Improving Voice Recognition Software With Pitch

- Research Paper

Project ID: SCM301

Table of Contents:

Abstract

Speech-to-text systems are commonly used in the modern world. While these systems are tremendously helpful, they come with drawbacks. One major drawback is being unable to speak naturally to input punctuation. For example, to write a period, you must say “period” instead of the system automatically detecting it.

Additionally, the emotional context of speech is unable to be conveyed with standard speech to text systems, disadvantageous to people with hearing loss or conditions that prevent them from recognizing emotion directly from one’s voice.

The objective of our project was to address these issues by developing a software that automatically detects punctuation and emotion by analyzing pitch. Using the CREPE model for pitch detection and Brain.js for emotion detection, we developed our program in JavaScript.

We tested it with 17 participants, each reading 23 items in 5 categories: Questions, Commas, Emotions, Sentences, and Combined. Our program distinguished between sentences and questions correctly 67% of the time on average ($p = 0$). It detected punctuation completely correctly 24% of the time (criteria for completely correct: in one item on the script, software detected the correct type, number, and location of punctuation). Emotion detection worked 25% of the time in the emotion category, but accurately detected “No Emotion” when participants were not encouraged to express a certain emotion.

Our study is small, but has promising indications that pitch, if understood, is a powerful tool which can be utilized to improve this technology and ultimately benefit society.

Introduction

Speech-to-text software is becoming more prevalent, with benefits of convenience and efficiency. While speech-to-text software has evolved, there are some fallacies in terms of lacking automatic punctuation and emotion detection.

Speech-to-text programs that can automatically add punctuations and emotions have various potential applications in the fields of medicine, education, and media, just to name a few.

For example, doctors who use speech-to-text programs on a daily basis would be able to perform their job much more efficiently. People with disabilities, such as hearing impairments, and the elderly would be able understand the emotional context behind voice through closed captions without hearing.

This idea came from observing one of our parents (a doctor) dictating punctuation verbally in a speech recognition program.

In the first year of this project, we hypothesized that pitch has a correlation with specific punctuation and emotions. We tested the hypothesis by writing a script, having participants read off of it into a pitch detection software, and analyzing the data to find patterns. The following year, we created a program that, by comparing the pitch of their voice with our previously identified patterns, could detect a user's emotions and punctuation (ex. a certain type of pitch change corresponds to a specific emotion). For the current project, we hypothesized that the implementation of pitch detection into speech-to-text systems would allow for punctuation and emotion detection.

Materials and Methods

Materials:

- The writtenscript for participants to read from (Table 1)
- A microphone for improved sound input
- The JavaScript program developed by us
- A laptop to run the program

In our first year of our experiment, we created a script with 23 total sentences in 5 categories: Questions, Commas, Sentences, Emotions, and Combined. There are 5 sentences in each of the first 4 categories and 3 sentences in the last.

In the next year, we developed a HTML, CSS, and JavaScript program that would utilize data from the first year to detect elements of sentences through pitch. We began testing the accuracy by having participants read off the same script.

Due to the cancellation of last year's fair, we decided to improve the program for this year. We replaced our former pitch detection software (an instrument tuner) with CREPE (NYU), one powered by AI. We then reworked sentence type detection (differentiating between sentences and questions) by accounting for the fact that questions often increase in pitch towards the end. We also reworked the punctuation detection by adjusting the pause required for certain punctuations to be detected. Finally, we utilized a JavaScript machine learning library called Brain.js to detect emotions.

