APPLYING AUTOMATIC SPEECH TO TEXT IN ACADEMIC SETTINGS FOR THE DEAF AND HARD OF HEARING

APPLYING AUTOMATIC SPEECH TO TEXT IN ACADEMIC SETTINGS FOR THE DEAF AND HARD OF HEARING

By CARLA WEIGEL, B.A. (Hons.)

A Thesis Submitted to the School of Graduate Studies in

Partial Fulfilment of the Requirements for the Degree Master of Science

McMaster University
MASTER OF SCIENCE (2021)
Hamilton, Ontario (Cognitive Science of Language)

TITLE: Applying Automatic Speech to Text in Academic Settings for the Deaf and Hard of Hearing

AUTHOR: Carla Weigel (Hons.) (McMaster University)
SUPERVISORS: Dr. Magda Stroinska, Dr. Daniel Pape
NUMBER OF PAGES: xi, 72

Lay Abstract

In hopes to encourage more D/deaf and hard of hearing (DHH) students to pursue academia, automatic captioning has been suggested to address notetaking issues. Captioning programs use speech recognition (SR) technology to caption lectures in real-time and produce a transcript afterwards. This research examined several transcripts created by two untrained speech-to-text programs, Ava and Otter, using 11 different speakers. Observations regarding functionality and error analysis are detailed in this thesis. The project has several objectives: 1) to outline how the DHH students' experience differs from other note-taking needs; 2) to use linguistic analysis to understand how transcript accuracy converts to real-world use and to investigate why errors occur; and 3) to describe what needs to be addressed before assigning DHH students with a captioning service.

Results from a focus group showed that current notetaking services are problematic, and that automatic captioning may solve some issues, but some types of errors are detrimental as it is particularly difficult for DHH students to identify and fix errors within transcripts.

Transcripts produced by the programs were difficult to read, as outputs contain poor punctuation and lack breaks between thoughts. Captioning of scripted speech was more accurate than that of spontaneous speech for native and most non-native English speakers; and an analysis of errors showed that some errors are less severe than others. In response, we offer an alternative way to view errors: as *insignificant, obvious,* or *critical* errors. Errors are caused by either the program's inability to identify various items, such as word breaks, abbreviations, and numbers, or a blend of various speaker factors. Both programs worked best with intelligible speech; One seemed to prefer fast speech from native English speakers and the other preferred slow speech; a

preference of male or female voices showed conflicting results. Some reasons for errors could not be determined, as one would have to observe how the systems were programed.

Abstract

In hopes to encourage more D/deaf and hard of hearing (DHH) students to pursue academia, speech-to-text has been suggested to address notetaking issues. This research examined several transcripts created by two untrained speech-to-text programs, Ava and Otter, using 11 different speakers in academic contexts. Observations regarding functionality and error analysis are detailed in this thesis. This project has several objectives, including: 1) to outline how the DHH students' experience differs from other note-taking needs; 2) to use linguistic analysis to understand how transcript accuracy converts to real-world use and to investigate why errors occur; and 3) to describe what needs to be addressed before assigning DHH students with a captioning service.

Results from a focus group showed that current notetaking services are problematic, and that automatic captioning may solve some issues, but some errors are detrimental as it is particularly difficult for DHH students to identify and fix errors within transcripts.

Transcripts produced by the programs were difficult to read, as outputs lacked accurate utterance breaks and contained poor punctuation. The captioning of scripted speech was more accurate than that of spontaneous speech for native and most non-native English speakers. An analysis of errors showed that some errors are less severe than others; in response, we offer an alternative way to view errors: as *insignificant, obvious,* or *critical* errors. Errors are caused by either the program's inability to identify various items, such as word breaks, abbreviations, and numbers, or a blend of various speaker factors including: assimilation, vowel approximation, epenthesis, phoneme reduction, and overall intelligibility. Both programs worked best with intelligible speech, as measured by human perception. Speech rate trends were surprising: Otter seemed to prefer fast speech from native English speakers and Ava preferred, as expected, slow

speech, but results differed between scripted and spontaneous speech. Correlations of accuracy and fundamental frequencies showed conflicting results. Some reasons for errors could not be determined without knowing more about how the systems were programed.

Acknowledgments

This project was a lot of fun and I hope that its results could be very beneficial. Heartfelt thank-you to everyone who helped me with this project over the last two years:

Magda Stroinska, thank you for remembering me as an undergraduate student many years before starting graduate studies and supervising me. Thank you for giving me this project, for trusting me with it and giving me the opportunity to work on other projects. Thank you for supporting me throughout my studies and always trying to find me things to do. It was a pleasure to work with you on this project as well as your other courses. Your guidance and your kind words will stay with me always. I wish you all the best for your future and that your retirement is relaxing and rewarding.

Daniel Pape, thank you for accepting me as your student, even though we did not know each other, and I had a lot to learn. Thank you for teaching me more than I ever thought I'd learn, giving me deadlines and for your honesty. Thank you for joining our silly social events; It meant a lot and I really enjoyed getting to know you and Sabastian through them. Thank you for your guidance and advice. It was a pleasure working with you on this project and I very much enjoyed being your TA. I wish you well as you continue your career, and I am honoured to be one of your first masters students.

Thank you to my husband, Robert Stea, who did everything for me while I spent our money on graduate school. Thank you for your patience, your support and all the little things I'll never be able to remember. I love you very much and I'm so grateful to have you in my life.

Thank you to my parents, Rita and Fred, for supporting my decision to return to school. Thank you for your support, your phone calls and all the help you've provided throughout the years. I know you'll continue to support me and my family as I continue my journey through life. I love you.

Finally, thank you to everyone else who helped with this project. The professors that taught me everything in graduate classes. I learned a lot from you. Thank you to the participants, my lab colleagues, and my friends who always had kind things to say and keep reminding me that I am pretty great at what I do. Thank you for providing some well-needed distractions.

Thank you to the ARiEAL Research Centre and the Linguistics and Languages Department at McMaster. Thank you, Chia-Yu, for being a great help and support navigating all the things I did not know as well as working with me as social coordinator. It was a great experience. Nanci Cole, for all the administrative work you did for me in both my undergraduate and graduate studies. And to Hyunji Shin, whose assistance helped keep this project. I wish you all the best in your future career.

Contents

List of Figures and Tables

List of all Abbreviations and Symbols

A – Additions

AAVE – African American Vernacular English

ADHD – Attention Deficit Hyperactive Disorder

AI - Artificial Intelligence

ASL – American Sign Language

DHH – D/deaf and hard of hearing

ESL – English as a Second Language

I – Insertions

O – Omissions

S – Substitutions

SAS – Student Accommodation Services

SR – Speech Recognition

SRN – speech rate normalization

VTLN – vocal tract length normalisation

WER – Word Error Rate

1. Introduction

This project aims to examine speech-to-text technology for use in academia with the intent to augment the learning experience for Deaf, deaf, and hard of hearing (DHH) students. The following review will outline why this study is relevant, background information on speech-to-text and accuracy levels, how it has been used, what previous research has found, and shortcomings from that research.

1.1 Literature Review

In this thesis, we distinguish between "Deaf", "deaf" and "hard of hearing" individuals. Deaf individuals (i.e., capitalized "D") are people who identify as Deaf and being part of the Deaf culture and community, who have their own traditions and communicate using Sign Language (we will refer specifically to American Sign Language, or ASL, as different countries have different sign languages with unique semantics and syntactic structures), regardless of physical hearing ability. As we capitalize the first letter when describing cultural groups (e.g., German, Japanese, English, etc.), we also capitalize the "D" to indicate that the individual identifies as being part of this specific cultural group. Alternatively, a deaf person is someone who has little to no hearing ability, medically described as severe to profound hearing loss. Someone who is hard-of-hearing has some hearing ability, medically described as mild to moderate hearing loss. Woodcock et al. (2007) stated that approximately 4-5% of the population of Canada has communication needs due to hearing loss. Many of these people benefit from accommodations such as hearing aids (which are inserted directly into the ear and amplify specific frequencies), FM systems (where a speaker wears a microphone and the listener has a receiver that picks up that specific speech signal to reduce background noise), or classroom amplification systems (where speakers wear a microphone which is connected to a loudspeaker system that amplifies and crisps sounds from that speaker); however approximately 1% of the Canadian population (close to 400,000) are "deaf" and, even with hearing aids, cannot converse through channels using speech alone (e.g., telephone).

Universities are accepting more students who require accommodations, including those who are D/deaf and hard of hearing. Accommodations may be expensive and technology, although available, does not always meet standards necessary for success. In response to this, Student Accommodation Services Office (SAS) at McMaster had subscribed to Ava as an

accessibility option in 2018 (see SAS Ministry Report, 2018). Tim Nolan, the Disability Services manager at McMaster, reached out to Magda Stroinska in 2019 in hopes to evaluate the program regarding how effective it is and whether this service would be something that students would want. This study began in 2019 with the intent to focus on Ava as an option for students who are D/deaf or hard of hearing to use in academia. The study was an opportunity to expand accessibility for the D/deaf and hard of hearing, however Tim Nolan retired in 2020 and the subscription to Ava was decidedly not renewed. The research question of whether automatic speech-to-text would benefit students as an accessibility option was intriguing and worth pursuing. The project also provided an opportunity to analyze speech-to-text using linguistics rather than algorithms and programing, which is something that has not been done to this research team's knowledge. Knowing from day-to-day experiences with speech recognition (SR), as well as numerous publications, the technology is not perfect. For instance, SR programmers admit that the quality of transcriptions decline when speakers have accents or if the program is untrained. These variables are very relevant to how the service would be used in a practical setting at university. Universities hire professors from around the world, so accents are expected. Professors are very busy throughout the year, and it would be unreasonable to ask them to go through the lengthy process of training speech-to-text programs. The current study aims to outline a practical solution to note-taking needs through automatic speech-to-text technology to accommodate the needs of Deaf, deaf, and hard of hearing students.

Because Deaf, deaf and hard of hearing (DHH) individuals encounter huge barriers while pursuing postsecondary education, many choose not to pursue undergraduate studies, and few consider graduate and doctoral studies. A review by Woodcock et al. (2007) outlines just how difficult academia is for DHH to pursue. This includes lack of Deaf role-models in higher education, and common misconceptions about the limitations of lip-reading and the effectiveness of hearing technology (for instance, hearing individuals may think hearing technologies work like glasses, which correct vision impairment up to normal vision, however, the experience of using hearing technology is more complicated: Hearing aids amplify specific frequencies up to normal hearing levels; however, background noises are also amplified, and some frequencies cannot be recovered). The researchers comment that hearing individuals easily forget to, or resent having to, make adjustments for a DHH student. DHH students have far less access to lecturers than hearing students, since ASL (visual) translation is often necessary, which requires

preparation; some ASL translators may be hesitant to book, since course-specific vocabulary may be unknown to the translator. Furthermore, staff and faculty do not fully understand what being deaf and hard of hearing is like – for instance, they may not realize that English is often a second language of the DHH person, with ASL being the first language. In addition to the social difficulties of being DHH in academia, accessibility services require a huge number of resources in order to fully meet the needs of these students. The available resources are limited and thus students' needs cannot be fully met.

According to the McMaster University Student Accommodation Services (SAS) Ministry report, there were 53 students that were Deaf, deafened, or hard of hearing during the 2018-2019 academic year. This made up 7.5% of students with physical disabilities. The same report, which outlines the new implementation of Ava (which gave impetus to the current study) and Echo 360, suggests that more support is needed for notetaking, video captioning and lecture transcripts. The report also describes a trend in closed-captioned video requests that has been increasing since 2011. The report also outlines various other physical and non-physical disabilities reported by students who require accommodation services, including notetaking.

Note-taking is the most requested service for accessibility services (Van Meter et al, 1994). Several previous studies suggest that note-takers, including computer-generated notes, benefit not only DHH students, but also other students with or without other disabilities, e.g., ESL students (Ranchal et al., 2013, Hwang et al., 2012; Wald, 2010; Wald et al., 2009; Wald & Bain, 2007; Ryba et al. 2006; Bain et al., 2002). Although it is possible to acquire paid note-takers, many classes request a volunteer to take notes for students who need them, sometimes for extra credit. However, these note-takers are usually not trained and are taking notes for their own studying needs. These notes are simply distributed to students who requested them. The result is that sometimes the notes are subjective or incomplete: the content is selective, based on what the writer thought as important and on their previous knowledge, missing information due to a number of reasons, or organized in a way that makes sense to the writer, but not necessarily the reader. In their 1994 study, Van Meter et al. interviewed undergraduate students regarding their note-taking process: every student reported to take notes during lectures and outlined the reasons for taking notes (e.g., review for tests or that taking notes help to stay awake or pay attention in class); they found that half of the students paraphrase lecture notes and half attempt to write

word-for-word what is said in lecture. They also found that students are selective about what they put into their notes and have different ways of indicating content they don't understand or when the professor emphasised a particular point, using their own shorthand method. Lecture notes are extremely varied between students. Any student may have a hard time getting the information they need from a classmate's notes.

In order to address notetaking issues, it was proposed that notes could be taken using online, automatic speech-to-text programs. However, in order for these services to be adequate, they must be accurate. Speech-to-text programs often advertise high accuracy levels: Google claimed a 95% word accuracy rate in 2017 (Abner, 2017), but other sources have determined it to be a more reasonable "it depends" (i.e. between approximately 9% and 95%, depending on the platform), with Google Enhanced showing between approximately 52% up to 95% accuracy (mean: approx. 85% accuracy) in June, 2020 (Jarmulak, 2020). Some programs showed improvements merely months later in October 2020 (Jarmulak, 2020). However, these accuracy levels are established using word error rates (i.e., number of errors compared to actual utterances in a recording). These are binary correct/incorrect measures that do not consider content, error types or overall readability. In addition, an accuracy rating may sound impressive, but an average accuracy percentage does not convey a user's experience with the program in the real world. There is a difference in the severity of errors that may cause a transcript to be legible or illegible. One must consider the severity of the error to consider if the sentence is understood, and, if the reader is unable to hear the original recording, how well can the transcription be understood? Such questions have not, to this author's knowledge, been asked, which begs the question, how accurate is "accurate" with respect to understanding the content? This study aims to answer this, and other questions, related to SR technology.

1.1a Speech Recognition Software - background

Speech recognition technology has its roots in 1952, with David and colleagues of AT&T Bell Labs creating a system that could recognize ten numbers in English (David, Biddulph & Balashek, 1952). It later became possible to identify vowels following research by Forgie & Forgie (1959) and further research in the technology started to develop in the 1960s. At the time, dynamic programming and linear predictive coding analysis paved the way into speech recognition research, allowing the 1970s to introduce theories such as dynamic time warping,

vector quantisation and hidden Markov models. The 1980s resulted in algorithms to use statistical models, rather than pattern-matching techniques, which lead into the 1990s when single-word transcription was starting to emerge (Cao & Guo, 2020). Today, speech-to-text technology has become an everyday application rather than a luxury, with Google Assistant (including Google Home), Siri, Alexa, and others, which illustrate only a fraction of how relevant speech-to-text research has impacted society. Speech-to-text is used in education, from elementary school to university, and in businesses. Though speech-to-text delivers impressive results, as anyone who has used it can attest to, it is also prone to errors.

1.1b How Speech-to-text works

As humans vary in size and shape, individual voices are also unique. This uniqueness is contained and expressed within variable information that humans receive when listening to one another: extra, unintentional information such as the speaker's age, sex, region, mood etc. This unintentional information is what makes human speech rich in variability. Therefore, a speech signal produced from a human is highly variable and individualized. Individual phonemes contain universal rules (such as formant values, voice onset times, etc.) contained within the voice-pulse shape, which makes the overall content of the message largely understood by another human, regardless of the speaker; in addition, the human cognitive system repairs broken speech signals and filters extraneous information subconsciously. Computers, however, rely on rigidity to work as intended. As speech recognition is concerned, computers rely on predictable patterns in order to function, so it follows that variance becomes a major problem for speech recognition software (Leonov & Sorokin, 2014). Makhoul and Schwartz (1992) described how speech recognition works using hidden

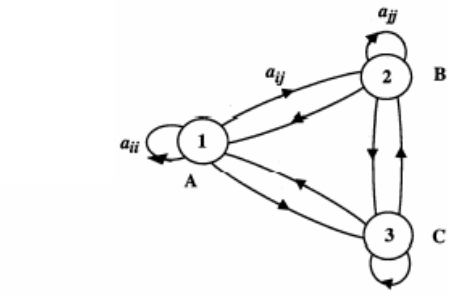

Figure 1-1 A three-state Markov Chain (from Makhoul, J. & Schwartz, R. (1995) State of the Art in Continuous Speech Recognition. Proceedings of the National Academy of Sciences of the United States of America, 92(22), 9956-9963.

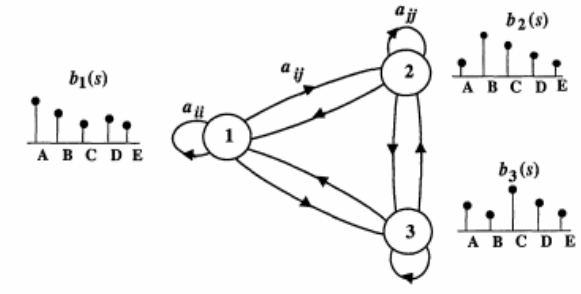

Figure 1.2 a three-state hidden Markov Model (from Makhoul, J. & Schwartz, R. (1995) State of the Art in Continuous Speech Recognition. Proceedings of the National Academy of Sciences of the United States of America, 92(22), 9956-9963.

Markov models. They developed software that analyzed the speech signal by considering each context of a phoneme to an allophone, which produce a value using these hidden Markov models, which were state-of-the-art at the time and are still used in modern programs.

A Markov Model begins with a Markov Chain (see figure 1.1). The figure shows three symbols: A, B and C, each related to a state, 1, 2 and 3, respectively. As a transition occurs from State 1 to State 2, for example, it generates the symbol, B, as output. The Markov chain continues to output symbols as transitions happen from state to state. The transition is difficult, as the output is determined by the transition. So, for example, the transition of states 1 2 3 2 1 creates a symbol sequence of A B C B A. This output is deterministic, i.e., the transition determines the state. However, the symbol output in hidden Markov Models is probabilistic.

