Amy Danoff, Nari Johnson, and Sarah Lucioni
CS105 Final Project
December 15, 2018

Analyzing Algorithmic Bias in Voice Recognition Technologies

"Everyone who speaks a language, speaks it with an accent"[1]

Abstract:

This project seeks to analyze algorithmic bias via regional accents in voice recognition technologies built by four prominent technology companies: Google, Amazon, Microsoft, and IBM. Given the increasing use of voice recognition technology in today's society, it is important to assess claims to universal accessibility. Thus, this paper explores the accuracy of four speech-to-text technologies with respect to English spoken by individuals with a variety of common international accents. Our analysis finds that IBM is the top-performing technology, while Microsoft's Bing speech-to-text consistently performs the worst. Additionally, we find clear differences in the accuracy of these technologies by accent, with three of the four technologies performing considerably better accuracy-wise on English spoken with a US American accent than on any other accent. Our analysis also finds particularly troublesome implications for use for those speaking English with Vietnamese or Spanish accents, as all four technologies perform poorly in these categories. These discrepancies have significant implications for the accessibility of hands-free and voice recognition technologies for individuals speaking English with a non-US American accent.

Introduction:

Over the past few years, promises of accessibility via hands-free voice control have played an increasingly large role in tech products. A study done by HubSpot in 2016 showed that AI personal assistants like Amazon's Alexa, Microsoft's Cortana, Apple's Siri, and Google Home are used by over 325 million users per month[2]. In addition, Google voice search queries have increased by over a factor of 35 throughout the past 10 years[3]. There are many promising implications of this increased use of hands-free technology, most notably an increase in accessibility for people from many walks of life. However, given the lofty promises of efficiency and accessibility from technologies that ended up negatively impacting specific communities, such as the COMPAS algorithm, it is important to have a certain level of skepticism regarding

[1] Weinberger, Steven. (2015). *Speech Accent Archive*. George Mason University. Retrieved from http://accent.gmu.edu

[2] HubSpot. "The Ultimate List of Marketing Statistics for 2018." *Hubspot*, www.hubspot.com/marketing-statistics.

[3] Sentance, Rebecca. "What Does Meeker's Internet Trends Report Tell Us about Voice Search?" *Search Engine Watch Search Marketing Guide to Naver Koreas Most Popular Search Engine Comments*, Search Engine Watch, 15 Nov. 2018, searchenginewatch.com/2016/06/03/what-does-meekers-internet-trends-report-tell-us-about-voice-search/.

universal promises of accessibility and efficiency. In particular, when evaluating machine-learning based technology like voice-to-text algorithms, it is important to examine algorithmic bias among different test samples. In the past, these vocal recognition algorithms have struggled to effectively detect female voices, and research has hinted that similar problems may exist for those with non-American English accents. Thus, our research seeks to examine how vocal differences caused by regional accents affect the accuracy of four different voice-to-text algorithms (Google Cloud speech-to-text, Amazon Web Services speech-to-text, Microsoft's Bing speech-to-text, and IBM Watson's speech-to-text). In short, our research question is: are speech-to-text algorithms biased?

Motivation:

Hands-free voice control not only leads way to fun add-ons such as Alexa bedtime stories, but also has enabled increased accessibility for people with visual or physical disabilities, enabling them to easily control their home, contact their loved ones, or even order groceries to their doorstep just through voice control. 77 percent of Americans currently own a smartphone with some voice-controlled digital assistant, and 46 percent of these people report that they "actively use" voice-controlled digital assistance every day[4]. The increased prominence of voice control also has made critical communication easier - the popularity of free Alexa Baby Monitor extensions have enabled tools that used to be available only through purchasing additional devices now free to use on popular personal devices. Many of these companies also encourage developers to build and release their own skills that utilize their voice-to-text APIs. The possibilities to coordinate new bluetooth or WiFi-connected products such as cars to home control systems with hands-free voice control gives anyone who knows how to use their APIs the opportunity to change the way we interact with the physical world.

But in a world where voice recognition offers the potential for more inclusive access to opportunities, who is it really that benefits from these new technologies? Are some voice recognition technologies harder to use for certain kinds of people? In instances when there may be high stakes when voice recognition may misunderstand or not recognize specific cues, are there some people where these algorithms are more likely to fail? Our project explores these questions of algorithmic bias in the accuracy of common voice recognition technologies for people who speak English with an accent.

Existing studies have demonstrated evidence of bias in voice recognition technologies: in some cases, popular technologies such as Alexa were dramatically more accurate and accessible to certain populations. Several studies have proved that some technologies were significantly

[4] Olmstead, Kenneth. "Voice Assistants Used by 46% of Americans, Mostly on Smartphones." *Pew Research Center*, Pew Research Center, 12 Dec. 2017, www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/.

better at understanding male voices than female voices[5]. The MIT Technology Review published “AI Programs are Learning to Exclude African American Voices”, where they examined how specific regional dialects were disproportionately misinterpreted. The Economist’s “In the wake of Voice recognition technologies, not all accents are equal” features interviews with linguists who attempted to understand why certain US regional accents were never understood. The “trapped in a voice recognition-controlled elevator” sketch, which chronicles a man with a strong Scottish accent whose voice simply can’t be recognized, has over 1.5 million hits on YouTube.

We chose particularly to examine accents in English, particularly foreign instead of US regional accents, because of the prevalence of foreign-born residents in the US. As of 2010, 25 percent of all US residents under the age of 18 were first generation immigrants[6]. Foreign-born US residents are particularly more likely to also have less socioeconomic mobility in comparison to native US residents, as ⅓ of all children living in poverty are first or second-generation immigrants[7]. Furthermore, when broken down by country of origin, traditional benchmarks of assimilation and mobility are historically higher for “whiter” immigrants - those from European countries - in comparison to immigrants from other countries. Are these historic inequities in access to opportunities perpetuated in voice recognition technologies? To truly achieve a vision of a more accessible and equitable world through voice recognition technology, we need to ensure that these particularly vulnerable populations of US immigrants are not left behind.

Experiment:

Our experiment to test the efficacy of the four different speech-to-text algorithms relied on a voice dataset and the four respective speech-to-text APIs. We wrote all code in Python 3 -- all code can be found at the public GitHub repository cited in Appendix A. Some code was adapted from pre-existing code found online, (i.e. to query speech-to-text APIs), and is documented as such in the comments of the corresponding code.

The first stage of the experiment was to identify relevant data. We chose to use George Mason University’s *The Speech Accent Archive*[8]. This dataset is an ongoing, publicly sourced project that aims to exhibit the differing speech accents from a variety of language backgrounds. We elected to examine this dataset using Amazon, Google, Microsoft, and IBM’s speech-to-text technologies due to the prevalence of these technologies in households (Amazon Alexa, Google

[5]Mailonline, Joe Pinkstone For. “AI Assistants Are Sexist and Understand Men Better.” *Daily Mail Online*, Associated Newspapers, 14 Mar. 2018, www.dailymail.co.uk/sciencetech/article-5499339/AI-assistants-sexist-understand-men-better.html.