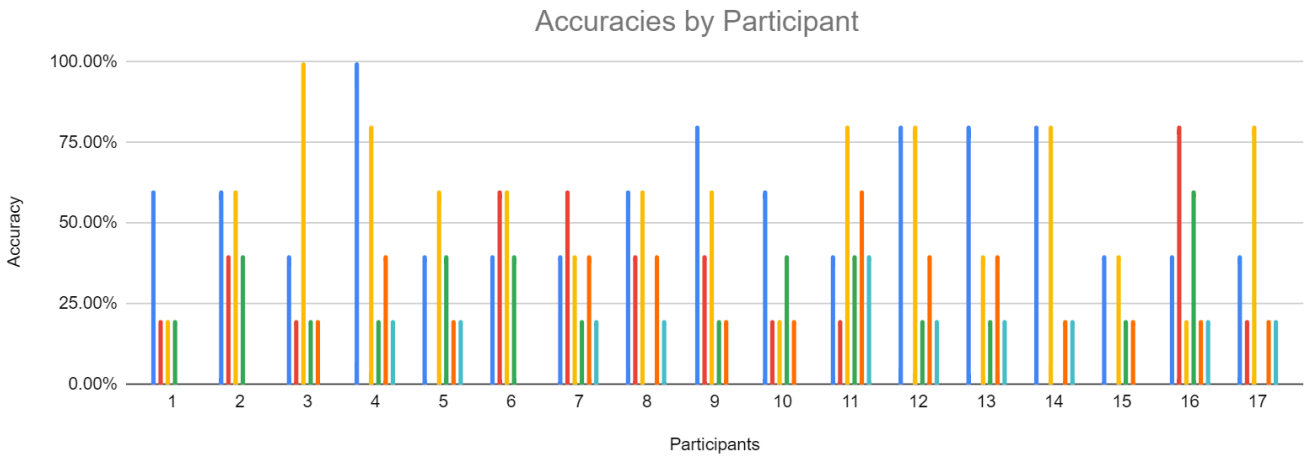
Using the same script as previous years, we had 17 adult participants either read the script into their microphone and send the recordings to us, or we recorded them in person (while wearing masks). Afterwards, we ran the recordings through our

program to analyze. We recorded the accuracy of the program being able to detect the correct sentence, emotion, and punctuation for each item on the script.

Table 1: The script participants read off of

Sentence Number	1	2	3	4	5
Questions	Do you like to read?	Do you enjoy riding bikes?	Are you going to come tomorrow?	How hard is it to learn how to make origami?	Do you like to eat pie?
Commas	I like pie, cheese, and cake.	After a long day, I felt very tired.	I like Science, but I don't really like Social Studies.	If you ever need help, I'll be there.	The Declaration of Independence was signed on July 4, 1776.
Emotions	(Excited) I can't believe my birthday is tomorrow!	(Tired) I only got three hours of sleep last night.	(Mad) You should have finished more than 2 pages in 2 months!	(Surprise) I can't believe you finished that in so little time!	(Sad) I only have 2 days of break before I have to go back to work.
Sentences	I do like to cook. I make very good pancakes.	Mark is very good at sewing. He also is a great runner.	Traveling to Africa is a great experience and fun for the whole family.	Those stuntmen face so much danger. I could never be like them.	I have a google account. It is very useful.
Combined	(Surprise) How do I still have 4 months of work left?! (Emotions + Questions)	(Sad) I wish I could go, but I have too much work. (Emotions + Commas)	Do you like engineering, science, or math? (Questions + Commas)	NA	NA

