Automatic Speech Recognition Services: Deaf and Hard-of-Hearing Usability

Abraham Glasser
Rochester Institute of Technology
Rochester, NY, USA
abraham.glasser@gmail.com

ABSTRACT

Nowadays, speech is becoming a more common, if not standard, interface to technology. This can be seen in the trend of technology changes over the years. Increasingly, voice is used to control programs, appliances and personal devices within homes, cars, workplaces, and public spaces through smartphones and home assistant devices using Amazon's Alexa, Google's Assistant and Apple's Siri, and other proliferating technologies. However, most speech interfaces are not accessible for Deaf and Hard-of-Hearing (DHH) people. In this paper, performances of current Automatic Speech Recognition (ASR) with voices of DHH speakers are evaluated. ASR has improved over the years, and is able to reach Word Error Rates (WER) as low as 5-6% [1][2][3], with the help of cloud-computing and machine learning algorithms that take in custom vocabulary models. In this paper, a custom vocabulary model is used, and the significance of the improvement is evaluated when using DHH speech.

KEYWORDS

Automatic Speech Recognition; Deaf and Hard-of-Hearing; Speech Usability.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

© 2019 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3308461

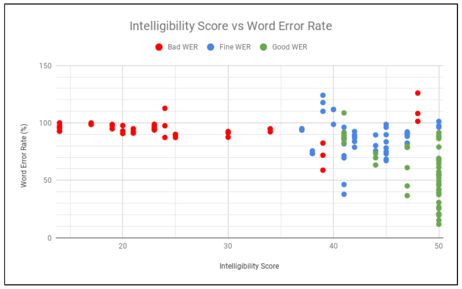


Figure 1: Scatterplot of intelligibility scores and WER for the audio database

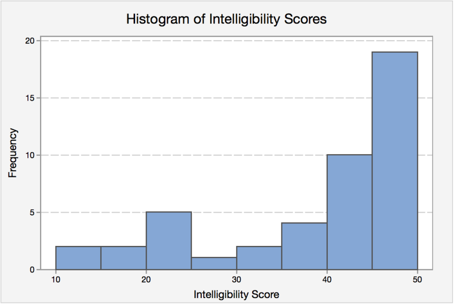


Figure 2: Histogram showing frequency of intelligibility scores

INTRODUCTION AND PREVIOUS WORK

The current and developing trend of speech interfaces can be seen in modern cars and home assistant devices. Such devices use Automatic Speech Recognition (ASR) to detect what is being spoken to them, and perform accordingly. Foreign accents and disfluencies in speech, historically, have had big impacts on the capability of ASR to understand human speech. However, current ASR technology is able to perform well, even if human speakers have these differences in speech. Current technology is able to do this with the help of cloud computing, machine learning, and sufficient datasets.

Glasser, Kushalnagar, and Kushalnagar did a preliminary study on using Deaf and Hard-of-Hearing (DHH) speech [6]. However, there have been significant advances in ASR since then. ASRs have been becoming more accurate, boasting Word Error Rates (WERs) as low as 5-6%[1][2][3]. Also, there have been advancements where an user can provide custom language models for the ASR to have context awareness and improve accuracy.

In this study, a deeper analysis is done than the preliminary work done in [6]. Also, an analysis of WER improvement when using an ASR engine with custom language models is performed.

Even though ASR technology has improved dramatically over the past few years, and is now being incorporated in everyday technologies, it still has an usability challenge when it comes to the DHH population. DHH speech generally sounds different from hearing speech, and varies greatly between DHH individuals. DHH speech is often so variable that there is difficulty in understanding, even among experienced and inexperienced human listeners [8].

METHODOLOGY

Audio Dataset

The dataset used in this study is a subset of a large speech dataset of 650 Deaf and Hard-of-Hearing (DHH) individuals at the National Technical Institute for the Deaf at Rochester Institute Technology, which has an enrollment of around 1100 DHH students. The dataset consists of DHH individuals who took the Clarke Sentences intelligibility test [7]. The test has 60 sentence lists, with 10 sentences per list. The sentences each have 10 syllables. The number of actual words varies across the sentences and lists. Each audio file has one DHH speaker reading one sentence list. The audio recordings were rated by a speech pathologist, who gave them an intelligibility score of 0 to 50. A score of 50 indicates that the speech is understood by the pathologist, while a score of 30 means difficult to understand, and a score of 0 means completely unintelligible.