In a hidden Markov Model (see figure 1.2), all symbols are possible at each state, each with its own probability and probability distribution. Thus, it is impossible to know what transition of states created a given output, thus, giving a *hidden* Markov Model.

Makhoul and Schwartz identified that variability is the cause of most errors in speech-to-text. Three main components that affect the speech signal were outlined: 1) Linguistic variability (phonetics, syntax, semantics and discourse); 2) Speaker variability (speaker-specific variables including co-articulation); and 3) channel variability (background noise and how the sound is transmitted, e.g., through microphone, telephone, or a room with high reverberation); however, these variables continue to be issues for speech recognition over the past several decades (Makhoul and Schwartz, 1992; Deng et al., 2006; Fu & Murphy, 2006). Modern speech recognition software adds more robust algorithms, complex probability models, new technology, and artificial intelligence (Babu et al. 2010; Bod, 2012; Leonov & Soroken, 2014; Cao & Guo, 2020; Ava, 2020). Yet even with these major improvements in recent decades, studies still show that more work is necessary.

1.1c Accuracy levels and how they are created and achieved

Errattahi et al. (2018) outline three types of errors in speech to text transcripts: substitution, deletion, and insertion. Substitution errors are errors that substitute one word for a different word or words (e.g., “for” to “four”, “bows” to “balls”). Omission errors are when a word or group of words that were spoken are not transcribed. Here, we counted each word as a

single omission error (e.g., if the program omitted "and", it would count as one omission error. If the program omitted "and he said", it would count as three omission errors). Insertion errors (which we refer to as "additions" in this paper) are words that were in the transcript that were not said by the speaker. Word Error Rate (WER) is a common tool used to determine the accuracy of automated speech-to-text. A higher WER indicates poorer speech recognition program performance. The definition of WER is outlined as follows:

$$\mathrm{WER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}}$$

Where S, D, and I are the number of word substitutions, deletions, and insertions, respectively, and N is the total number of words (Errattahi et al. 2018). As mentioned previously, this accuracy equation is binary and does not qualify types of errors, nor does it consider the readability of a text. The final calculation groups all error types together, thus eliminating a categorized inventory of errors.

Accuracy can be improved in a number of ways, which have not changed much in the past 30 years. Fluent speech (i.e., speech without natural errors and pauses) gives more accurate transcriptions than spontaneous speech, which contain natural disfluencies. Fluent speech can be achieved by using read, practiced, or scripted speech. WER on disfluent speech can be improved by training the system on a specific speaker. The same results are found across the board, with the same issues such as room acoustics, speech disfluencies and grammatical errors, low-frequency words, etc. still causing accuracy to fall (Jarmulak, 2020), along with speaker differences such as fundamental frequency and speech rate (Pfau et al, 2017, Sorokin & Leonov, 2019). Because of this, SR systems report their WER by first training a system on a speaker, and then using scripted speech. Current research shows even more specific variables that also contribute to errors.

Sorokin and Leonov (2014 and 2019) performed studies that focused on the glottis and speech source pulses. Their 2019 study indicated that speech recognition accuracy goes down as fundamental frequency goes up. Fundamental frequency (f_0) is, put simply, the lowest pitch of a person's voice. Males generally have an f_0 of 100Hz. Females generally have higher voices than males, around 200Hz. Children have even higher voices, around 400Hz. Li & Russell (2002) found that SR accuracy is worse for children than adults (however the study attributed lower

accuracy to poor pronunciation as well as formant frequencies). It follows then, that female voices may result in a lower accuracy rate than male voices, as females have higher voices than males. Bajorek (2019) wrote an article discussing the SR discrepancy between male and female voices. She describes how SR programmers tend to have more data from males, which contributes to the problem. She outlines that Google's SR technology is 13% less accurate for female voices than male voices and how that might affect someone using the technology in the real world.

Koenecke et al. (2020) performed a study using Microsoft, IBM, Google, Apple, and Amazon systems that compared white and black speakers. They found that each system had higher WER for black speakers than white speakers. This was attributed to two main differences: different grammatical structures in the African American Vernacular English (AAVE) (for example, deletion of the verb "to be" in phrases such as "he a pastor") and differences in pronunciation and prosody of black speakers versus white speakers. Studies show that speech recognition relies on a predictable grammatical structure; as the study was comparing "standard" English to AAVE which does not align with the system's programmed predictability models, the results were showing lower scores for AAVE.

Jones et al. (2017) performed three case studies to determine readability in speech-to-text programs: one in English and two in Arabic; here we will focus only on the findings of the English study, as the other two involved translation and are not relevant to the current work presented here. The researchers compared human-generated transcripts, which included proper punctuation and were accurate, to machine-generated speech-to text. Participants answered content questions after reading a text, transcribed using Speech-to-text by a program. Researchers measured answer accuracy, time taken to read the content and answer question, and a score based on the participants' opinion of the transcript. Results were compared to a control group, which included the same tasks, but with a human-created transcription of the same texts. The WER for the computer-generated English transcript ranged from 8.6% to 58.2%, averaging at 30.4%. The computer-generated text also showed much slower processing speed, by approximately 20% (from .52 seconds per word for the control group to .65 seconds for the computer-generated text); however, the comprehension test lowered only from 90% to 85%. The researchers concluded that the readability of the English transcript did cause readers to slow

down and misunderstand content. The higher-than-expected accuracy rate was attributed to the transcripts not being "rich in information content".

Pfau et al. (2017) have been studying how to reduce errors caused by speaker-specific variation and fast speech rate. They outline that fast speech causes an increase in processing load, which leads to errors. They found that WER could be reduced by 15% by using vocal tract length normalisation (VTLN) and speech rate normalization (SRN) procedures, which they developed. In their 2017 study, the developers compared output WER to recordings that were run through again using their processes, which essentially slowed fast speech (SRN) and normalized acoustic models (VTLN), comparing outputs from one or the other system as well as both systems. While comparing "fast", "medium" and "slow" speech, Pfau et al. found that "slow" speech caused fewer errors. Furthermore, WER decreased after using one or the other system, but the most effective results were produced when speech was run through both systems. One can conclude from this study that, although the technology continues to evolve, speech rate affects the accuracy of speech recognition programs, with slow speech having better results than fast speech.

1.1d The use of Speech-to-Text transcription in educational settings

Everyone benefits from automated transcriptions in academic settings, even if they do not have a disability or impairment. Bain et al. (2002) noted that students without disabilities are able to cross-reference their own notes to the transcript to check for accuracy, missed content and ultimately improve their notes. Of course, these students also have the benefit of perceiving the lecture auditorily.

Ranchal et al. (2013) performed an experiment which compared post-lecture transcription (an audio recording of the lecture with transcripts embedded into the PowerPoint presentation available online after class) and real-time captioning (using a client-server application which was viewed during class either using a projected screen or on students' personal laptops). The study showed that post-lecture transcription resulted in better student performance than real-time captioning, however lectures with both showed better student performance than lectures without any transcription software. This was supported with recorded audio and video lectures which were available to students after in-person lectures in a science class. This study shows the value of captioning services, however the study used robust training methods for each instructor

involved in the study; a process that was likely time-intensive and not necessarily feasible for a large-scale university.

The Liberated Learning Project, as outlined in Bain et al. (2002), started at Saint Mary's University in Halifax, Nova Scotia, in 1998 and attempted to increase flexibility in the classroom for students with disabilities by developing automatic speech recognition technology. The project involves training the system on lecturers' voices that teaches software the nuances of the voice, which helps reduce errors. The study by Bain et al. outlined the accuracy, production, and readability of transcripts. Bain et al. outline that accuracy varies in significance and give the following example:

> Actual words spoken: I went to the store
>
> Scenario # 1, SR transcribed: I went to the **door**
> Scenario #2 SR transcribed: I went **too** the store
>
> "Both scenarios return an accuracy rate of 80% (4/5 words recognized correctly; word transcription error bolded). However, in the absence of audio cues to aid understanding, for example as experienced by a person with a hearing disability, the difference between the two transcriptions directly affects comprehensibility of the phrase." (Bain et al., 2002)

Although the text does not go into further detail on how to approach the different scenarios, the authors outline that it is something that future researchers should consider (as the current project attempts).

Regarding note production and editing, Bain et al. suggest that a 1-hour lecture would take 3 hours to edit, even with an accuracy of 80%; it follows that a 3-hour lecture might take up to 9 hours to edit. Regarding readability, the authors acknowledge the difficulty of reading continuous text and point out that technologies are working on making transcripts easier to read by adding punctuation or displaying transcripts differently. As the technology in 2002 was unable to meet these requirements, the authors point out that this is an area that needs improvement. Bain et al. argue that, even with limitations, the potential of their system is profound, having already found increased success in 2002.

Paez et al. (2002) attempted to find out if students with disabilities improve their academic performance if they are given a transcript of lectures. The study included students who were D/deaf, hard of hearing, or had medical, physical, and learning disabilities, including

ADHD. Students had varying use of digitized screen, as some found it difficult to focus on both reading and listening at the same time and found errors distracting. The D/deaf student found that transcript availability enabled full participation, however, the hard-of-hearing student (as well as others from the study) found that there was too high an error rate, but would use the transcript if it were more accurate.

Ryba et al. performed a study to assess students' perception of the display, how much would students use the Liberated Learning program, and the advantages and disadvantages of the program. The study involved 160 students, approximately half of which were L2 learners, in four lectures. The lecturer was trained on the system and, using a wireless microphone, the voice was recorded and transcribed using Viascribe™. As the lecturer spoke, a transcript was projected onto a screen, and after the lecture, the transcript was edited and posted online. Then the students were given a survey asking about how they felt about the experience. Researchers asked about students' first impressions, their thoughts on accuracy of the transcript, what problems they found. and if the visual display was distracting. After the final lecture that used the system, students were given another lecture without the system and were invited to answer another survey. Students provided positive reactions, although only ten students responded to the survey; however, students reported the screen was sometimes distracting. Students generally reported that increased accuracy would be beneficial. Another survey was given to students that consisted of agreement statements, which resulted with mixed responses. The researchers addressed some ideas on how to resolve some of the issues, such as having arranged seating for those that wanted view of the screen, more participation from students, giving students advice on how to split attention effectively and increase accuracy. Not only does this study outline the benefits of improving speech-to-text, but it also highlights how varied individual needs are. This demonstrates how any approach may be very beneficial to some students, but detrimental to others. Thus, when working on improving accessibilities, it is important to be mindful that, while any service offered must be individualized, it simultaneously cannot encroach on classmates.

Every study researched suggested that software needs improvement and relied heavily on lecturers taking the time to train systems. In a real-world university setting, professors would not have time to train a system and one student may use the same system for many lecturers as well as for personal communication. In order to address these shortcomings, we are interested in how

a system works “out of the box”, without lecturers first training the system, thus reducing the amount of time necessary to set up a system for a student.

Furthermore, no previous study observed considered how errors affect how well the information is absorbed by students. Wald et al. (2007) suggested that simultaneous speech, even with errors, is understood by most students if accuracy is over 85%, but acknowledged that transcript readability is important, focusing on punctuation and edited transcripts after the lecture is finished. In most cases, students had access to both audio and visual channels to absorb information, meaning that errors observed could be cross-referenced and fixed by students themselves. Studies that included Deaf and hard of hearing students often had edited transcripts, which is a lengthy process, or did not check for accuracy of understood information. Ideally, a transcription system would allow students to have full control over the transcription device so that students can retrieve information quickly and easily and increase participation during lectures without involving too many external channels. With this in mind, we are interested to find out if the technology is reliable enough for a student to be confident that the system will work for them.

Speech-to-text software products that claim high accuracy are tested in ideal conditions, with noise-controlled rooms and scripted texts that, where typical human speech behaviours such as dysfluencies, repetitions, self-correcting, varied speech rate, etc., are reduced, all of which can cause lower accuracy (Bain et al. 2002). The producers of Otter even mention in its website that words such as “um” are deleted from transcripts, even if the word is added manually to its dictionary. As far as this author knows, there has not been any analysis of speech-to-text using linguistic parameters that judge the severity of an error; Bain et al. allude to an accuracy sub-test of the test of automated speech recognition readability by Dr. Ross Stuckless (Bain et al. 2002), however efforts to find this test were proven fruitless.

As Koenecke et al (2002) explain, it is important to note that speech-to-text struggles with certain vernaculars. The following study focuses on “standard English” because generally, academic language is standard English, rather than other vernaculars. Standard English is also used in the government, media, and education in Canada. Although an interesting topic worthy of more study, vernaculars that differ from standard English are not pertinent to academic study at McMaster University and thus is not within the scope of this study.

1.2 Research Questions

The current study aims to analyze speech-to-text as an accessibility service for DHH students. Although, as previous research suggests, SR technology is beneficial to all students, regardless of their individual needs; however, DHH students have different experiences and barriers than hearing students. This project was designed to investigate SR as an accommodation service with the barriers of DHH students in mind, i.e., that errors cannot necessarily be detected through hearing the lecture. Therefore, we must imagine how a student would use the technology with ideal expectations; specifically, the service should be convenient and accurate, as if these two expectations are not met, the student may not enjoy their experience and decide it is not suitable. With this in mind, we recognize a few issues: 1) that professors and students often have busy schedules and likely will not be able to take the time to train an SR system; 2) students have many different professors, so the system must work with many different speakers; and 3) although professors are knowledgeable in their disciplines, they usually do not script their lectures and will speak freely during lectures. For this reason, we question how an advertised accuracy fares when used with speakers who are not trained on the system and using spontaneous speech. This leads to the following question: how much would an accented speaker affect the accuracy of a transcript? McMaster hires experts from around the world, so students would expect to have lectures with professors with accents. For this reason, we selected participants with different accents and observed how the systems handled accented speech. Given that errors are expected, we ask, what does an advertised accuracy mean in the real world? Certainly, humans do not speak with 100% accuracy; however, humans have the innate ability to repair broken speech signals and perceive complex information from a speech signal. This allows humans to speak to each other even if information is missing. A computer, however, relies on a clear speech signal to deduce what was said and outputs its analysis. So, what does an "accurate" transcript look like? Are we able to say that an accuracy of 90% is good enough for a written document to be understood? Then we also ask, are there errors that are more severe than others? Certainly, people make mistakes in written documents, as spelling mistakes and homophone confusion are common; as such documents created by humans, even if they contain errors, are generally legible. Finally, we ask if the technology is reliable enough from the viewpoint of a student trying to gather information. A student must trust that an accommodation will benefit

them, but if the student finds that errors are causing confusion, they may doubt that the information they are gathering is correct. This could cause anxiety, for fear that what they read may not be what was said.

We are also interested in the linguistic analysis of speech-to-text. Although not directly relevant to the DHH population, programmers may be interested in viewing errors from a perspective in which they may not have expertise. Analyzing errors using linguistic knowledge may lead to advancements in how the technology is programmed, ultimately improving existing SR systems. For this reason, we ask the following questions: How do speaker-specific variables, such as accents, speech rate, fundamental frequency, and intelligibility affect the program's accuracy? Most pertinently, would we be able to use linguistics to examine what went wrong from the speech signal and why an error occurred? By addressing these questions, we hope to initiate discussions about how speech is recognized by computers and how it differs from human speech perception in hopes that the technology may be improved.

1.3 New Developments

In 2020, due to the COVID-19 pandemic, the general public were forced to find alternative ways of completing day-to-day tasks, as work-from-home mandates and social distancing caused a sudden upheaval in daily life. Many industries migrated from communal workspaces to work-from-home environments. This included elementary, high school, college, undergraduate, and graduate classes. The transition has been challenging for many reasons; however, it highlighted the importance of accessibility and captioning. In January 2021, the Disabilities Act made captions mandatory for recorded materials posted online; however, the captions are not required to be edited. Although McMaster has already been working towards having captioning available for all online video content, the importance of accurate captions has increased significantly. Otter has been used to caption Zoom meetings, and its use skyrocketed once lockdowns were in place, including a partnership with Google Meet (Collins, 2021). Though it is difficult to know whether the increase of Otter's use has improved its algorithms, it is important to note that Otter quickly transformed from being fairly unknown to a staple in captioning options in educational settings at many universities, including McMaster.

2. Research Methodology

2.1. Introduction

As explained in chapter 1, the research described in this thesis was originally planned before the COVID 19 pandemic closed the university. The research ethics clearance that we applied for assumed that the researcher would attend selected lectures in person and would record the lecturers live, using a Wi-Fi lapel microphone attached to a receiver and recorded onto a computer, which would later be sent through the SR systems using an Irig pro to an iPad. The closure of the university in March 2020 required significant changes to this plan.

With all lectures having moved to online delivery, the researcher proceeded with using already recorded class presentations that instructors agreed to share. This change in research approach eliminated the ability to study differences in classroom acoustics and introduced variance in recording quality and thus, original speech signal channels. Segments were selected from the samples with clear audio and student voices, if present, were removed. Recordings maintained to be run through an audio interface and adjusted for gain.

2.2. Participants

Thirteen McMaster University lecturers originally agreed to participate in the study, however two declined partway through the research. The remaining 11 lecturers (6 male, 4 female) participated in the study. Participants were either native English speakers (North American or British) or non-native English speakers and taught different areas and levels of study. Participant accents include French, Japanese, Polish, Korean, Chinese, and German, as well as Canadian English, British English, and American English. “Accent” here is used broadly, as the scope of study was for an overall understanding of how accents affect speech recognition, rather than detailed regional differences. Participant ages ranged from under 40 to 65+. The disciplines of study included different subfields of linguistics, anthropology, literature, mathematics, Korean culture, Japanese cinema, and Chinese language.

A focus group was organized by the Student Accessibility Services (SAS) office at McMaster University. All students who identified as DHH were invited via email to attend. Four participants accepted the invite. Three were hard-of-hearing, one was deaf. All were born to hearing parents and reported English as their first language. The group consisted of four

participants (2 male, 2 female), one of whom was late, near the end of the discussion. Two were senior citizens, all were undergraduate students. Three of the participants were hard of hearing and used hearing aids: two of whom did not use a hearing aid during the group. One of the three hard of hearing participants was deaf on the left side and thus sat so that the right ear was facing the leader. The remaining participant was deaf, used a cochlear implant and grew up in a hearing family. All participants had English as their first language, were proficient in reading and had limited, if any, ASL ability. All participants had various proficiency in lip reading, one relied solely on lip reading until age 3, when hearing aids were introduced. One senior participant reported that most of their life they relied on lip reading, but proficiency has decreased with age. All participants reported that lip reading is difficult, inaccurate, and context-driven; accuracy was improved by the listener requesting more information or clarification from the speaker.

