The Efficacy of Speech-to-Text Synthesis in Diagnosing Phoneme-Level Pronunciation Deficiencies

Daniel White
dwhite317@gatech.edu

***Abstract*—The goal of this research was to assess how effectively speech recognition systems can differentiate between native and non-native speaker pronunciation based on the accuracy and confidence of the speech recognition system's interpretation of the word that was spoken. 55 participants including 18 native speakers and 37 non-native speakers were evaluated using a web app designed specifically for this study. The results showed that speech synthesis accuracy and minimal pair errors could determine with reasonable accuracy whether or not a user is a native speaker or non-native speaker. However, whether or not this distinction can be meaningfully applied to create a pronunciation diagnosis tool is yet to be determined.**

1 INTRODUCTION

AI driven speech synthesis has become a prevalent feature in smart devices, such as Siri in Apple products, Alexa in Amazon's Echo Smart Speaker, and Google Assistant, which can be used across many different devices. The accuracy of these tools has been steadily improving over the past several years, with the most accurate Speech-to-text tools having an 88% accuracy rate as of September 2020 (Jarmulak, 2020). Because these tools are most often used by native speakers and are AI driven to increase accuracy based on that input, it would stand to reason that they would be most accurate when used by a typical native speaker. The question this study attempts to answer is whether or not there is a substantial enough difference in accuracy between native and non-native speakers, that the inaccuracy of these tools may actually be able to diagnose pronunciation problems in non-native speakers.

2 AMERICAN ENGLISH PRONUNCIATION

2.1 Native-like pronunciation

While the primary goal of second-language pronunciation should, first and foremost, be about intelligible communication, often, non-native speakers have aims towards reaching "native-like pronunciation" (Simon, 2005; Sung, 2013; Uchida, 2020). While this is a very challenging and often frustrating goal, students nonetheless are driven towards it. Developing native-like pronunciation requires a lot of practice and retraining of the muscles in the mouth, as well as the keen ear of a pronunciation professional to perceive pronunciation deficiencies and offer ongoing assessments. It is generally impractical financially and logistically for a student to receive the one-on-one feedback needed in order to continuously assess their pronunciation. This research attempts to build a foundation which may lead to an automated, accessible, ongoing pronunciation assessment tool.

2.2 Phoneme-level pronunciation

A phoneme is a unit of sound that is used to distinguish one word from another. For example, the words *bat* and *hat* are separated by a /b/ and an /h/ sound, meaning that those sounds are individual phonemes, as they are both able to give separate meanings to a word. Phonemes are both vowel sounds and consonant sounds, and are not one-to-one with spelling (particularly in English). In fact, though there are 26 letters in the English language, some often having overlapping sounds (i.e. *k* and *c*), most scholars agree that English has at least 36 distinct phonemes (Bett, 2002). Depending on the background of the English leaner, some of these phonemes will be much more challenging than others. However, generally among consonants, the American English *r* sound (/ɹ/) is considered very challenging because of its unusual tongue position requirements. Among vowel sounds, the *short i* sound (/I/) is considered difficult to perceive and produce because of its close proximity to the *long i* (/i/) sound.

It is important to acknowledge that phoneme-level pronunciation is only a small part of pronunciation development, especially for English as a Second Language learners. Word stress, intonation, and rhythm are also vital for developing overall English pronunciation (Celce-Murcia, 1996), but for the scope of this research, the focus will be specifically targeting phoneme-level pronunciation.

2.3 Minimal pairs

Minimal pairs are two words that are separated by a single phoneme. The phoneme can be a vowel or a consonant. For example, *arrive* and *alive*. These words are only separated by the /ɹ/ (American English *r)* and /l/ sound. With vowels, heat and hit is an example of the /i/ and /I/ minimal pairs.

Minimal pairs are often used for phoneme-based pronunciation instruction, both for student awareness and assessment (Barlow and Gierut, 2002). This seems like a fairly obvious choice, as minimal pairs effectively isolate phonemes in the context of words and flawed pronunciation of these phonemes can cause the word to be mistaken for the other word in the pair. For example, an ESL learner may be instructing a taxi driver to "go to the right", wanting the driver to turn right, but the taxi driver may hear "go to the light" thinking they want to be let off at the traffic light. Though minimal pairs that can be exchanged in context, such as the previous example are exceedingly rare, phoneme isolation can be effectively accomplished by using any minimal pairs.