[6]“Immigration to the United States.” *Wikipedia*, Wikimedia Foundation, 9 Dec. 2018, en.wikipedia.org/wiki/Immigration_to_the_United_States.

[7]Zong, Jie, et al. “Frequently Requested Statistics on Immigrants and Immigration in the United States.” *Migrationpolicy.org*, 27 Feb. 2018, www.migrationpolicy.org/article/frequently-requested-statistics-immigrants-and-immigration-united-states.

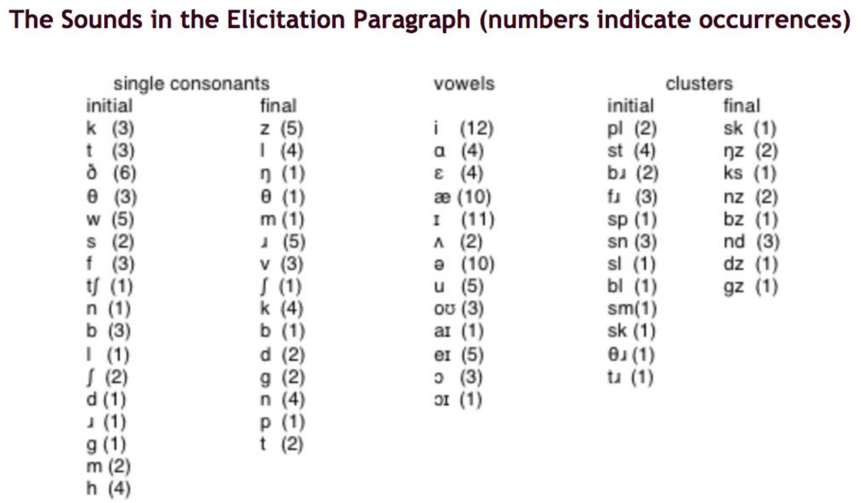
[8] Weinberger, Steven. (2015). *Speech Accent Archive*. George Mason University. Retrieved from http://accent.gmu.edu

Home)[9] and the varied uses from Microsoft and IBM. Using these four technologies, we ran over 19,000 seconds of English speech with various regional accents from top-spoken languages from the GMU dataset and measured their accuracy and precision at translating vocally different speech to text.

We selected GMU's *The Speech Accent Archive* dataset for the consistency, quality, and quantity of the data provided, as well as the breadth of the data. All the samples in this dataset record the same English paragraph. This 69-word paragraph is as follows:

> *"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."*[10]

While the content of the above paragraph may not make logical sense in English, it is a great sample for linguistic testing, as it contains most of the consonants, vowels, and clusters in American English in order to capture almost all the sounds of English. The following illustrates the distribution of these sounds (English consonants, vowels, and clusters) in the elicitation paragraph:



The Sounds in the Elicitation Paragraph (numbers indicate occurrences)

single consonants		vowels	clusters	
initial	final		initial	final
k (3)	z (5)	i (12)	pl (2)	sk (1)
t (3)	l (4)	ɑ (4)	st (4)	ŋz (2)
ð (6)	ŋ (1)	ɛ (4)	bɹ (2)	ks (1)
θ (3)	θ (1)	æ (10)	fɹ (3)	nz (2)
w (5)	m (1)	ɪ (11)	sp (1)	bz (1)
s (2)	ɹ (5)	ʌ (2)	sn (3)	nd (3)
f (3)	v (3)	ə (10)	sl (1)	dz (1)
tʃ (1)	ʃ (1)	u (5)	bl (1)	gz (1)
n (1)	k (4)	oʊ (3)	sm(1)	
b (3)	b (1)	aɪ (1)	sk (1)	
l (1)	d (2)	eɪ (5)	θɹ (1)	
ʃ (2)	g (2)	ɔ (3)	tɹ (1)	
d (1)	n (4)	ɔɪ (1)		
ɹ (1)	p (1)			
g (1)	t (2)			
m (2)				
h (4)				

Having a uniform prompt among all the sample's allowed us to create a stronger measure of accuracy, as we could compare each sample to the same transcript. In addition, GMU's dataset contains almost 3,000 distinct recordings where each recording includes the speaker's

[9] Perez, Sarah. "47.3 Million U.S. Adults Have Access to a Smart Speaker, Report Says." *TechCrunch*, TechCrunch, 7 Mar. 2018, techcrunch.com/2018/03/07/47-3-million-u-s-adults-have-access-to-a-smart-speaker-report-says/.

[10] Weinberger, Steven. (2015). *Speech Accent Archive*. George Mason University. Retrieved from http://accent.gmu.edu

corresponding demographic and linguistic background (including country and language of origin, how the individual learned English, and how long s/he has spoken English). This information allowed us to categorize the accent of each sample. We selected the top five languages in the US (English, Spanish, Chinese, Tagalog, Vietnamese, Arabic)[11] and in the world (Chinese, Spanish, English, Arabic, Hindi)[12] to analyze (seven accents total). For each of these languages, we randomly selected a sample of 15 recordings to test on the four speech-to-text technologies.

Our four voice recognition contenders are Amazon, Google, Microsoft, and IBM. We selected Amazon Web Services Speech-to-text[13] and Google cloud speech-to-text[14] due to their widespread use. As one would expect, Amazon leads the voice recognition market with a reported dominance of 71.9 percent of the smart speaker trade[15]. Google follows with an 18.4 percent share of the market[16]. We chose Microsoft's Bing speech-to-text[17] in order to put their mission of accessibility[18] to the test (and because Microsoft is commonly compared with Amazon and Google). Finally, we picked IBM's Watson Speech-to-text[19] technology as our fourth option because Watson has a more professional intention versus a home based/consumer intention[20]. With our final selection of technologies, we each dove in and explored how to use the respective API. We wrote our programs so that we could pass each technology a WAV file (each accent audio file) and receive back the transcribed text which we then stored in a CSV file with their corresponding accuracy and precision measures.

We wanted more than one measurement from each technology for our analysis in order to conclude a result about the algorithmic bias present, so we collected metrics of both accuracy (the percentage of words transcribed correctly) and precision through edit distance (the minimum

[11] "Languages of the United States." *Wikipedia*, Wikimedia Foundation, 13 Dec. 2018, en.wikipedia.org/wiki/Languages_of_the_United_States.

[12] Lesson Nine GmbH. "What Are The 10 Most Spoken Languages In The World? | Babbel Magazine." *The Babbel Magazine*, www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world/.

[13] Hunt, Randall, et al. "Amazon Transcribe – Accurate Speech To Text At Scale | Amazon Web Services." *Amazon*, Amazon, 30 Nov. 2017, aws.amazon.com/blogs/aws/amazon-transcribe-scalable-and-accurate-automatic-speech-recognition/.