Four volunteers from the phonetics lab from the McMaster ARiEAL Research Centre completed a comprehension rating scale for each speaker from 60-second excerpts of both the rainbow passages and lectures. Participant recordings were coded using a numbered system and the volunteers rated how easy the speech excerpt was to understand using a rating scale from 1 (very easy) to 10 (very difficult). These subjective ratings were used to calculate a comprehensibility score for each participant for both read and spontaneous samples.

2.3. Materials and procedures

Participants were asked to provide a recording of both a prepared passage containing all phonemes from the English language[1] as a *read condition*, and a lecture as a *spontaneous condition*, either via email or through shared drive. As mentioned above, due to restrictions caused by the COVID-19 pandemic, recordings could not be created by the researcher, so participant data were collected using microphones owned by the participants using platforms familiar to each participant. The prepared passage was 330 words long; however, some participants omitted words, or added words or phrases. Our samples were between 329-346 words (μ=333.82, s=5.88). Most participants gave the researcher access to their Avenue to Learn online platform used by McMaster University. Some shared with the researcher a single recorded lecture. The formats of the recorded lectures ranged from recorded audio mp3 or video mp4,

[1] "The Rainbow Passage" from Fairbanks G. (1960). Voice and articulation drillbook, 2nd edition. New York. Harper and Row. Pp124-139

recorded from the participant's computer or Bluetooth microphone, Zoom (an online portal used for real-time meetings), and MacVideo (a recording option for McMaster University professors to pre-record lectures to post online). Full lectures provided by the participants were then cut to three- to five-minute sections. These sections were selected by the researcher based on overall complex language used. The researcher also aimed for sections with as much English as possible, as some lectures contained many foreign words. The lectures varied in number of words, as speech rate varied between participants and content was sometimes filled with silence while the lecturer provided demonstrations or was forming the next thought. Samples were run through a Focusrite Scarlett Solo and an Irig pro to an iPad equipped with free versions of Ava and Otter.

The current study uses Otter and Ava as transcription software. Otter was chosen because it is emerging as a reliable captioning tool, used by universities, including Columbia University and UCLA. It is also used by many organizations and companies, including Google Meet and Zoom, having become very successful due to the online environment caused by the COVID-19 pandemic. Otter has both free and paid versions available to consumers, with paid services allowing more transcribed minutes and individualized experiences. Otter claims on its website to have "the most accurate automated transcription for meetings, interviews, lectures, and other long-form conversations ", however a specific percentage is not available on the website as of when this was written. Ava was chosen because it was developed by Deaf people, for Deaf people. Ava is also used in universities, including Cornell University and Universite Paris II. Ava has many features that are appealing, such as transcriptions that identify the speaker in group conversations, language settings, and the ability to request live, trained scribes that merge artificial and human intelligence, improving accuracy, punctuation, and readability. Ava also has free and paid versions; free versions currently claim 90% accuracy with unlimited caption time and paid versions offer better quality transcriptions (95% accuracy) plus punctuation and 30 minutes of scribed captions (99% accuracy), as well as other features. At the time this study began, Ava's free version claimed 80% accuracy, which was the expectation for the duration of the study. This researcher did not notice any major differences in the accuracy in transcriptions provided by Ava between when the study started and ended.

This research team decided to use the free versions of the programs because we wanted to study the most accessible versions available. Otter's paid service allows for more transcription

time, claiming that the free and paid versions have the same accuracy. Ava claims that the paid version is slightly more accurate and provides punctuation. The current study was not intended to study punctuation and, with the idea of functionality and fairness at the forefront, it was decided to compare free versions of both systems.

Previous researchers in this area and companies that use the software typically use a training process which is time-consuming and unrealistic in a university setting. Since we were interested in the ease of using these programs, Otter and Ava were not trained on any participant. This decision to use untrained programs was to showcase the effectiveness of the programs "out of the box" with as little time required from professors as possible. The same device and programs were used for all passages and lectures. This mimics how a student might use the same device between classes in a real-world situation.

Transcriptions produced by both programs were then exported from the iPad via an email option imbedded in the program, to be viewed on a computer. The entirety of the Rainbow passage and the lecture excerpt were separated into utterances. Using an Excel form, three categories of transcripts were created for each speaker: the first was a typed transcript of what was actually said by the speaker; the second was raw form of the transcript, produced by either Otter or Ava; the third was an edited version of the transcript, which highlighted errors in red and contained corrections marked in green.

Errors were then counted and categorized as caused by either Substitution (S), Omission (O) or Addition (A). Word Error Rate (WER) was calculated by adding together all errors (S+O+A), divided by the total number of words spoken. This provided an accuracy level and a comprehensive index of errors. Accuracy levels were calculated and discussed in two ways: 1) using overall mean from all speakers and 2) a mean calculated from all five native English speakers and a mean calculated from the six non-native English speaker groups. Accuracy was correlated with mean speech rate of the speaker (calculated as a number of syllables per second), mean fundamental frequency of the speaker, and comprehensibility scores, mentioned in section 2.2. Segments of speech where both Ava and Otter systems showed similar errors were investigated in detail using Praat to view the speech signal and spectrogram.

Otter and Ava are programmed to avoid disfluencies like "um" and "ah". Ava seems to delete other disfluencies, for instance, if a participant says "right" or "okay" often, as well as

consecutive repetitions of the same word or phrase. However, meaningful omissions were inconsistent; because of these omissions that seemed intentional were not counted as errors (i.e., if the omission increased readability), since they could be considered a feature of the AI, rather than a shortcoming. However, disfluencies that might have been justifiably omitted, but were transcribed differently than the speaker's intent (e.g., "right" into "write") were counted as errors.

Speech rates were calculated for each speaker in both read and spontaneous speech samples. This was accomplished by counting syllables per second over a 60 second sample. Some spontaneous speech samples contained long pauses. These pauses were usually caused by either the lecturer writing while speaking, changing slides, or preparing for the next sentence. Excerpts with pauses longer than one second were corrected by deleting the pause and replacing it with subsequent speech until 60-second segment was complete.

Since Previous research showed that SR systems are more accurate for males than females, the overall mean fundamental frequency (f_0) was collected for each participant using Praat with standard settings.

Comparisons of the read passage to the lecture by the same speaker were made regarding error rate, readability, and speech rate. Findings were then compared across participants. The results of the detailed analysis is presented in Chapter 3.

3. Results

This chapter summarizes results of the study in terms of accuracy scores, word error rate, speech rate, fundamental frequency, and the general ease of understanding.

3.1 Accuracy

Accuracy scores were produced from the transcripts generated by Ava and Otter (see figure 3.1). Generally, read speech was more accurate than spontaneous speech for both programs; however, spontaneous speech from the Chinese accented transcript was higher than the read passage for both programs. Otter generated overall more accurate transcriptions than Ava in both conditions. Otter generated a mean accuracy of 95.86% (s=3.14) for read speech and 94.83% (s=3.09) for spontaneous speech (a difference of 1.03%). Ava generated a mean accuracy of 80.41% (s=6.33) for read speech and 65.91% (s=13.49) for spontaneous speech (a difference of 14.87%).

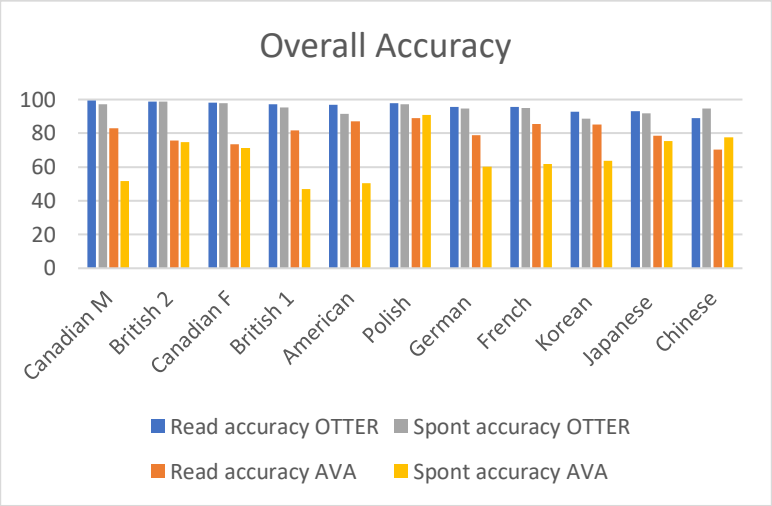

Figure 3-1 Overall accuracy for native and non-native speakers in read and spontaneous speech

Native English speakers generated more accurate transcripts than non-native English speakers for Otter, while Ava showed the opposite (see table 3.1). Otter generated an average of 98.12% accuracy (s = 1.01) from read speech and an average of 96.18% (s=2.85, n=5) from spontaneous speech for native English speakers and 93.97% (s=3.08) for read speech and 93.7% for spontaneous speech for non-native English speakers (a difference of 4.15% for the read condition and 2.48% for the spontaneous condition). Ava generated transcripts from native English speakers with 79.43% (s=6.38) from read speech and 59.03% (s=17.38) from spontaneous speech; transcripts from non-Native English speakers showed an average of 81.22% (s=6.76) from read speech (a difference of 0.97%) and 71.63% from spontaneous speech (a difference of 12.6%).

Mean	Read		Spontaneous	
	Otter	**Ava**	**Otter**	**Ava**
Overall	95.86	80.41	94.83	65.91
Native English	98.12	79.43	96.18	59.03
Non-Native English	93.97	81.22	93.70	71.63
European-accented	96.34	84.43	95.66	70.95
Asian-accented	91.61	78.00	91.73	72.31

Table 3-1 Mean accuracy for read and spontaneous speech

Averages of non-native English speakers were further categorized into European L1 speakers (German, French and Polish) and Asian L1 speakers (Chinese, Japanese, Korean). Otter generated transcripts from read speech with a mean accuracy of 96.34% (s=1.38) from European-accented English, and 91.61% (s=2.25) from Asian accented English (a difference of 4.73%). The same trend was found in spontaneous speech, where European-accented English resulted in a mean of 95.66% (s=1.31) and 91.73% (s=3.13) for Asian-accented English (a difference of 3.93%). Ava showed a mean accuracy from European-accented English of 84.43% and 70.95% for Asian-accented speech in the read condition (a difference of 6.42%). However, the trend was different in spontaneous speech, where European-accented speakers generated transcripts with a mean accuracy of 70.95% and Asian-accented speakers generated transcripts with a mean accuracy of 71.63% (a difference of 1.4%). The differences between read and spontaneous speech are outlined in the bale below (see table 3.2).

Differences in Read vs Spontaneous Speech

	Native	Non-Native	European	Asian	Overall
Otter	1.94	0.28	0.68	-0.12	1.03
Ava	20.40	9.58	13.47	5.69	14.87

Table 3-2 Differences in read versus spontaneous speech

3.2 Word Error Rate

To view the samples from a different point of view, one could present the data in terms of word error rate (WER) instead of accuracy level (see figure 3.2). Here the data show the same trend as figure 3.1, however the lower bars indicate more accurate transcripts (with fewer errors). This figure shows a more noticeable difference between each condition for each program.

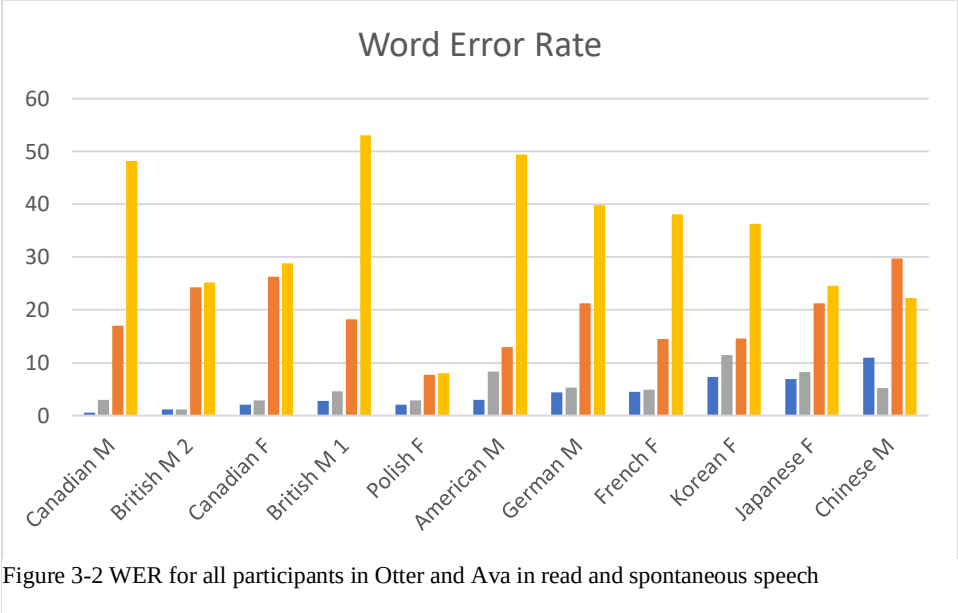

Figure 3-2 WER for all participants in Otter and Ava in read and spontaneous speech

Previous research suggests that there is a correlation between accuracy and speech rate and fundamental frequency (f_0). This study was interested whether the complexity or understandability of a lecture would affect accuracy levels of a transcript. The next section outlines results in correlations with accuracy scores (see table 3.3) to speech rate, f_0 and subjective understandability ratings from graduate students.

Participant (native)	Read speech accuracy (%)		Spontaneous speech accuracy (%)		Participant (non-native)	Read speech accuracy (%)		Spontaneous speech accuracy (%)	
	OTTER	AVA	OTTER	AVA		OTTER	AVA	OTTER	AVA
Canadian M	99.39	82.72	97.06	51.76	Polish	97.93	92.31	97.16	91
British 2	98.79	75.76	98.88	74.76	German	95.62	78.72	94.74	60.19
Canadian F	98.19	73.64	97.94	71.21	French	95.47	85.5	95.09	61.67
British 1	97.26	81.76	95.4	46.9	Korean	93.31	85.71	88.56	63.99
American	96.99	87.05	91.62	50.54	Japanese	93.69	79.58	91.81	75.44
					Chinese	89.02	70.23	94.82	77.74

Table 3-3 Accuracy of transcripts in % for native and non-native participants

3.3 Speech Rate

Correlations were gathered from accuracy to speech rates for both read and spontaneous speech for Otter and Ava (see table 3.4).

In the read condition, overall, the results showed a positive correlation with speech rate and accuracy for Otter (.592) and a moderate negative correlation for Ava (-.258). However, when participants were divided into native and non-native groups, correlations were strengthened for native speakers, but not for non-native speakers. Transcripts generated from native English speakers showed a strong positive correlation for Otter (0.839) and a negative correlation for Ava (-0.57). Non-native speakers showed weak positive correlation for Otter (0.331) and no correlation for Ava (-0.085).

Participant (native)	**Read speech rate (SPS)**	**Spontaneous speech rate (SPS)**	**Participant (non-native)**	**Read speech rate (SPS)**	**Spontaneous speech rate (SPS)**
Canadian	4.25	3.67	Polish	3.25	3.3
British 2	4.45	4.01	German	4.48	4.8
Canadian F	4.18	4.51	French	3.62	2.98
British 1	3.3	3.6	Korean	3	3.8
American	3.68	3.13	Japanese	2.75	3.2
			Chinese	3.3	3.6

Table 3-4 Speech rates from read and spontaneous audio speech samples in syllables per second

In the spontaneous condition, overall results showed a weak positive correlation with speech rate and accuracy for Otter (0.292) and no correlation for Ava (-0.02). When participants were divided into native and non-native groups, a conflicting result emerged for both systems. Otter's transcripts generated from native English speakers showed the same positive correlation as the read condition for Otter (0.839), while showing no correlation for non-native speakers, trending toward the negative (-0.093). Ava showed a strong positive correlation for native speakers (.792), but a moderate negative correlation for non-native speakers (-0.422).

3.4 Fundamental Frequency

Correlations were gathered from accuracy to fundamental frequency for both read and spontaneous speech for Otter and Ava (see table 3.5).

In the read condition, overall, the results showed no correlation with f_0 and accuracy for Otter (-0.08) and a moderate positive correlation for Ava (0.407). However, when participants were divided into native and non-native groups, conflicting correlations were found. Transcripts generated from native English speakers showed a very weak negative correlation between f_0 and accuracy for Otter (-0.163) and a moderate positive correlation for non-native speakers (0.462). Ava showed a stronger negative correlation for native English speakers (-0.520) and a strong positive correlation for non-native English speakers (0.817).

Participant (native)	**Read Speech f_0 (Hz)**	**Spontaneous speech f_0 (Hz)**	**Participant (non-native)**	**Read Speech f_0 (Hz)**	**Spontaneous speech f_0 (Hz)**
Canadian	95.84	105.96	Polish	165.61	182.39
British 2	90.19	84.78	German	97.73	87.61
Canadian F	162.47	149.86	French	186.05	183.27
British 1	105.8	88.07	Korean	176.17	200
American	105.62	92.6	Japanese	163.9	166.7
			Chinese	84.9	89.72

Table 3-5 fundamental frequencies in Hz for read and spontaneous speech

For the spontaneous condition, overall results showed weak negative correlation with f_0 and accuracy for Otter (-0.34) and weak positive correlation for Ava (0.396). When participants were divided into native and non-native groups, conflicting results emerged for Otter and different trends for Ava. Transcripts generated from native English speakers showed a weak positive correlation between accuracy and f_0 for Otter (0.311), while non-native speakers showed the opposite: a weak negative correlation (-0.346). Ava showed a similar positive correlation to Otter for native English speakers (.385), but a very weak positive correlation for non-native speakers (0.104).

3.5 Ease of Understanding

Correlations were gathered from accuracy to subjective opinions of intelligibility for both read and spontaneous speech for Otter and Ava (see table 3.6).