So, for assessment, if a non-native speaker is attempting to say a word and it is mistaken for a minimal pair of that word, it could be possible that the speaker is struggling to clearly articulate the phoneme separating the two words. For example, if the user is instructed to say *bit* and they are understood by the listener as *beat*, it could be an indication of the speaker's trouble articulating /i/ and /I/.

3 SPEECH RECOGNITION

3.1 History

Though computational speech recognition has been developing since the 1950s, the first practical application came with the use of an n-gram probabilistic language model in the 1980s, which uses probability to make intelligent guesses to more accurately interpret human utterances (Huang, 2014).

In the early 2000s, deep learning methods, such as the long short-term memory neural network model. This intelligent speech recognition has continued to grow and in 2017, Microsoft researchers claimed that their speech recognition software had reached human-level parity, meaning their speech recognition application was benchmarked and found to be as accurate or more accurate than average human transcribers (Huang, 2017).

3.2 Currently Available Systems

As of the time of this research, there are many commercial speech recognition APIs, including Google Speech, IBM Watson, Siri, Wit, RealSpeaker, DeepSpeech, and Web Speech API. These choices all vary in accuracy, implementation, and most importantly cost. Among these options, the only ones that were truly free to implement and use were DeepSpeech, which is an open-source speech recognition API, and Web Speech, which is a JavaScript-based API currently supported only by Google Chrome using Google's speech-to-text cloud services. After narrowing the choices, for this project, Web Speech API was chosen. Despite the limited browser support, the ease of implementation and accuracy are both acceptable for the scope of this project.

3.3 Web Speech API Implementation

Web Speech API is built-in web API, meaning the functions can simply be called in JavaScript code without any additional implementation steps. Web Speech API, importantly, includes both speech-to-text and text-to-speech functionality.

This allows for an implementation where the target speech is presented both visually (written) and auditorily (spoken). By giving both of these prompts, the ideal environment for the user to utter the target speech is created.

The program was designed to present target speech (word and sound), then prompt the user to repeat what they read/saw. Once the program captures their attempt, it moves on to the next word automatically and repeats.

3.4 Intentional hinderance

One of the primary reasons for improved accuracy of speech-synthesis from deep learning is sentence-level context-driven word prediction. Much how human listeners may be able to fill in the gaps to understand meaning based on context, speech recognition applications have done the same. Because the scope of this project is focused specifically on phoneme-level pronunciation improvement, context should not be used to make up for the speaker's phonemic deficiencies. This means that the rather than presenting the user with a sentence or phrase, the user is only presented with a word. This forces the speech recognition to attempt to figure out the target speech without any contextual clues, relying solely on the sounds it hears spoken in that word. This will naturally reduce the

accuracy, but maximized accuracy is not the goal of this research. Simply differentiating between native and non-native speakers is the goal.

4 PROCEDURE

4.1 Participant recruitment

Participant recruitment was performed through social media, promotion to graduate students in the Educational Technology course at Georgia Tech's Online Master's of Computer Science program, and face-to-face with ESL learners in Korea. Both native and non-native speakers were recruited in order to discover patterns and find meaningful, significant separations between native and non-native speech recognition analysis.

4.2 Word selections

For the scope of the initial research, only two minimal pair groups were selected. One was the consonant pair: /ɹ/ /l/, and the other was the vowel pair: /i/ /I/. These were chosen because the non-native speaker group that would mostly be available for participation were native Korean learners of English, and these pairs are generally considered particularly challenging for these learners. Other pairs such as /t/ /θ/ (voiceless *th* sound), syllable final /p/ /b/, and several others may also be relevant based on the background of the English learners. However, one consonant pair and one vowel pair should give plenty of meaningful data to evaluate.

4.3 Programming

An interface was created to allow users to guide themselves through the test. The first page is a consent form with important information about how their voice will be used. The user also self-reports their English pronunciation level between English beginner and native speaker. They also report their nationality.

Figure 1 - Screenshot from the evaluation: https://www.dan-teacher.com/pronunciation

After completing the consent form, the user's microphone and speakers are tested to ensure they gave the appropriate permissions and everything is in working order before the evaluation begins. Once the survey and technical requirements are satisfied, the evaluation begins.

The users are presented with a word and asked to repeat, the API sends their utterance to Google Cloud speech recognition services which returns an attempt to analyze their voice and return the word that they said in written form. A confidence rating (from 0-100) is returned as well. This confidence represents how certain the speech recognition is in its analysis.

