software used to transcribe recorded speech into text.

We selected five STT applications for this study: (a) Google Docs voice typing tool, (b) Apple Dictation, (c) Windows 10 Dictation, (d) Dictation.io (a website service), and (e) “Transcribe” (an app on iOS). The reason for choosing these five tools is based on the need to provide a review of free and accessible tools available to anyone, yet which are based on different platforms. Google, Apple, and Microsoft products represent three major platforms. Dictation can be accessed anywhere through a browser. “Transcribe” (although available only on iOS devices) can be used on mobile devices. Below is a brief description of each service with the features available as of spring 2020.

**Google Docs Voice Typing**

Google Docs voice typing tool is a speech-to-text feature available in cloud-based Google Docs and Google Slides services using a Chrome browser (Figure 1). In Google Docs, voice typing is accessed by opening a Google Docs document in a browser, selecting the “Tools” tab in the top menu, and then “Voice typing” in the drop-down menu (Google, n.d.-a).
Currently, voice typing supports 62 languages (or 119 language varieties, including 13 varieties of English; Google, n.d.-b). Google uses its own speech recognition engines, algorithms, and AI technology to process speech. As speech recognition is conducted in the cloud, the service should be used with an Internet connection. The voice typing feature is free of charge and can be used by anyone.

It is worth noting that Google also offers a separate speech recognition service called speech-to-text, aimed at larger commercial projects, in which it continuously shows improvements in language support, models for speech recognition, speech adaptation, speaker diarization, automatic punctuation, and more (Barnes, 2020). Considering that voice typing uses Google’s speech recognition system, it is speculated that voice typing features benefit from the progress Google has achieved with its speech recognition models. This feature has a few troubleshooting elements. For one, it is possible to see a message on the screen saying, “We’re having trouble hearing you,” suggesting that there is too much noise around or the level of input sound is insufficient, which might influence speech recognition accuracy. Additionally, while dictating, the voice typing tool suggests possible transcriptions of words if the pronunciation is not very clear (Figure 2).
**Apple Dictation**

Dictation is a speech-to-text tool available on iOS devices in any document where typing is possible. To use Dictation on a laptop, it is necessary to open “System Preferences” in laptop settings, choose the “Keyboard” category, and in the “Dictation” tab, activate the feature and select the language of input (Figure 3). After that, once a text document (e.g., Microsoft Word or Notes) is open, it becomes possible to start speech recognition by pressing a key shortcut specified in the settings.

**Figure 3**

*Apple Dictation in Settings*

![Apple Dictation in Settings](image)

Apple Dictation supports 32 languages (or 62 varieties) within its regular dictation service. Apple also requires an Internet connection, as the data are analyzed using Apple servers and AI (Apple, n.d.). Moreover, according to Apple Support, Apple Dictation has the potential to improve its accuracy with repeated exposure to a person’s speech. Dictation is available free of charge across Apple devices (Apple, 2015).

Using Apple Dictation during the experiment, it was noticed that continuous dictation was a challenge, and it was necessary to dictate only a sentence or two at a time. When too much text was dictated at a time, the dictation tool wrote a portion of the text dictated and then erased it. Apple’s speech recognition performance seems to be congruent with the reports on Siri’s speech recognition ability, as misinterpretation of numerous users’ requests has been noticed, especially when questions were longer (Mossberg, 2016). While Siri’s functions may be limited by developers (Velde, 2019), the issues with the interpretation of requests pose a question about Apple’s speech recognition accuracy.

**Windows 10 Dictation**

Windows 10 Dictation is similar to Apple Dictation, in that this tool is free of charge and pre-installed with the device (in this case, a device that uses a Windows operating system). Windows 10 Dictation is a newer addition to the speech recognition system built into Windows OS and requires an Internet connection (Microsoft, n.d.). To use it, one should start with the settings in Windows OS and search for the “Speech” menu, where it is possible to set the language settings and even opt for “Recognize nonnative accents for this language” (Figure 4).
This option is a useful consideration when it comes to NNS learning a foreign language. It is also possible that the system will first ask to set up speech recognition and a microphone. When speech recognition is set up, Windows 10 Dictation is prompted by hitting the Windows logo key + H key on the keyboard. It currently supports only seven languages (or 12 varieties), and if the necessary language is not available, Microsoft Support suggests using original speech recognition (Microsoft, n.d.).

A useful feature of the Windows 10 Dictation is its troubleshooting. When the text is spoken too fast or in too large a portion, it shows a message saying, “Hang on, we need a moment to catch up” (Figure 5). With this feature, Windows offers suggestions for better input and, therefore, helps improve the accuracy of speech recognition.