In [6], 45 audio files were chosen by a naive listener. 15 samples were rated "good", 16 samples were "fine" and 14 samples were "bad". These were determined by the naive listener who categorized the audio files in these three categories. The terms "good", "fine", and "bad" are used in this paper to refer to these categories. The average intelligibility score for the audio files in the "bad" category was 25, 43 for the "fine", and 48 for the "good".

The Waveform Audio File Format (filename extension .wav) container format was used, and the audio itself was encoded using PCM 16-bit little-endian encoding. This is high fidelity, uncompressed digital audio.

ASR engines

For this study, the modern and widely used Automatic Speech Recognition (ASR) engines are the best fit, since they are generally being incorporated in everyday technology, and are also freely available for public use. Of these, the Microsoft Translator Speech API and the IBM Watson Speech to Text service were used. The web version of Microsoft Translator was used for the "base" model of the ASR engine, while the Presentation Translator for Microsoft PowerPoint plug-in was used for the "custom" model of the ASR engine.

For MSPPT, the entire list of Clarke Sentences was used as the keywords for the ASR engine to learn from. In theory, this should not decrease the accuracy of the MSPPT ASR compared to MS. MS and MSPPT are essentially the same ASR, with the exception that MSPPT has some "training" from vocabulary keywords that are given, and learns from it to improve accuracy. Throughout this paper, "customization" refers to this Context awareness.

IBM Watson Speech to Text was selected because it is available as a demo in their website, easily accessible, free to use, and was developed by a well-known large corporation.

Links to the ASRs used are provided in Sidebar 1. They are all available for public use. They are also commercially available and continue to be improved on by the corporations.

MS
https://translator.microsoft.com/

MSPPT
https://translator.microsoft.com/help/presentation-translator/

IBM
https://speech-to-text-demo.ng.bluemix.net/

Sidebar 1: Links to the ASRs services used in this study

WER Analysis

Word Error Rate (WER) is a standard measure of how accurate an ASR engine is. In this study, the National Institute of Standards and Technology Speech Recognition Scoring Toolkit (SCTK) Version 2.4.0.4 was used. This is freely available for people to use [9]. The SCTK compares the reference ("truth script") and the hypothesis (output from ASR engine) transcripts, and calculates the WER. The transcripts are aligned, and the number of word substitutions, deletions, and insertions are found. The total of these divided by the total number of words in the reference transcript is the WER.

During the process of WER analysis, both the hypothesis and reference transcripts were converted to lowercase. Also, the newlines were stripped, and the entire transcript is on one line with a new line at the end. All punctuation marks except for apostrophes were removed. All this was manually done before WER analysis to eliminate things that may be different but should not be penalized across the reference and hypothesis transcripts.

Also, in the recordings themselves, there are sentences that were spoken by the person recording the session. These sentences consisted of the date, and prompts for the number of the sentence in

Table 1: Bonferroni correlation results for each ASR engine's performance across each audio category.

T.P = 2-sample T-test P-values

B.P = Bonferroni Correlation P-values (* means significant)

ASR	T.P	B.P
IBM		
Bad-Fine	.1927	1
Bad-Good	.0005	.0045 *
Fine-Good	.0045	.0405 *
MS		
Bad-Fine	.161	1
Bad-Good	.0001	.0009 *
Fine-Good	.0005	.0045 *
MSPPT		
Bad-Fine	.0629	.5661
Bad-Good	.0001	.0009 *
Fine-Good	.0005	.0045 *

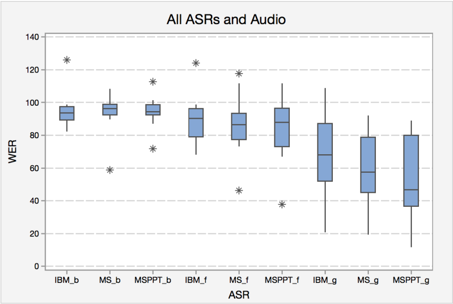


Figure 3: Side by side boxplots of WER for all ASRs and audio categories Asterisks (*) denote outliers

the list. For example: "Today is 1/5/2009. Reading Clarke Sentences list number 43. Number one.", "Number two," and such. These were removed from the transcripts, and the reference transcripts did not contain these prompts. These "cleanings" of the transcripts were done to ensure that we did not account for information that was spoken by the non-DHH individual.