Results



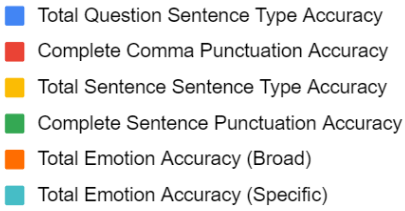
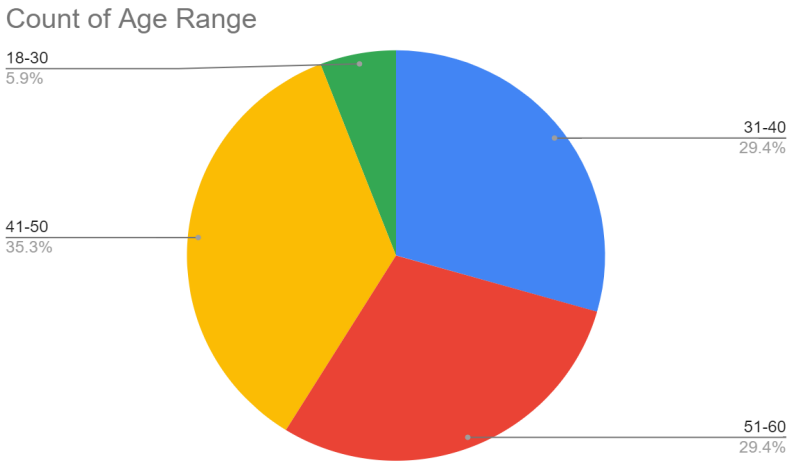
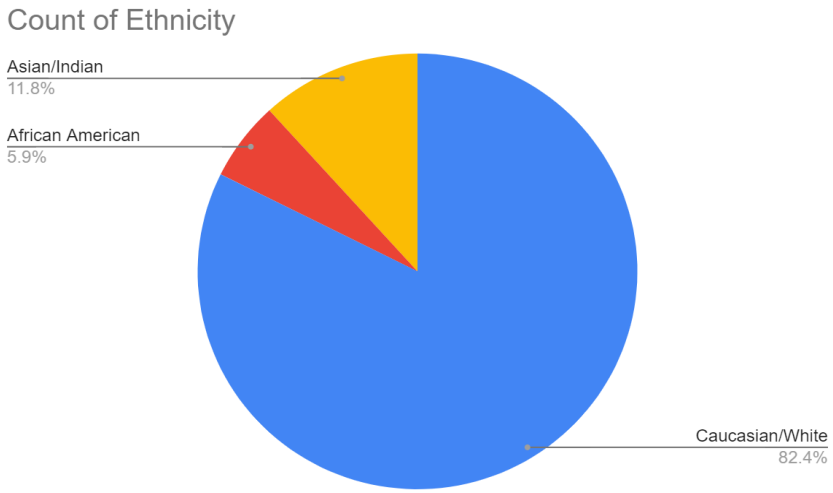
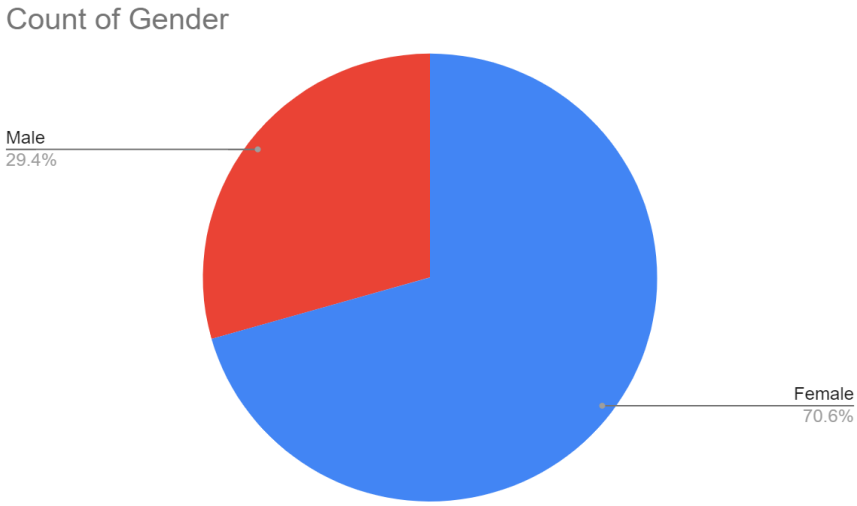


Figure 1: The graph depicts the combined accuracies in all categories for each individual participant.







