"Examining the Impact of Dialectal Variation on Speech-to-Text Algorithm Fairness."

Abhinav Akkiraju
Carnegie Mellon University

Introduction

History of Speech Recognition Systems

The first speech recognition system was invented in 1952 by Bell Laboratories, called Audrey. Taking up the space of a 6-foot-high rack, Audrey could recognize spoken numbers from one to nine with nearly 90% accuracy. This high accuracy (for the time) could only be achieved when the numbers were uttered by the system's inventor, HK David, foreshadowing the major conflicts that have succeeded the use of an exclusive dataset in training speech recognition algorithms in today's world (Pieraccini, slide 3). When other speakers used Audrey, the accuracy would drop sharply, with the potential to increase over time as the speakers continued to use Audrey and let it adjust to their voice and speaking traits, habits, and characteristics.

Following Audrey, IBM created a machine called the Shoebox that could recognize 16 spoken English words. Shoebox was physically more compact than Audrey, fitting on a desk rather than requiring a full equipment rack, though it still relied on external computational resources. It could recognize all the digits between zero to ten, as well as six words including "plus," "minus," and "total." Shoebox took the recognized words and split them into phonemes–the smallest units of sound that distinguish meaning between words in a language (e.g., /p/ and /b/ distinguish 'pat' from 'bat'). It then worked to understand the phonemes before combining them back into full words. It relied on two models that could recognize the phonemes and string the phonemes back into words, respectively (Reid). William C. Dersch, one of the main engineers for Shoebox, believed that human responses to speech patterns like pitch should not be considered in the approach for machines. Rather, he took a novel approach to word recognition that involved recognizing words based on patterns that the human ear doesn't often recognize. Shoebox used proprietary terminology, classifying voiced sounds originating in the vocal tract as 'machine vowels and fricative sounds as 'machine consonants ("Speech Recognition").

In the same decade, Soviet researchers created an algorithm, based on pre-recorded speech, that could recognize more than 200 words. Shortly

after this, the development of these algorithms proliferated across the world. In 1972, Carnegie Mellon University, funded by the US Department of Defense, invented a device called Harpy that could recognize not just individual words, but entire sentences. By 1976, Harpy had developed an extensive vocabulary of over 1000 words (The Harpy Speech Recognition System). By the 1980s and 90s, speech recognition algorithms could detect and recognize over 20,000 words. From IBM's new voice-activated typewriter called Tangora to Dr. James Baker's The Dragon Dictate, which was the first consumer-grade text-to-speech product, these algorithms continued to grow stronger with an increasing reliance and number of users.

Modern Conflicts with this Technology

Dialectal variation refers to systematic differences in pronunciation, grammar, and vocabulary associated with social, regional, and cultural factors within a language community. These variations are socially constructed rather than biologically or environmentally determined. Dialectal variation reflects our world's linguistic diversity but has also slowed speech-to-text development. The influence of variation in dialects on the fairness of speech-to-text algorithms is a highly nuanced matter that has garnered notable attention in recent years. The rapid development and reliance of speech recognition technology have caused these algorithms to be used at an increasingly rapid rate across commercial, medical, and accessibility contexts. These applications have a wide range, from accessibility technology for disabled individuals to multimedia content transcription services. Despite their growing importance, the existence of dialectal variation continues to pose challenges for these algorithms. Dialects reflect our world's linguistic diversity, highlighting the many human variables: geography, culture, education, social standing, etc. For speech-to-text algorithms, which are commonly provided with training data that only partially reflects the complete range of existing dialects, this diversity can lead to considerable disparities in factors such as pronunciation, vocabulary, and grammar.

The implications of this "human barrier" are extensive and have notable effects on the fairness and accuracy of speech-to-text algorithms. However, it is important to note that the barrier is not linguistic variation itself, but the mismatch between linguistic diversity and the narrow distributions represented in ASR training data. For example, erroneous transcriptions in voice recognition systems used by disabled individuals might impair these people's capacity to communicate effectively, which could lead to physical and/or mental harm. The impairment of certain groups' ability to communicate is crucial to note because it presents an issue of human rights due to inequitable access to these technologies. Exclusive access to fair speech-to-text technologies will lead to certain groups and demographics falling even more disadvantaged as these technologies progress and grow in reliance. Inaccurate transcriptions for

multimedia content can lead to the loss of crucial information and negatively affect the content's accuracy and completeness. This will have negative implications in applications of these technologies, such as automated captioning, virtual assistance, court transcription, and speech recognition in operating rooms.

