From Voice to Data: How Lack of Transparency Shapes Speech Recognition Performance

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Madison Sullivan

Spring 2025

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

Advisor

Caitlin D. Wylie, Department of Engineering and Society

Introduction

In 2012, a jury awarded $140 million in damages to the family of a patient who died due to an excessively high dosage of insulin that was a result of an improperly dictated discharge summary using speech recognition software (The Joint Commission, 2022). Speech recognition (SR), "a digital dictation system combined with a mathematical model of word recognition" (Poder et al., 2018, p. 1), has been "shown to struggle with speech variance due to gender, age, speech impairment, race, and accents" (Feng et al., 2021, p.1), demonstrating that SR is not being designed to consider all use cases and user profiles. This showcases a lack of inclusivity in the communication sphere and leads to miscommunication and errors, some of which have been as severe as the one described above. The relational view theory (Leonelli, 2020) states that we need to understand the "data journey" which describes the context and process of how data is collected and prepared in order to understand how it can be used. I use this theoretical framework in my analysis to highlight the importance of data transparency when developing and utilizing SR algorithms. In this paper, I investigate why there are varying success rates using these technologies by exploring the composition, utilization, and evaluation of SR software. I will be looking at training data and its selection process for composition, case studies showing examples of errors for utilization, and metrics used to measure success for evaluation.

Composition

The past few decades have seen a significant increase in the use of SR software in the professional world, with applications ranging from healthcare, customer service, and education. More often than not, your call to customer service is not answered by a human being but rather by an automated voice recognition system that fails to connect you to the right department. Or you're at the doctor's office and the physician is sticking a device in your face to record and

transcribe your entire conversation. While these advancements seem great for productivity and cost savings, SR is not up to par in terms of accuracy. In a traditional setting, a doctor would record a conversation with a patient, pass it on to a transcriber, the transcriber would write down the conversation as they heard it, and then the doctor would review it for accuracy. With the implementation of SR software, the middle step is taken out, as the conversation is transcribed in the moment. However, using SR software with limited training data can result in a multitude of errors and mistranslations. This leads to the doctor having to go back and review the transcription for issues and make adjustments, negating the time-saving benefit of SR in the first place.

The composition of SR software is not just designing the algorithm itself, but also the selection of data used to train the model and the person making that selection. It is crucial that the person or people designing the SR software consider how it will be used and who will be using it. This requires careful deliberation of use cases and diversification of data so that all potential parties are represented and at the correct proportions. Additionally, sharing the data or metadata used to construct the SR software, as Leonelli (2020) suggests, can optimize the performance of the software. For example, when building a model that will be used globally, it is important to include training data from across the globe to encompass different accents, ages, genders, and socioeconomic backgrounds. However, if you are designing a model to be used in the medical field, there should be plenty of medical terminology included in the training data and that fact should be explicitly stated so that the database can be used properly. Google's SR software, the basis for generating YouTube's automatic captions, has proprietary training data, which goes against the relational view theory's concept of a "data journey" (Leonelli, 2020). Without knowledge of how the data behind this SR software was created and processed, one can

not use it on the right demographics. If it was known that an algorithm used training data that only contained 10% female data, it would not be appropriate to use it on a group that is majority female. It is not feasible to assume that all training data will have an equal representation of all demographics, but it is important to understand the composition of the data so that it can be accurately used. Specific use cases of the SR being developed need to be established before gathering training data so that it can be selected to meet the needs of all users.

Another example comes from Dragon Medical Software by Nuance Communications Inc., which was a popular SR software used in the medical field despite having been found to exhibit errors that limit interpretation and understanding. Today, Nuance Communications Inc. has been purchased by Microsoft and Dragon Medical Software works in tandem with Dragon Ambient eXperience (DAX), an AI-powered tool by Microsoft to help with medical documentation and physician productivity. The training data used to build Dragon Medical Software is proprietary, but Microsoft (2024) has released that DAX has been trained on a "rich clinical data set anchored in more than 1B+ minutes of medical dictation annually and 15M+ ambient encounters" (p. 2). This supports Leonelli's (2020) relational view theory as Microsoft shares some metadata about the composition of the algorithm so that it can be properly used. They also include a vast collection of data overall, which can improve documentation and translations, because the greater the amount of data fed into the model, the greater the diversity and recognition ability.