Figures 2 A, B, C: Demographics of the 17 participants

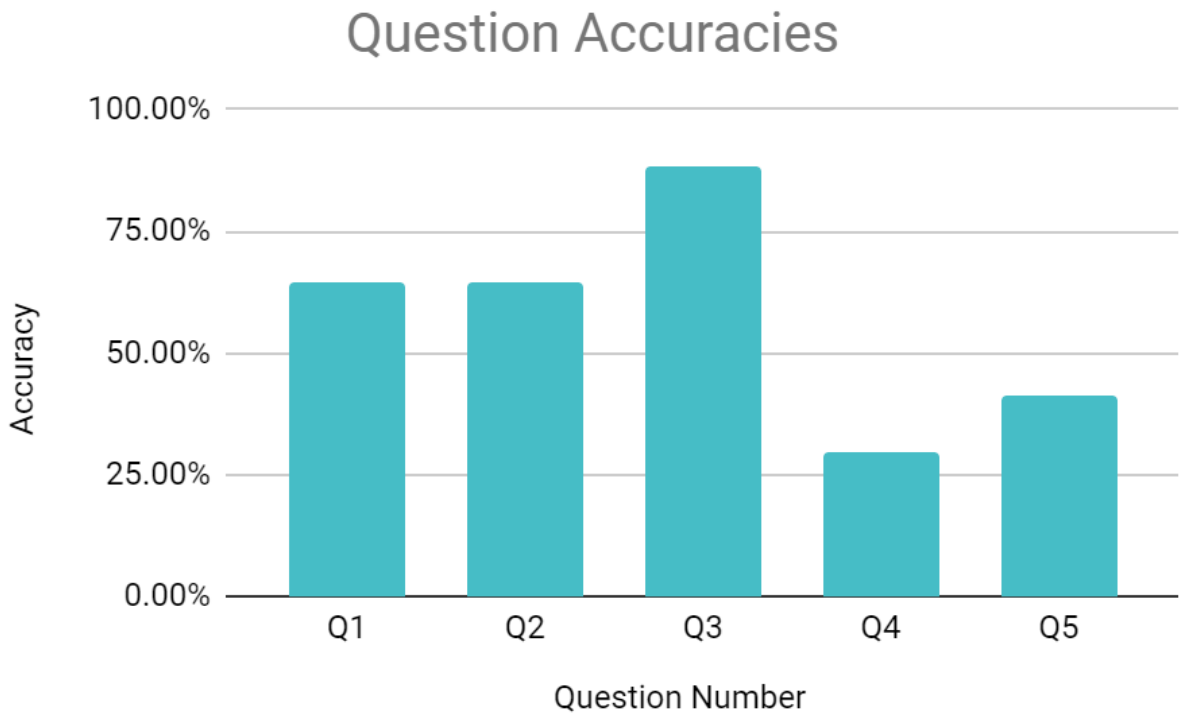


Figure 3: The graph depicts the combined accuracy of all participants for sentence type in the question category of the script.