The Website “Dictation.io”

The “Dictation.io” website is a very straightforward, easy-to-use speech-to-text service that can be accessed.
on Google Chrome regardless of the computer operating system and Google account availability. Once the website is opened, the speech language is selected, and the person starts speaking, the website displays continuous transcription of the speech (Figure 6).

**Figure 6**

*The Website “Dictation.io”*

The Website “Dictation.io”

Dictation uses the Google Speech recognition system and requires an Internet connection to send the audio to Google servers. It offers speech recognition for 69 languages (or 134 language varieties), and additionally, the website allows the use of the enhanced spell check feature provided by Google (Figure 7).

**Figure 7**

*The Website “Dictation.io” Enhanced Spell Check Feature*

**iOS Application “Transcribe”**

“Transcribe,” an application available on iOS devices, offers a speech recognition service by transcribing an uploaded voice recording, as well as by offering a dictation feature (a new feature added in September 2020). The application supports over 120 language varieties, as it utilizes Google’s language database (M. Grushin, personal communication, October 9, 2020). “Transcribe,” however, makes use of several different providers of audio recognition depending on the type of audio (live or pre-recorded) and language. The application is free of charge for the first 15 minutes of transcription and requires a premium subscription option if the recordings exceed the free option.
An interesting feature of “Transcribe” is that it offers time stamps for recording as well as accuracy stamps within the transcribed text (Figure 8). In other words, the application predicts the transcription accuracy. Additionally, it is possible to listen to the audio and follow the highlight that identifies the spoken section. This seems to be a particularly useful feature for language learners as it allows them to notice the words that were misunderstood by the application and, therefore, mistranscribed.

Figure 8
Accuracy and Timestamp Features in iOS Application “Transcribe”

These five widely used applications, four of which are free to use and one which has a free trial, were examined in this study.

The Significance and Research Questions of the Study

Analyzing the accuracy of modern STT software and knowing its function is essential because teachers can examine whether it can help students practice their pronunciation and find their weak spots. This is especially helpful for learners who have strong accents but cannot receive immediate feedback from their teachers. Working on their own pronunciation will also help them become more independent and autonomous in their learning (McCrocklin, 2015, pp. 126–127). STT software can also help teachers save a lot of time in assessing individual students’ pronunciation.

This study is especially useful in the context of COVID-19 when access to face-to-face practice with NS is limited. However, the nature of STT software has not yet been investigated, especially regarding how it can cope with NNS’ English speeches under various speech conditions. Thus, this study aims to investigate how accurately these five STT applications transcribe EFL learners’ speeches for classroom use and self-practice. To examine the features of these applications, we prepared four tasks that elicit different types of speech, such as prepared or spontaneous speech, simple or complicated speech, and speech containing loanwords and katakana words. Accordingly, this study examines the following four research questions:

RQ1. To what extent do the five STT applications transcribe EFL learners’ speech?

RQ2. How does speech type influence the transcription accuracy of the five STT applications?

RQ3. How do STT applications transcribe loanwords and katakana words when embedded in English speech?

RQ4. What EFL learners’ words are not transcribed accurately?
Method

Participants

Thirty university students in Japan with the following first languages (L1s) participated in the study: 18 Japanese, 5 Chinese, 3 Korean, 1 Urdu, 1 Czech, 1 Hungarian, and 1 French. Regarding their English language proficiency levels based on the questionnaire, 4 students were at the A1 English level, 4 at the B1 level, 15 at the B2 level, and 7 at the C1 level, according to the Common European Framework of Reference for Languages (CEFR) framework. In addition, 13 of them reported having had a prolonged communication experience abroad.

Having students of varied linguistic backgrounds was deemed necessary, as the goal of this study was to observe the ASR’s ability to access the intelligibility of nonnative English speech in a broader context. In other words, as the study aimed to test the accuracy of speech recognition for English practice in EFL classrooms as well as self-practice, it was important to use the speech data of participants who could potentially benefit from these STT applications. Additionally, it is assumed that the application’s speech recognition accuracy depends on the quality of the input in terms of grammatical and lexical clarity. Therefore, using the speech of EFL learners with different levels of language proficiency can help identify whether this assumption is true.

Materials

Five Speech-To-Text Applications

Speech analysis was conducted using five STT applications: (a) Google Docs voice typing tool, (b) Apple Dictation, (c) Windows 10 Dictation, (d) www.Dictation.io, and (e) the iOS mobile application “Transcribe.”