There are some other speech data sets that are open-source, which support the relational view theory concept of a "data journey" but still showcase flawed training data. Databases such as Librevox, TIMIT, Switchboard, Numbers corpus, and the AMI meeting corpus showcase gender imbalances or a lack of demographic information in general (Tatman, 2017). For

example, TIMIT is a popular open-source speech corpus used for the "development and evaluation of automatic speech recognition systems" (Garofolo et al., 1993, Introduction), but it is composed of approximately 70% male data and 30% female data. However, TIMIT does not explain why this is the case or how this information should be used. Since the documentation states that the database is used for development and evaluation, it suggests that females will likely face more mistranslations if using a model trained on this data since they are underrepresented. This example highlights that although it is important for a database to be transparent, transparency alone does not fix underlying discrimination, such as the gender imbalance in this case. To improve inclusivity, the creators of the database should strive to balance the gender distribution or, if this was a purposeful choice, provide rationale so that designers of SR software can know if this database is appropriate for their use cases.

Similar to my STS research, my capstone project is investigating biases in transmitted voice over the internet, a different sector of communication. Previous studies have found that females are harder to understand than males, demonstrating another example of a lack of inclusivity in the communications sphere. This was inspired by a personal experience of my advisor who misunderstood women more often than men when connecting with colleagues over Zoom. When researching data that could be used for our project, my capstone team was alarmed with how difficult it was to find a highly diverse, equally representative, open-source database. For the purposes of our study, we were looking at differences between sexes. Hence, despite having an approximately equal representation of males and females in our chosen data set (Weinberger, 2025), there are 229 native languages represented unequally. Out of 3,031 speakers, there are 4 languages that each have over 100 native speakers. These include English, with 658, Spanish, with 242, Arabic, with 201, and Mandarin, with 157. There are also only 5

languages that have between 50-100 native speakers, and 86 languages that have only 1 native speaker. This shows the imbalance of the database by language and the inherent biases in SR software that uses this for training purposes. To become more ethical, the speech corpus should gather data from more speakers of each language to improve inclusivity or disclose the current metadata to promote proper utility and transparency.

Utilization

Even common utilizations of SR, such as YouTube automatic captions, have been prone to errors, demonstrating that those using SR need to be cautious with their choice of software and how they are using it. SR software has helped increase translation efficiency and decrease the time spent on tedious tasks, but there remain many legal and ethical concerns behind their use. Both deaf people and those learning a new language rely on video captions to understand the displayed content. When these captions are generated automatically through SR software, they must be properly translated to avoid errors and misinterpretation. One common example is YouTube's automatic captions, which were released in 2009 and use SR software to generate captions for uploaded videos. However, there are many errors reported by those who use them, to the point where Google, who owns YouTube, released a statement saying, “automatic captions might misrepresent the spoken content due to mispronunciations, accents, dialects, or background noise” (YouTube Help, 2025, Note). Coming from the owners of the feature, this statement infers that there have been many complaints filed about YouTube’s automatic captions and that they may be trying to avoid liabilities. Mispronunciations and background noise may not be the fault of YouTube, but ranging accents and dialects can be accounted for in the training database chosen for the software. Additionally, including a “data journey”, as supported by the

relational view theory, to document how the algorithm was created and should be used could help to ensure alignment and transparency between the software and the user.