Participant (native)	Mean understandability scores		Participant (non-native)	Mean understandability scores	
	Read speech	Spontaneous speech		Read Speech	Spontaneous speech
Canadian M	9.5	9	Polish	10	9.5
British 2	10	10	German	9.5	9.5
Canadian F	9.5	9.75	French	8.75	7
British 1	10	8.25	Korean	7.75	5.25
American	8.5	7.5	Japanese	6.75	7
			Chinese	6.5	7.25

Table 3-6 Subjective understandability scores for read and spontaneous speech

In the read condition, overall, the results showed a strong correlation with intelligibility and accuracy for Otter (0.894) and a weak positive correlation for Ava (0.289). When participants were divided into native and non-native groups, Otter showed weaker positive correlations for native speakers (0.418) and stronger positive correlations for non-native speakers (0.914). Ava showed conflicting correlations for native and non-native speakers. Native English speakers showed a negative correlation for native speakers (-0.622) and an equally strong positive correlation for non-native speakers (0.694).

For the spontaneous condition, overall results showed similar positive correlation with understandability and accuracy as the read condition for Otter (0.879) and a very weak positive correlation for Ava (0.18). When participants were divided into native and non-native groups, positive correlation patterns emerged for both Otter and Ava. Otter showed a strong positive correlation for both native English speakers (0.964) and non-native English speakers (0.82). Ava also showed a strong positive correlation for native English speakers (0.869) and a moderate correlation for non-native speakers (0.356).

The results obtained from the study are interpreted and discussed in the next chapter.

4. Discussion

The current study was developed to investigate whether online speech-to-text applications could be utilized in a university context. In this chapter, we will discuss the findings from the focus group and the results of our experiments and how they relate to the research questions outlined in chapter one.

4.1. Focus Group

The focus group was organized with the assistance of McMaster Student Accessibility Services on January 2, 2020. The findings from the focus group largely confirmed what the report from Woodcock et. al (2007) discussed, as well as some insights that had not been considered. Most of the participants reported that McMaster University was trying to accommodate their needs, but the effects are not perfect and there are struggles. One student reported a bad experience from a professor with a heavy New Zealand accent. The professor did not want the student to use Ava in class because he didn't like the idea that there would be a transcript of his lecture. This class was particularly difficult for the student since the professor did not seem to want to accommodate any of the student's accessibility requests. The requests included: keeping projected notes up longer, facing the student while speaking, having full information on slides, and using Avenue to Learn. Another participant agreed with this specific experience. This person took this course at the same time, but withdrew since they felt that not doing so would result in a failing grade. This seems to be an isolated incident with an individual professor and was atypical of the other attendee's experiences at university and professors' general willingness to accommodate students with impaired hearing. However, this anecdote shows how some DHH students struggle with stigma and equity, as outlined in the report by Woodcock et. al.

a. Language proficiency and lip-reading

English was the first language for all the participants, and although some knew limited American Sign Language (ASL), none felt proficient enough for an ASL translator to be useful. Since English was the first language for all the participants in the focus group, this study lacks valuable insight from the Deaf community regarding ASL and their experience with language and lip-reading. The focus group reported they felt proficient in lip-reading, but lip reading is difficult, inaccurate, and context-driven and relies on the speaker to avoid speaking while turning

their back (something that is very difficult for a professor who is writing on a board). Accuracy is improved if the listener has the opportunity to request more information or clarification from the speaker. In an academic setting, relying solely on lip-reading could be an issue, especially if accuracy would require the listener to ask questions. This would result in a lecture with many interruptions and potentially anger pointed at the student, so it is not unreasonable for students to feel that they cannot repeatedly ask questions in large lectures.

b. Note-takers

All the participants in the focus group relied on note-takers. Participants reported that most note-takers are volunteers; if volunteers cannot be found, a paid person could do it. This is an issue for students who are senior; as they do not pay student fees, paid services are limited for them. Students found that, in all cases, whether a note-taker is paid or a volunteer, what the note-taker writes is subjective. They use their own ideas about what is important and rely on their own previous knowledge, which affects the notes. Notes are essentially the note-taker's perceptions of the class and what is important, and so they simply do not write everything down; one participant reported that notes were sometimes different than the textbook. One participant explained that they relied on their hearing friends both during and after class for clarification. This student did not have accommodations beyond note-takers, but also said that most of their classes were small and acknowledged that more accommodations would be necessary for larger classes.

c. Experiences at other universities:

One of the participants discussed their experience at a different university. In this case, the student was given access to a stenographer during class. The student had the ability to chat with the stenographer in real time to fix any errors. In these cases, the stenographer would go back and fix mistakes. The final transcript would be edited and extremely accurate. However, this participant discussed how they believe that courses such as science would be difficult and a stenographer might not be helpful, due to equations, technical information, etc. One of the major issues with the use of stenographers is that they are expensive. This student was able to have stenographer services because they were in a small program within a larger university. Another issue reported is that there is a lag between what is said and what is written. This was something to be expected, as some stenographers have more experience than others.

d. Current services at McMaster:

As well as providing note-takers, McMaster SAS Office offers other accessibility options, including FM systems in some (but not all) lecture rooms and Echo360, which is another automatic captioning system for video-recorded lectures. One participant used Echo360 as their main accessibility support, however none of the other students knew about it. One student did not know that automatic live transcription existed. In 2018-2019, McMaster had offered students Ava as an accommodation for DHH students to enable them to acquire transcripts as lectures were being presented. These services will be outlined in the sections below. Another important finding was the fact that participants did not immediately realize that most accessibility services do not extend to a third party. For instance, if a fellow student asks a question in class, that question cannot be heard and transcribed by the service. The DHH student hears the answer only and misses important information. The focus group also discussed divided attention (e.g., between reading the transcript and paying attention to the lecturer), which was an issue in some previous studies (e.g., Bain et al). The participants in the focus group did not find divided attention an issue in some classes; however, it is not an ideal solution in some specific courses (engineering was named). Nevertheless, the students admitted that having the transcript after the lecture is better than dividing attention between watching the lecture and reading captions.

e. FM systems:

One participant really liked FM systems (specifically T-switch with a hearing aid), but the school did not have the budget to provide one, as seniors do not pay for school and new regulations restrict budget for accessibility services. Others in the group said that FM systems make audio louder, which the participant reported isn't helpful. The other issue is that FM systems create very crisp sounding speech. Participants reported that this crisp speech is distracting, and they would prefer if FM systems were more like natural speech.

f. Echo360:

One student from the focus group had access to Echo360 transcripts. Others in the group did not know about Echo360 as it was not offered to them by SAS. The student with experience with Echo360 reported that access to the program was great. The program allows the viewer to

have both the transcript and video; it allows a user to highlight a sentence in the transcript and it then jumps to that point in the video. It is also beneficial since a student can catch up on missed course content. However, the Echo360 has issues as well. The lecturer and/or TA must be mindful of what they record and post. In a specific example, all students in a chemistry class had access to the Echo360 notes. This class had several different tutorials to accommodate the large class size. All the posted transcripts and videos were from a specific tutorial; however, the DHH student was in a tutorial different from the posted tutorial. This caused problems because the individual's experience is missing from the final document. The content presented is from an experience the individual did not have, which made relating to the content, or remembering specifics (e.g., a missed question in class which was not asked in a different tutorial), difficult. The actual output is not ideal. The student described transcripts as "just words", as there is no punctuation nor utterance breaks, making the user rely heavily on the video. The student reported that the program would glitch sometimes and not work properly and that it was heavily affected by accents (e.g., the program wrote "chicken" in a math class and the participant could not decipher the correct word based on context). For error mediation, the student can review the video and edit errors, however, the student found this option to be less effective than one might imagine. They found the content no clearer while reviewing the video, since the listener is pre-exposed to the error and thus can't process the original word. The participant had suggested that Echo360 would be perfect if a human could edit and break transcriptions into sentences and add punctuation.

g. Ava:

One participant had experience with Ava, but decided it was not suitable for their needs. They reported the following: "if the professor isn't Canadian, spelling is not correct", citing that "ready" was written as "reddit". When asked if there was anything the student liked about Ava, the student responded that that they didn't use it enough to assess anything positive about the experience. They found that Ava was heavy and cumbersome. At the time, Ava was supplied using an iPad, owned by SAS. With the case, the iPad is 25cm x 18cm and weighs 976g (2lb, 2.4oz). In order for Ava to work, the student is supplied with a lapel microphone hooked up to a Wi-Fi transmitter, a transponder, and an Irig that connects the transponder to the iPad, which weigh approximately 550g (1lb 3.4oz). Finally, students are supplied a carrying bag that, when

empty, weighs approximately 275g. Together with the iPad, the kit weighs 1.8kg (3lb 15.5oz). For comparison, a 634-page hardcover textbook weighed 1255g (2lb 12.3oz). It would be very difficult for a student to carry around an extra 1.8kg of school supplies, as well as navigating the cords and being mindful not to break the kit. It was a very challenging experience for the student and was largely the reason the student gave the equipment it back.

This researcher asked the SAS office why they use the cumbersome Wi-Fi option, as opposed to a Bluetooth system. The reason given was that Wi-Fi is reliable and is capable of a much longer range. Bluetooth might cut out unexpectedly or may be interrupted by conflicting Bluetooth signals if many students are accessing unrelated Bluetooth systems. Also, the lecturer may walk out of range of the Bluetooth signal, which would result in no transcript. So, although Bluetooth is more convenient, there are more risks. Wi-Fi is more reliable. As the bottom line for SAS is lecture accuracy, the Wi-Fi microphone system is what is supplied by the SAS office.

h. Otter:

One student was aware of Otter, though at the time it was not an option for SAS accommodation services. The student downloaded the program for personal use on their own phone. The student reported that Otter is only good for short-range communication and that accuracy is affected by accents. The participant did not attempt to use Otter in lecture settings. The student liked that Otter could identify individual speakers and has decent accuracy. The student recognized that the AI enables Otter to learn with more use to become more accurate, and that Otter self-corrects.

i. Regarding mistakes:

All participants had experience of a time when a mistake occurred in either the notes provided or in a computer-generated transcript that was not realized until much later. In such cases, errors lead students to form erroneous beliefs about the course content, which could result in poor grades on tests and exams. All the participants believe that there are mistakes in their notes and transcripts that are not noticed, which can cause anxiety.

As DHH students are sensitive to transcript errors, it was important that we ask what experience these students had with transcript errors. We asked what kind of errors they thought would be worse than others. Participants agreed that discipline specific jargon might be an issue

for speech-to-text systems and expected that classes with many low-frequency words would result in inaccurate transcripts. The participants discussed how content word mistakes (i.e., words that hold meaning) seem to be most noticeable and easy to describe, but function word mistakes (i.e., words and morphemes that hold grammatical information, such as tense and plurality, including determiners and negation) affect flow and understanding. The group had decided that both types of mistakes are equally problematic; however, some mistakes are worse than others: Any mistake slows down the student's ability to grasp content. In social situations, participants reported that mistakes affect human interaction, for instance missing a joke.

4.2 Speech-to-text

4.2.1 Transcript Readability

Intentional omissions

Both Otter and Ava attempt to increase readability by omitting disfluencies such as "um" and "ah". Ava seems to attempt to omit repetitions and habitual filler words (e.g., "like" and "okay"), but does not always succeed. Sometimes these omissions lead to more errors as these deletions do aid in the legibility of a transcript. It would be interesting to observe a transcript that did not delete these words, since sometimes they can help the reader identify when a speaker misspeaks or needed to reword something, as in the following example:

Original:	first	ways	in	which	languages	were	type	uh,	typologized,	if	you	like
Transcript:	first	ways	in	which	languages	were	type		pologized	if	you	like

The original utterance, with the disfluency, illustrates that "type" is not a separate part of the intended content, which may be more legible to some readers.

Output presentation

Both Otter and Ava offer online captions, that is, captions that are created as the speaker is speaking. Observations of real-time captions are discussed here. Technological changes happen so quickly, especially as lectures were pushed to online format due to the COVID pandemic; it is difficult to comment on all aspects of this ever-changing situation. This project did not include the analysis of the video of software as it captioned a speaker, but it would be interesting to involve video screenshots as a more robust observation. Generally, captions were created with very little lag. Words that had already been written were being adjusted as the speaker continued, making it difficult to follow when errors occurred. Though both programs did

this, Ava's post-analysis was particularly confusing. Sometimes the post-analysis would correct an error. However, on more than a few occasions, an utterance would be written correctly, but would change into an error in post-analysis. For example, a participant said, "Dark Ages". Initially, the program wrote, "Dark Ages", but changed it to "Dog Cages" in post-analysis. Sometimes Ava would transcribe a phrase and then delete it completely. Otter also performs online post-analysis; however, it also takes time after the recording is finished to do another post-analysis cycle. This process can take anywhere from a few seconds to several minutes.

Once a lecture is finished, Ava and Otter output a large quantity of text that may be difficult to read. The two programs have different approaches to combat readability. Both attempt to insert punctuation, however, it is not always appropriate and could change the meaning of a sentence. Errors in punctuation were too numerous and varied to examine in this paper, so errors caused by omitted or misinterpreted punctuation were ignored.

Ava seems to attempt to separate output into utterances. However, the utterance breaks are often inappropriate and are surrounded by errors. Ava seems to create utterance breaks between spoken words somewhat randomly. Pauses in speech do not always cause utterance breaks, as sometimes a speaker may pause for a long time and the program does not perform an utterance break. Sometimes the software will create a break when there is no appropriate pause and then omits one or skips several words, which makes it seem like the speaker stopped mid-sentence.

Otter creates a solid block of text. Sometimes it will give timestamps, but this seems to happen either between very long pauses or when it identifies a different speaker. However, Otter creates an audio recording of the speaker as it transcribes, as well as a list of keywords that can be searched within the document. Users can navigate through the transcript and listen to the audio while reading. This seems like an excellent way to battle readability and error mitigation, however one of our participants in the focus group pointed out that this was not ideal (see section 4.1f).

4.2.2 Accuracy

Accuracy ratings and programmer expectations

Previous research indicates that read speech results in more accurate speech recognition than spontaneous speech. Our results provide evidence that supports this claim. The drop in

accuracy for Otter was less significant than for Ava, which suggests that Otter is more reliable than Ava. If a program is claiming a certain accuracy level, it is not unreasonable for the public to desire (or assume) that the advertised accuracy level would be for everyday, spontaneous speech. What was surprising was that the transcript from spontaneous speech was more accurate than the read speech for our Chinese L1 participant. Specifically, 5.8% more accurate for Otter and 7.51% for Ava. It's difficult to reason why this might happen. Perhaps in read speech, there is more pressure to produce English phonemes that do not exist in Chinese, which results in more inaccuracies. Conversely, in spontaneous speech, it is difficult for a speaker to concentrate on pronunciation and content at the same time, which results in more fluent speech. This was evident in our understandability experiment. The volunteers generally rated read speech as more understandable than the spontaneous speech for all lecturers, with three exceptions. Spontaneous speech for the Chinese speaker was rated on average, 0.8 "points" more understandable than read speech. To a lesser degree, the volunteers rated one Canadian and the Korean lecturers as 0.3 "points" more understandable during their spontaneous speech than their read speech. It would be interesting to try this experiment again with more volunteers to help validate the findings. The correlations between understandability and transcript accuracy were shown for Otter; generally, the same correlation was found for Ava, except the negative correlation for native speakers during read speech suggests that less clear speech, under certain circumstances, may provide a more accurate transcript than one might expect.

Both Ava and Otter claim to have very high accuracy rates, but the actual number is no longer available on the Otter website. Currently, on Ava's website (August 2021), there is a graphic that shows "Free & Unlimited Automatic Captions" with "90% accuracy" below, demonstrating "Ava uses AI to transcribe instantly what people saytwo[2] further improve quality algorithms then add punctuation speakers and vocabulary from your diction airy[3]" [sic]. The site then describes premium captions (i.e., paid) as having 95% accuracy, seeming to improve the caption to "Ava uses AI to transcribe instantly what people say, then, to improve quality, algorithms add punctuation, speakers, and vocabulary from your diction airy[4]" [sic]. Ava then

[2] This error is as shown on the Ava website to exemplify what 90% accuracy may look like
[3] This error is as shown on the Ava website to exemplify what 90% accuracy may look like
[4] This error is as shown on the Ava website to exemplify what 95% accuracy may look like, along with the addition of punctuation

describes their scribe service, which is another paid service that uses humans to further improve the captions.

In 2020, Ava claimed to have over 80% accuracy for the free version of their program. This held true in our experiment for speech averages overall in the read condition; however surprisingly, transcriptions from native English speakers achieved an average just under 80%. To understand more clearly, we can observe the standard deviations from Ava's output. Ava showed wide standard deviations between participants, which indicates that accuracy levels are inconsistent. This means that, in the real world, a given read transcript may not achieve the promised accuracy level. Pertinently, our findings showed that Ava was unable to meet an 80% average for spontaneous speech. We have no reason to believe that the paid version cannot achieve what the company claims; however, the free version showed a huge decline in accuracy for spontaneous speech. It stands to reason that even if the paid version would achieve high accuracy levels for read speech, a similar decline may be expected for spontaneous speech. This indicates an issue for live lectures and meetings. For this reason, one should take accuracy claims with a grain of salt because they are tested under ideal circumstances with read speech, trained speakers, and little to no background noise. For the purposes of speech-to-text in academic settings, the read speech rate is not necessarily relevant to our purposes, but does offer some valuable comparison observations between speakers. Based on the data, accuracy ratings are highly individualized, as seen with our two British speakers. Both were males with South-Eastern accents but yielded different results, albeit the difference between the two speakers was more apparent with Ava than Otter.

Overall, Otter produced more accurate transcriptions than Ava. The standard deviations were lower, meaning that the Otter's average accuracy was more consistent than Ava's and finally, the difference in read vs. spontaneous speech was much lower. This could be due to the use of Ava's free version, as the creators state that the paid version is more accurate than the free version, however the discrepancy between users, combined with the fact that Ava seems to perform worse for some native accents than the advertised 80% indicates a potential issue when relying on this program for academic notetaking. The results could be very different depending on the lecturer, which might make a student more frustrated with the automatic service than volunteer or paid note-takers.