Furthermore, dialectal diversity influences speech-to-text algorithms far beyond just reading accuracy. It also significantly impacts how inclusive and fair these algorithms are. In contrast to people who speak dialects that are more well-represented, immigrants and members of minority groups who speak dialects that are under-represented in the training data used by speech-to-text algorithms are likely to experience more biases in these algorithms. Marginalized groups who use these technologies will only become more marginalized as these technologies' prevalence and reliance grow. This divide highlights the need for further research into how dialectal variance affects speech-to-text algorithms and the requirement for the creation of algorithms that are inclusive and respectful of everyone, no matter their background or dialect.

In this regard, researchers and practitioners alike must take dialectal variation into much more consideration when developing speech-to-text algorithms. A multidisciplinary approach that incorporates knowledge from areas like linguistics, speech recognition, and machine learning will be necessary to address these concerns. It will also necessitate a dedication to designing algorithms that are inclusive and equitable, regardless of their background or dialect. Only by adopting this strategy will it be possible to develop voice-to-text algorithms that are respectful of all people and are capable of reliably and accurately transcribing speech.

Literature Review

The proliferation and increasing dependence on speech-to-text algorithms necessitate the examination of dialectal variation and how it affects the fairness of speech-to-text algorithms. This literature review evaluates the existing research that has been conducted on this matter, the theoretical and methodological foundations of those studies, and the implications of their findings for the development of fair and accurate speech-to-text algorithms that serve a more diverse range of communities.

The concern about the algorithms used by speech recognition technology perpetuating inaccuracies due to variation in dialects has been an issue for many individuals. Various studies have highlighted that these algorithms may not perform equally well for dialects, including African American Vernacular English (AAVE), which is often underrepresented in the training data. This can lead to lower accuracy and higher error rates when processing speech from these dialects. The underrepresentation of dialects in training data can be attributed to factors such as limited availability of data, insufficient resources for data annotation, and biases in data collection. Furthermore, these dialects possess characteristics and phonological variations that are not adequately accounted for by

conventional models. AAVE, for example, possesses linguistic features–such as copula absence, the habitual "be", and familiar markers–that are uncommon or nonexistent in other dialects (Lauture 169). Additionally, regional differences result in variations of English, such as British English, Australian English, and American English.

One study concluded that speech-to-text algorithms trained on Standard American English exhibited lower accuracy and higher error rates when processing speech from AAVE speakers, particularly those who were classified as older adults (Dorn 18). This emphasizes the need to train speech-to-text algorithms on a multitude of dialects and demographic groupings to ensure fairness. Age, gender, and education level are among demographic factors that have an impact on speech patterns, which in turn regulate the performance of speech-to-text algorithms. Additionally, dialectal variation can be influenced by socio-economic factors, such as class and occupation, which affect the way people speak and are spoken to.

Bringing the effects of gender to light, previous research has concluded that female speakers have better average recognition results than male speakers. In an experiment conducted by Decker et al, the researchers found that, in conversational telephone speech (CTS), male English speakers had a WER of 15.7% while female English speakers had a WER of 13.7%. Additionally, male French speakers had a WER of 45.2% while female French speakers had a WER of 37.9%. They claimed that a possible reason for this drastic difference in word error rate could be the traditional role of women in language acquisition and education. Regardless, these results highlight the need to consider gender and gender roles in certain cultures/societies when developing equitable speech-to-text algorithms (Decker 2205).

Numerous studies have suggested employing dialectal feature extraction and dialectal language models to address these issues. These methods enhance the performance of speech from various dialects by enabling the depiction of dialectal variance in speech-to-text algorithms (Nigmatulina 18). Furthermore, several studies have recommended adopting active learning techniques to broaden the diversity of training data, leading to more equitable and accurate speech-to-text algorithms (Riccardi 1). Dialectal variances in speech data can be recognized with the use of active learning. The model's performance on various dialectal variants within a language may also be assessed using these criteria.