Minimizing errors in automatic captions is important as they are often used in the higher education setting. Both online classes and those who offer lecture captioning for review and attendance substitution take advantage of automatic captioning to save the professor time on transcription. According to Anastasopoulos and Baer (2013), "an institution's communications with persons with disabilities must be as effective as the institution's communications with others" (para. 3) to be in compliance with the U.S. Department of Education's Office for Civil Rights policy. This is backed by Section 504 of the Rehabilitation Act 1973 and Title II of the Americans with Disabilities Act 1990. These two laws state that a disabled person cannot be denied or excluded benefits from a college or the services it offers (Parton, 2016). However, neither explicitly state rules regarding accuracy in captions, or using SR software to increase inclusivity. I think these laws need to be updated to include specific accuracy rates or a number of errors that are allowable in SR software used by federally-funded programs to reflect the increased use of technology by the disabled. Without this, professors have no measure to assess whether their students who are deaf or hearing-impaired have equal learning opportunities as their other students.

An example of a case where the accuracy rate of automatic captions was not considered comes from Parton (2016), who looked at the accuracy of YouTube's automatic captions on videos created by a professor for the online courses he taught. The study analyzed 68 minutes of video in 21 segments and found a total of 525 phrase errors (not including grammatical errors, misspelled words, and minor word changes) (Parton, 2016). This comes out to an average of 7.7 phrase errors per minute, which can completely alter the message of the video and the

individual's comprehension. I acknowledge that only 68 minutes of video is not enough for a full comprehensive study, and that only looking at captions generated by one person is not a precise reflection of the accuracy rate of YouTube's automatic captions holistically. This study could be improved with a larger sample size of both audio content and speaker diversity. However, this high error rate would limit accessibility and inclusivity for students with disabilities in the class. The captions would likely lead to misunderstandings and difficulties with comprehension, meaning that those who rely on the captions are not offered the same access to learning benefits as those without a disability, and showcase that YouTube's SR software may not be up to par with legal standards.

In addition to having many errors in general, YouTube's automatic captions have been proven to exhibit biases against certain genders and dialects. Tatman (2017) looked at the word error rate (WER) for males and females across five different dialects of native English speakers when using YouTube's automatic captions. The data analyzed came from individuals participating in the accent tag, a popular internet trend where an individual uploaded a video introducing themselves and their linguistic background, followed by reading a list of words designed to elicit dialect differences (Tatman, 2017). The results found statistical differences between groups for both gender and dialect, with women and those from Scotland performing worse than their counterparts. A limitation of this study is the small sample size (8 men and 8 women from each of the five dialects) as well as the isolation of spoken words rather than integration within a sentence to help improve SR ability. If YouTube were to release metadata on its SR algorithm and the data used to build it, this could potentially explain the varying success rates by gender and dialect. For example, if the training data consists of mostly male data, or if the algorithm was only trained on sentences of many words strung together rather than in

isolation, these results would make sense. However, there is a clear lack of transparency regarding the composition of the algorithm, which in turn affects the utilization process.

Despite the limitations of this study, it is important to highlight that all participants are native English speakers and that the algorithm is producing statistically significant results solely based on the individual's gender and dialect. English is spoken in 186 countries around the world (International Center for Language Studies, 2024), so it is vital that a SR software designed for English speakers is based on global data. Based on Tatman's (2017) results, those from California performed the best, with those from New England not far behind. This suggests that most of the training data used for YouTube's automatic captions at the time of data collection came from individuals in those areas of the United States. This could be an example of convenience bias as California is the state with the largest population in the United States according to the 2020 census (Tikkanen, 2024).

SR software mistakes are detrimental in all contexts, but they are especially severe in the medical field when a small error could lead to a tragic mistake. According to Poder and coauthors (2018), major errors in transcription for medical reporting occurred three times more when using SR compared to a human transcriber. This results in a need to review the notes and recall the correct context. In Poder and coauthors' systematic review of systematic reviews (2018), they found that "the error rate with SR was 0.05 to 6.66 errors per report, compared with the error rate with a transcriptionist of 0.02 to 0.4" (p. 4). Poder et al. share a compelling statistic, as having almost 7 errors within one report could change the message entirely, or be completely misinterpreted by someone who does not know the context. For example, Goss and coauthors (2016) shared a case study detailing the mishap that occurred when the SR software interpreted the word "period" for a ".". A female had come to the emergency department with an issue

regarding her arm and mentioned that she had missed her last period. SR transcribed this as ".", and when the woman came back the following day and saw a different doctor, they prescribed her an antibiotic that is advised not to take when pregnant. This is just one example of how a small mistake can lead to a serious translation issue if not caught. Including more data in the model so the software can decipher between "period" and "." depending on the context will reduce the likelihood of this error.