Four Speech Tasks

Four different speaking tasks were prepared. Task 1 (T1_ReadS) required participants to read two short sentences with Japanese words (“haiku” and “bukatsu”), eliciting prepared short speech, but that which contained presumably unfamiliar words for speech recognition tools. Task 2 (T2_ReadL) asked students to read a short passage with a basic vocabulary to help them produce a balanced prepared speech. Task 3 (T3_Retell) asked students to retell the short paragraph from Task 2, thus producing a relatively long semi-spontaneous speech. The difficulty of this passage for EFL learners and its use for a retelling task has been validated in a different study (Hirai & Koizumi, 2009). Finally, Task 4 (T4_QA) required students to read three questions and answer them freely and spontaneously, using the vocabulary and grammatical competence corresponding to the students’ English proficiency level. One of the questions also involved using Japanese words (“senpai” and “kouhai”) to check how the STT applications would manage to transcribe unfamiliar words or loanwords. Both Tasks 1 and 4 were checked by a native English speaker to determine whether the loanwords in the sentences were naturally used.

Procedure

The experiment consisted of two stages. In the first stage, the participants were asked to execute the four tasks with their voices recorded via a voice recording application on an iPhone. The recordings were performed in a quiet room using a microphone to eliminate external noise. Additionally, each participant was asked to complete a short questionnaire explaining their English language ability (i.e., proficiency level, experience living or studying abroad, and cultural background).

The next stage involved the researchers playing the recordings against each STT application and letting them produce the participants’ speech transcriptions within their native word-processing element of the application, later transferring the transcriptions to a word document. The audio recordings were also played in a quiet room to prevent external noise from interfering with the speech recognition process.

For logistical reasons, voice recordings were used rather than live voice feed. It is impossible to have each participant perform four speaking tasks five times (feeding the speech into five different STT applications).
to specifically compare different speech types. Additionally, considering that audio recordings have both prepared and spontaneous speech, it was assumed that the difference between them would be noticeable.

It is also important to note which language settings were used when transcribing the audio files because each of the five STT applications has multiple language recognition options, especially when it comes to English. Google Docs voice typing tool, Apple Dictation, www.Dictation.io, and the “Transcribe” application contain 13 English language varieties, and the U.S. English language option was utilized. The Windows 10 Dictation has only three English language varieties, and the U.S. English language option was selected in addition to ticking off the menu option to “Recognize non-native accents for this language.”

Once the transcriptions were saved, each application-produced transcription text was analyzed against the original text (Tasks 1 and 2) and human-written transcription (Tasks 3 and 4) to count the number and percentage of correctly transcribed words. The study adopted repeated-measures ANOVA tests for RQ1 and 2, followed by a qualitative method in which the researchers calculated and examined incorrectly transcribed words for RQ3 and 4.

Results

Transcription Accuracy of the Five STT Applications

The transcriptions of each participant’s four tasks were completed using the five STT applications. Table 1 shows the descriptive statistics regarding the means and SDs of the transcription accuracy rates of participants’ speech. Depending on the STT applications and the four tasks, the accuracy rates ranged widely, but mostly ranged between 50% and 70%.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (%)</td>
<td>SD</td>
<td>M (%)</td>
<td>SD</td>
</tr>
<tr>
<td>Google</td>
<td>64.46</td>
<td>19.90</td>
<td>64.28</td>
<td>19.53</td>
</tr>
<tr>
<td>Apple</td>
<td>45.38</td>
<td>18.38</td>
<td>52.44</td>
<td>17.45</td>
</tr>
<tr>
<td>Windows 10</td>
<td>60.97</td>
<td>17.59</td>
<td>69.75</td>
<td>13.87</td>
</tr>
<tr>
<td>Dictation</td>
<td>58.42</td>
<td>19.41</td>
<td>57.39</td>
<td>19.23</td>
</tr>
<tr>
<td>Transcribe</td>
<td>65.97</td>
<td>18.16</td>
<td>68.11</td>
<td>16.72</td>
</tr>
</tbody>
</table>

Note. ReadS (reading short sentences), ReadL (reading a passage), Retell (retelling the passage), and QA (answering questions)

To examine the interaction between the applications and speech conditions, a two-way (application with five levels by task with four levels) repeated-measures ANOVA and Mauchly’s test of sphericity was conducted. As the sphericity was not met on two factors of Application and Application × Task, Greenhouse-Geisser adjustment was applied to these factors. There was a significant interaction between application and task: $F(7.36, 213.56) = 4.42, p < .001, \eta^2_p = 0.13$. This means that the type of speech influences the accuracy of transcription differently. In other words, the applications have strengths and weaknesses regarding the types of speech. In addition, the main effects of the application and task factors were both significant: $F(2.78, 80.58) = 52.85, p < .001, \eta^2_p = 0.65$; $F(3.00, 87) = 7.60, p < .001, \eta^2_p = 0.21$.