Existing literature in the field has demonstrated the cruciality of considering dialectal variance in the development of speech-to-text algorithms. Recent work has also documented systematic ASR performance disparities across low-resource languages and diverse accents, including evaluations of large-scale systems such as OpenAI's Whisper (Graham & Roll, 2024; Nakatumba-Nabende et al., 2025). Studies have indicated that dialectal variance can impact how fair and accurate these algorithms are, especially for speakers from dialects and

underrepresented demographic groups. Research has suggested the use of techniques including dialectal feature extraction, dialectal language models, active learning, and dialect-aware evaluation metrics to address these issues. To fully comprehend the effects of dialectal variation on speech-to-text algorithms and to develop more equitable and accurate algorithms for a wide range of dialects, particularly in under-resourced languages and demographic groups, additional study is nonetheless necessary. Research is furthermore required to understand the ethical implications of these models.

Methodology

Qualitative and Quantitative Data

My research methodology for examining the impact of dialectal variation on speech-to-text algorithm fairness involves a multifaceted approach that incorporates both qualitative and quantitative elements and analyses. Algorithmic fairness will be operationalized as a relative transcription accuracy across dialect groups rather than a comprehensive ethical or legal standard.

Firstly, a thorough literature review will be conducted to gather existing research on dialectal variance in speech and how it influences speech-to-text algorithms. In addition to highlighting the topic's relevance, this literature review will search for any gaps in the body of current knowledge and, most significantly, lay the theoretical groundwork for the study.

Secondly, a comprehensive quantitative investigation will be conducted to gather data regarding how well speech-to-text algorithms function on a wide range of dialectal variances. Samples from various speaking cohorts will be used in the study, and each speaking cohort's participants will use identical phrases but in their unique dialects. Using ElevateAI's automated speech recognition services, the audio recordings will be transcribed, and the results will be compared to the original phrases to determine the accuracy of the algorithms on each dialect using the following factor: word error rate. Word error rate (WER) was calculated using the standard formula: WER = (substitutions + deletions + insertions) / total number of words in the reference transcription. ElevateAI was selected as a representative commercial speech-to-text system due to its accessibility and use of contemporary neural ASR architectures. While it is not identical to systems developed by Google, Apple, or Amazon, it serves as a proxy for modern, data-driven ASR pipelines.

Speech Selection

These speech samples will be drawn from the Speech Accent Archive, which features a comprehensive collection of vocalic renditions from globally dispersed linguistic communities. Speakers were categorized by dialect based on the Speech Accent Archive's metadata, which identifies

speakers by region of origin and linguistic background. 20 speech samples were chosen from each linguistic background that was analyzed in the study. Given the limited sample size, this study focuses on descriptive differences in WER rather than formal statistical significance testing. Future work with larger datasets would allow for inferential analyses such as ANOVA or mixed-effects modeling. The subjects of the speech samples were tasked with saying the following phrase: "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 on Wednesday at the train station." An equivalent number of speech samples will be drawn from each of the following five dialects: Australian English, Southern American English, Pacific Northwest English, New York English, and British English. The 'Please call Stella' passage is a phonetically balanced elicitation paragraph widely used in accent and dialect research (Jiao et al, 2019). It contains all English phonemes in a naturalistic context.

Analysis and Implications

Finally, a mixed-methods approach will be used to integrate and evaluate the data from the experiment and the literature review to identify patterns and trends in the impact of dialectal diversity on the fairness of the speech-to-text algorithm. This will involve the analysis of distinct linguistic traits found in dialects and of the solutions and concepts suggested by prior researchers. The findings of these analyses will be used to suggest solutions for the development of fairer, more diverse speech-to-text algorithms that can accurately transcribe a wide range of dialectal variations.

This methodology adopts an interdisciplinary approach that considers both the technical and human elements of the problem, allowing us to fully comprehend the effect of dialectal variation on speech-to-text algorithm fairness. Additionally, using a representative sample, consisting of five distinct dialects of the English language, and conducting a sizable quantitative study will improve the generalizability of the results to a wider population complete row without any missing values.