The speech analysis can be set by language and dialect, and for this research, the language and dialect were set to American English, as most of the native-speaker participants would be American. The target word, returned word, and confidence rating were sent to the database along with the user's survey info and IP Address. The results are sent one by one using jQuery.ajax API for synchronous database storage. Each of these individual results will be referred to as *tokens*. The participant is free to quit whenever they would like. However, the interface recommends completing 50 tokens.

4.4 Evaluation

On the backend of the app, the information in the database is evaluated. The number if participants is determined by the number of unique IP addresses. This could be a problem as two participants may use the same computer, but there is no other method that could have been used with any more effectiveness. The amount of unique IP addresses indicates that there were at least 55 participants, but based on the total number of tokens received, it is more likely that there were over 70 participants. Unfortunately, due to the method of data collection, the true number of participants cannot be accurately ascertained. So, when presenting

the results, the number of participants will be prefaced with "at least", as there were at least that many participants.

Each token stored in the database can be evaluated in several ways. First, the returned word is compared to the target word, and if they match, it is deemed correct. If it doesn't match, it is deemed incorrect. However, if the returned word matches the minimal pair, then it is considered a minimal pair error.

Target Word	**Returned Word**		
still	*still*	*silk*	*steal*
	Correct	Incorrect	Minimal Pair Error

Table 1 - Demonstration of possible interpretations of speech recognition results for a target word.

5 RESULTS

At least 55 participants were evaluated. 18 were native speakers, and 37 were non-native speakers. 3681 tokens were gathered, which represents over 70 participants. It is impossible to know the exact number of participants as explained in 4.4. The tokens were evaluated in several different ways. First, each token was separated into native speaker and non-native speaker. Then each token's target word was compared to the speech recognition analyzed word. If the words matched, it was considered correct. If not, it was considered incorrect. The confidence rating of each token was also documented.

Figure-2 is a visual representation of the accuracy and confidence of native and non-native participants in this evaluation. The overall accuracy of native speakers was 73% and non-native speakers were 41%, which was a difference of 32%. The overall analysis confidence of native speakers was 87% and non-native speakers was 82%, which was a difference of 5%. This shows that accuracy is a good indicator of native versus non-native speaker pronunciation, while confidence was not a good indicator, as there was simply not enough separation.

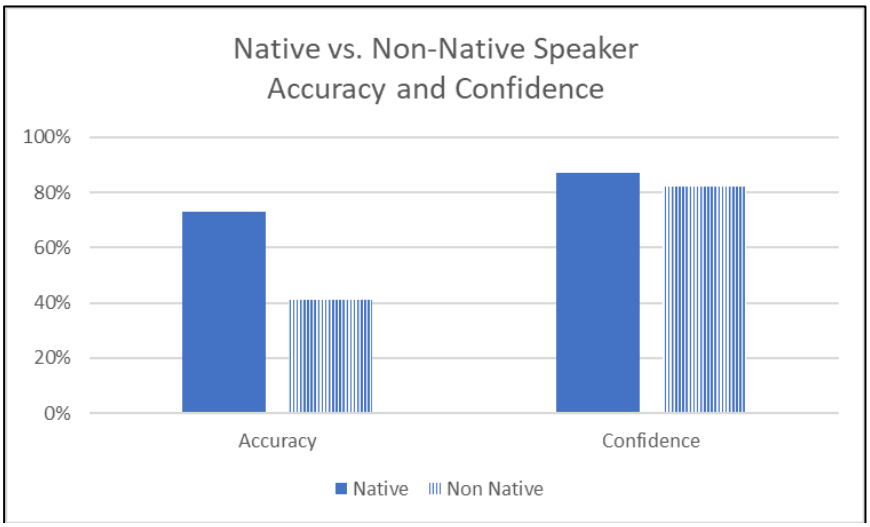

Figure 2 - Accuracy and Confidence evaluation of Native and Non-Native Speaker Tokens

Figure-3 shows how often native and non-native speakers made minimal pair errors, where the speech recognition interpreted their word is its minimal pair. Native speakers did not have any tokens where they swapped the pronunciation of /ɹ/ (American English *r* sound) and /l/. Non-native speakers swapped these 1.28% of the time. Native speakers swapped /i/ and /I/ sounds 1.59% of the time, while non-native speakers swapped these two sounds 9.59% of the time. This demonstrates that minimal pair swaps are much rarer for native speakers than they are for non-native speakers. This is a good indicator that targeted phoneme-level assessment is possible.