[14] "Cloud Speech-to-Text - Speech Recognition | Cloud Speech-to-Text API | Google Cloud." *Google*, Google, cloud.google.com/speech-to-text/.

[15] Perez, Sarah. "47.3 Million U.S. Adults Have Access to a Smart Speaker, Report Says." *TechCrunch*, TechCrunch, 7 Mar. 2018, techcrunch.com/2018/03/07/47-3-million-u-s-adults-have-access-to-a-smart-speaker-report-says/.

[16] Ibid.

[17] "Speech to Text API | Microsoft Azure." *A Beginner's Guide | Microsoft Azure*, azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/.

[18] "The Ability Hacks Book." *Software Asset Management – Microsoft SAM*, www.microsoft.com/en-us/accessibility.

[19] "Speech to Text." *The Analytics Maturity Model (IT Best Kept Secret Is Optimization)*, IBM Corporation, 28 Nov. 2016, www.ibm.com/watson/services/speech-to-text/.

[20] Nay, Chris (September 6, 2011). "Putting Watson to work: Interview with GM of Watson Solutions Manoj Saxena". *Smarter Planet Blog*. IBM. Retrieved November 12, 2013.

number of operations (insertions, deletions, one character flips) needed in order to change one string into another).

Our "accuracy" function looks at the number of words in each 69-word clip transcribed correctly by the respective speech-to-text technologies. In the accuracy function, we initialized a dictionary using the GMU dataset elicitation paragraph described above with words for the keys and corresponding counts as the values. For each word in the transcribed text that matches a value in the dictionary (i.e. a word that had been correctly transcribed from the input audio file), the function decrements the corresponding value in the dictionary. In the end, we took the absolute value of each value in the dictionary and totaled them. This gave us the raw number of *incorrect* words, which we then used to calculate an accuracy percentage (= 69 - incorrect / 69).

To calculate edit distance, we used an edit distance Python library.

With these two measurements, we ran each of the four technologies on an average of 15 audio clips per language analyzed. For three of the four technologies (IBM, Google, and Amazon), we analyzed 14 different accents, so our total accuracy percentages are based on 210 audio clips total. For Bing, we analyzed 11 different languages, for a total of 165 audio clips. The aggregate averages for the four technologies represents averages over this full data output. In the second part of our final analysis, we focused on the results from a subset of this collected data, namely on the data that represents the 5 top accents based on top-spoken languages worldwide and domestically, with a distinction between English spoken with a UK accent and English spoken with a US accent.

Results:

Using the "Accuracy" rating function, as described above as our primary metric for analysis, we broke down the overall results from the four speech-to-text APIs into overall accuracy (as a percentage out of 100, representing the average percentage of words from the 69-word clip that were transcribed correctly). Our results will focus primarily on the 10 top unique accents, as determined by world and US top languages, with a distinction between "English" with a US accent and with a UK accent. On average, our results represent running 15 distinct clips from each accent group, with some exceptions with only 14 clips.

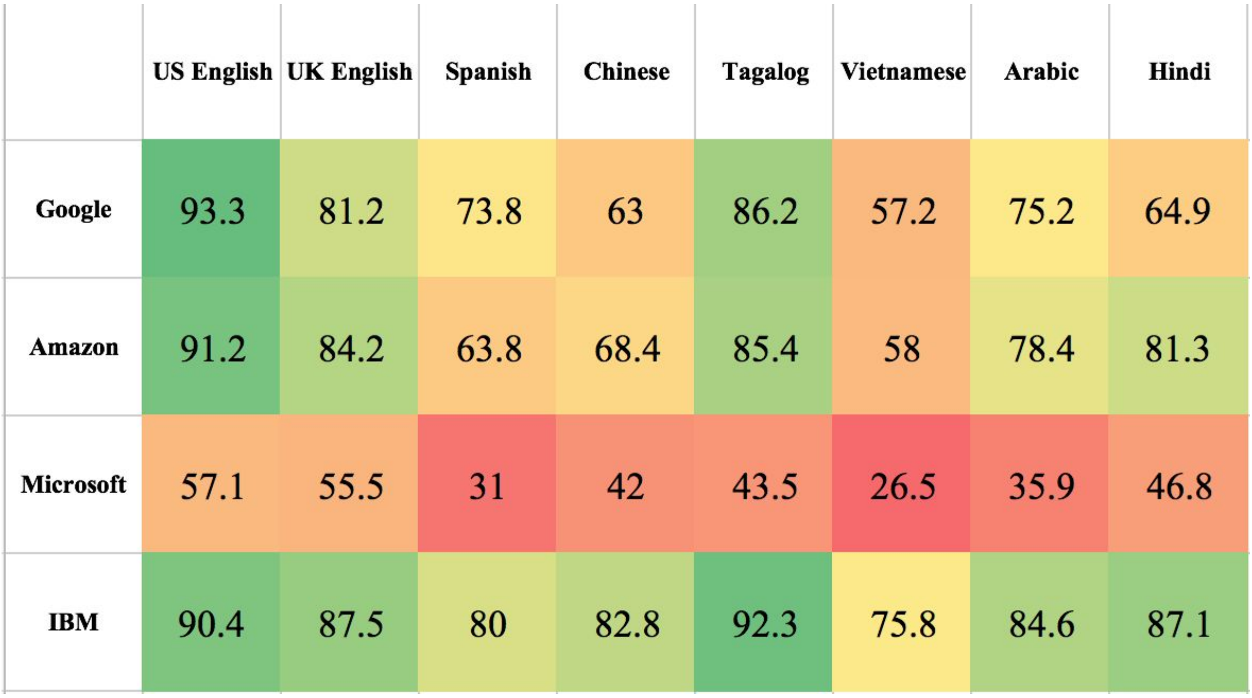
According to the overall output, IBM had the best overall accuracy, with a rate of 85.29 percent accuracy from all clips combined. Microsoft's Bing speech-to-text performed the worst, with overall accuracy rates of only 43.5 percent. It is worth noting that Microsoft's overall accuracy rate is based on fewer clips than the other APIs, based on limits to free use of the API (Bing's overall output is based on 165 audio clips, whereas the other APIs are based on 210 audio clips on average). However, Microsoft's overall standard deviation for accuracy is 8.11 words, the second lowest standard deviation (with IBM's being the lowest), which suggests that Microsoft's speech-to-text algorithm is consistently inaccurate. This low rate is less likely due to statistical anomalies, at least on the clips used as input data. Additionally, while Amazon's overall accuracy rate was relatively high (76.15 percent), it also had the highest standard deviation for overall accuracy, with a standard deviation of 27.18 words. This represents a very high standard deviation, given that the clip is only 69 words long.