RESULTS

Three different ASRs ("IBM", "MS", and "MSPPT") and three different audio categories are used ("bad", "fine", and "good"). See the respective sections for in-depth explanations.

WER for Deaf and Hard-of-Hearing Speech

Figure 3 shows a side by side comparison of all the boxplots for each ASR engine in each audio category. IBM_b refers to the "IBM" ASR engine in the "bad" audio category, MS_f refers to the "MS" ASR engine in the "fine" audio category, and so forth. As seen in these boxplots, all of the ASR engines in the "bad" and "fine" audio category had a high WER and a small variance. The "good" audio category had a large variance but an overall lower WER. This shows that for all the "bad" and "fine" audio, all the ASRs resulted in a very high WER, and a lower WER for the "good" audio, although not consistent. With voices of the non-DHH population, ASRs have improved drastically over the years and resulted in WERs as low as 5-6% [1][2][3].

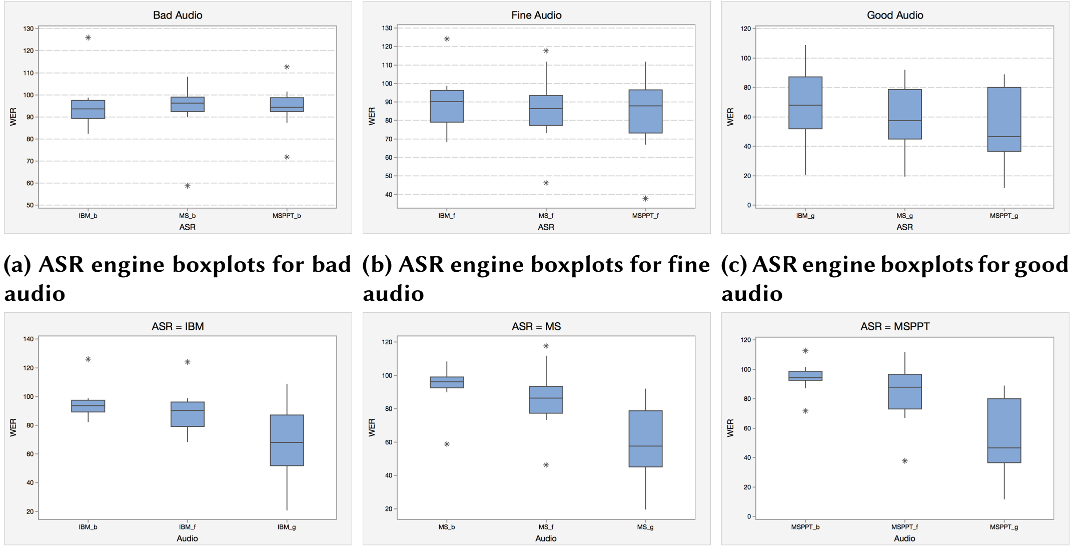
Even if a DHH voice is clear and sounds "good" to a naive listener, ASRs still do not always perform as expected. It was expected for the "good" audio to have a low WER but the variance in the data was very high, and the median of the WERs was above 45% for all the ASRs in that audio category.

A one-sample t-test was performed for the WERs for each audio category. The 95% confidence interval was found to be (91.338, 97.443) for "bad", (82.109, 91.316) for "fine", and (51.288, 66.068) for the "good" audio category. This shows that, with these ASRs, DHH speech will most likely result in a very high WER whether the audio sounds "bad" or "fine" to a naive listener and/or was assigned a intermediate intelligibility score by a speech pathologist.

Improvements in WER for Deaf and Hard-of-Hearing Speech with Context Awareness

Modern ASRs have reached low WERs with the help of customization, where they are provided a vocabulary/keyword list so they have Context awareness [5]. Context awareness helps the ASR engine better predict and identify what was said, and generally improves the accuracy of ASRs significantly.

However, there was not a significant improvement between the MS and MSPPT results. Even though the median WER for the "good" audio category improved by a little more than 10%, the standard deviation was more than 20%, since the results were very various. 2-sample T-tests were performed for all the MS vs. MSPPT results in all the audio categories, and the lowest P-value was .5472 from the "good" audio category. The results from the "bad" and "fine" showed there was basically no improvement between MS and MSPPT, as visualized in Figure 4.