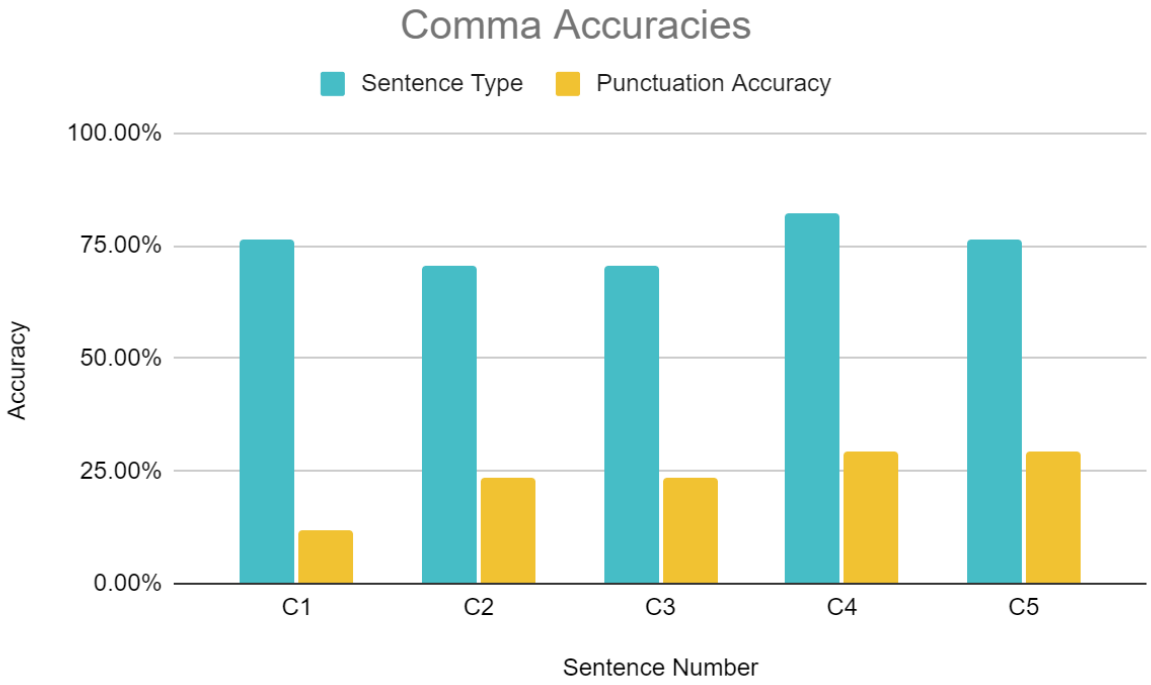


Figure 4: The graph above depicts the combined accuracy of all participants for sentence type and punctuation in the comma category of the script.

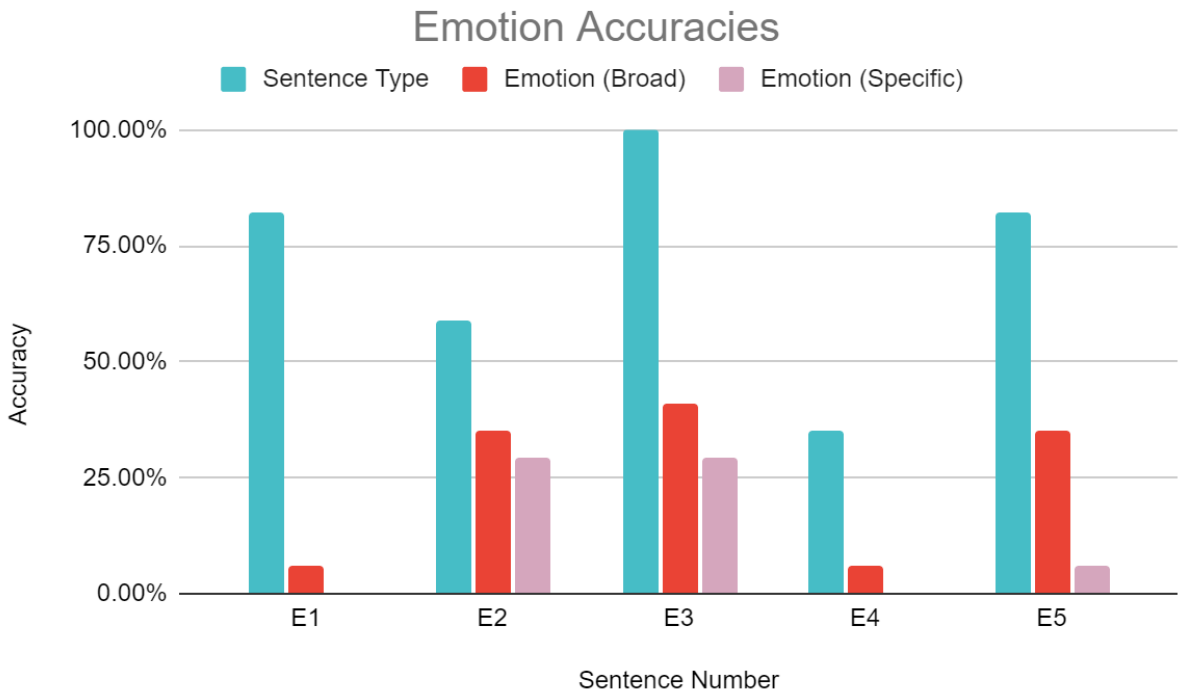


Figure 5: The graph depicts the combined accuracy of all participants for sentence type and emotion in the emotion category of the script. Note: Emotion (Broad) refers to the emotion categories: No Emotion, Subdued, and Excited, while Emotion (Specific) refers to specific emotions within the broad categories: No Emotion in its own category, Sad and Tired in the Subdued category, and Surprised, Excited, and Mad in the Excited category.