Lectures were captioned over the span of 18 months. Speech recognition claims to consistently be improving and there's a possibility that some speakers were captioned using a different version than what is currently available; however, this researcher ran trials of the same lecture from 2020 in 2021 (which are not part of the data set) and found that, although some errors were fixed, new errors emerged, and omissions were still an issue for Ava. Not much difference was found in overall accuracy over the span of a year.

Accuracy in the real world

It is difficult to define accuracy and describe what that means using numbers. "Over 90% accuracy" seems like a great system, but, broken down into what that means is often underestimated. For instance, we can compare transcripts from the read speech samples (see table 4.1). Accuracy levels varied between 92.71-99.39%, however, errors varied between 2-24 for the same read script.

Speaker	Accuracy	Words	Substitutions	Omissions	Additions	Total
Canadian	99.39	330	2	-	-	2
British	98.8	330	2	2	-	4
Canadian	98.18	331	6	1	-	7
Polish	97.93	338	7	-	-	7
British	97.26	329	8	1	-	9
American	96.99	332	8	2	-	10
German	95.62	343	11	3	1	15
French	95.47	331	13	1	1	15
Japanese	93.09	333	17	2	2	21
Korean	92.71	329	20	1	3	24

Table 4-1 Number of errors within transcripts in the read condition with over 90% accuracy

It is important to note that word count is very significant for these accuracy levels. Even though participants were given the same script, some participants omitted or added words or phrases, which may skew the data when comparing samples.

These samples show utterances with many errors and others with no errors. The accuracy is averaged out for the whole sample. Therefore, specific areas of a sample could, hypothetically, be selected by a company to demonstrate the software at its best. To show what "over 90%" might mean, samples of utterances with notable errors were collected in the appendix. Here we will demonstrate a few of them. Below is an utterance from two transcripts, one is a transcript

with 99.39% accuracy and the other with 93.31% accuracy. The first utterance below is the former, which contained both substitution errors in the same utterance, whereas the rest of the transcript was errorless:

Original:	the	actual	primary	rainbow	observed	is	said	to	be	the
Transcript:	the	actual	primary	rainbow	observed	is	said	to	be	in
Original:	effect	of	super-imposition	of	a	number	of	bows		
Transcript	effect	the	super imposition	of	a	number	of	bows		

We can compare this to an utterance from the sample with 93.31% accuracy. This transcript had errors in other utterances as well as in this example:

Original:	since	then,	physicists	have	found	that	it	is	not	reflection,
Transcript:	since	then,	please he says to	have	found	that	it	is	not	reflection,
Original:	but	Refraction	by	the	raindrops	which	causes	the	rainbows	
Transcript:	but	we fraction	by	the	raindrops,	which	causes	the	rainbows	

Even though over 90% accuracy suggests an impressive transcript, that number does not tell the whole story for two reasons: Firstly, specific errors are not shown. It is possible that the only error in the transcript completely alters the meaning of that sentence and potentially of the entire message (which will be discussed in more detail below); Secondly, accuracy is judged based on errors that differ from what was said. In the example above, “refraction” (one word), is substituted by two words “we fraction”, but it is still considered a single error. The same can be applied to “please he says to” from “physicists” also only counts as a single error. To demonstrate the issue, 90% average can be interpreted as “one error for every 10 words”. The example above shows 19 words spoken, but 6 words from the transcript are incorrect. This would make the accuracy of the computer-generated transcript 68.42%; however, since the accuracy is based on word error rate, the transcript has a reported accuracy of 89.47% (i.e., 2 errors in 19 words). It would be unfair to report that the program is 68.42% accurate for this speaker, as this describes neither what happened, nor the accuracy of the rest of the document. To gain full understanding of accuracy, we analyzed what types of errors were made and how they affect the overall readability of a transcript.

4.2.3 Error Analysis

4.2.3a Error types

As indicated above, some errors are worse than others. I have classified three types of errors, based on their effect on overall understanding. They are as follows: Insignificant errors, Obvious errors, and Critical errors. Insignificant errors are errors that do not affect understanding at all and will not affect comprehension of the message, whether they are noticeable or not.

Insignificant errors

Insignificant errors are errors that make no changes to the overall meaning of the transcript and do not require much, if any processing power to understand.

Example 1

Original:	…which	otherwise	doesn't	use	morphology	for	inflectional	purposes	at	all
Transcript:	…which	otherwise	doesn't	use	morphology	for	inflectional	purposes	-	-

Here, the application omitted the final phrase, “at all”. That utterance is extraneous to the listener’s understanding the meaning of the sentence, thus losing it in the transcript is not an issue; however, in real-time, the student may notice that something was said and not transcribed, which could lead to confusion or anxiety.

Example 2

Original:	where	people	started	to	think	about	languages
Transcript:	where	people	started		thinking	about	languages

Here the program replaces “to think” with “thinking”. The temporal and overall meaning is preserved in the erroneous transcript. In real-time, this error might be noticed by a student, but with the context, the error may not pose a threat to processing speed.

Example 3

Original:	you	have	to	remember	what	the	definition	of	closed	is
Transcript:	you	have	to	remember	what's	the	definition	of	closed	-

Here the AI seems to take over the entire sentence, changing the syntax, but not the overall meaning. It is considered erroneous because the words spoken do not line up with what was transcribed. The transcribed version is more awkward than what was said and it

does not line up with what was said in real-time, which could lead to processing issues if students are lip-reading.

Obvious Errors

Obvious errors are clearly mistakes. Students might be able to mitigate misunderstandings caused by these errors. For instance, they can ask for clarification either during of after class, or in some cases, may use reasoning to figure out what was meant.

Example 1

Original:	one	cause	of	Depopulation	was	migration
Transcript:	one	cause	of	D population	Was	migration

Here, the error is clear. “D” in “D population” clearly does not belong in the sentence, but a student might be able to figure out that it was supposed to mean “depopulation”. However, some students may have an easier time with this than others. Errors like this take extra processing time, even if they may be easy to recover.

Example 2

Original:	Japan's	first	animation	studio	*Kitayama*	*Eiga*	*Seisakusho*	was	founded
Transcript:	Japan's	first	animation	studio	Kitayama	a	say sex shop	was	founded

Here, the occurrence of an error is also clear. A lecturer is unlikely to say “a sex shop” during a lecture about animation. In a lecture on a foreign culture, it is likely that the foreign phrase “Kitayama Eiga Seisakusho” is posted elsewhere for reference, and the error was caused by the original foreign word; however, the error still takes processing time to correct.

Example 3

Original:	sometimes	we	talk	about	first	language	and	second	language
Transcript:	sometimes	we	talked	about	first	language	and	second	language

Here, the syntax does not quite fit, but the overall message remains intact. The transcript might require extra processing time, as the past participle changes the interpretation of the sentence, but the error is not detrimental to overall comprehension.

Critical Errors

Finally, critical errors are errors that are undetectable syntactically or semantically. The overall message makes sense, but the content is erroneous. These are often cases where negation is added or omitted, or when an error changes a significant part of speech. In these cases, errors can drastically change the meaning of a sentence.

Example 1

Original:	Until	the	mid	1930's	Japanese	animation	used		cut-out	animation
Transcript:	until	the	mid	1930s	Japanese	animation	used	to	cut out	animation,
Original:	instead	of	cel	animation						
Transcript:	instead	of	cell	animation						

Here, the past participle + adjective + noun “used cut-out animation” was changed to the infinitive form of the verb + noun “to cut out animation”, suggesting something very different than the intended meaning. Without the original for reference, a student may be led to believe that Japanese animators used to physically cut out their animations, rather than using the “cut-out” style. The second error, “cell”, which should have been “cel”, from “celluloid film” would be an example of an insignificant error, as this is an error that a student might make listening to the lecture, and it probably wouldn’t affect their overall grade. It should be mentioned that later within the same transcript “cel” is written correctly. In this case, the student may realize the error while reviewing the transcript.

Example 2

Original:	So	the	idea	here	with	number	5	is,	"Why	would	we	make	mistakes
Transcript:	So	the	idea	here	with	number	5	is	Why	do women		make	mistakes

Here, the substitution goes wrong, which not only changes the meaning of the sentence, but indicates a controversial opinion which was never expressed in the lecture. Not only

is the student receiving an erroneous message, but an error like this could be detrimental to the student's view of the lecturer.

Example 3

Original:	Dialectics	implies	a	thesis,	antithesis,	and	what	is	called	an	*Aufhebung*
Transcript	dialectics	implies	a	thesis	and criticism,	-	what	is	called	an	alpha table

Here, completely erroneous information is presented in the transcript, and the error might easily go undetected. "Dialectics implies a thesis and criticism" is syntactically and semantically plausible; however, not only does it not convey the intended message, but also presents an erroneous definition. The error, *alpha table* is also a significant error, but it is possible that the correct word is presented somewhere else, thus fitting more into the "obvious error" type.

Classifying errors in this way is highly subjective. Not all errors easily fall into these categories, for instance, large chunks of omitted speech. It may be clear to the student that information is missing, but finding that information without a recording or a friend could be difficult.

It is worth mentioning that the error "alpha table" from "Aufhebung" and "a say sex shop" were not counted as an error in the accuracy rate presented earlier in this paper. As mentioned in the methodologies chapter of this paper, foreign words that resulted in errors were not counted as errors. However, in academia, non-English words are often used to describe or label concepts or objects. In these cases, transcripts must be edited, because these terms may be crucial for comprehension and, in turn, may affect student's performance on tests or assignments. If transcripts cannot be edited and a student were to rely on electronically generated notes, it may reduce the types of classes this student could enroll in.

4.2.3b Causes of Errors

Jargon, abbreviations, and foreign words

It is very difficult to assess how a program will perform with low-frequency words and jargon. The data collected showed inconsistencies with some words and effects of AI learning with others. This is demonstrated in tables 4.2 to 4.4. Table 4.2 shows low-frequency words that were uttered several times in the same lecture in the leftmost column. The middle column shows

how Otter transcribed that instance of the word and the rightmost shows how Ava transcribed the same instance of that word. Words in red show errors. Words in orange are noteworthy. A dash indicates an omission.

Otter was able to correctly transcribe 12 of the 25 occurrences of the words, however, errors are inconsistent. The word, pharynx, was uttered 11 times; 3 of those times Otter transcribed it incorrectly, as “fairing” or “parents”. Otter was able to approximate “laryngopharynx” and “nasopharynx” with arguable accuracy but failed at “oropharynx”. Ava, conversely, was able to correctly transcribe only 4 of the 25 words; however, it was able to transcribe “pharynx” and “larynx” correctly once but was unable to do it in other contexts. This shows that the word is in Ava’s vocabulary, but the system does not access it appropriately. Puzzlingly, Ava was able to correctly transcribe “laryngopharynx” and “oropharynx” and was able to approximate “nasal pharynx”.

Similarly, abbreviations are difficult for SR programs to transcribe (See table 4.3). Here the speaker says “L1” and “L2”. Otter is never able to get it quite right, but learns to approximate into “L to”, “L one” and “L/l two”. Ava was able to transcribe “L2” correctly once, but then changes their choice to an error.

Original	Otter	Ava
pharynx	pharynx	parent switch
pharynx	pharynx	fairest
pharynx	pharynx	parents
pharynx	pharynx	pharynx
pharynx	fairing	parents
pharynx	pharynx	parents
pharynx	pharynx	parents
pharynx	parents	parents
velum	Vilem	elem
pharynx	parents	parents
velum	venum	wheel on
pharynx	pharynx	-
oral tract	arbitrator	electric
oral tract	arbitrator	electric
auditory	auditory	Play Toby
larynx	larynx	larynx
larynx	larynx	lyrics
larynx	larynx	-
larynx	learnings	lyrics
pharynx	pharynx	parents
laryngopharynx	laryngeal pharynx	laryngopharynx
oropharynx	our pharynx	oropharynx
velum	wheel on	building
velum	bottom	-
nasopharynx	nasal pharynx	nasal pharynx
	12/25	4/25

Table 4-2 Original word spoken VS transcribed word produced by Otter and Ava

As expected, foreign words are an issue for speech recognition, but is also inconsistent (see table 4.4). This table shows utterances from a lecture with many Japanese names and words. Even though foreign words were not counted in the accuracy level, it is prudent to view what might be transcribed in their place. The table below shows some surprizing abilities for the speech-to-text programs observed. “Kanto”,

which is a foreign word, but not unrecognizable by English readers, could not be transcribed by Otter, but Ava was successful. “Murata”, a Japanese name, was transcribed accurately by Otter. Kenzo Masaoka, another Japanese name, was transcribed by Ava; Otter was able to transcribe “Kenzo”, but created an approximation of “Masaoka” into “masa oka”. Otter was able to approximate “Touki” (pronounced [to:ki]) into Toki, omitted only two syllables from “Chikara to onna no yono naka”, and created a word break in “Chaggama”. Considering foreign words should be incredibly difficult to transcribe, the speech-to-text technology shows surprising potential.

Original	Otter	Ava
L1	-	alarm
L2	-	L2
L1	a one	everyone
L2	L to	L2
L1	one	a one
L1	L one	everyone
L2	L two	02
L2	l two	02

Table 4-3 "L1" and "L2"spoken VS transcribed word produced by Otter and Ava

Original	Otter	Ava
Eiga Seisakusho	a say sex shop	C stock show
Great Kanto earthquake	great condo us quake	great Kanto earthquake
Yasuji Murata	CG Murata	yesterday
Noburo Ofuji	no Budo off is he	I
Kenzo Masaoka	Kenzo masa Oka	Kenzo masaoka
Michio Seo	neatly yourself	I need co
Masaoka	Master, Olga	-
Touki	Toki	turkey
Chikara to onna no yono naka	Chikara on No, yo no naka	Y'all know that got
Chagama	Chagga ma	chair
Seo	cell	Phil

Table 4-4 Foreign word spoken VS transcribed word produced by Otter and Ava

4.2.3c Speaker-Specific Habits and Language Features

Accents

It is widely accepted that accents affect transcript accuracy. One possible reason for errors being more prevalent in accented speech is phonotactics. Every language has rules governing how phonemes work together within a language. Some combinations are legal in some languages, but illegal in other (e.g., /pf/ is legal in German, but not English). Similarly, some languages do not contain sounds that are in English and vice-versa. For instance, the /Θ/ and /ð/ sounds are very common in English words, however they are rare sounds world-wide. Many non-native English speakers may replace the voiceless and voiced interdental sounds with a variety of other sounds that are prevalent both in English and in their mother tongues (e.g., /z/, /d/, or /t/). The speech-to-text programs observed seemed to be able to mitigate these and other common mispronunciation of English phonemes. Rather, it is often a combination of slight mispronunciations that cause errors.

Before examining samples from the data, one should consider the complex nature of vowels. While every language contains vowels, some languages contain more vowels than others. English distinguishes approximately 14-20 vowels. Japanese, however, contains only 5. Vowels are produced using different areas of the mouth, called vowel space. For example, /i/ (e.g. [iʧ]; “each”) is produced in the high-front area of the mouth. Conversely, /u/ (e.g. [but]; “boot”) is produced in the high-back area of the mouth and /a/ is produced in the low-central part of the mouth. Humans perceive sound using categorical perception, which is a term that describes how a human might hear an unfamiliar sound that is located between two familiar sounds but will subconsciously choose to hear either one of the familiar sounds or the other. This is exemplified in the sound /ɪ/ (e.g. [fɪt]; “fit”), which, like /i/, is a high-front vowel, but lax, rather than tense. If a listener’s language does not contain the /ɪ/ sound, they may hear that sound as /i/, making [fɪt] sound like [fit] (“feet”); when speaking, that same listener will produce the word “feet” instead of “fit” because their perception of the /ɪ/ sound is /i/. Speech-to-text may be able to use formant frequencies to detect which vowel is being said, and then modify it using AI and probability to choose which word to transcribe. Again, speech-to-text samples studied seem to be able to mitigate many errors caused by vowel confusion using probability and AI; errors

arise when several of these mispronunciations happen within the same word or cluster of words. For example:

original:	with	its	path	high	above
transcript:	with	This[1]	path	how you bob[2]	

This example shows the following errors: 1) “its” substituted with “this” and 2) “high above” substituted with “how you bob”. On the surface, it seems that error 1 was a result of the program inserting a /ð/ into the onset, and the /t/ was deleted. Error 2 looks very strange. However, upon closer inspection, one discovers a different pattern (see figure 4.1)

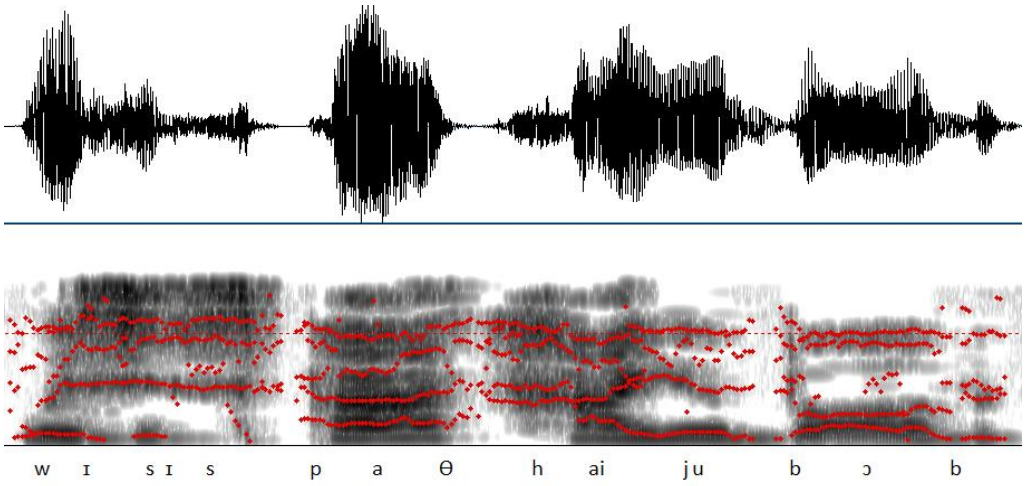

Figure 4-1 Speech signal and spectrogram of "with its path high above" by non-native English speaker

The /Ɵ/ in “with” sounded like /s/ and merges into “its” sounding like “wissis” and no evidence of a /t/. This caused the substitution error “this”; the second error was caused largely by the /v/ in “above” sounding like /b/, resulting in [bɔb], rather than [bɔv]. Just before this, the /u/ sound was preceded by /j/ when the speaker went from /i/ to /u/, resulting in the addition, “you” in the transcript. Instead of interpreting [hai ubɔb] as “high above”, the AI seems to take over to make sense of what happened, leaving with “how you bob”, even though the vowels in “high” vs “how” are very different.