The overall accuracy rates and standard deviations among all clips analyzed for all four technologies are as follows:

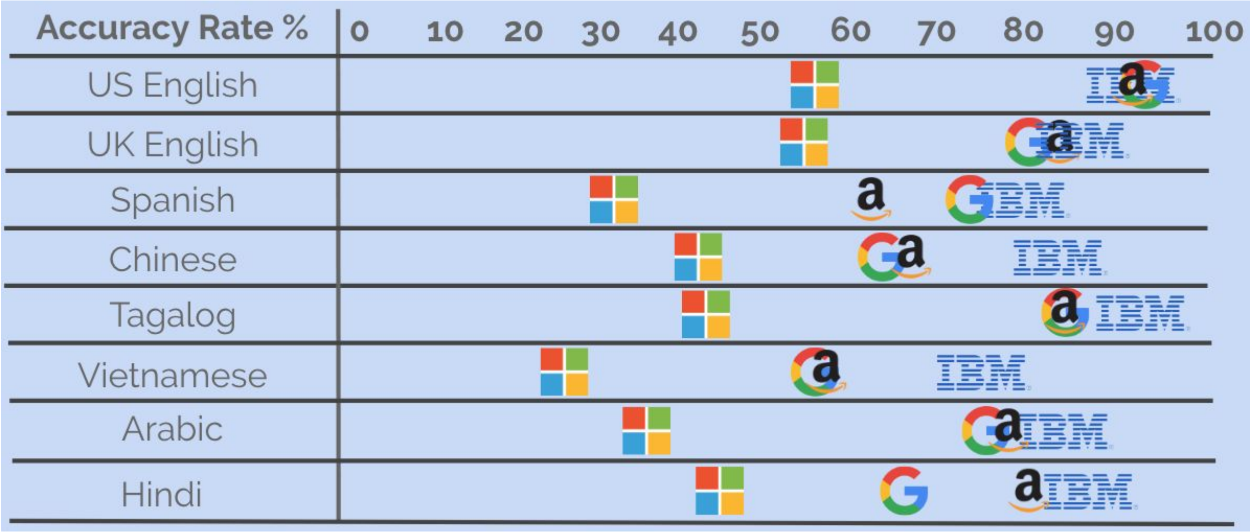
Voice Recognition Technology Company	Overall Accuracy (Mean)	Overall Accuracy (Standard Deviation)
Google	74.65%	13.97
Amazon	76.15%	27.18
Microsoft	43.50%	8.11
IBM	85.29%	6.81

An accent-by-accent breakdown of accuracy for the four technologies revealed that Microsoft's performance for each accent was consistently the worst, whereas IBM performed the best for each accent with the exception of US English, for which Amazon and Google performed better. Perhaps unsurprisingly, US English had the highest performance of accuracy of all the languages among all four technologies. Tagalog (a Filipino dialect) had the 2nd highest overall performance, and UK English had the 3rd highest.

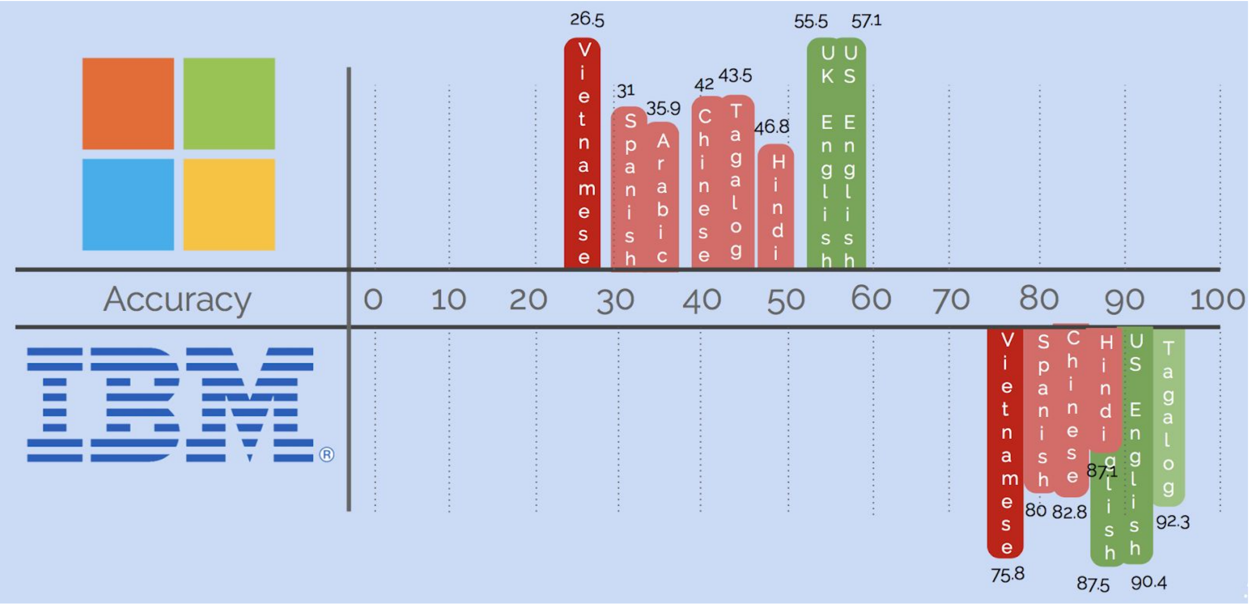
A breakdown of the accuracy average percentages for each of the four technologies is as follows:



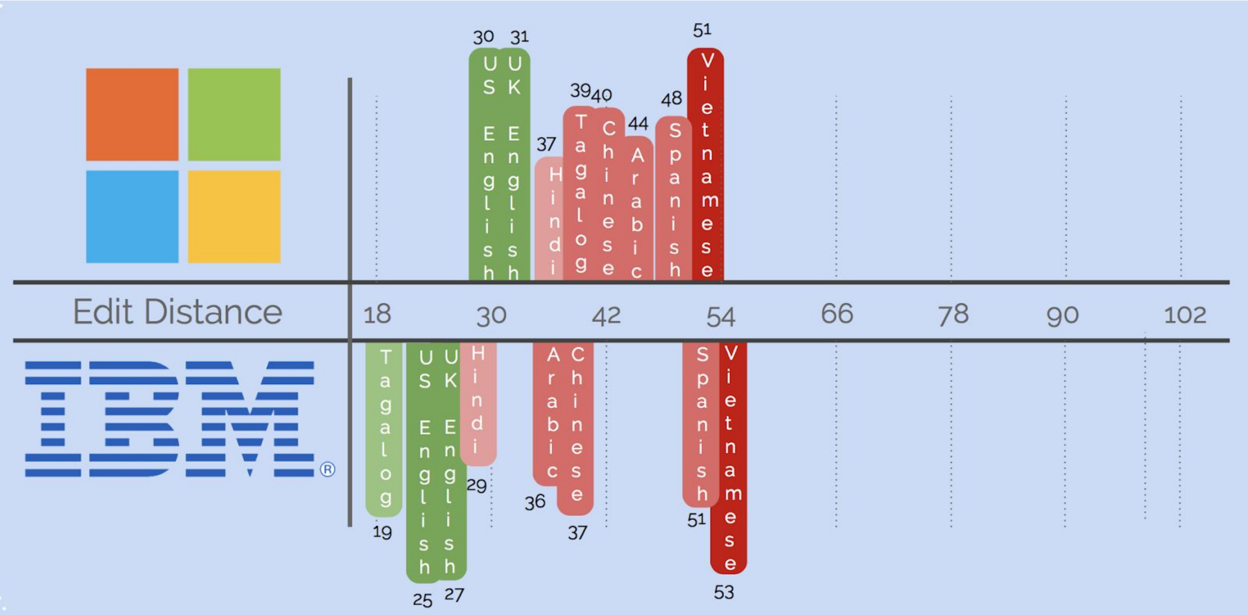
	US English	UK English	Spanish	Chinese	Tagalog	Vietnamese	Arabic	Hindi
Google	93.3	81.2	73.8	63	86.2	57.2	75.2	64.9
Amazon	91.2	84.2	63.8	68.4	85.4	58	78.4	81.3
Microsoft	57.1	55.5	31	42	43.5	26.5	35.9	46.8
IBM	90.4	87.5	80	82.8	92.3	75.8	84.6	87.1