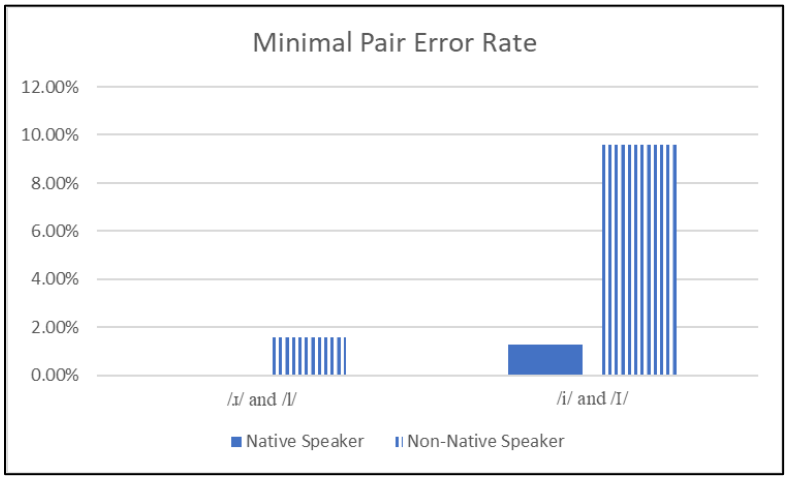

Figure 3 - The rate at which words are mistaken for their minimal pair by speech recognition.

6 CONCLUSION

The research demonstrates that it is feasible to differentiate the pronunciation of native and non-native speakers using existing speech recognition APIs. This differentiation can be seen in both overall accuracy, and phoneme-level accuracy.

6.1 Concerns of methodology

The primary concern is the integrity of the collected data, as this was done openly online without anything beyond self-reporting. This means that some speakers may evaluate their pronunciation level and even whether or not they are a native speaker in a way that is different from the intended definitions. Intentional manipulation of data is also largely uncontrollable. If a user had wanted to skew the results, it would be quite easy to do so. Though, inspecting the inbound data, there are no clear indications that anyone intentionally misrepresented themselves or intentionally tried to skew data.

Technical differences may also impact the results, as microphone quality and ambient noise could impact the speech recognition's ability to give an appropriate evaluation.

Additionally, while having at least 55 participants is fairly impressive for the scope of this project, for truly statistically significant results, it would be much better to have more participants and more overall tokens.

6.2 Concerns of benefit

The primary concern with the benefit of this research is whether or not it the speech recognition's struggles to interpret non-native speakers align with actual human hearing. It can be speculated that as the accuracy of speech recognition is close, and in some cases surpasses human listening, that a human listener's struggles would align with speech recognition. However, without a separate trial comparing a native listener's interpretation of a non-native speaker's pronunciation versus the speech recognition's interpretation, the true benefit of this research cannot fully be determined.

Another concern is simply over the importance of phoneme-level pronunciation. As mentioned previously, though phoneme-level pronunciation is a big part of language instruction, it is only a small part of overall pronunciation

development. With English, stress, intonation, rhythm, syllable timing, and several other factors play a major role in overall comprehensibility when speaking.

6.3 Concerns of privacy

There are a few privacy concerns that must be addressed any time that voice recording and storage is occurring. First, it is important that it is very clear when the voice is being recorded, to prevent users from accidentally giving sensitive information. Second, it is important that the user's voice is not used in any way beyond what is absolutely necessary. The app developed for this project does not in any way record the user's voice. The voice is sent to Google's speech-to-text services where it is analyzed and the results are sent back. In the Google Chrome documentation about the privacy of Web Speech API, it is stated:

> "Chrome supports the Web Speech API, a mechanism for converting speech to text on a web page. It uses Google's servers to perform the conversion. Using the feature sends an audio recording to Google (audio data is not sent directly to the page itself), along with the domain of the website using the API, your default browser language and the language settings of the website. Cookies are not sent along with these requests." (Google, n.d.)

"Audio data is not sent directly to the page itself" refers to the fact that the app developer using the API does not have access to the voice recording. The app developed for this project simply keeps the information returned by Google in text form. This could still lead to sensitive data being stored on our servers, which means users need to be aware and careful about leaking sensitive information during the evaluation.

There were some examples where random words were picked up seemingly unrelated to the evaluation. Tokens with obviously extraneous words need to be deleted to ensure that nothing private is being stored in the app's databases.