Another issue with vowels is epenthesis. Some languages do not allow certain consonant clusters, so speakers will add a vowel between phonemes in their L2 to match the phonotactics of their L1. The following example is from a Japanese L1 speaker, where words cannot end in a consonant, with the exception of a nasal. The original utterance was “and without detail”, but the

program transcribed “and without the detail” due to the speaker’s epenthesis between “without” and “detail”. See figure 4.2 for the spectrogram.

Original:	and	without		detail
Transcript:	and	without	the	detail

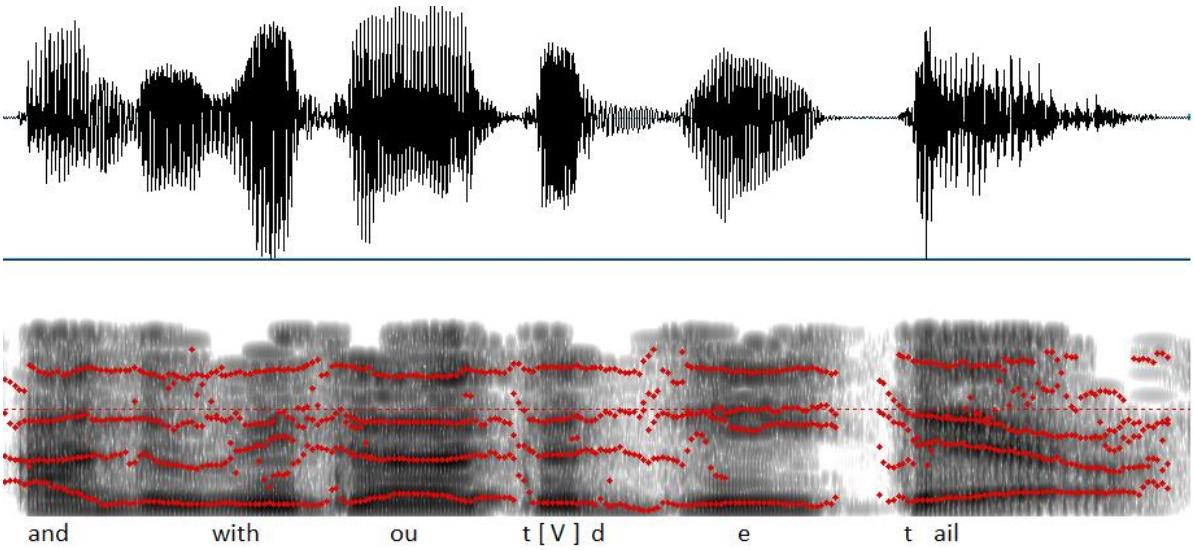

Figure 4-2 Speech signal and spectrogram of "and without detail" by non-native English Speaker

The spectrogram shows the epenthesized vowel ([V]) between “without” and “detail”, which caused the program to interpret a “the” between the two words. If this were a native English speaker, the vowel could easily be meant as a determiner, however since the speaker is Japanese, the vowel in question is a feature of the accent, rather than a deliberate sound. There is also an epenthesized vowel between “and” and “with”, that seemingly went undetected by the software. It seems this is because the utterance is highly assimilated, where the vowel in question is distinct from the sounds around it. The issue of whether errors caused by epenthesis should be considered further. Though it is not unreasonable for a program to have interpreted an inserted vowel as a word, that word was not intended by the speaker; nor was it the speaker’s fault for uttering a vowel caused by their subconscious phonotactics created by their L1. Questions like this should be addressed when classifying and counting errors by SR programs and made transparent when outlining accuracy and program expectations.

Native Speakers

Just as foreign accents can affect the accuracy of a transcript, mastery of the language may affect accuracy as well. Sometimes native speakers are so fluent that they make errors that go undetected by human perception, for an example of this, see figure 4.3.

Original:	The	first	one	it	mentions	is
Transcript:	–	first	one	is	maintenance	–

Above is an example from a British speaker who says, “the first one it mentions is”. The transcript shows four errors: an omission of “the”; two substitutions: 1) “it” to “is” and 2) “mentions” to “maintenance”; and the omission of “is”.

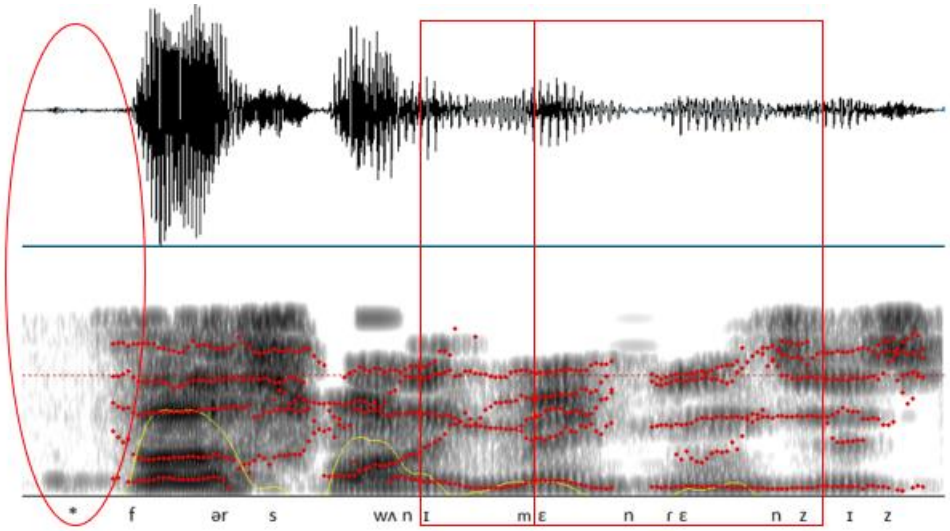

Figure 4-3 Speech signal and spectrogram of "the first one it mentions is" by native English speaker

The initial deletion is caused by a feature in British English called a silent breath pulse. The spectrogram shows a small amount of energy where a listener might repair this speech signal and hear “the”, even though it’s barely there. The machine, however, did not interpret this energy as speech – there’s simply not enough information in the speech signal. The second error happened because the speaker deleted the /t/ going from the high vowel /ɪ/ to the nasal /m/. It seems the software’s AI guessed what was said based on the words around it. Naturally after “first one”, “is”, is more frequent than “it”, even though there’s no indication of frication in the soundwave. The error is then solidified by the third error, substitution of “maintenance” from “mentions”. The speaker performed a clitic, which is another feature of some British accents where vowels or consonants are severely reduced. Here, the /ʧ/ in “mentions” is reduced to an

approximated flap, resulting in speech that sounds like “mendens”. Similar to the other errors, a human would possibly repair the speech signal and hear what was intended, especially in the context of the overall lecture, but the same cannot be expected for the software.

The final omitted “is” could be caused by a number of things. Word boundaries can be an issue for speech recognition, as seen in previous examples. However, in this case, we are unsure if it a pure omission of the final two phonemes, or if it substituted “mentions is” with “maintenance”, since both phrases are three syllables.

The data showed that native speakers sometimes neglect to say every word in a sentence or reduce certain sounds. A human listener will interpret the missing or reduced word without recognizing it was missing. In these cases, SR programs are at a disadvantage: there is no cue from the speech signal to insert the missing word and programmers, being human, repair the speech signal subconsciously, interpreting the result as an error without cause. As transcripts are prone to errors caused by clippings, contractions, or reductions (which are common for native English speakers who have mastered the language), we compare non-native speakers with highly accurate transcripts.

It is possible that the higher-than-expected accuracy produced by some non-native speakers could be that these non-native speakers are more aware of their accents and try to enunciate every consonant when possible, unlike a native speaker. This was evident in the results from Polish accented samples, as the participant spoke with clear enunciation, even though the accent was apparent. Volunteers rated this participant’s read speech as 10/10, or perfectly understandable, while spontaneous speech was rated 9.5/10 on average, which were the highest intelligibility scores from non-native English speaker samples.

4.3.2d Correlations

Transcriptions created by Ava contained many errors, but most seemed to be caused by various levels of omission, i.e., part of a word, whole words, or phrases. On observation, as Ava transcribed speech, it seemed to have difficulty transcribing utterances during post-analysis, at the same time failing to transcribe what was being said while it focused on deciding on what had already been said. Previous research had identified speech rate as being a cause for SR errors.

Speech rates were calculated to observe whether there was a correlation between faster speech and error rate.

Correlation of speech rates and accuracy

Previous research suggests that SR programs are more accurate for slow speech than fast speech. For this reason, speech rates were correlated with accuracy levels for read and spontaneous speech. Results are more evident in native-English vs non-native English speaker results.

Otter

See figure 4.4. Otter showed positive correlations for both native and non-native speakers in the read condition, but conflicting outcomes for non-native English.

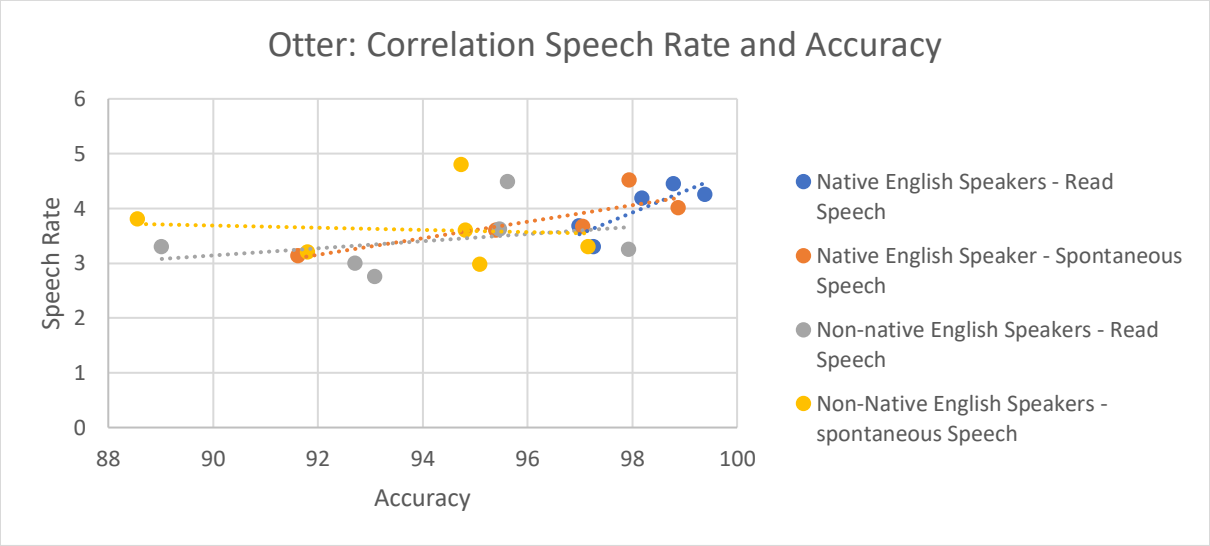

Figure 4-4 Correlation of Speech Rate and Accuracy for Otter

Correlations were stronger for native speakers (0.839) than non-native speakers (0.331) in the read condition and the same correlation for the spontaneous condition for native English speakers. This suggests that Otter consistently performs better with faster speech than slow speech when the speaker is a native English speaker, which contradicts both our expectations and previous research. However, in spontaneous speech, there seems to be no correlation between non-native English and speech rate; in fact, it tended toward the negative (-0.093). It seems that Otter prefers fast speech from native English speakers, but there is no definite pattern for non-native English speakers. It is interesting that Otter seems to prefer fast speech with native

speakers, as fast speech often results in disfluencies, and, as we show later in this chapter, predictable errors in transcription.

Ava

See figure 4.5. Ava showed conflicting results for native English speakers. The data showed a negative correlation in read conditions (-0.57), which indicates that Ava works better with slow speech; however, it showed a positive correlation in the spontaneous condition (0.79), which indicates that faster, spontaneous speech leads to more accurate transcriptions. However, previous research that showed more accurate transcriptions for slow speech were done in ideal situations with read speech. It is possible that spontaneous speech shows a different trend.

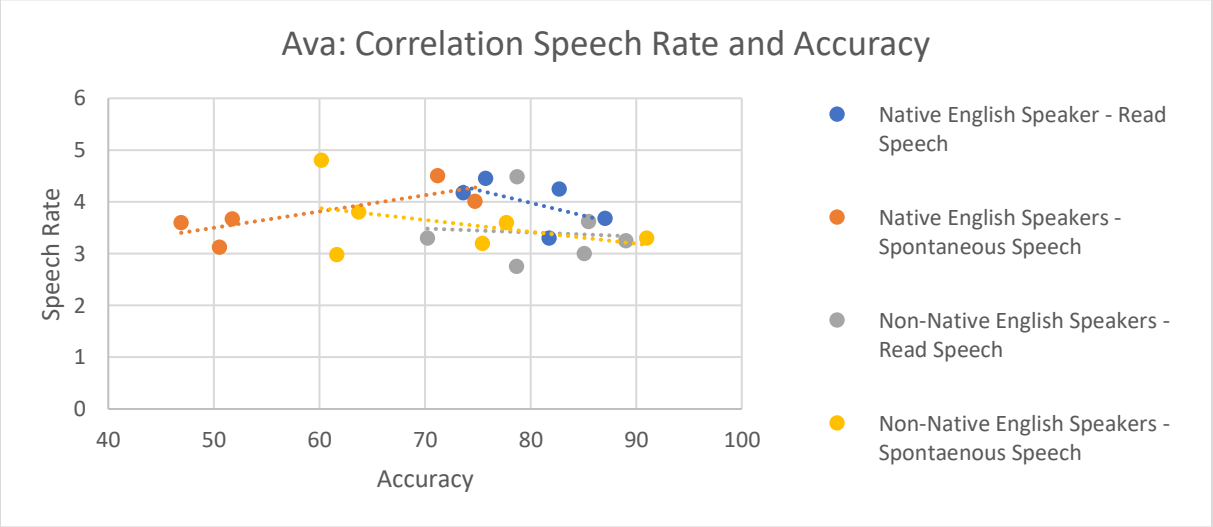

Figure 4-5 Correlation of Speech Rate and Accuracy for Ava

Transcripts generated from non-native English speakers tended toward the negative. Though Ava showed no correlation in the read condition (-0.085), there was some indication that slow speech led to more accurate transcription in the spontaneous condition (-0.422). The discrepancies between all conditions and participants seems to reflect of the wide arrays of accuracy between participants. It is very difficult to rationalize how correlations work with accuracy for Ava without knowing the inner workings of the system. It is very difficult to find a pattern with a system that varies so much between participants.

Correlation of fundamental frequency and accuracy

Previous research suggests that SR programs work better for men than women. For this reason, fundamental frequency was correlated with accuracy for read and spontaneous speech. Results are more evident in native-English vs non-native English speaker results.

Otter

See figure 4.6. Otter did not show strong correlations between f_0 and accuracy, and any correlations it did show were inconsistent between groups and conditions. For instance, native English speakers showed a weak negative correlation in the read condition, and a moderate positive correlation in the spontaneous condition. Non-native speakers showed a positive correlation in the read condition and a negative correlation in the spontaneous condition.

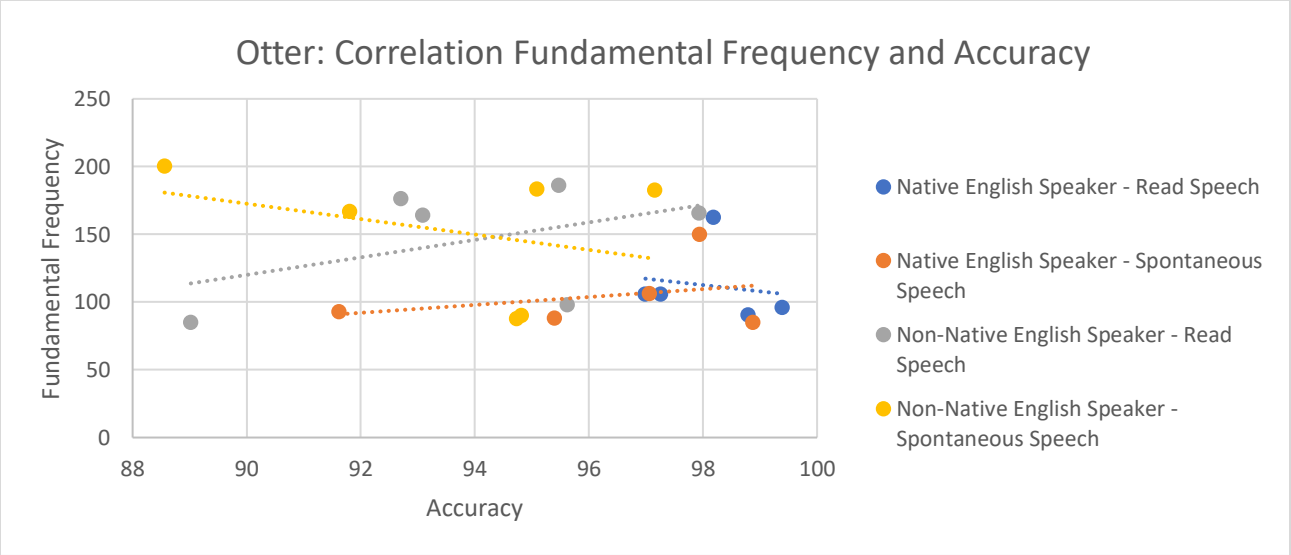

Figure 4-6 Correlation of Fundamental Frequency and Accuracy for Otter

Ava

See figure 4.7. Ava showed a strong positive correlation between accuracy and f_0 in the read condition for non-native English speakers (0.817), which suggests that Ava performs better for participants with a higher f_0 (and, therefore women should have more accurate transcriptions); however, it showed a negative correlation between f_0 and accuracy in the read condition for native speakers (-0.52), which illustrates the opposite effect.