As the interaction was significant, it was further analyzed by one-way ANOVA and multiple comparisons. As a result, there was an interaction between Windows 10 and other applications, as shown in Figure 9.
Windows 10 transcribed speech most accurately on T2_ReadL, but was relatively lower on T1_ReadS, in
which Japanese loanwords were embedded. In this regard, Windows 10 Dictation handled syntactically
simple predictable utterances the best, but it was more affected by the short speeches that contained
loanwords than the “Transcribe,” Google, and Dictation applications.

“Transcribe,” in contrast, performed better with tasks that elicited more spontaneous speech (Tasks 3 and
4). As shown in Table 2, overall, the quality of “Transcribe” was the highest, followed by the Google,
Dictation, and Apple applications. However, with the Bonferroni adjustment for multiple comparisons, the
differences between “Transcribe” and Windows 10 (p = .30, d = 0.10, 95%CI [0.01, 0.19]) and between
Google and Windows 10 (p = .14, d = 0.20, 95%CI [0.09, 0.32]) were not statistically significant, and the
magnitudes of the effect size were small.

Figure 9

Comparison of Five Applications

Note. ReadS (reading short sentences), ReadL (reading a passage), Retell (retelling the passage), and QA (answering
questions)

Table 2

Means and 95% CIs of Accuracy Rate of the Five Applications

<table>
<thead>
<tr>
<th>Application</th>
<th>M (%)</th>
<th>SE</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>63.09</td>
<td>3.39</td>
<td>56.16</td>
<td>70.01</td>
</tr>
<tr>
<td>Apple</td>
<td>48.97</td>
<td>3.06</td>
<td>42.72</td>
<td>55.22</td>
</tr>
<tr>
<td>Windows 10</td>
<td>66.96</td>
<td>2.79</td>
<td>61.26</td>
<td>72.66</td>
</tr>
<tr>
<td>Dictation</td>
<td>56.72</td>
<td>3.26</td>
<td>50.05</td>
<td>63.39</td>
</tr>
<tr>
<td>Transcribe</td>
<td>68.57</td>
<td>2.70</td>
<td>63.05</td>
<td>74.09</td>
</tr>
</tbody>
</table>
As for the comparison across the four tasks for RQ2, T4_QA was the highest, even though the task required learners’ unplanned spontaneous utterances (see Figure 10 and Table 3). This is because the two questions in this task were quite easy, and the learners could answer them smoothly. T1_ReadS (a task of reading short sentences that included Japanese loanwords) was transcribed differently by each application. The accuracy of retelling speech (T3_Retell) was the lowest, as it included the most utterances affected by the elements of natural speech such as false starts, repetitions, and self-corrections. T3_Retell uses the same reading passage as T2_ReadL, but the task of retelling was found to be significantly more difficult than the task of reading aloud (p = .032, d = 0.25, 96%CI [0.81, 0.42]).

**Figure 10**

*Comparison of Four Speaking Tasks*

![Comparison of Four Speaking Tasks](image)

**Table 3**

*Means and 95%CIs of Accuracy Rates by Tasks*

<table>
<thead>
<tr>
<th>Task</th>
<th>M (%)</th>
<th>SE</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1. Reading sentences</td>
<td>59.04</td>
<td>3.12</td>
<td>52.66</td>
<td>65.42</td>
</tr>
<tr>
<td>T2. Reading a passage</td>
<td>62.39</td>
<td>2.97</td>
<td>56.33</td>
<td>68.46</td>
</tr>
<tr>
<td>T3. Retelling the passage</td>
<td>57.60</td>
<td>3.57</td>
<td>50.31</td>
<td>64.90</td>
</tr>
<tr>
<td>T4. Answering questions</td>
<td>64.41</td>
<td>2.41</td>
<td>59.48</td>
<td>69.34</td>
</tr>
</tbody>
</table>
Words Wrongly Transcribed by the STT Applications

Next, to examine how STT applications transcribe loanwords embedded in English speeches, four Japanese loanwords were embedded into our speaking tasks: “haiku,” “bukatsu,” “kouhai,” and “senpai,” which participants had to use when completing the tasks. “Haiku” appears in English dictionaries and is in common use. “Senpai” and “kouhai” are common in English-Japanese cultural contexts and are therefore transliterated into English. “Bukatsu,” however, is a Japanese word that is not common outside of the Japanese language.