Now, let's look into comparisons of specific technologies. First, let us compare the worst performing technology (Bing speech-to-text) to the best performing technology (IBM Watson speech-to-text). The below visualization shows the relative accuracy rates for our top eight accents:

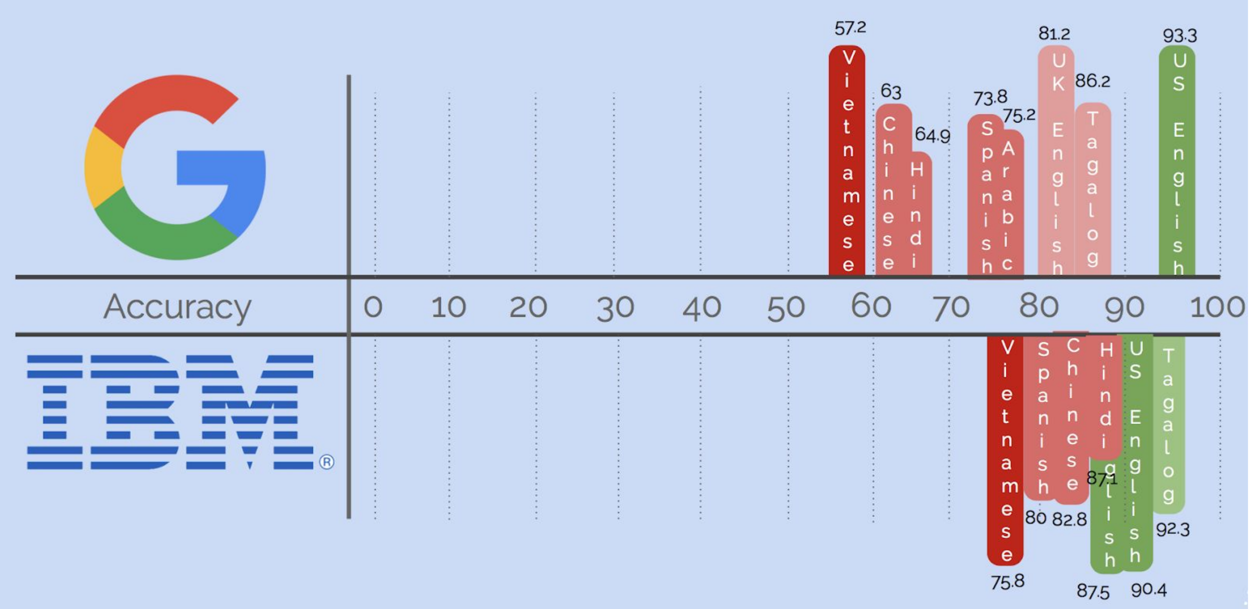


It is strikingly apparent how much better IBM performs for every single accent from this visualization. The spread of the data is also interesting because IBM is centralized between 75 - 95 percent (about a 20 percent spread) while Microsoft is spread between 25 - 60 percent (about a 35 percent spread). This tells us not only that IBM is more accurate, but it also has the best reliability for transcription disregarding accent (least accent bias). Some more interesting observations from the above data is that Vietnamese performed the worst on both technologies (in fact, Vietnamese performed the worst on *all* four technologies despite being the fifth top language in the U.S.). U.S. English unsurprisingly performed the best out of all the accents for Microsoft. However, Tagalog produced the most accurate result on IBM's technology. We can see similar conclusions by comparing the edit distance of Microsoft vs. IBM:

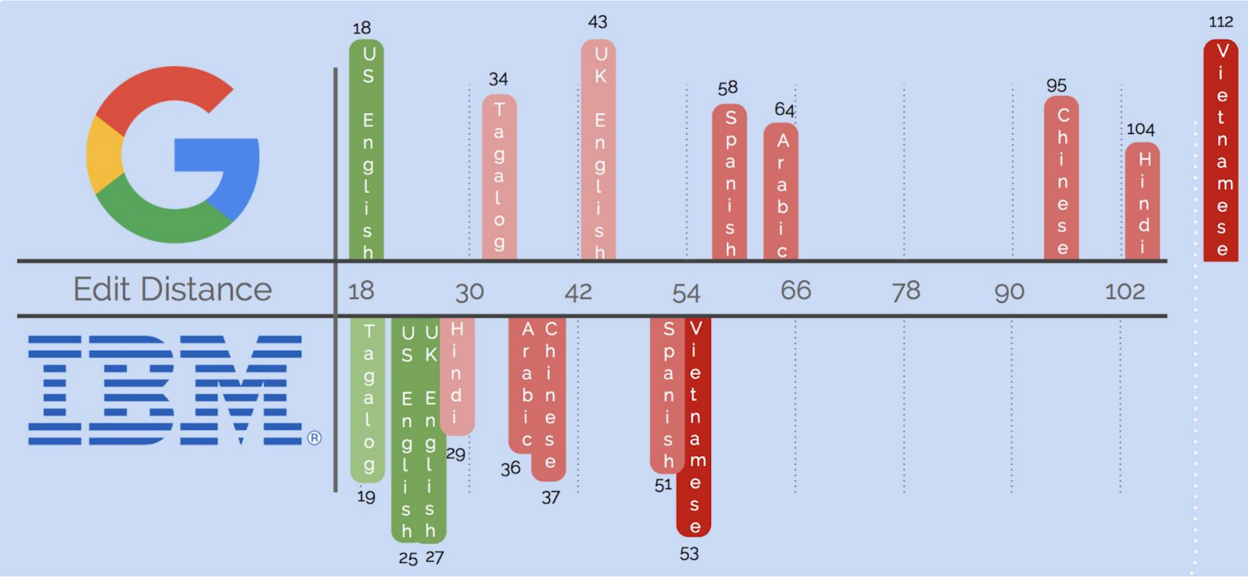


A lower score on edit distance correlates to a better transcription. It is interesting to see that Microsoft had little variation compared to IBM regarding edit distance. While IBM still transcribes most of the accents better from a measure of edit distance, Microsoft slightly beats IBM with Spanish and Vietnamese. This is quite different from what we saw with accuracy because IBM outperforms Microsoft by at least 35 percent with every single accent. This tells us that Microsoft's technology might perform much better than we initially thought. For example, our accuracy test may have been a poor test for Microsoft because the words were split up more than they should have been. This would result in a terrible accuracy score (because they barely match any words), but a good edit distance score because only space deletions need to be made. Let us next observe Google vs. IBM to further explore two technologies.