Results

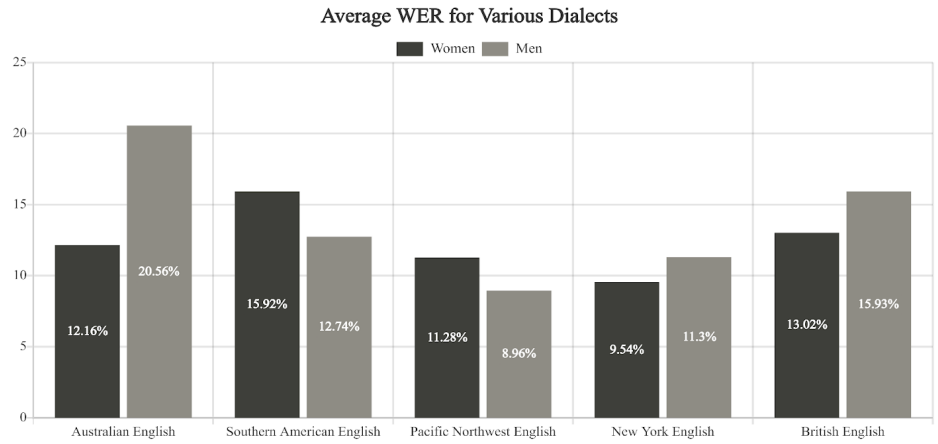

FIGURE 1. Average WER for Various Dialects

Geographical Bias

Figure 1 showcases the outcomes of my analysis. As depicted from the data, it is evident that the word error rates for Australian English, British English, and Southern American English are noticeably higher compared to Pacific Northwest English and New York English. This disparity in performance shows how speech-to-text algorithms are more likely to encounter difficulties in transcribing speech from underrepresented speaking cohorts, such as immigrant and minority groups.

It is crucial to note that Southern American English, despite being a dialect within the United States, also experiences a higher word error rate. This observation highlights the persistence of linguistic bias in technology even within a certain geographical area. This higher word error rate for Southern American English speakers could be attributed to the location where this model is being developed and/or the demand for speech-to-text technology in the southern region of the United States.

Demographic Bias

These findings also reveal that demographic factors like gender should be heavily considered in the development of speech recognition systems. Male speakers showed a WER difference of 11.6 percentage points between Pacific Northwest English (lower WER) and Australian English (higher WER), compared to only 0.88 percentage points for female speakers across the same dialects. This pronounced disparity between the word error rates of different genders implies that gender should also be considered when designing speech-to-text algorithms. This aligns with the findings of Decker et al, which were discussed previously in the literature

review, as female speakers had an average WER of 12.38 percentage points while male speakers had an average WER of 13.9 percentage points in my experiment.

Impact of Unique Linguistic Traits

The findings from my experiment also show the difficulties that arise when trying to transcribe words and phrases due to variations in language traits between dialects. Australian and British English, for instance, presented challenges for this algorithm in this regard. One common mistake was the conversion of the word "call" to "cool," as well as "peas" to "pace" and "meet her" to "made a."

This phenomenon can be attributed to the unique linguistic characteristics that are prevalent in these dialects, such as T-glottalization and non-rhoticity. T-glottalization refers to the replacement of the voiced dental plosive (also known as the "t" sound) with a glottal stop. Non-rhoticity refers to the absence of the post-vocalic "r" sound in certain words. These linguistic traits are not commonly found in other dialects, resulting in difficulties for speech-to-text algorithms that are trained on data from other regions.

Regional pronunciation (RP) refers to the “regionally neutral accent” that is spoken by middle-class speakers in the United Kingdom. The use of these linguistic traits mentioned above is affected by socioeconomic classes within the region where the dialect is spoken. T-glottalization, for example, is considered a linguistic invention by the middle-class in the UK (Barrera B. B. 33). It is primarily used in casual conversation by the middle-class, and its prevalence is dependent on socioeconomic factors such as education. The most elitist sectors of RP speakers show a greater hostility to linguistic movements such as T-glottalization, and these sectors are led by upper-class citizens “who have attended the major public (private) boarding schools” (Barrera B. B. 75). This highlights how the environment, in terms of socioeconomic factors, can affect the usage of certain linguistic traits, which can lead to the underrepresentation of unique dialects within a language.

One consistent error produced by the speech-to-text algorithm in the transcriptions for speech samples from New York English was the substitution of the phrase “meet her” with “meter.” Additionally, in the Southern American English speech samples, the algorithm frequently misinterpreted the word “frog” for “fraud.” These inaccuracies may be caused by the distinctive vowel sounds and pronunciation patterns that are typical of these dialects. For example, the common mistake found in the results of the speech samples from New York English can be attributed to the linguistic trait known as h-dropping. An h-drop occurs when the letter h at the beginning of a word is not pronounced when it is supposed to be. In the case of the transcription from "frog" to "fraud" in Southern American English, this inaccuracy can be due to the cot-caught merger,

which occurs when two vowel sounds, typically the "short o" (as in "cot") and the "short a" (as in "caught"), are pronounced the same way.