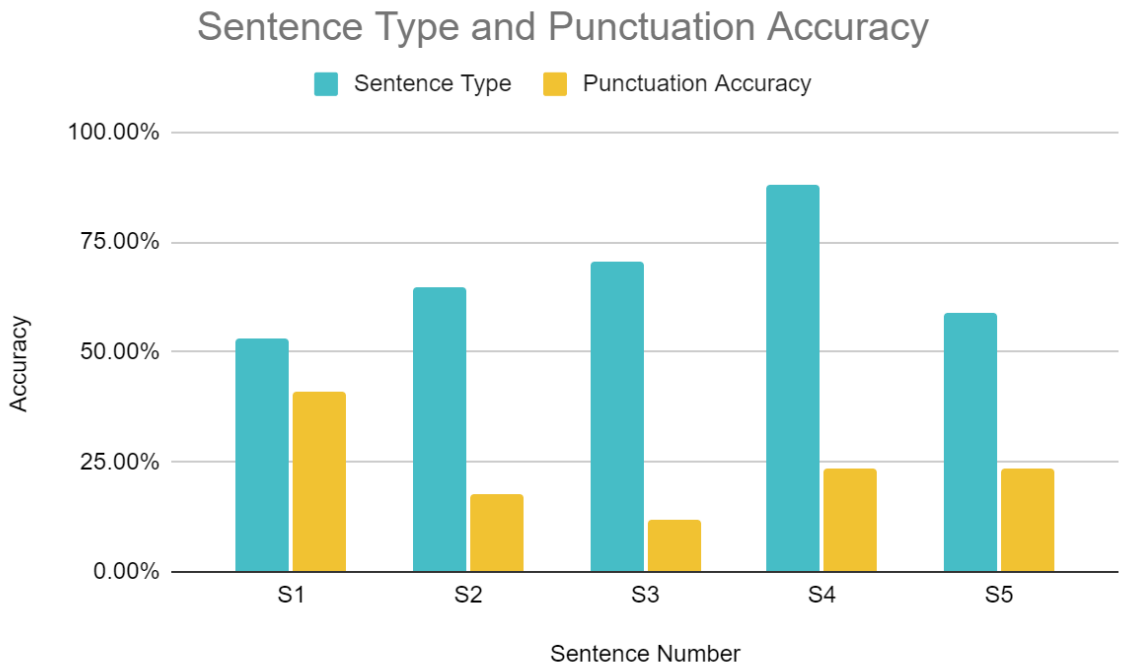


Figure 6: The graph depicts the combined accuracy of all participants for sentence type and punctuation in the sentence category of the script.