Table 4

Accuracy Rate of Transcription of Loanwords and Foreign Words

<table>
<thead>
<tr>
<th>Word</th>
<th>Correctly Transcribed</th>
<th>Most Often Transcribed as:</th>
<th>Best App</th>
</tr>
</thead>
<tbody>
<tr>
<td>haiku</td>
<td>34.67%</td>
<td>hiker (17), hike (9), Ohio (13)</td>
<td>Windows (13)</td>
</tr>
<tr>
<td>kouhai</td>
<td>31.21%</td>
<td>go high (9), call hi (5), cool I (5), go hi (4), co high (4)</td>
<td>Transcribe (30)</td>
</tr>
<tr>
<td>senpai</td>
<td>30.00%</td>
<td>empire (30), sent by (17), simpi (11), sempy (5), standby (4)</td>
<td>Transcribe (45)</td>
</tr>
<tr>
<td>bukatsu</td>
<td>2.00%</td>
<td>concert (13), two cats (6), Chicago (3)</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4 shows the results of the transcription accuracy rates for the four Japanese words. These words were only occasionally transcribed correctly with an approximately 30.00% accuracy rate, except for “bukatsu.” At other times, they were transcribed with words or phrases phonetically similar to the original words. For example, “haiku” was transcribed as “hiker” or “hike” (which is the homophone) and “kouhai” as “go high.” The Windows 10 software performed the best for the word “haiku,” while “Transcribe” did the best for the words “kouhai” and “senpai.” However, “bukatsu” had an accuracy rate of only 2.00%.

Transcription analysis revealed some words, in addition to loanwords, that were continuously transcribed incorrectly. Considering that the study participants were all NNS of English, it is possible to notice how some of their unique pronunciation features influence the accuracy of transcriptions.

Katakana Japanese (a Japanese writing system used for borrowed words) also affects the transcription of Japanese speakers’ pronunciation. As mentioned above, katakana Japanese utilizes a syllabary (or mora) system for transliteration. For instance, the word “etiquette /ˈɛtɪket/” is often pronounced as /ˈɛtıʃketto/ by Japanese speakers with a vowel attached after every consonant. When a speech recognition program hears /ˈɛtıʃketto/, this phonetic transcription does not match English “etiquette,” and so the speech recognition system searches for another closest phonetic match, offering a different word in the end (see Table 5).

Similarly, the Japanese name “Kenji” (/kendʒi:/) that appeared in the original text was wrongly transcribed as “can she” (/kən/ and /ʃi:/) or “candy” (/ˈkændi/) as shown in Table 5. Thus, when it comes to transcribing words not included in the language model of the speech recognition system, STT applications produced English words phonetically similar to the word.

Table 5

Transcription Errors Related to Katakana Pronunciation

<table>
<thead>
<tr>
<th>Original</th>
<th>Wrong Transcriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>etiquette</td>
<td>a ticket, educate, adequate</td>
</tr>
<tr>
<td>Kenji</td>
<td>can she, Kensi, candy, Kenzie, change it</td>
</tr>
</tbody>
</table>
Another case of errors attributed to Japanese learners’ L1 was observed in words that included the English sounds /l/ and /r/. As /l/ and /r/ are represented in the Japanese language by only one phoneme /r/ (within a set of syllables: ระ /ra/, ริ /ri/, รุ /ru/, รี /re/, รอ /ro/), Japanese speakers often do not distinguish between /l/ and /r/ in their pronunciation. Thus, speech recognition systems make mistakes when recognizing intended sounds. For instance, as shown in Table 6, it was often the case that “play” was transcribed as “pray,” and “reading” as “leading.” Moreover, because of Japanese speakers’ mixed weak pronunciation, the speech recognition system misrepresented some words such as “lunch” as “branch” or “ranch” and “culture” as “Carter.” Thus, it was clear that Japanese speakers had difficulty pronouncing English words whose sounds did not exist in their L1.

Table 6
Examples of Transcription Errors in Words with /l/ and /r/ Sounds

<table>
<thead>
<tr>
<th>Original</th>
<th>Wrong Transcriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>lunch</td>
<td>branch, ranch</td>
</tr>
<tr>
<td>play</td>
<td>pray</td>
</tr>
<tr>
<td>school</td>
<td>scooter</td>
</tr>
<tr>
<td>reading</td>
<td>leading</td>
</tr>
<tr>
<td>sleepy</td>
<td>sweet</td>
</tr>
<tr>
<td>culture</td>
<td>Carter</td>
</tr>
<tr>
<td>remember</td>
<td>new member</td>
</tr>
</tbody>
</table>

Similarly, as shown in Table 7, “three /θriː/” was transcribed as “free” or “tree.” This is a typical mispronunciation of Japanese speakers because the fricative /θ/ does not exist in the Japanese language. In addition, errors due to the weak pronunciation of consonants in unstressed syllables were also common. For example, “bag” was transcribed as “back” or “bank.” When a sound is weak or mispronounced and the speech recognition system cannot find a perfect phonetic match, it selects another possible combination of phonemes. In most instances, the wrongly transcribed word differs from the original only partially, retaining the syllabic structure of the word and some phonetic elements.