Despite its wide use in the medical field, Dragon Medical Software has been described in many case studies as producing errors affecting communication and understanding, as mentioned earlier. This software was used in the study carried out by Goss and coauthors (2016), and was also one of the softwares studied in the systematic review done by Poder and coauthors (2018). Despite being so widely used, Goss and coauthors (2016) found that in their study of 100 dictated notes, "71% of the notes contained errors. There were 128 errors in total or 1.3 errors per note" (p. 4). Articulation errors were the most commonly found, which could be improved by both the user and the software if trained on a more diverse set of data and pronunciations. After articulation, deletions and additions were the next most common errors, which are direct results of the software used. Deleting and adding words can change the overall message of the note, which if passed around to multiple individuals, can lead to a potentially detrimental communication issue similar to the "period" example. Goss and coauthors (2016) were published by the National Institutes of Health which shows that they uphold scientific integrity and are fact-checked by experts in the field.

Evaluation

Evaluating and iterating on SR algorithms is just as important, if not more, than the composition and utilization steps since there is always room for improvement. Additionally, this

step involves both the makers and users of SR software. The makers need to evaluate and test the model they developed, while users need to report issues and errors with the technology for improvement. One method for improvement, specifically when developing SR software for a low-resource language, is pre-training the model on a high-resource automatic speech recognition (ASR) system. Bansal and coauthors (2018) conducted a study that showed if you are developing a model to be used for a language with limited available training data, you can add in data from another language regardless if it is the source or target of the SR algorithm. SR with a different input and output language requires the transcribed audio to show the model the correct results of the spoken input. However, languages that lack this written documentation and corresponding translation are hence difficult to model. Basal and coauthors (2018) found that "the main benefit of pre-training arises from the transfer of the encoder parameters, which model the input acoustic signal" (p. 2) showcasing that the actual spoken language does not matter, but rather feeding a voice into the model. Pre-training with high-resource data allows the model to normalize the differences in speaker and channel variability better, leading to higher accuracy rates after it is fine-tuned with the low-resource language (Bansal et al., 2018). Based on this concept, it is unreasonable for the developers of SR algorithms to claim that the flaws in their model stem from a lack of accessible data since there exist many open-source speech corpora that can be used to pre-train the model for improved performance, and this can be documented in the "data journey" (Leonelli, 2020) to offer transparency.

While conducting my research, I came across multiple different metrics that were used when accessing SR accuracy and performance. Some of these included the Bilingual Evaluation Understudy (BLEU) score, word error rate (WER), dialect density measure (DDM), and out-of-vocabulary (OOV) words. Bansal and coauthors (2018) assessed the improvements of pre-

training based on the BLEU score, which takes into account word choice and word order when assessing the accuracy of the generated text compared to the reference translation (Zhang et al., 2004). The study by Bansal and coauthors (2018) found that if they pre-trained on a high-resource ASR and fine-tuned with only 5 hours of the low-resource language, the model produced a BLEU score of 9.1, compared to a similar score of 10.8 from a model trained on 20 hours of just a high-resource language. This shows that improving the BLEU score of a model with a low-resource language is relatively simple. There are plenty of large English ASR models that can be used to fine-tune less popular languages to improve inclusivity and performance. Although Bansal and coauthors' study was published on ArXiv, it was accepted at the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics, which is a peer-reviewed conference that requires multiple experts in the field to review papers before they are accepted for publication, demonstrating credibility.