(a) ASR engine boxplots for bad audio

(b) ASR engine boxplots for fine audio

(c) ASR engine boxplots for good audio

(d) IBM ASR engine boxplots for all audio

(e) MS ASR engine boxplots for all audio

(f) MSPPT ASR engine boxplots for all audio

Figure 4: Comparison of WER results boxplots for each ASR service and each audio category Asterisks (*) denote outliers in these boxplots

CONCLUSION

The results for the 2-sample t-tests between the same ASR engine's performance in two different audio categories are shown in Table 1. A Bonferroni correlation test was done with these results, and all of the "bad" vs "fine" results were not significant. This verifies that the WER performance was not significantly different for any ASR engine between the "bad" and "fine" audio.

What this means is that DHH individuals would not be able to achieve equal WERs as the non-DHH population if their voice fell within the "bad" or "fine" audio categories. Even if their voice is "good", is still likely that they will get unpredictable results from the ASRs, as patterns in speech are very various within the DHH population, as pointed out by [4].

As seen in Figure 4, the WER for the audio did not vary much between the ASR engines for the "bad" and "fine" categories, even when Context was provided for MSPPT. For the "good" audio category, where the DHH speech was rated very intelligible by a speech pathologist and put in the "good" category by a naive listener, the WER improved slightly between IBM and MS to MSPPT, albeit not significantly. The variance in the WER was much larger for all the ASR engines in the "good" audio category. This shows that you cannot yet use general DHH speech with ASRs.

FUTURE WORK

Our research shows that with enough data, it should be possible for ASRs to achieve consistent results with DHH speech, whether or not those results achieve low WERs. DHH speech is very various between individuals, and is even sometimes various within a specific DHH individual. If a DHH individual is able to achieve consistent results with an ASR, they might be able to use an acoustic model in addition to Context awareness to tailor the ASR to work with their speech.

ASR relies on having datasets to train with and learn from. Without enough data, ASR would not be able to achieve such low WERs as seen with voices of the non-DHH population. Companies and developers of modern ASRs have had access to large datasets of non-DHH speech, but not as much data for DHH speech. If a sufficiently large dataset of DHH speech is obtained and organized, then it is possible that ASRs will improve over time with this data and perform better than it has been with voices of DHH individuals.

REFERENCES

[1] 2017. Making sense of Google CEO Sundar Pichai's plan to move every direction at once. (2017). https://www.cnbc.com/2017/05/18/google-ceo-sundar-pichai-machine-learning-big-data.html

[2] 2017. Microsoft researchers achieve new conversational speech recognition milestone. (2017). https://www.microsoft.com/en-us/research/blog/microsoft-researchers-achieve-new-conversational-speech-recognition-milestone/

[3] 2017. Reaching new records in speech recognition. (2017). https://www.ibm.com/blogs/watson/2017/03/reaching-new-records-in-speech-recognition/

[4] Jeffrey P. Bigham, Raja Kushalnagar, Ting-Hao Kenneth Huang, Juan Pablo Flores, and Saiph Savage. 2017. On How Deaf People Might Use Speech to Control Devices. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '17*. ACM Press. DOI: http://dx.doi.org/10.1145/3132525.3134821

[5] G. E. Dahl, Dong Yu, Li Deng, and A. Acero. 2012. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 1 (jan 2012), 30–42. DOI: http://dx.doi.org/10.1109/tasl.2011.2134090

[6] Abraham T. Glasser, Kesavan R. Kushalnagar, and Raja S. Kushalnagar. 2017. Feasibility of Using Automatic Speech Recognition with Voices of Deaf and Hard-of-Hearing Individuals. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '17*. ACM Press. DOI: http://dx.doi.org/10.1145/3132525.3134819

[7] Marjorie E. Magner. 1972. *A speech intelligibility test for deaf children.* Technical Report. Clarke School for the Deaf, Northampton, MA.

[8] Nancy S. McGarr. 1983. The Intelligibility of Deaf Speech to Experienced and Inexperienced Listeners. *Journal of Speech Language and Hearing Research* 26, 3 (sep 1983), 451. DOI: http://dx.doi.org/10.1044/jshr.2603.451

[9] National Institute of Standards and Technology. SCTK. (????). https://github.com/usnistgov/SCTK