Looking at Google vs. IBM, we can also make some interesting observations. We can also highlight apparent trends regarding accent bias via voice recognition technologies. Let us look at the relative accuracy rates for our top eight accents between Google and IBM:



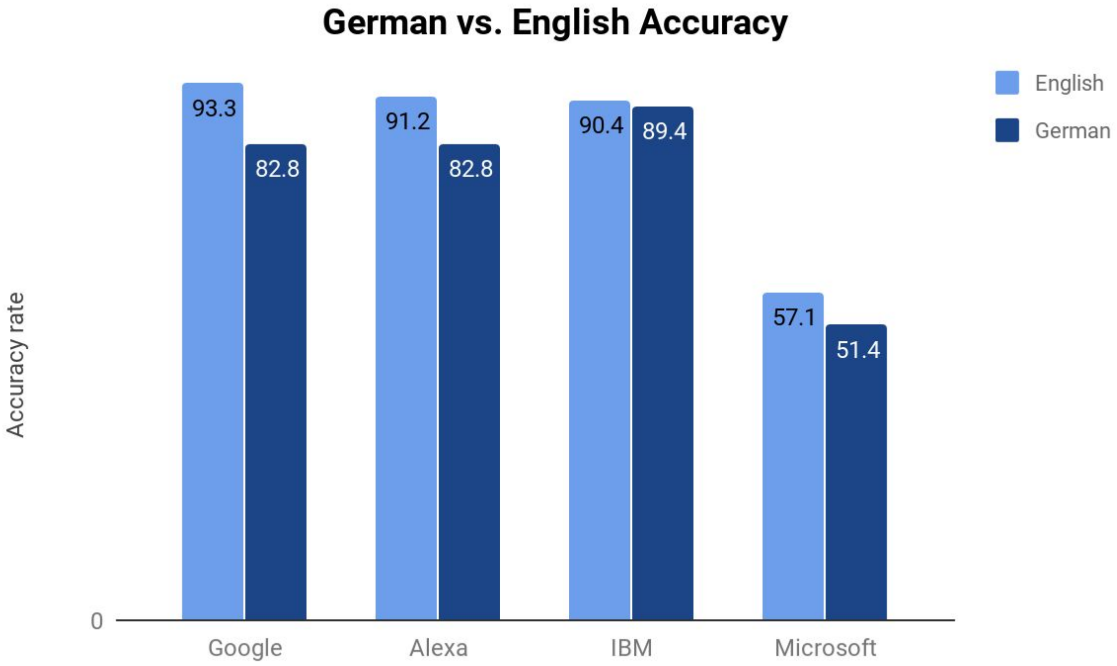
Once again, we see that IBM's spread of data is centralized between 75 - 95 percent (about a 20 percent spread) while Google is spread between 55 - 93 percent (about a 40 percent spread). As with the Microsoft comparison, this highlights IBM's accuracy and relatively low accent bias. Interestingly, Google's data is slightly more spread than Microsoft's. However, Google overall performs much better than Microsoft. In this visualization, we continue to see the trend that Vietnamese performs the worst. We also observe that Google has the best performing U.S. English detection. Google's U.S. English was roughly three percent more accurate than IBM's detection. We can see similar conclusions by comparing the edit distance of Google vs. IBM:



Again, a lower score on edit distance correlates to a better transcription. This visualization corresponds with our observations above about the accuracy of Google vs. IBM. Edit distance further highlights the differences between the spread of Google and IBM's data. Again, Vietnamese is the worst transcribed accent. We can corroborate U.S. English's performance as the best accent for Google and as the best transcribed (via Google) out of all accents and technologies. These comparisons clearly visualize the existence of an accent bias (albeit varying based on technology) among voice recognition technologies.

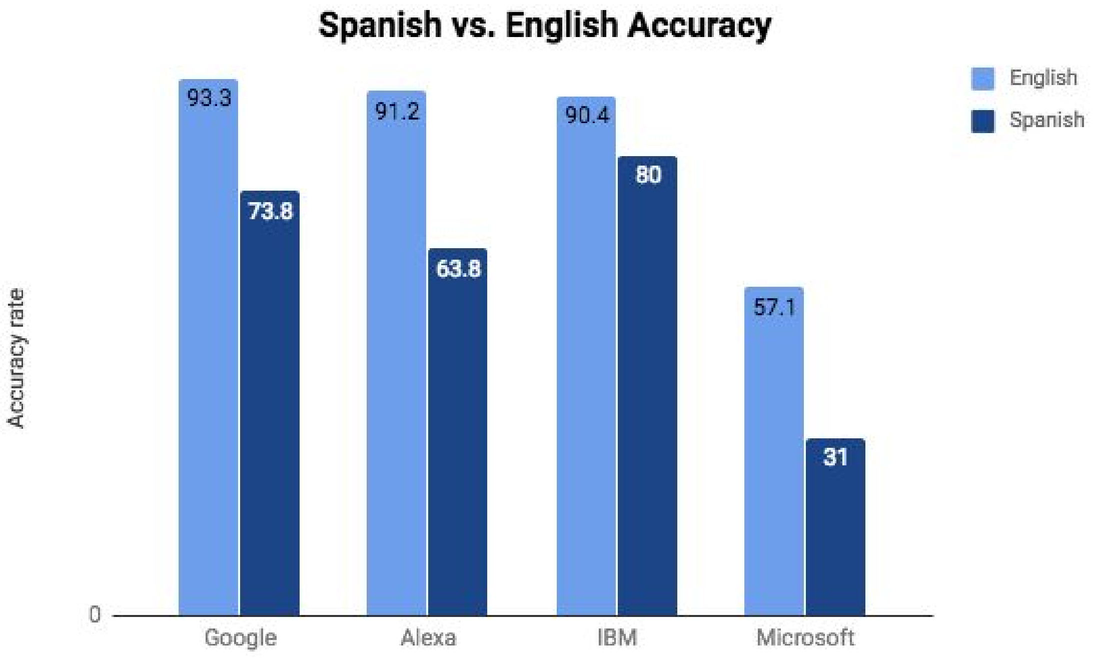
Now, let's examine accuracy rates for a few specific accents: German, Spanish, and Chinese. We'll compare the different accuracy rates for these accents to a "baseline" of US English to better visualize how user experience may vary for people with or without an "un-American" accent.