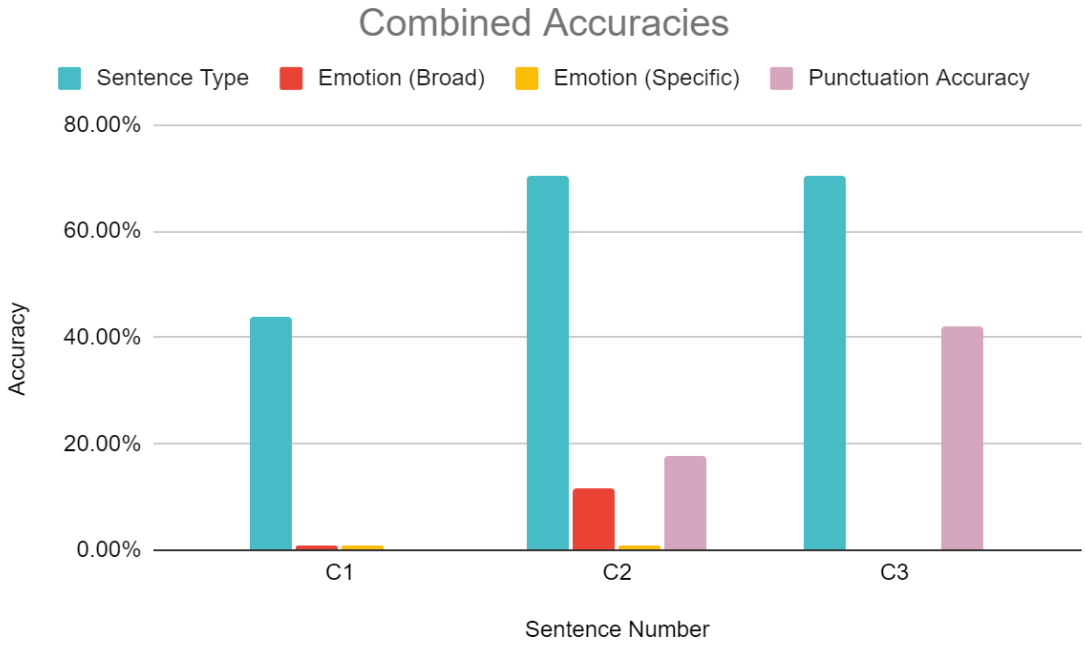


Figure 7: The graph depicts the combined accuracy of all participants for sentence type and punctuation in the combined category of the script.

Table 2: Shows the accuracy of each individual sentence when tested by our program.

Accuracy of Our Program							
Categories	**Sentence Numbers**	#1	#2	#3	#4	#5	Total Acc.
Questions	Sentence Type	11/17	11/17	15/17	5/17	7/17	58%
Commas	Sentence Type	13/17	12/17	12/17	14/17	13/17	75%
	Punctuation Accuracy	2/17	4/17	4/17	5/17	5/17	24%
----------> Breakdown of Punctuation Accuracy ---------->	# of Punctuation	4/17	6/17	4/17	7/17	6/17	32%
	Type of Punctuation	2/4	4/6	4/4	5/7	5/6	74%
Emotions	Sentence Type	14/17	10/17	17/17	6/17	14/17	72%
	Emotion (Broad)	1/17	6/17	7/17	1/17	6/17	25%
	Emotion (Specific)	0/17	5/17	5/17	0/17	1/17	13%
Sentences	Sentence Type	9/17	11/17	12/17	15/17	10/17	66%
	Punctuation Accuracy	7/17	3/17	2/17	4/17	4/17	24%
----------> Breakdown of Punctuation Accuracy ---------->	# of Punctuation	11/17	8/17	2/17	9/17	12/17	48%
	Type of Punctuation	7/11	3/8	2/2	4/9	4/12	49%
Combined	Sentence Type	7/16	12/17	12/17	NA		62%

	Emotion (Broad)	0/17	2/17	NA		6%
	Emotion (Specific)	0/17	0/17	NA		0%
	Punctuation Accuracy	NA	3/17	7/17	NA	29%
	# of Punctuation	NA	5/17	8/17	NA	38%
	Type of Punctuation	NA	3/5	7/8	NA	77%

Discussion

This study is unique, for there is a lack of similar studies that we know of. Our study indicates that pitch detection is a viable solution to improve voice recognition software.

Our results indicated that for almost every participant, the highest accuracies were in sentence type, while the rest of the accuracies varied. Each participant had very different results (Fig. 1). Most of our participants are females aged 31-60 years as expected, since most are our teachers (Fig. 2).

We found that our program worked in identifying questions and sentences (sentence type) at 58% and 66% in the Question and Sentence categories, respectively (Figs. 3 & 6, Table 2). However, it detected the correct sentence type at 75%, 72%, and 62% for the Comma, Emotion, and Combined categories, respectively (Figs. 4, 6 & 7, Table 2). These accuracy rates are noticeably higher than those of the Question and Sentence categories, possibly because those three categories primarily consist of singular sentences opposed to multiple sentences in

the sentence category which could be harder to detect for the program. Sentence type detection only has two possible outcomes: Sentence or Question. Therefore, because the accuracy is above 50% (true random), our program has a degree of accuracy.

We found the punctuation accuracy was 24% for both the Comma and Sentence categories (Table 2, Figs. 4 & 6). However, the lower accuracy is understandable because of the extremely specific criteria for having correct punctuation in a sentence: the punctuation has to be detected at the correct time, have the correct amount, and be of the correct type (comma or period). These specific criteria prevent calculating a baseline to judge whether the program is any more accurate than chance.

Our program detected emotion in the broader categories correctly with 25% accuracy, and detected specific emotions with 13% accuracy (Table 2, Fig. 5). Since there are three broad categories, our program performed worse than random probability. And because there are 6 specific emotions, our program underperformed here as well.