Implications of Results

The results of my experiment provide valuable insights into the current limitations of speech-to-text algorithms and the need for continued improvement in language modeling technology. The disparities in performance between various dialects and genders highlight the importance of creating inclusive and equitable language models that accurately transcribe speech from a diverse range of individuals. This is necessary in the prevention of underrepresented speaking cohorts being misunderstood in the growing number of real-life applications using speech recognition algorithms. While no universal WER threshold defines usability failure, prior ASR research suggests that even single-digit percentage increases in WER can significantly degrade user experience in accessibility and transcription-dependent contexts (Wang et al, 2003). This study, therefore, interprets relative differences as indicative of potential usability disparities rather than definitive failures. It is important to note that this study examines a very limited set of five English dialects using a single elicitation passage and one ASR system. As such, the findings should be interpreted as indicative rather than exhaustive. Broader claims about global dialectal variation or immigrant populations require larger, multi-system, multi-language analyses.

Conclusion

The findings from this study illustrate the significant challenges that dialectal variation can pose for speech-to-text algorithms. Most of these issues stem from the fact that these algorithms are trained on large amounts of data, which do not adequately represent the multitude of dialects spoken worldwide, even for just one language.

This is particularly problematic given the growing use of speech-to-text algorithms in real-world settings where accuracy and dependability are crucial. For instance, inaccurate transcriptions in voice recognition systems for people with disabilities might cause these people serious problems and hinder their ability to communicate effectively. Students with motor and cognitive skills impairments may experience difficulties trying to convey thoughts and ideas in written expression. In this regard, many have turned to speech recognition software as it allows them to focus less on mechanics and more on the development of text/idea generation. Currently, however, speech recognition technology isn’t accurate enough for this to be a foolproof solution to the issues caused by learning disabilities. Prior research has found that individuals with normal speech patterns have much more accurate results with speech recognition software than those with dysarthric speech (Hux 1). Additional research has shown that factors such as cognitive disabilities and voice problems

also lead to inaccuracies in the results of speech recognition algorithms (Koester H. H 1).

Another example is how these exact transcriptions can lead to multimedia content experiencing a loss of crucial information and depreciation of validity and comprehensiveness. For example, speech recognition technology is increasingly being used in nursing and the documentation of nursing reports. While this technology is faster and more efficient than traditional methods of documentation, factors such as slang, accent, environmental noise, hardware/software reliability problems, etc., have a sizable impact on the accuracy of these technologies (Moulaei 4). Understanding that the loss of these barriers will lead to improved productivity and higher efficiency in hospitals, the motivation for developing more equitable and accurate speech recognition algorithms is increased.

It is imperative to consider dialectal variation when developing speech-to-text algorithms to overcome these difficulties. This can be accomplished by gathering and employing training data that accurately represents the variety of dialects worldwide and by creating algorithms that can generalize to new speech samples from a variety of dialects. This requires the ability not only to use common, easily retrievable data but also to produce new data from different regions, genders, socioeconomic classes, ages, etc., to account for the numerous combinations of human factors that lead to different dialects.

Developing algorithms uniquely suited to each dialect is one method for overcoming the difficulties caused by dialectal variation. These algorithms can be trained to consistently and accurately transcribe speech while considering the unique linguistic characteristics of each dialect. This can be accomplished by collecting and using dialect-specific training data. However, significant data and processing resources are required for this, which may only be practical for some languages. This approach would also complicate the use of these speech-to-text technologies, as one would have to fully understand their dialect and linguistic features to receive personalized services. Despite these drawbacks, this method has the potential to dramatically improve the accuracy and reliability of speech-to-text algorithms.

An alternative approach is to develop algorithms that can recognize and translate a variety of dialects without being specially adapted to each dialect. This can be accomplished by employing inclusive and diverse training data and creating algorithms that generalize to new speech samples from various dialects. This strategy has the benefit of being more scalable and economical. However, it also calls for sophisticated machine-learning methods that can recognize and transcribe speech from various dialects.