Table 7
Examples of Transcription Errors Connected to Weak Pronunciation

<table>
<thead>
<tr>
<th>Original</th>
<th>Wrong Transcriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>three</td>
<td>free, tree</td>
</tr>
<tr>
<td>shocked</td>
<td>shot, choked</td>
</tr>
<tr>
<td>bag</td>
<td>back, bank</td>
</tr>
<tr>
<td>a dictionary</td>
<td>Addiction</td>
</tr>
<tr>
<td>train</td>
<td>chain</td>
</tr>
<tr>
<td>one</td>
<td>wrong</td>
</tr>
</tbody>
</table>
While many of the transcription errors have a logical connection to pronunciation features, sometimes the speech recognition program fails to correctly identify the whole utterance. In the cases shown in Table 8, it completed the sections that were phonetically clear and arranged the rest of the utterances according to the statistical model or sometimes even omitted some words altogether. Overall, however, when each word is not pronounced intelligibly, especially when it comes to pronunciation features not included in the available language models, STT transcription stops and fails to correctly produce the rest.

**Table 8**

*Examples of Mistranscribed Utterances*

<table>
<thead>
<tr>
<th>Original</th>
<th>Wrong Transcriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>His mother said you are lucky. A kind man brought your bag.</td>
<td>Mother, you are a key. You are back.</td>
</tr>
<tr>
<td>… and a dictionary were in the bag</td>
<td>… and it is not even in the back</td>
</tr>
<tr>
<td>… but in fact, after graduation in working place I have already not used the English</td>
<td>… bed scene fact graduation in working place note is the English</td>
</tr>
</tbody>
</table>

**Discussion**

This study investigated the transcription accuracy of American English in five available STT applications, using the speech of NNS of English to seek a new way of utilizing them for pronunciation practice and assessment. The experimental results offer useful findings for the following RQs:

Regarding RQ1, when analyzing the transcription accuracy of English learners’ speech, the “Transcribe” iOS application (with an accuracy rate of 68.57%) and Windows 10 Dictation (66.96%) were found to be the most successful at transcribing NNS’ English speech, followed by Google Docs voice typing (63.00%). In contrast, Apple Dictation (with the lowest accuracy rate of 48.97%) showed a particular weakness in transcribing short sentences that contain loanwords and speech with frequent pauses and hesitation (see Figure 9). Dictation.io, with the second-lowest accuracy rate of 56.72%, also showed lower results in speeches with frequent pauses and hesitations.

Provided that all five STT applications were subjected to the same experimental conditions, the reason “Transcribe” and Windows 10 applications succeeded most lies in their make-up. Of the five STT applications, Windows 10 is the only application that offers to “recognize nonnative accents,” which means that its acoustic model has some additional algorithms that allow recognition of a broader range of pronunciation variations within the same language mode. However, compared to Google, the performance of Windows 10 was relatively lower for short speech or speech in which some loanwords were embedded (see Figure 9). The advantage of Google documentation on STT recognition is that it differentiates between the three methods of speech recognition: (a) synchronous recognition for audio recordings of less than one minute, (b) asynchronous recognition for audio recordings of more than one minute, and (c) streaming recognition for capturing real-time speech (Google, n.d.-a).

In contrast, “Transcribe” differs from other applications in that it adopts an asynchronous transcription model in which an audio recording must be uploaded fully before the transcription is conducted. According to M. Grushin (personal communication, October 9, 2020), “Transcribe” uses an enhanced speech recognition model for English language recognition, which is very effective for analyzing speech from a speaker—a method used in our study (M. Grushin, personal communication, October 9, 2020). Owing to this synchronous model and the enhanced speech recognition model, “Transcribe” was fairly successful in transcribing speech of various speech types.

Thus, it was found that each application has different capacities, strengths, and weaknesses. In addition, the
average recognition accuracy rates of NNS speech were far below the industry accuracy rate of 90%–99%, as claimed by various ASR applications (Hachman, 2017; Worthy, 2019). In other words, this is the gap that learners can see and narrow by practice.

