Bias in Digital Voice Communication

The Effectiveness and Biases of Speech-to-text Algorithms due to their Flawed Training Data Sets

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Systems Engineering

By
Madison Sullivan

November 8, 2024

Technical Team Members:
Catherine Nguyen, Elizabeth Recktenwald, and Lucas Vallerino

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Caitlin D. Wylie, Department of Engineering and Society

Matthew Bolton, Department of Systems and Information Engineering

Introduction

During my first meeting with my technical capstone group, our advisor shared a personal annocdote with us that inspired his research. He had described this issue in a paper, stating, “in my own subjective experience using VoIP [Voice Over Internet Protocol] to communicate with my wife, students, and colleagues, I have had more trouble understanding female voices (for people I had no trouble understanding in person) over VoIP than male voice” (Bolton, 2022, 227). With the rise of internet communication usage, especially following the Covid-19 pandemic, Professor Bolton decided that this needed to be investigated more to see if this was the case for others as well, since this could then portray a sexist performance in VoIP (Bolton, 2022). VoIP is the technology that allows voices to be transferred over the internet, for both phone and video calls, which needs to meet the needs of all users regardless of their demographic, or else it faces the problem of being discriminatory. In my technical report, my team will share our findings testing VoIP at different settings to see if the biases described by Professor Bolton hold true.

For my STS thesis, I decided to research biases that exist in another communication software, speech-to-text algorithms. Speech-to-text algorithms are based on artificial intelligence (AI) and use automatic speech recognition (ASR) to form words and sentences based on what is spoken. The ASR used in speech-to-text technologies has been “ shown to struggle with speech variance due to gender, age, speech impairment, race, and accents” (Feng et al., 2021, 1). Feng’s claims inspired me to explore this concept further, because if similar biases occur in speech-to-text as in VoIP, this is a major problem for inclusivity in communication field. The accuracy of speech-to-text algorithms is dependent on the data used to train the model, including the range of gender, language, accent, pace, and age of the speaker. Training data is typically composed of

native speakers of a “standard” variant of a language. Unsurprisingly, this causes higher error rates for non-native speakers or those who speak a regional dialect. I aim to study possible mitigations to this problem in my STS paper, focusing on underrepresented groups in training data that leads to flaws in speech-to-text algorithms.

Technical Topic

My technical topic focuses on biases that exist in VoIP communication, specifically looking at sex biases between males and females. VoIP is dependent on a specific codec, which is a software algorithm that encodes voices when inputted and decodes them on the other end, for quicker transmission. Codecs range in settings and quality, resulting in different vocal interpretations depending on the codec that is used. Depending on the vocal frequency of the user, their voice can become distorted when passing through the codec, which causes them to be misunderstood on the other end, just as Professor Bolton had more difficulty understanding his wife compared to male collegues. I agree with Bolton’s preliminary claims and I am encouraged to investigate this problem more, as if it is found that popular VoIP codecs are biased against higher vocal frequencies, this could lead to a diminished use or lawsuits against technologies using such codecs, including Zoom or Microsoft Teams. These technologies are often used in the workplace, which is particularly important to have equal access to communication for all individuals, no matter their vocal frequency.

Professor Bolton’s research sought to investigate these biases only in Opus, a codec used in common VoIP applications including Zoom, Skype, Microsoft Teams, Discord, and many others (Bolton, 2024). Bolton’s 2024 study provides the basis for our capstone project, as it found significant differences between the results for males and females in six measures, all of which will be included in our list of measures so that we can accept or reject his findings. For our

research, we are continuing to investigate Opus, as well as other codecs, including AMR and Codec-2 to see if they also reflect sex biases. I will be working with my teammates and Professor Bolton to analyze data from the Speech Accent Archive in tandem with Opus-6, AMR narrow-band, and Codec-2 against measures including but not limited to: Perceptual Evaluation of Speech Quality, Virtual Speech Quality Objective Listener, Signal to Noise and Distortion Ration, and Short-term Objective Intelligibility to compare to the results of previous research (Bolton, 2024). Our team will be running the audio files from our dataset through a MATLAB script we have developed that outputs quantitative values for these measures that can be used for statistical analysis. Unlike Bolton (2024), who only used the minimal quality settings of Opus, we will choose a range of settings for each of our codecs so that we can compare the differences across voices and settings. As a team, we hope to pinpoint specific vocal frequencies and at which settings in these codecs that cause this proposed sex bias in VoIP.

STS Topic

My STS research focuses on the problem of biases in speech-to-text technologies, due to inadequate training data sets. Speech-to-text is a common feature now integrated in phones, computers, tablets, and even cars, but relies on a diverse set of training data for peak effectiveness. In order to be indicative of society at large, the training data fed to the algorithm has to equally represent all demographics. If this is not the case, then the algorithm is considered to have under-representation bias, which "is the disproportionately low representation of a specific group" (Sun et al., 2019, 2). Sun and co-authors cite one of my other sources (Tatman, 2017), which found that "Automatic speech detection works better with male voices than female voices" (Sun et al., 2019, 2). Tatman (2017) analyzed the WER for males and females of five different English dialects using Youtube's automated captions, and used a linear mixed-effects

regression model to backup her claims. Training data sets often underrepresent the voices of women, those with speech impediments, and those with unique regional accents. As a result, there tends to be more inaccuracy in the translation for these groups. One study even found that an average word error rate (WER) was almost double for black speakers compared to white (Ngueajio & Washington, 2022). This study is critical for my research as black people are both a minority as well as a marginalized due to systematic racism and U.S. social constructs. Therefore, a discrepancy in WER could further isolate minorities and marginalized groups from using these technologies. Additionally, those with disabilities that require them to utilize speech-to-text to communicate or rely on automated captions generated by this technology should be able to without the fear of their message becoming misconstrued. To make speech-to-text more inclusive and ethical, it is vital to look at the creation and sourcing of training data sets, to uncover biases that are being fed into the algorithms.