Whichever method is used, it is evident that considering the impact of dialectal variance is essential for the creation of fair and reliable speech-to-text algorithms. It is essential to consider how other aspects, including

gender, may affect the effectiveness of these algorithms in addition to dialectal variation. According to my findings, the average word error rate for one gender was significantly higher than the other for dialects, such as Australian English. This demonstrates the necessity of additional study into the elements influencing the fairness and accuracy of speech-to-text algorithms and emphasizes the significance of creating algorithms that are inclusive and respectful to all people.

The outcomes of my study show the significant challenges that dialectal variation can present for speech-to-text algorithms. However, it is feasible to construct algorithms that are accurate, fair, and respectful to all individuals, regardless of their origin or dialect, by considering dialectal diversity when developing these algorithms. This will improve the precision and dependability of voice recognition systems for people with disabilities as well as the standard of speech-to-text algorithms used to transcribe multimedia information.

References

Adda-Decker, M., & Lamel, L. (2005*). Do speech recognizers prefer female speakers?* https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8bcf842ec0d73cbdc6d08d95a898eb3d5bc6199f

Barrera, B. B. (2015). *A sociolinguistic study of T-glottalling in young RP: Accent, class and education* (Master's thesis). https://core.ac.uk/reader/74374200

Dinari, F., Rahimi, M., Asgari, S., & Khademi, A. (2023). Benefits, barriers, and facilitators of using speech recognition technology in nursing documentation and reporting: A cross-sectional study. *Health Science Reports,* 6(6), e1330. https://doi.org/10.1002/hsr2.1330

Dorn, R. (2019). Dialect-specific models for automatic speech recognition of African American Vernacular English. *Proceedings of the Student Research Workshop Associated with RANLP* 2019, 21–27, Incoma Ltd. https://doi.org/10.26615/issn.2603-2821.2019_003

Graham, C., & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *Cambridge Open Engage.* https://doi.org/10.1121/10.0024876

Hux, K., Rankin-Erickson, J., Manasse, N., & Lauritzen, E. (2000). Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative and Alternative Communication,* 16(3), 186–196. https://doi.org/10.1080/074346100123312790 44Jiao, D., Watson, V.,

IBM. (n.d.). *Speech recognition*. https://www.ibm.com/history/voice-Recognition

Jiao, D., Watson, V., Wong, S. G.-J., Gnevsheva, K., & Nixon, J. S. (2019). Age estimation in foreign-accented speech by non-native speakers of English. *Speech Communication,* 106, 118–126. https://doi.org/10.1016/j.specom.2018.12.005

Koester, H. H. (2001). User performance with speech recognition: A literature review. *Assistive Technology*, 13(2), 116–130. https://doi.org/10.1080/10400435.2001.10132042

Lauture, C. (2020). African American Vernacular English: Categories of

necessity in a language that refuses to be standard. *Undergraduate Review*, 15, 166–183. https://vc.bridgew.edu/cgi/viewcontent.cgi?article=1469&context=undergrad_rev

Nakatumba-Nabende, J., Kagumire, S., Kantono, C., & Nabende, P. (2025). A systematic literature review on bias evaluation and mitigation in automatic speech recognition models for low-resource African languages. *ACM Computing Surveys*, 58(4), 1–24. https://doi.org/10.1145/3769089

Nigmatulina, I., Scherrer, Y., & Samardžić, T. (2020). ASR for non-standardized languages with dialectal variation: The case of Swiss German. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 12–21. Association for Computational Linguistics. https://aclanthology.org/2020.vardial-1.2.pdf

Pieraccini, R. (2012). *From AUDREY to Siri: Is speech recognition a solved problem?* International Computer Science Institute. https://www1.icsi.berkeley.edu/pubs/speech/audreytosiri12.pdf

Reid, K. (2023, July 5). Are you listening to me? The inner workings of speech recognition. *ANU School of Cybernetics*. https://cybernetics.anu.edu.au/news/2023/07/05/are-you-listening-to-me/

Riccardi, G., & Hakkani-Tür, D. (2005). Active learning: Theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing,* 13(4), 504–511. https://doi.org/10.1109/TSA.2005.848882

U.S. Department of Commerce. (1976). *The Harpy speech recognition system*, 1–125. Pittsburgh, PA.

Wang, Y.-Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy? *Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding,* 577–582. IEEE. https://doi.org/10.1109/ASRU.2003.1318504