RQ2 is concerned with how speeches elicited under four different tasks influence the transcription accuracy of the five STT applications. Task 4 (QA) followed by Task 2 (reading a passage) was transcribed more accurately (64.41% and 62.39%, respectively). The result of the highest accuracy rate of Task 4 (QA) was rather unexpected because the task required EFL learners’ spontaneous utterances, not prepared utterances. Velde (2019) also mentioned that speech recognition software can provide more accuracy under simple, prepared utterances in a quiet environment.

Considering that the questions in the QA task were easy and required learners to produce syntactically simple utterances with rather basic vocabulary, STT applications probably managed to predict the next words in the speech relatively accurately. Therefore, regardless of whether it is spontaneous or prepared, speech is clearly uttered with simple, predictable words and less hesitation; it can be transcribed fairly accurately. This makes sense because STT applications have been developed for transcribing spontaneous speech.

Related to the predictability of words, the accuracy rate of Task 1 (ReadS) was significantly lower than Tasks 2 and 4, even though Task 1 was also a read-aloud task like Task 2. There are two possible explanations for this result. One is that STT applications cannot predict words coming out next in such a short speech. Another reason is that these sentences embedded loanwords, which may interfere with STT applications searching for matching words. Thus, the length of utterance and whether unpredictable words are embedded may affect the accuracy of transcription.

Another pedagogically informative point is that there was a significant gap in the accuracy rate between Task 2 (ReadL with an accuracy rate of 62.39%) and Task 3 (Retell with 57.60%). In other words, when EFL learners were required to retell the passage read in Task 2, the quality of their utterances significantly deteriorated by an average of 5%. This means that EFL learners cannot pronounce words properly under semi-spontaneous conditions, even though they know how to do so. Further, as what to say was already decided in the retelling, learners may not use an avoidance strategy, that is, to avoid using words and phrases they are not familiar with, which can be used in Task 4 (QA). Therefore, an instruction to narrow this gap by practice can be an interesting way to encourage students to improve their speaking skills.

Regarding RQ3, when transcribing Japanese loanwords embedded in English speech, STT applications also showed varied transcription abilities. Table 4 shows that the three loanwords (“haiku,” “senpai,” and “kouhai”) on average were transcribed correctly only approximately 30% of the time, with “haiku” the most accurately transcribed by Windows 10 Dictation, and “senpai” and “kouhai” transcribed best by “Transcribe.”

However, the Japanese word “bukatsu” was transcribed with an accuracy rate of only 2%, which shows that the word is not used in the English-speaking world and is treated as a pure foreign word in the English mode of the application. The same rule applies to the Japanese name “Kenji,” which appears in the text of Tasks 2 and 3 but was not transcribed correctly most of the time. Thus, it was found that loan and foreign words were transcribed very differently, depending on the words and applications, and generally at much lower accuracy rates than English words.

Keeping in mind that automatic speech recognition systems recognize phonemes and words based on a specific language model (i.e., a dictionary with a list of words, limited to a specific language), regardless of how these loanwords are pronounced, they have a lower chance of being recognized because these words are not part of that language model. In this regard, the pronunciation of these words may not be of concern.

Lastly, RQ4 deals with EFL learners’ words that are not transcribed accurately. The results confirmed that the pronunciation features of learners’ L1 can affect the pronunciation features of L2 to a noticeable degree (Vaughn et al., 2019). Specifically, the absence of /θ/ and the absence of a clear distinction between
phonemes /l/ and /r/ in the Japanese language result in common pronunciation errors in English-transcribed words such as “three” as “tree” and “lunch” as “ranch.”

Additionally, Japanese learners tend to transfer *katakana* pronunciation (or reading) patterns into English, which is the phenomenon of adding a vowel after each consonant, such as “school” as /suku:ru/, which matches a similar English word such as “scooter.” As Koon (2018) notes, katakana reading, which is often used to assist Japanese learners with pronunciation of English words, hinders learners’ ability to pronounce English words correctly, thus making it impossible for native speakers to understand the original English word. This explains why ASR systems misrepresent the original English words pronounced with katakana pronunciation.

When katakana English is used at the beginning or end of unstressed consonants, transcription errors occur either by substituting phonemes or omitting the consonant. Smith (2012) indicates that in the Japanese language, consonants /d/, /t/, and /k/ are always followed by a vowel, unlike in English, where those consonants can often be in the final position of a word. As observed in the example of “shocked,” frequently pronounced and transcribed as “shot,” the last consonant is reduced or replaced with another consonant. In other words, this illustrates how speech recognition software reveals weak spots in learners’ pronunciation. However, it was found that when heavy-accented katakana English continues, word recognition systems cannot follow them word by word and produce meaningless chunks of sentences, as shown in Table 8.