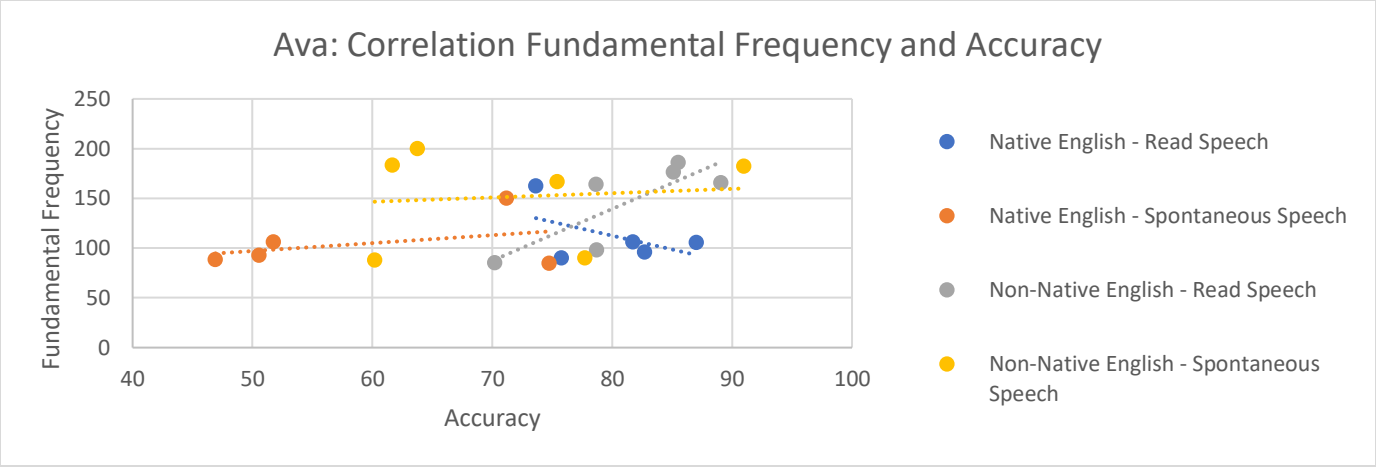

Figure 4-7 Correlation Fundamental Frequency and Accuracy for Ava

Overall, correlations between fundamental frequency and accuracy are inconclusive. We did not have enough participants for this analysis. Participants were selected based on availability and subjects taught. Fundamental frequency could not factor into the selection process or there would have been too many limitations.

Correlation of intelligibility and accuracy

See figure 4.8. This team was interested to find out if intelligibility, as rated by a human, would correlate with accuracy levels in speech recognition. As mentioned in the first section of this chapter, this seemed a plausible reason for the Chinese speaker having a more accurate transcript with spontaneous speech than read speech. We decided to investigate further.

Otter

Otter showed positive correlations between intelligibility and transcript accuracy in all groups and conditions. This suggests that the more intelligible the person is, the more accurate their transcription will be. This includes clear speech and easy-to-understand content.

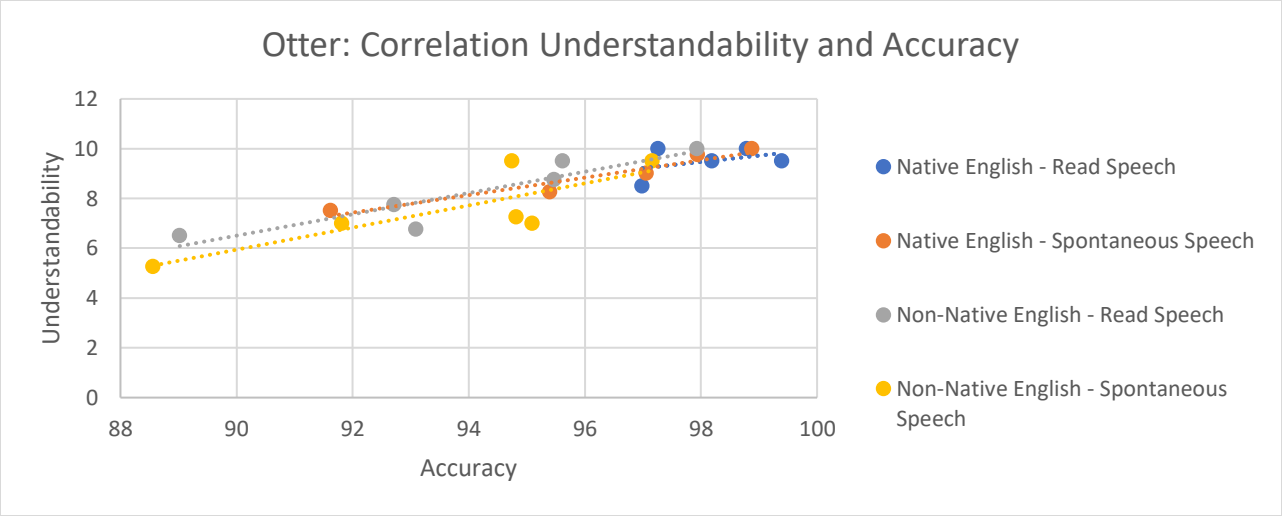

Figure 4-8 Correlation Understandability and Accuracy for Otter

Ava

See figure 4.9. Ava showed a strong negative correlation between intelligibility and transcripts produced from native English speakers during the read condition. This suggests that Ava is able to transcribe unintelligible speech under more ideal conditions; however, transcripts generated from non-native English speakers during the read condition showed an equally strong positive correlation, which suggests that non-native English speakers may show better results under ideal conditions with clear speech. During spontaneous speech, it seems that participants who were judged more intelligible produced more accurate transcripts from Ava.

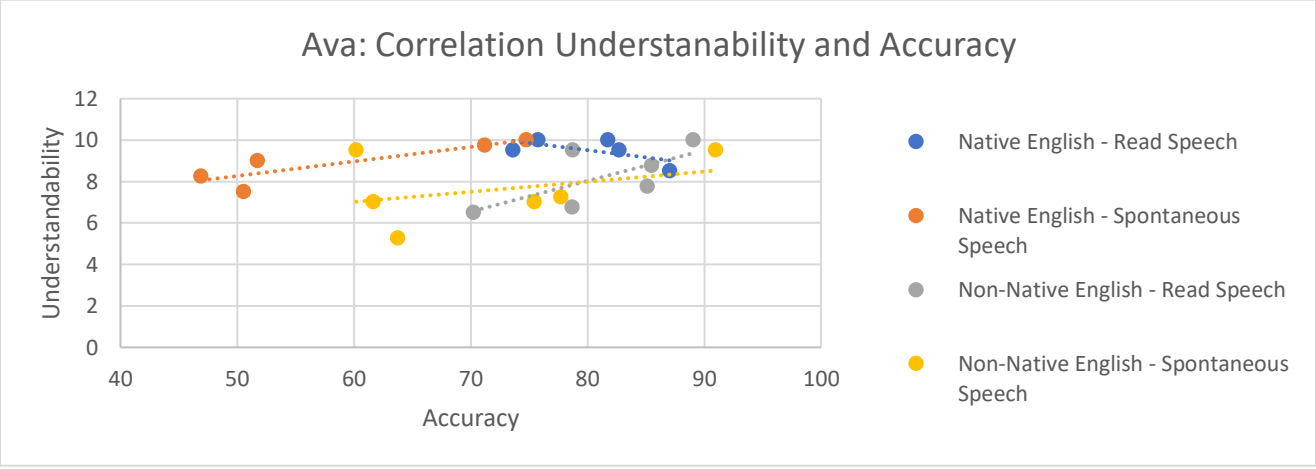

Figure 4-9 Correlation Understandability and Accuracy for Ava

Combining all findings, it seems that, during read speech, Otter prefers fast, clear speech. During spontaneous speech, Otter performs best with fast, clear speech from native English speakers and slow, clear speech from non-native English speakers. Conversely, in read speech, Ava seems to prefer slow, unintelligible speech from native English speakers and slow, intelligible speech from non-native English speakers. Ava seems to perform best with fast, clear speech from native English speakers and slow, clear speech from non-native English speakers.

4.2.3e Word breaks and syllables

Word breaks seem to be an issue for SR programs. Here we identify three examples that illustrate how SR technology may misinterpret word breaks

Example 1

Original:	not	an	academically	supported	statement
Transcript:	not	an	academically	support his	statement

Here, the program separated "supported" to "support his". The speaker assimilated [tə] from "supported" to the first phonemes of the following word [s], resulting in a clipped /d/. The program heard a syllable after "support", but interpreted the nucleus /ə/ as /ɪ/, resulting in the addition error, "his".

Example 2

Original:	universal	processes	affecting	SLA
Transcript	universal	process is	dsla	

To demonstrate what might have happened, see figure 4.10, which shows the speech signal in "processes" as well as the following two phonemes. In this case, we have four vowels separated by frication. The vowels (/o, ə, ɪ, a/) can be seen in the large waves and the fricatives (/s, s, s, f/) in the smaller waves. Humans do not speak using word breaks, which can be exemplified by the sound wave. i.e., the wave continues between [s] and [a]. In this case,

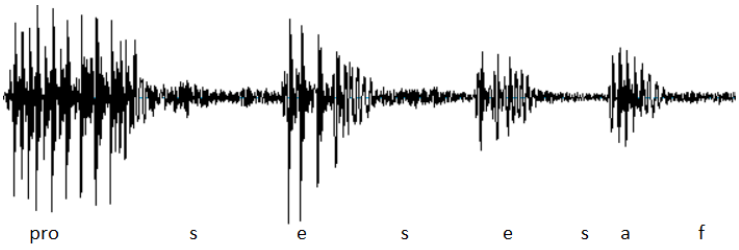

Figure 4-10 Speech Signal "processes af-" cut from "processes affecting"

the program had to identify where the word break is. “Processes” and “process is” sound very similar, so the program must “guess” at which was meant. The sample shows that the program proceeds to delete the following word, “affecting”. This could be due to load. If such is case, it is possible that the load required to decide which option to use caused the program to stall recognizing the next word. Still realizing a word was present, the program transcribed “dsla”, possibly interpreted a voiced /t/ in “affecting” but was unable to pick out the syllables surrounding it.

Example 3

Original:	it	doesn't	go	before	the	verb
Transcript:		isn't		before	the	verb

This example seems to indicate that the program was either unable to pick up or to identify /tdʌ/ from /ɪtdʌznt/, resulting in /ɪznt/. Essentially the program omitted an entire syllable, specifically the nucleus /ʌ/, resulting in two errors.

4.2.3f Unknown reasons

Sometimes the reasons for errors are unknown. The speech signal seems clear, both to the human ear and on a spectrogram, but the transcription is erroneous. The following example is a strange occurrence:

original:	but	one	of	the	characterizations	of	a	closed	set	is	that	every	limit	point	of	k	must	belong	to	k
transcript:	Edwin	-	-		-		-	-	-	-	-	-	-	Luna	por que		musculo today			

The transcription looks like Spanish, rather than English. It looks as if the Spanish and English lexicons became mixed up, resulting in a transcript that made little sense.

4.2.3g Speaker errors

Sometimes, especially during spontaneous speech, a speaker can misspeak, which might cause confusion for a reader, but may not necessarily be counted as an error by the software.

Example 1:

Intent:	When	the	sunlight	strikes	raindrops		in	the	air
Original:	When	**a**	**song like**	strikes	rainbow	drops	in	the	air
Transcript:	When	**a**	**song like**	strikes	rainbow	drops	in	the	air

This example shows two points to examine. The first is “a song like”. Though the intent was “When the sunlight”, as is written in *The Rainbow Passage*, the participant’s utterance truly sounded more like “a song like” than “the sunlight”. It is very difficult for a researcher, or technician, to decide whether this should be counted as an error. Someone who is trying to showcase their SR technology may argue that the speaker uttered those words, so it is not an error. Another may argue that, not only was the intent “the sunlight”, the phrase, “When a song like strikes” is nonsensical, so these should count as errors. In this case, this research team chose the latter argument, as the purpose for the study was to investigate content, rather than word count. Secondly, other utterance “rainbow”, should have been “rain”, but the participant misspoke and said “rainbow”. This cannot be counted as an error in regard to SR technology, but human errors like it may affect readability.

There are many reasons for errors in SR systems and why accuracy varies between speakers. It is a daunting task to discover why each error was made and how the system might be improved to mitigate them; sometimes errors are quite obvious, but many errors are caused by a number of issues that are not always apparent at the surface level. Causes of errors goes beyond lack of vocabulary, so even a system with enhanced vocabulary will still fail to reach 100% accuracy. All the factors discussed above, as well as factors beyond what is discussed in this chapter, contribute to the effectiveness of a program in real life. How these details reflect how SR can be used presently in academia will be addressed in the next chapter.

5. Conclusions

Based on the findings from the focus group and issues outlined by Woodcock et al. (2007), automatic transcripts are a superior choice to notes provided by volunteers, but only if they are edited. Automatic transcripts eliminate the subjective nature of notes when they are written by others and allows the student the freedom to select relative information, review the lecture and be assured that the information they are receiving is correct.

Most surprisingly, it seems that awareness of available services is an issue for DHH students. Even though there were only four participants in the focus group, each one was unaware of the services offered to the others. There are many DHH students at McMaster that were not part of this focus group. It is difficult to know if services exist that DHH students are unaware of. Given that participants in our group of four were unaware of what was offered to the others, it is possible that other DHH students could benefit from services that are available, but are not being used. A better way of advertising available services should be the first step towards making education more accessible for DHH students. Pertinently, in order to offer either Ava or Otter to students, the SAS office would have to re-evaluate the burden: benefit ratio of a Wi-Fi lapel microphone vs Bluetooth microphone. Bluetooth microphones are able to be connected to a device without a receiver or Irig, which alleviates issues with weight, cords and setup time. Bluetooth offers convenience at the expense of reliability; however, students may prefer it.

Regarding captioning lectures in real-time, all the participants of the focus group emphatically stated that they would be interested in a service that captures lectures in real time. Ideally, students would like accurate captioning in the moment, although the concern remains that, in science courses, a live transcript might be confusing. The focus group agreed that this would be best solved by combining a stenographer and a better version of ECHO360 (i.e., with human rather than machine edits, as mentioned previously). Students seemed to agree that ECHO360 appears to be the best option for now because of the video that goes along with the notes, but the transcript output from ECHO360 still should be edited before being sent to the student. If automatic speech-to-text were reliable, it might be an ideal accessibility option, as it is instant, user-friendly and can increase independence. However, errors are huge problems. If students do not receive accurate information, it could lead to misunderstandings and poor grades. Equally important, if DHH students are unable to verify that they are receiving correct

information, it could lead to anxiety. In addition, lecturers and teaching assistants should be mindful of which classes are uploaded online. If a given class has several tutorials or labs, the selected video should be the one that was experienced by the most students with note-taking needs. Information regarding why a student has a specific need is protected by SAS, and it would be problematic to choose the posted class based on the needs of a specific student, thus choosing a video through statistics is more objective and fairer to other students with note-taking needs, even if they are not DHH. In this way, transcripts could be available for all students online, if steps had been taken to create them.

All previous research suggests that all students benefit from captions and access to transcripts, regardless of whether they require accessibility services. Captions and transcripts, if available, should be accessible to all students, not only the ones who request them. Lectures posted online, as well as access to transcripts, would benefit every student at McMaster. Particularly students who do not have access to SAS services, for instance those who may have undiagnosed learning disabilities or language barriers. ESL students, for instance, would be able to better understand content from a speaker they find difficult to understand. However, this researcher acknowledges that having lectures and transcripts posted online may deter students from coming to class.

Transcripts should in no way be seen by students as a replacement to attending lectures. Lecture attendance offers benefits such as social interactions with fellow students; opportunities to ask questions; extra information or examples that may not be included in the recording; and networking with professors which can lead to references or lab assistant positions. These types of benefits should be outlined for students if lectures are posted online. If professors are finding that classroom attendance dwindles as a result of accessible transcripts, they may wish to re-evaluate whether they are offered. However, captions are necessary for video and audio content posted online.

Captions that are already being posted on videos, as mandated by McMaster University, should be edited. Even though studies showed that erroneous captions are still beneficial, students require accurate information to do well in class; DHH students have a much different experience with inaccurate notes than students who can hear and they should have access to error-free notes.

In order to ensure that lectures and videos are edited, this research team have a few suggestions. Lecturers could ask students to volunteer to edit transcripts. This would be a benefit because students who volunteer to edit transcripts would gain valuable reinforcement of the lecture. However, editing can be time-consuming. As mentioned in chapter 1, a one-hour class could take up to three hours to edit; therefore, a three-hour class might take up to nine hours to edit. It is vital that DHH students have access to the transcripts well before the next class, so it might be a difficult task for students to volunteer for and it may be hard to rely on them to finish the edits in good time. As a result, it might be difficult for professors to find volunteers. To combat this, students could be given extra credit for volunteering. Some departments at McMaster University offer up to 2% bonus for participating in experiments; some professors give credit for providing written notes (for one class) if there are students in class who require this accommodation, and they may post them online for all students to use. It is easier to find a volunteer for one class than for the full term. Students in departments where experiment participation is not an option might benefit from gaining credit in exchange for editing an hour or two of lecture transcripts.

Another alternative would be to hire someone to edit lecture transcripts; however, this option seems unfeasible. Most classes are at least 3 hours per week. If it takes three hours to edit a one-hour lecture, a single person, given a 40-hour work week, could only be expected to edit 13-14 courses per week as a full-time job. McMaster would have to hire too many people to accommodate the number of courses McMaster offers in each term. Ava offers scribe captions for $1.25 per minute. If the scribe is as accurate as advertised, a student taking five 3-hour classes per term over a 12-week period would cost SAS approximately $180 USD per term for that student, on top of the monthly fee necessary to use the program.

Speech recognition errors are caused by more factors than a program simply not knowing a word, i.e., not having it in its lexicon. Services that allow a user to add words to the lexicon seem superfluous for general purposes, as these programs seem to have a good vocabulary out of the box. However, in specialized lectures, this may still be a helpful tool. Errors are not always predictable, and they are not consistent. Therefore, human involvement seems necessary to achieve accuracy high enough for academia. Currently, it seems that Otter might be suitable for some classes without edits, but only if the program can reliably achieve over 98% accuracy for

that speaker, which is something that would have to be tested. The benefit of Otter is that it records the audio as it transcribes, allowing students to review the transcript. Ava would have to improve accuracy and deviations between speakers in order to achieve the accuracy necessary to be a reliable option for students. The higher accuracy promised from a paid subscription may address some of these issues, but the deviations found between speakers indicates that the system may not work across the board for all lecturers.