These results show that our program was able to detect elements of speech at a moderate accuracy. In addition, other studies have shown that pitch-based voice recognition software is a viable solution. For example, EmoVoice, a study where students from Denmark made an algorithm that could detect emotion using pitch information, proved successful. Another example of pitch detection software improving voice detection is shown in the study, “Using Pitch Frequency Information in speech recognition”. This study shows how pitch detection significantly improved the accuracy of voice recognition programs. In addition, UC Berkeley says that voice reveals more emotion than faces, showing that

voice based emotion detection is a very effective way of detecting emotion. The paper, “Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies” also acknowledges the limitations of current speech recognition systems, such as lack of punctuation detection. However, it does not provide pitch detection as a solution.

A limitation of our study is the small number of participants. Our software underperformed in the category of emotions. We feel that it is because of the shortcomings of the Neural Network: specifically, it didn’t accommodate our data format properly. Also, there is a lot of variability amongst the participants in the Emotion category. Punctuation detection also underperformed due to the mismatch between the set values of our program and the actual duration of the participants’ pauses as well as speed, loudness, and enunciation sometimes being misinterpreted. There could be some differences between the in-person recordings with microphones (with masks) and in-home recordings (without masks) sent to us, but it is unclear to us. Finally, a possible reason why the program sometimes failed to differentiate between sentence types could be because it based detection of the types on generalized rules.

A potential solution for these problems and any others is implementing user learning and personalization, similar to the way Siri has to be trained to understand the user’s voice the first time it is used. This will allow for better accuracy as the recognition is suited more to a specific person.

Conclusion

In conclusion, speech-to-text systems are emerging technologies with room to improve, especially in emotion and punctuation detection. Our study is small, but has promising indications that pitch, if understood, is a powerful tool which can be utilized to improve this technology and ultimately benefit society.

Appendix

Acknowledgements:

We appreciate our parents for proofreading our documents.

We also appreciate our school science fair coordinator for his support.

Bibliography:

Skorobogatov, Yana. “What's Up with Upspeak.” *UC Berkeley Social Science Matrix*, 21 Sept. 2015. Retrieved November 25 from matrix.berkeley.edu/research/whats-upspeak

André, Elisabeth, et al. “EmoVoice - Real-Time Emotion Recognition from Speech.” *University of Augsburg - Institute for Computer Science*, 1 Jan. 2005. Retrieved November 25, 2018 from www.informatik.uni-augsburg.de/lehrstuehle/hcm/projects/tools/emovoice/

Seppala, Emma. “Does Your Voice Reveal More Emotion Than Your Face?” *Berkeley Greater Good Magazine*, 19 June 2017. Retrieved November 26, 2018 from greatergood.berkeley.edu/article/item/does_your_voice_reveal_more_emotion_than_your_face

“Describing Voice.” *BBC Bitesize*. Retrieved on November 26, 2018 from www.bbc.com/bitesize/guides/zqtgq6f/revision/2

Magimai-Doss, Matthew, et al. “Using Pitch Frequency Information in Speech Recognition.” *Semantic Scholar*, 2003. Retrieved November 26, 2018, from pdfs.semanticscholar.org/8e0e/4bf4ad54919d543f32517821e13bfdde520e.pdf

Varga, Imre, and Imre Kiss. “Speech Recognition in Mobile Phones.” SpringerLink, Springer, London, 1 Jan. 1970, link.springer.com/chapter/10.1007/978-1-84800-143-5_14#citeas

Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies - IEEE Journals & Magazine, ieeexplore.ieee.org/abstract/document/1677974 Hanžl, Jan Bartošek Václav.

[PDF] Comparing Pitch Detection Algorithms for Voice Applications - Semantic Scholar, 1 Jan. 1970, https://www.semanticscholar.org/paper/Comparing-Pitch-Detection-Algorithms-for-Voice-Hanzl/a284f18ff0edf3d3ceaace6d3cd76a063e582ed8

A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition - IEEE Conference Publication, https://ieeexplore.ieee.org/abstract/document/6854049.

Effects of Feature Type, Learning Algorithm and Speaking Style for Depression Detection from Speech - IEEE Conference Publication, https://ieeexplore.ieee.org/document/7178877.