This study has some limitations. The first limitation is the absence of a native speaker control group. We discuss the results of NNS’ speech rates, referring to native-speaker accuracy rates reported on the websites of the STT applications or relevant literature. If we included a native-speaker control group, the results would be more comparable. The second limitation is the relatively small number of participants and their broad background, restricting this study from recognizing larger trends. Narrowing down the sample size to one L1 with a specific proficiency level should be attempted in future studies to further validate the outcomes proposed in this study. Finally, it should be noted that the pre-recorded audio samples, used for the STT applications to transcribe, may differ in quality from a live dictation, commonly used in other studies. Since this difference can potentially affect the results, the effect of recorded and live voice feed on transcription quality needs to be further investigated.

**Conclusion**

The current study aimed to evaluate the transcription accuracy of American English in five STT applications, assessing voice recordings of NNS speech (L1: Japanese, Chinese, Korean, Urdu, Czech, Hungarian, and French) for English pronunciation practice and feedback. It provides a better understanding of the elements that affect the transcription accuracy of adult nonnative speech provided by ASR technology.

This study established a few important outcomes. First, the ability of the five STT applications to comprehend nonnative English speech depends on the type of speech and the number of prosodic errors that arise due to the speech tasks. The more difficult the tasks that learners had to perform, the more prosodic errors with frequent false starts and fillers they produced, which resulted in lower accuracy rates. In general, the accuracy rates were approximately 50%–70% for NNS speech transcription. Second, the accuracy rate of the loanwords embedded in English sentences further declined to approximately 30%, except for one uncommon loanword with an accuracy rate of only 2%. Furthermore, STT applications were also sensitive to pronunciation errors, attributed to L1. This means that the same text read by two nonnative speakers with different degrees of pronunciation errors can result in transcriptions with varying degrees of accuracy, although this can be pedagogically informative for pronunciation practice.

Knowing these characteristics of STT applications, both teachers and EFL learners can adjust the conditions of pronunciation practice to be able to work in particular areas of their pronunciation and improve the intelligibility of their speech. As each STT application has a slightly different strength of accuracy assessment, low-proficiency learners with heavily accented pronunciation may be advised to choose STT applications that can produce relatively higher transcription accuracy and adaptation ability. In contrast,
high-proficiency learners who wish to improve to a near-native speaker’s pronunciation level may choose more rigid STT applications. In both cases, the students will receive individual feedback in the form of mistranscribed words, which will be more suitable for their abilities. As for the teachers, if they decide to integrate STT applications into the pronunciation practice during class, they may design different tasks depending on the students’ language proficiency. Low-proficiency students may be offered more controlled tasks, such as reading short texts, while higher-proficiency students can test their speech with short spontaneous utterances.

Thus, instead of the normal use of STT applications for the dictation of meetings or jotting down oral thoughts, they can be an excellent tool for NNS to identify weak points in their pronunciation. Moreover, they can help teachers give individual feedback to students, which is normally very time-consuming, and can also be used for assessment purposes. This study contributes to the growing field of automatic speech recognition, addressing the needs of NNS. We hope that teachers and students who need additional support with pronunciation practice will benefit from the information and guidance provided in this study.

References


https://www.researchgate.net/publication/267263862_Effects_of_Speech-to-Text_Recognition_Application_on_Learning_Performance_in_Synchronous_Cyber_Classrooms

https://www.wired.com/2016/04/long-form-voice-transcription/


https://doi.org/10.1073/pnas.1915768117


https://doi.org/10.1017/S0266078404001087


https://doi.org/10.1075/jslp.16034.mcc


https://support.microsoft.com/en-us/windows/use-dictation-to-talk-instead-of-type-on-your-pc-fec94565-c4bd-329d-e59a-a0f33fa5689f


https://www.theverge.com/2016/10/12/13251068/walt-Mossberg-apple-siri-is-dumb


About the Authors

Akiyo Hirai is a professor in the English Language Education program at University of Tsukuba, Japan. Her research interests include language testing, the assessment of critical thinking skills, and AI-based language learning. She is the corresponding author of this article.

E-mail: hirai.akiyo.ft@u.tsukuba.ac.jp
ORCiD: https://orcid.org/0000-0002-2541-7799

Angelina Kovalyova is a lecturer in Academic Communication courses at the Center for Language Education and Research (CLER) at Sophia University, Japan. She holds a PhD in English Education from Tsukuba University, Japan, where she researched the use of mobile applications (focusing on chatbots) for text-based EFL conversation practice. Her current research interests include the use of chatbots for language learning and automatic speech recognition (ASR).

E-mail: angelina.kovalyova@gmail.com
ORCiD: https://orcid.org/0009-0007-0344-4597