First, let's compare German to US English:



As illustrated above, while German accents still do considerably worse in comparison to a US English baseline (particularly with Google and Microsoft's APIs), German accents are still generally well understood - while 9 out of every 10 words spoken with a US English accent were correctly translated on average for all APIs except for Microsoft's, approximately 8 out of every words spoken with a German accent were correctly translated. The IBM API does particularly well with German accents, with only a 1 percent difference in accuracy.

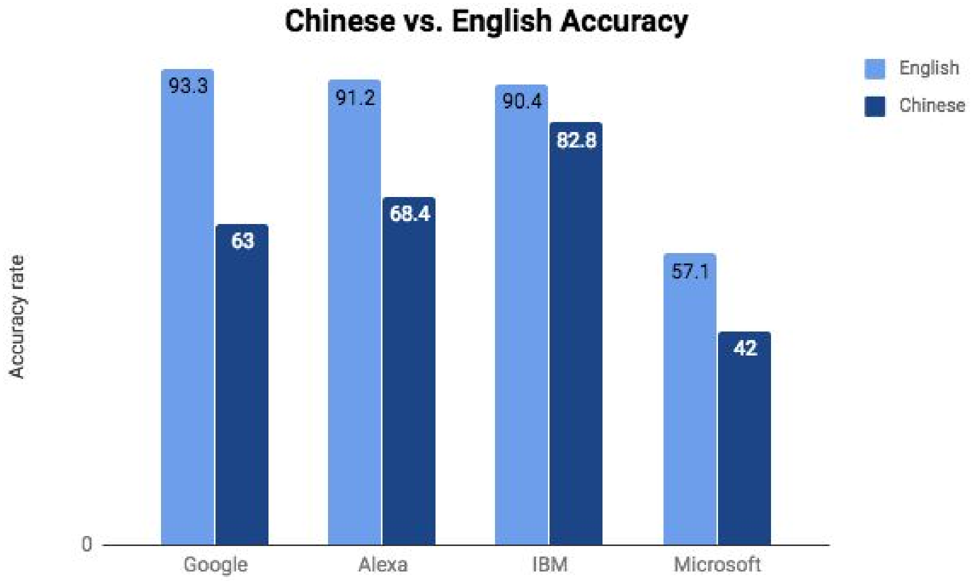
In contrast to Spanish, another European language for which when compared to US English accents does considerably worse than German accents:



We chose to examine Spanish accents because over 20 percent of the US population speaks Spanish, and over 45 million Americans report speaking Spanish at home. Of these people, only half of them reported that they speak English “very well” while the other half reported speaking English with some accent that isn’t always commonly understood[21]. The chart above shows drastic disparities between US English and Spanish accuracy rates; while IBM’s API did only 10 percent worse, Microsoft, Google, and Alexa all did considerably worse. This disparity is most prominent for Alexa, where fewer than 2 out of every 3 words spoken by a Spanish accent were understood. This has drastic implications for the approximately 25 million Americans who report speaking English with some form of Spanish accent and their ability to fully utilize voice-powered technologies. In future studies, we also hope to further explore how different dialects of Spanish perform; the sample of audio clips with Spanish accents contained a variation of speaker’s native countries from South America to Central America to Spain which would be interesting to separately examine the different dialects.

[21] Burton, James. “The Most Spoken Languages In America.” World Atlas, Worldatlas, 30 Dec. 2015, www.worldatlas.com/articles/the-most-spoken-languages-in-america.html.

Finally, we chose to examine Chinese accents not only because Chinese is the most spoken language in the world, but also because 3.4 million Americans speak Chinese at home[22], of which only around 41 percent reported speaking English "very well"[23].



With the exception of the IBM API, for which Chinese accents were still able to be correctly translated over 80 percent of the time, the other APIs did considerably worse; while Alexa performed better on Chinese accents than Spanish accents, Google did 10 percent worse. For both Google and Alexa, only around 2 out of every 3 words spoken with a Chinese accent were correctly understood. Similarly to how the Spanish language itself has many different dialects that potentially influence Spanish accents when speaking US English, it would be interesting to further examine accuracy rates for different Chinese language dialects too (such as Cantonese, Mandarin, and Taiwanese).

Technical Challenges/Limitations:

A few relevant technical challenges and limitations with our data collection and analysis techniques, all of which may have affected the accuracy and robustness of our results, are worth noting. First of all, only a certain amount of data was allowed to be processed for free for each

[22] LanguageLine. "Census: More than 20 Percent of U.S. Residents Speak a Language Other than English at Home." *LanguageLine Blog*, Mar. 2017, blog.languageline.com/limited-english-proficient-census.

[23] "One in Five U.S. Residents Speaks Foreign Language at Home, Record 61.8 Million." Center for Immigration Studies, Center for Immigration Studies, 2015, cis.org/One-Five-US-Residents-Speaks-Foreign-Language-Home-Record-618-million.

API, so the number of overall samples ran for each dataset was limited to 165 for the Bing API, and 210 for all the other APIs. This represents an average of 15 clips of data for each language we analyzed.

Additionally, it is worth noting that we did not remove significant outliers before analyzing our overall data. Given the relatively small sample size (~15 clips) that we had for each language, it is possible that outliers in the dataset threw off the average accuracy for certain languages.

For Bing specifically, we encountered a few additional technical challenges that may have affected our ultimate output. The Bing speech-to-text REST API, which we used to process our audio data, is limited to processing clips of 15 seconds. All audio clips were longer than 15 seconds; on average, the full clips were around 25 seconds, with some as long as 45 seconds long. Therefore, we split the audio files before running them through the Bing API and concatenated the results for each clip before running them through our accuracy checker. We did not adjust for words that may have been inaccurately analyzed because they were split in two by the audio file splitting. On average, this certainly brought down the accuracy of the Bing output.

Potential Future Research:

Potential future extensions of our research project include a deep dive into which words were most commonly misinterpreted and what those words were incorrectly translated to. Better understanding when the technologies fail may give insight into what the algorithms are missing. As mentioned above, we are also interested in understanding how different regional dialects of a foreign accent are able to be translated, as some regional dialects that determine accents speaking English are strongly associated with socioeconomic status. Our experiment also only ran audio clips of people speaking one specific script; while that script was chosen to capture almost all phonetic sounds in the English language, there's no way that one script itself can be indicative of the most commonly made requests and uses of voice recognition technology. An interesting future experiment would be to run larger samples of audio clips of people reading a variety of longer and more differing scripts. Finally, the script that we chose isn't a common request that's made to one's Alexa or Google Home; running audio files reading common requests to personal devices would make our experiment more directly applicable to existing uses of voice control.

In addition to examining bias in the context of foreign accents speaking scripts in English, we are also interested in exploring how US regional accents and dialects characteristic to regional communities would perform for each of the most popular APIs. Finally, we also want to explore how people with speech impediments are understood by voice recognition technologies. This is a really important experiment, as many people who are physically disabled and therefore have the most to gain through innovations in hands-free voice control are more likely to have speech impediments.