Though the documentation specific to Web Speech API does not indicate whether or not Google actually keeps the voice data, the paid Google speech-to-text service claims that:

> "By default, Speech-to-Text does not log customer audio data or transcripts. To help Speech-to-Text to better suit your needs, you

> can opt into the data logging program. The data logging program allows Google to improve the quality of Speech-to-Text through using customer data to refine its speech recognition service. As a benefit for opting in, you gain access to discounted pricing." (Google, n.d.)

The use of the Web Speech API is free, so it is uncertain if data logging is automatically opted in or not.

6.4 Moving forward

From here, there are several steps that can be taken. First, from an academic research perspective, the data needs to be collected in a more controlled environment to ensure consistency between participants. Human evaluators also need to be used to see how well human listening aligns with speech recognition listening when it pertains to non-native pronunciation.

Larger groups, better proctoring, and ESL professional comparisons would give much more definitive results to the primary question of this research, which is whether or not existing speech recognition systems can be used to evaluate second-language learner pronunciation at the phoneme level.

From a commercial product perspective, the interface should be shifted to be more friendly and welcoming, as well as easier to use. The words used need to be curated based on the trial to ensure that every word is doing its part to differentiate. For example, the word *ease* was never interpreted correctly by speech recognition regardless of the speaker's English proficiency. The low accuracy of this word means it will not assist in differentiating between correct and incorrect phoneme production.

There was also an issue with words always being mistaken for their homonyms. For example, of the 52 tokens targeting the word *fry*, it was never interpreted correctly by the speech recognition API. 27 times (51.9%) it was interpreted as the proper noun *Frye*. Both of these words have the same pronunciation, so although the app deemed these incorrect because they did not match, with some adjustment to the database to include all homonyms as correct matches, this word could still have value in determining phoneme-level pronunciation deficiencies.

Once the words are curated, a system needs to be placed to efficiently assess the speaker's pronunciation level both overall and for each minimal pair group. For example, if a speaker gets 10 /i/ versus /I/ tokens correct consecutively, it is unlikely that they will need to continue having their /i/ versus /I/ accuracy evaluated.

Additionally, more minimal pair groups need to be added to evaluate beyond the two groups that are currently here. This will require more research into the most commonly mistaken minimal pair groups across a wide variety of non-native speaker backgrounds.

Finally, instructional resources need to be provided to users to give them a path towards improvement. There are plenty of great websites and YouTube videos with thorough guides to improving the articulation of challenging phonemes. ESL professionals can compile and curate these resources so that when students complete their assessment, they can follow the recommendations and actually improve their capabilities. This could also be combined with gamification elements such as scoring and trophies for making improvements.

7 REFERENCES

1. Barlow, J. A., & Gierut, J. A. (2002). Minimal pair approaches to phonological remediation. In *Seminars in speech and language* (Vol. 23, No. 01, pp. 057-068). Copyright© 2002 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel.:+ 1 (212) 584-4662.
2. Bett, S. (2002). The number of phonemes in English. *In Memory of Ken Ives (1917–2002)*, *30*, 1.
3. Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge University Press.
4. Google Cloud Speech-to-Text Documentation Data Logging (n.d.). Retrieved November 29, 2020, from https://cloud.google.com/speech-to-text/docs/data-logging
5. Google Chrome Privacy Whitepaper. (n.d.). Retrieved November 29, 2020, from https://www.google.com/chrome/privacy/whitepaper.html
6. Huang, X., Baker, J., & Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, *57*(1), 94-103.
7. Huang, X. (2017). Microsoft researchers achieve new conversational speech recognition milestone. *Microsoft, August*.
8. Levis, J., & Cortes, V. (2008). Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching. *Towards adaptive CALL: Natural language processing for diagnostic language assessment*, *197208*.
9. Newman, H. & Joyner, D. A. (2018). Sentiment Analysis of Student Evaluations of Teaching. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom. Springer.
10. Simon, E. (2005). How native-like do you want to sound? A study of the pronunciation target of advanced learners of English in Flanders. *Moderna Sprak*, *99*(1), 12-21.
11. Sung, C. C. M. (2013). 'I would like to sound like Heidi Klum': What do non-native speakers say about who they want to sound like?. *English Today*, *29*(2), 17.
12. Uchida, Y., & Sugimoto, J. (2020). Pronunciation Goals of Japanese English Teachers in the EFL Classroom: Ambivalence Toward Native-like and Intelligible Pronunciation. *LANGUAGE TEACHER*, *44*, 3.