In order to investigate this problem more, I will analyze studies carried out by scholars that use metrics such as WER, dialect density measure (DDM), and success rate of automated captions. One group studied biases in British-English varieties by comparing measures such as speech rate, social status, and repertoire (Hung et al., 2022). The authors results showed that speech rate seemed to have the biggest impact on proper translation, but this was not consistent across all tested speakers. Hung et al. (2022) claimed that factors such as the accents of marginalized communities are not commonly included in training data sets, leading to a significant impact on speech-to-text efficiency. The authors study was peer-reviewed, demonstrating credibility.

Aftering analyzing these studies, I hope to find the specific measures that have the highest error rates and see which groups this effects. I will then review case studies illustrating

the lack of inclusivity of speech-to-text software on these groups. I hope to find examples of raw data sets used to train these algorithms that I can dissect to see if the problem truly is underrepresentation of these groups, or rather a flaw in the algorithm itself.

Through my analysis, I plan to explore the flaws in the construction of training data sets from the relational view theoretical framework, which claims that "the power to represent and thus document specific aspects of the world is not intrinsic to data in and of themselves, but rather derives from situated ways in which the data are handled" (Leonelli, 2020, 15). Essentially, this framework claims that data and evidence only make sense within the context of how they were made and used. I am choosing to frame my problem in this way as we can only analyze the accuracy of a speech-to-text model if we know the source and categorization of the data, as underrepresentative data should be treated as such.

For my argument, I plan to take an ethical approach to defend the need for inclusive training data in speech-to-text technology. Speech-to-text technology can be used to create automated captions which are useful for people with disabilities or those learning a new language. One study (Parton, 2016), which looked at mistakes made by Youtube's automated captions found an error rate of 7.7 phrase errors per minute. This could significantly alter the content of the video, and decrease comprehension for those that rely on these captions. Specifically in acadamia, there are laws that require insitiutions from denying disabled individuals any benefits (Parton, 2016). With an error rate as high as 7.7 phrases per minute, this inaccuracy is more than distracting; it is a barrier to communication and unethical.

Conclusion

Within both my technical and STS papers, I hope to uncover a greater understanding of why biases exist within audio technologies, and how we can tweak the process of gathering

training data to produce better results. My technical research will focus on collecting data that results from both female and male voices being processed through popular codecs and analyzing the results to see if sex-biases do exist. If my team and I can pinpoint at what settings of each codec these biases appear, we can recommend the optimal usage to avoid any discrepancies between sex. My STS research will focus on the problem of biases in speech-to-text technologies that result from faulty data sets. These data sets often underrepresent minorities and marginalized groups, hence resulting in more errors and less efficiency when used by a wider audience. I hope to identify common patterns in these data sets along with possible mitigation strategies to help improve minorities' usage of speech-to-text features to promote inclusivity and equality in the communication sphere.

References

Bolton, M. L. (2022). Preliminary Evidence of Sexual Bias in Voice over Internet Protocol Audio Compression. In M. Kurosu (Ed.), *Human-Computer Interaction. Technological Innovation* (Vol. 13303, pp. 227–237). Springer International Publishing. https://doi.org/10.1007/978-3-031-05409-9_17

Bolton, M. L. (2024). *HCC: Small: Sounding Fair: Investigating Voice Frequency Distribution Bias in VoIP Codecs* [Unpublished grant proposal]. National Science Foundation.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., & Goldwater, S. (2019). *Pre-training on high-resource speech recognition improves low-resource speech-to-text translation* (No. arXiv:1809.01431). arXiv. http://arxiv.org/abs/1809.01431

de Almeida, André Luís Monforte,Neves Azenha. (2021). *Impact of Vocal Traits Distribution on Speech Applications' Performance and Bias* (Order No. 30852902). Phd Dissertation. *ProQuest*. https://www.proquest.com/openview/d21c5fe27cfb4763159187f0f4a7f288/1?pq-origsite=gscholar&cbl=2026366&diss=y.

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). *Quantifying Bias in Automatic Speech Recognition* (No. arXiv:2103.15122). arXiv. http://arxiv.org/abs/2103.15122

Hung, K., Cardoso, A., Sharma, D., & Levon, E. (2023). Biases and Speech-to-Text Efficacy for British English Varities. *International Congress of Phonetic Sciences.*

Leonelli, S. (2020). Learning from Data Journeys. In *Data Journeys in the Sciences*, eds. Leonelli, S., & Tempini, N. Springer (Berlin).

Ngueajio, M. K., & Washington, G. (2022). Review of *Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review*. *Human-Computer Interaction 2022 – Late Breaking Papers*, 421–440.

Parton, B. (2016). Video Captions for Online Courses: Do YouTube's Auto-generated Captions Meet Deaf Students' Needs?. *Journal of Open, Flexible, and Distance Learning*, *20*(1), 8-18. Distance Education Association of New Zealand. Retrieved October 28, 2024 from https://www.learntechlib.org/p/174235/.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640. https://doi.org/10.18653/v1/P19-1159

Tatman, R., & Kasten, C. (2017). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. *Interspeech 2017*, 934–938. https://doi.org/10.21437/Interspeech.2017-1746

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. *Proceedings of the First Association for Computational Linguistics Workshop on Ethics in Natural Language Processing*, 53–59. https://doi.org/10.18653/v1/W17-1606