Conclusion:

Are speech-to-text algorithms biased? Our analysis of four top speech-to-text algorithms demonstrates that yes, they are heavily biased against particular foreign accents when translating requests made in English.

This analysis revealed significant discrepancies in accuracy between English spoken with a US American accents and English spoken with a different regional accent. These four technologies performed particularly poorly on English spoken with Vietnamese and Spanish accents, as well as on accents from a range of Asian dialects. Of the four technologies, IBM's performed the best overall, while Microsoft's performed the worst. When performance is examined for each individual accent, European accents such as French, German, and Portuguese perform significantly better than Asian accents. The one exception to this trend is Spanish accents, which misses approximately 33 percent more words in comparison to audio clips with speakers who have no foreign accent. These disparities are incredibly wide, as approximately only 2 out of every 3 words spoken in most of the Asian accents and Spanish can be correctly interpreted by most popular technologies, which has drastic implications for how easy it can be for these users to use new voice-enabled services every day. Some popular API technologies, such as IBM's Speech-to-Text, do much better in understanding all of the accents, but even for the IBM API itself these disparities and common biases towards European accents still exist.

While these disparities in accuracy rates may be explained by a number of factors - we hypothesize that the training datasets for the APIs themselves failed to include a wide array of people speaking English with foreign accents - we believe our results illustrate a reality that reinforces common struggles of foreign-born people living in the US: a struggle to communicate and be understood. We also believe that there are likely linguistic or phonetic justifications as to why certain accents sound more or less close to US English accents, and that companies continuing to improve natural language processing algorithms could greatly benefit from understanding how these linguistic underpinnings may inherently bias their algorithms towards certain patterns of speech characteristic to specific populations. But regardless of how we got here, we strongly believe that in order to build more inclusive products where innovations can be for the good of everyone instead of just those who are traditionally privileged, companies should do all they can to bridge the wide gap in how their technologies are significantly harder to use for people with foreign accents.

Appendix A

Github repository: https://github.com/njohnson99/CS105-Speech-Algorithms

Relevant files:

For collecting data:

- IBM_speech_to_text.py - querying IBM API
- Bing_stt.py - querying Bing API
- Google-translate.py - querying Google API
- AWS-transcribe.py - querying Amazon API
- mp3toWAV.py - formatting audio files to correct .wav format
- Split_wav_files.py - splitting .wav files into 15 second increments for analysis by the Bing API (15 second limits)

For analyzing data:

- Calc_edit_distance.py - for calculating edit distance between output text and true text.
- Accuracy_checker.py - for calculating word-by-word accuracy of output text

Data:

- IBM_summary.xlsx
- Bing_output.csv
- Google_summary.xlsx
- Amazon-summary.xlsx
- Overall_rates.xlsx
- summary.xlsx

Bibliography

Burton, James. “The Most Spoken Languages In America.” World Atlas, Worldatlas, 30 Dec. 2015, www.worldatlas.com/articles/the-most-spoken-languages-in-america.html.

“Cloud Speech-to-Text - Speech Recognition | Cloud Speech-to-Text API | Google Cloud.” *Google*, Google, cloud.google.com/speech-to-text/.

HubSpot. “The Ultimate List of Marketing Statistics for 2018.” *Hubspot*, www.hubspot.com/marketing-statistics.

Hunt, Randall, et al. “Amazon Transcribe – Accurate Speech To Text At Scale | Amazon Web Services.” *Amazon*, Amazon, 30 Nov. 2017, aws.amazon.com/blogs/aws/amazon-transcribe-scalable-and-accurate-automatic-speech-recognition/.

“Immigration to the United States.” *Wikipedia*, Wikimedia Foundation, 9 Dec. 2018, en.wikipedia.org/wiki/Immigration_to_the_United_States.

“Languages of the United States.” *Wikipedia*, Wikimedia Foundation, 13 Dec. 2018, en.wikipedia.org/wiki/Languages_of_the_United_States.

LanguageLine. “Census: More than 20 Percent of U.S. Residents Speak a Language Other than English at Home.” LanguageLine Blog, Mar. 2017, blog.languageline.com/limited-english-proficient-census.

Larson, Selena. “Research Shows Gender Bias in Google's Voice Recognition.” *The Daily Dot*, 15 July 2016, www.dailydot.com/debug/google-voice-recognition-gender-bias/.

Lesson Nine GmbH. “What Are The 10 Most Spoken Languages In The World? | Babbel Magazine.” *The Babbel Magazine*, www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world/.

Mailonline, Joe Pinkstone For. “AI Assistants Are Sexist and Understand Men Better.” *Daily Mail Online*, Associated Newspapers, 14 Mar. 2018, www.dailymail.co.uk/sciencetech/article-5499339/AI-assistants-sexist-understand-men-better.html.

Nay, Chris (September 6, 2011). "Putting Watson to work: Interview with GM of Watson

Solutions Manoj Saxena". *Smarter Planet Blog*. IBM. Retrieved November 12, 2013.

Olmstead, Kenneth. “Voice Assistants Used by 46% of Americans, Mostly on Smartphones.” *Pew Research Center*, Pew Research Center, 12 Dec. 2017, www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/.

“One in Five U.S. Residents Speaks Foreign Language at Home, Record 61.8 Million.” Center for Immigration Studies, Center for Immigration Studies, 2015, cis.org/One-Five-US-Residents-Speaks-Foreign-Language-Home-Record-618-million.

Perez, Sarah. “47.3 Million U.S. Adults Have Access to a Smart Speaker, Report Says.” *TechCrunch*, TechCrunch, 7 Mar. 2018, techcrunch.com/2018/03/07/47-3-million-u-s-adults-have-access-to-a-smart-speaker-report-says/.

Sentance, Rebecca. “What Does Meeker's Internet Trends Report Tell Us about Voice Search?” *Search Engine Watch Search Marketing Guide to Naver Koreas Most Popular Search Engine Comments*, Search Engine Watch, 15 Nov. 2018, searchenginewatch.com/2016/06/03/what-does-meekers-internet-trends-report-tell-us-about-voice-search/.

“Speech to Text.” *The Analytics Maturity Model (IT Best Kept Secret Is Optimization)*, IBM Corporation, 28 Nov. 2016, www.ibm.com/watson/services/speech-to-text/.

“Speech to Text API | Microsoft Azure.” *A Beginner's Guide | Microsoft Azure*, azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/.

“The Ability Hacks Book.” *Software Asset Management – Microsoft SAM*, www.microsoft.com/en-us/accessibility.

Weinberger, Steven. (2015). *Speech Accent Archive*. George Mason University. Retrieved from http://accent.gmu.edu

Zong, Jie, et al. “Frequently Requested Statistics on Immigrants and Immigration in the United States.” *Migrationpolicy.org*, 27 Feb. 2018, www.migrationpolicy.org/article/frequently-requested-statistics-immigrants-and-immigration-united-states.