Besides BLEU, other metrics used in the industry are WER, DDM, and OOV. WER is a calculation based on the number of deletions, insertions, and substitutions of words based on the total number of words in the transcript (Hung et al., 2023) showing that a lower WER means a higher performance. DDM is used to calculate the presence of accent features and OOV are words that were not in the dataset that trained the model, which leads to errors as they are not recognized (Hung et al., 2023). Throughout my research, the studies I have encountered have used varying combinations of these metrics to measure their success rates, which makes it difficult to compare algorithms and studies when looking for common trends. I propose always having one universal metric used in all SR software evaluations, such as a simple WER, and others can be included if needed.

Conclusion

It is important to strive for inclusivity and provide full transparency throughout the development and utilization of SR software to provide similar success rates for the designated users. Without a clear "data journey" it is unclear whether the software is appropriate for the chosen task or meets the criteria for the targeted audience. Examples from both the educational and medical settings showcase how developers of these algorithms need to consider all possible use cases and tailor the training data to capture these, and users need to be cognizant that the software they are using may not be appropriately constructed for their use case. Expanding the breadth and depth of training data will allow SR algorithms to minimize errors and decrease fatal flaws. Since much of the training data used to develop popular SR algorithms is proprietary, I was limited in the ability to conduct my own study to measure success rates when varying demographic factors. Future work could hone in on a specific demographic factor such as gender or native language by conducting a study to test how success rates differ based on the distribution of that demographic factor in the training data. Additionally, it would be interesting to do an updated version of Tatman's (2017) accent tag study to see if YouTube's automatic captions have improved over the years by including more training data and experience.

References

Anastasopoulos N., Baer A. M. (2013). *MOOCs: When Opening Doors to Education, Institutions Must Ensure that People with Disabilities Have Equal Access.* New England Board of Higher Education. https://nebhe.org/journal/moocs-when-opening-the-door-to-education-institutions-must-ensure-that-participants-with-disabilities-have-equal-access/

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

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

Garofolo, J., Lamel L. F., Fisher W. M., Fiscus J. G., Pallett D. S., Dahlgren N. L., Zue V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium. https://doi.org/10.35111/17gk-bn40

Goss, F. R., Zhou, L., & Weiner, S. G. (2016). Incidence of speech recognition errors in the emergency department. *International Journal of Medical Informatics*, *93*, 70–73. https://doi.org/10.1016/j.ijmedinf.2016.05.005

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

International Center for Language Studies. (2024). Most spoken languages in the world. https://www.icls.edu/blog/most-spoken-languages-in-the-world

The Joint Commission (2022). Speech recognition technology translates to patient risk. *Quick Safety,12*. https://www.jointcommission.org/-/media/tjc/newsletters/quick-safety-12-update-5-3-22.pdf

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

Microsoft. (2024). Automatically document care with DAX Copilot. https://www.nuance.com/asset/en_us/collateral/healthcare/data-sheet/ds-ambient-clinical-intelligence-en-us.pdf

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. https://www.learntechlib.org/p/174235/

Poder, T. G., Fisette, J.-F., & Déry, V. (2018). Speech Recognition for Medical Dictation: Overview in Quebec and Systematic Review. *Journal of Medical Systems*, *42*(89). https://doi.org/10.1007/s10916-018-0947-0

Tatman, R. S., (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

Tikkanen, A. (2024). list of U.S. states by population. Encyclopedia Britannica. https://www.britannica.com/topic/largest-U-S-state-by-population

Weinberger, S.H. (2025). Speech accent archive. George Mason University. http://accent.gmu.edu

YouTube Help. (2025). *Use Automatic Captioning.* https://support.google.com/youtube/answer/6373554?hl=en

Zhang, Y., Vogel, S., & Waibel, A. (2004). Interpreting BLEU/NIST Scores: How Much Improvement Do We Need to Have a Better System? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/pdf/755.pdf