Accents remain an issue for SR technology, as subconscious insertions, vowel perception and consonant production cannot be altered for the speaker. However, systems may be able to mitigate some errors by eliminating impossibilities. Vowels can be identified as high/low and front/back by using formant frequencies. If a system can identify a vowel using its formants, it can eliminate options produced by AI. For instance, the example, “high” to “how”. As the formant frequency indicates the central vowel, /a/, the second formant increases as the speaker moves from /a/ to /i/; however, the example in figure 4.1 showed that the system decided that the vowel was /u/, which implies that the second formant would have to decrease from /a/. This selection produced by the AI is an impossibility, given the speech input. Vowels that share a similar space, such as /ɪ/ and /i/ would remain difficult for a program to decipher, but if programs could be improved to eliminate certain selections given a speech input, vowels that are far away from each other, such as /i/ and /u/, should not be competing for word selection.

Consonant selection can be improved in a similar manner. Obstruents are more susceptible to reductions or omissions by a speaker than sonorants. In these cases, the speech signal should carry more weight for an AI that is selecting a word. For instance, if you have a vowel, for example /i/, followed by a reduced consonant /?/, but the speech signal does not show a sonorant, one could reason that the unknown consonant was more likely an obstruent than a sonorant; therefore, a sonorant should not be a likely option to be selected for the AI.

As AI is not always able to understand an utterance, speakers may be able to improve SR by being mindful of their speech. Native English speakers may not need to slow their speech, as this is difficult for speakers who habitually speak quickly. However, enunciating consonants clearly, even with fast speech, may improve SR accuracy. This study did not find that slower speech led to more accurate transcriptions for native speakers, however it did for non-native

speakers. As our evidence shows that SR improves as intelligibility increases, non-native speakers may reduce errors by speaking slowly, which increases clarity.

Finally, this study showed conflicting results about whether accuracy is affected by male or female voices. Many of our female participants had lower than average f_0. In order to confirm previous research, this study would require a more diverse group of participants and more controlled comparisons for accent variables.

Speech-to-text samples make fascinating case studies. However, without knowing exactly how speech-to-text software has been programmed, much of the analysis is speculation; nevertheless, from a linguistic perspective, it seems that some errors can be explained through thorough linguistic analysis in which programmers may not have specialized knowledge. In addition, skilled linguists can analyze an error that seems trivial in order to deduce why such error might have occurred. This could lead to extraordinary improvements on the system. Not only can linguists comment on errors, but they can also explain how different types of errors affect understanding in the real world, which may be valuable information both to the public and to software designers.

We hope that the findings not only help students succeed, but also give helpful insights on how linguists can be valuable assets to emerging projects and technologies.

References

Abner, Li. *Google's speech recognition is now almost as accurate as humans.* 9TO5Google. Retrieved from https://9to5google.com/2017/06/01/google-speech-recognition-humans/ 2017.

Ava (2020). https://ava.me. Accessed August, 2021.

Babu, C. G., Vanathi, P. T., Ramachandran, R., Rajaa, M. S., & Vengatesh, R. (2010). *Performance analysis of voice activity detection algorithm for robust speech recognition system under different noisy environment.* Journal of Scientific and Industrial Research, *69*(7), 515–522.

Bain K., Basson, S. & Wald, M. *Speech Recognition in University Classrooms: Liberated Learning Project, Proceedings of the Assets 2002,* The Fifth International ACM SIGCAPH Conference on Assistive Technologies, Edinburgh, Scotland, 2002. From: http://Eprints.Ecs.Soton.Ac.Uk/9089/1/Waldasset.Pdf.

Bajorek, J. *Voice Recognition Still Has Significant Race and Gender Biases*. Harvard Business Review, 2019. Accessed August 18, 2021. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

Bod, R. *Probabilistic Linguistics.* The Oxford Handbook of Linguistic Analysis. 1st Edition, 2012. From: https://doi.org/10.1093/oxfordhb/9780199544004.013.0025

Butzberger, J., Murveit, H., Shriberg, E., & Price, P. *Spontaneous speech effects in large vocabulary speech recognition applications*. 339, 1992. https://doi.org/10.3115/1075527.1075607

Cao, D. and Guo, Y. *Algorithm research of spoken English assessment based on fuzzy measure and speech recognition technology*, Int. J. Biometrics, Vol. 12, No. 1, pp.120–129, 2020.

Collins, B. *Google Meet Gets One of Zoom's Best Features: Live Otter.ai Transcription.* Forbes, 2021 January 21. Retrieved from:

https://www.forbes.com/sites/barrycollins/2021/01/21/google-meet-gets-one-of-zooms-best-features-live-otterai-transcription/?sh=54180200700d

Davis, K., Biddulph, R., & Balashek, S. *Automatic Recognition of Spoken Digits.* Journal of the Acoustical Society of America, 24, 637-642, 1952.

Deng H., Ward R. K., Beddoes M. P., and Hodgson M. *A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds*. IEEE Trans. Speech Audio Process. 14 (2), 445–455, 2006.

Errattahi, R., El Hannani, A., & Ouahmane, H. *Automatic Speech Recognition Errors Detection and Correction: A Review.* Procedia Computer Science, 128, 32–37, 2018. https://doi-org.libaccess.lib.mcmaster.ca/10.1016/j.procs.2018.03.005

Fairbanks, G. *Voice and articulation drillbook*, 2nd edition. New York. Harper and Row. Pp 124-139, 1960.

Forgie J. W., & Forgie C. D. *Results Obtained from a Vowel Recognition Computer Program.* Journal of the Acoustical Society of America, 31(11), 1480-1489, 1959.

Fu Q. and Murphy P. *Robust glottal source estimation based on joint source-filter model optimization,* IEEE Trans. Audio Speech Lang. Process. 14 (2), 492– 501, 2006.

Haug, K. N., & Klein, P. D. *The Effect of Speech-to-Text Technology on Learning a Writing Strategy.* Reading and Writing Quarterly, 34(1), 47–62, 2018. https://doi.org/10.1080/10573569.2017.1326014

Jarmulak, J. *Speech-to-text Accuracy Benchmark – June 2020 Results.* Voiceagain. 2020, June, 25. Retrieved from https://www.voicegain.ai/post/speech-to-text-accuracy-benchmark-june-2020-results

Jarmulak, J. *Speech-to-text Accuracy Benchmark Revisited.* Voiceagain. 2020, September, 13. Retrieved from https://www.voicegain.ai/post/speech-to-text-accuracy-benchmark-revisited

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., & Weinstein, C. *Measuring human readability of machine generated text: Three case studies in speech*

recognition and machine translation. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, V, 1009–1012, 2005. https://doi.org/10.1109/ICASSP.2005.1416477

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. *Racial disparities in automated speech recognition.* Proceedings of the National Academy of Sciences of the United States, 117(14), 7684, 2020. https://doi-org.libaccess.lib.mcmaster.ca/10.1073/pnas.1915768117

Leonov A. S. and Sorokin, V. N. *Two parametric voice source models and their asymptotic analysis*. Acoustical Physics. 60 (3), 323–334, 2014.

Li, Q., & Russell, M. *An analysis of the causes of increased error rates in children's speech recognition.* 7th International Conference on Spoken Language Processing, ICSLP 2002, 2337–2340, 2002.

Liberated Learning - Neil Squire Society. 2010. Accessed February, 2021. https://www.neilsquire.ca/liberated-learning-consortium/

Leitch D. and MacMillan, T. *Liberated Learning Initiative Innovative Technology and Inclusion: Current Issues and Future Directions for Liberated Learning Research, Year III Report,* Saint Mary's University, Nova Scotia, Canada, 2003.

Leitch, D. *GIFT Atlantic Liberated Learning High School Pilot Project: A Study of the Transfer of Speech Recognition Technology from University Classrooms to High School Classrooms, Phase III Report,* Saint Mary's University, Nova Scotia, Canada, 2008.

Makhoul, J. & Schwartz, R. *State of the Art in Continuous Speech Recognition.* Proceedings of the National Academy of Sciences of the United States of America, 92(22), 9956-9963, 1995.

McMaster University *Student Accessibility Services (SAS) report to the Ministry of Training, Colleges and Universities (MTCU),* 2019. Retrieved from: https://sas.mcmaster.ca/ministry-reports/ in January, 2020.

Otter (2021). https://otter.ai Accessed August, 2021

Paez, D., Bain, K., Makeham, K., and Burns, D. *Widening Participation Using Speech Recognition: Can it Deliver? Liberated Learning Project*. Pathways 6 Conference, 2002. Australia. https://www.adcet.edu.au

Pfau, T., Falhauser, R., & Rushe, G. *A Combination of Speaker Normalization and Speech Rate Normalization for Automatic Speech Recognition*. Institution for Human-Machine-communication, Technische Universität, München, (2017).

Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. P., & Duerstock, B. S. *Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom.* IEEE Transactions on Learning Technologies, 6(4), 299–311, 2013. https://doi-org.libaccess.lib.mcmaster.ca/10.1109/TLT.2013.21

Ryba, K., McIvor, T., Shakir, M., & Paez, D. *Liberated Learning: Analysis of University Students' Perceptions and Experiences with Continuous Automated Speech Recognition.* E-Journal of Instructional Science and Technology, 9(1), 1–19, 2006.

Sorokin, V. N., & Leonov, A. S. *Multisource Speech Analysis for Speaker Recognition.* Pattern Recognition and Image Analysis, 29(1), 181–193, 2019. https://doi.org/10.1134/S1054661818040260

Stan, A., Bell, P., & King, S. *A grapheme-based method for automatic alignment of speech and text data.* 2012 IEEE Spoken Language Technology Workshop (SLT), 286–290, 2012.

Van Meter P., Yokoi L., and Pressley M. *College Students' Theory of Note-Taking Derived from Their Perceptions of Notetaking.* J. Educational Psychology, vol. 86, no. 3, pp. 323-338, 1994.

Wald, M., & Bain, K. *Universal access to communication and learning: The role of automatic speech recognition.* Universal Access in the Information Society, 6(4), 435–447, 2007. https://doi.org/10.1007/s10209-007-0093-9

Woodcock, K., Rohan, M.J., & Campbell L. *Equitable representation of deaf people in mainstream academia: Why not?* Higher Education 53:359-379, 2007.

Appendix

Program	Speaker	Accuracy	Words	S	O	A	Total	Sample																				
Otter	Canadian M	99.39	330	2	0	0	2	Original	the	actual	primary	rainbow	observed	is	said	to	be	the	effect	of	super-imposition	of	a	number	of	bows			
								Transcript	the	actual	primary	rainbow	observed	is	said	to	be	in	effect	the	super imposition	of	a	number	of	bows			
Otter	Korean	92.71	329	19	1	2	22	Original	since	then,	physicists	have	found	that	it	is	not	reflection,	but	refraction	by	the	raindrops	which	causes	the	rainbows		
								Transcript	since	then,	please he says to	have	found	that	it	is	not	reflection,	but	we fraction	by	the	raindrops,	which	causes	the	rainbows		
Otter	British 2	98.8	330	2	2		4	Original	this	is	a	very	common	type	of	bow,	one	showing	mainly	red	and	yellow	with	little	or	no	green	or	blue
								Transcript	this	is	-	very	common	type	of	bow,	one	showing	me in the	red	and	yellow	with	little	or	no	green	or	blue
Otter	German	95.62	343	11	3	1	15	Original	Aristotle	thought	that	the	rainbow	was	caused	by	reflection	of	the	sun's	rays	by	the	rain					
								Transcript	I was totally	thought	that	the	rainbow	was	caused	by	reflection	of	the	sun's	rays	by	the	way					
Otter	British 1	97.26	329	8	1		9	Original	there	is,	according	to	legend,	a	boiling	pot	of	gold	at	one	end								
								Transcript	There	is	according	to	legend	a	boiling	pot	of	gold		wanting									
Otter	American	96.99	332	8	2		10	Original	Some	have	accepted	it	as	a	miracle	without	physical	explanation											
								Transcript	Some	have	accepted	it	as	immortal		without	physical	explanation										
Otter	French	95.47	331	13	1	1	15	Original	These	take	the	shape	of	a	long	round	arch,	with	its	path	high	above,	and	its	two	ends	apparently	beyond	the	horizon
								Transcript	These	take	the	shape	of	a	long	round	arch	with	its	path	How you bob		and	his	two	ends	apparently	beyond	the	horizon
Otter	Polish	97.93	338	7			7	Original	The	Norsemen	considered	the	rainbow	as	a	bridge	over	which	the	gods	passed	from	Earth	to	their	home	in	the	sky	
								Transcript	The	norseman	considered	the	rainbow	as	a	bridge	over	which	the	gold	passed	from	Earth	to	their	home	in	the	sky	
Otter	Japanese	93.09	333	17	2	2	21	Original	These	take	the	shape	of	a	long	round	arch,	with	its	path	high	above,	and	its	two	ends	apparently	beyond	the	horizon
								Transcript	This	takes	the	shape	of	a	long	round	arch,	with	its	path	high	above.	And	it's	to	end	up our entry	beyond	the	horizon
Otter	Canadian F	98.18	331	6	1		7	Original	the	actual	primary	rainbow	observed	is	said	to	be	the	effect	of	super-imposition	of	a	number	of	bows			
								Transcript	the	actual	primary	rainbow	observed	is	said	to	be	the	effect	of	super-imposition	of	a	number	of	balls			

Program	Speaker	Accuracy	Words	S	O	A	Total	Sample																												
Otter	British 2	98.88	626	6	1	0	7	Original:	one	cause	of	depopulation	was	migration																						
								Transcript:	one	cause	of	D population	was	migration																							
Otter	German	94.74	741	37	2	0	39	Original	so	your	air	stream	basically	has	two	possibilities	of	leaving	your	body																
								Transcript	so	you		asked him	basically	has	two	possibilities	of	leaving	your	body																
Otter	Canadian	97.06	510	12	2	1	15	Original	you	look	at	the	balls,	centered	at	zed,	radius	one	over	n,	you	take	their	closure,	take	their	compliment									
								Transcript	you	look	at	the	balls	Sandry	that	said	radius	one	over	n,	you	take	their	closure,	take	their	compliment									
Otter	British 1	95.4	435	15	4	1	20	Original	and	in	fact,	morphology	was	one	of	the	first	sub-disciplines	to	be	employed	by	typology													
								Transcript	and		effect	morphology	was	one	of	the	first	sub disciplines	to	be	employed	by	topology													
Otter	Chinese	94.82	521*	24	2	1	27	Original	Okay,	so	if	you	look	at	some	quotes	from	monks,	which	is	a	student	of	Confucius												
								Transcript	Okay,	so	if	you	look	at	some	posts	from	monk,	which	is	a	student	of	Confucius												
Otter	American	91.62	370*	29	0	2	31	Original	So	have	two	verbs	together	"to	write"	and	"to	lie",	produce	the	word	"book",	which	is	a	noun,	which	is	exocentric							
								Transcript	So	have	two	verbs	together	to	write	into		ly	produce	the	word	"book",	which	is	a	noun,	which	is	extra centric							
Otter	French	95.05	347*	17	0	0	17	Original	Dialogism	is	a	dialogical	life	and	every	life	is	dialogical																		
								Transcript	back or just miss		a	dialogical	life	and	every	life	is	dialogical																		
Otter	Polish	97.16	422	10	2		12	Original	And	the	language	that	the	interpreter	translated	into	L2																			
								Transcript	And	the	language	that	the	interpreter		translates to																				
Ava	Polish	91	422	28	7	3	38	Original	the	interpreter	waits	until	the	speaker	finishes,	and	then	translates																		
								Transcript	they	interpreted	weight	Feel	the	speaker	finishes	and	then	translate																		
Otter	Canadian	97.94	535	6	5		11	Original	So	the	idea	here	with	number	5	is,	"Why	would	we	make	mistakes	in	a	second	language?"	[unintelligible]	that	one	is	really	referring	to	interference			
								Transcript	So	the	idea	here	with	number	5	is	Why	do women		make	mistakes	in	a	second	language,	and	that	one	is	really	referring	to	interference			
Otter	Japanese	91.81	281*	13	4	6	23	Original	So	sound	and	technology,	such	as	multiplane	camera																				
								Transcript	so	some of the		technologies	such	as	multiplane	camera																				

Program	Speaker	Accuracy	Words	S	O	A	Total	Sample																							
AVA	Polish	89.06	338	18	9	0	27	Original	since	then,	physicists	have	found	that	it	is	not	reflection,	but	refraction	by	the	raindrops	which	causes	the	rainbows				
								Transcript	seems	-	Physicists	have	found	that	it	is	not	reflection	but	3 faction	by	the	raindrops	which	causes	the	rainbows				
AVA	Canadian M	82.72	330	23	33	0	56	Original	since	then,	physicists	have	found	that	it	is	not	reflection,	but	refraction	by	the	raindrops	which	causes	the	rainbows				
								Transcript	since	then	-	-	-	-	-	-	-	-	-	Action	by	The	raindrops	which	causes	the	rain bones				
AVA	Koran	85.11	321	41	5	1	47	Original	since	then,	physicists	have	found	that	it	is	not	reflection,	but	refraction	by	the	raindrops	which	causes	the	rainbows				
								Transcript	Winston		physicist	haven't	found	that	it	is	not	reflection	but	the fraction	by	the	raindrops	which	colors	are	rainbows				
AVA	British	81.76	329	29	31	0	60	Original	there	is,	according	to	legend,	a	boiling	pot	of	gold	at	one	end										
								Transcript	Who's the cooling			to	Legend,	a	Boiling	Pot	of	Gold	-	one	in										
AVA	Chinese	89.02	346	37	1	0	38	Original	since	then,	physicists	have	found	that	it	is	not	reflection,	but	refraction	by	the	raindrops	with	which	causes	the	rainbows			
								Transcript	Things	then,	physicists	have	found	that	it	is	not	reflected	by	refraction	by	the	raindrops	with	which	causes	the	rainbows			
AVA	American	87.05	332	27	12	4	43	Original	The	rainbow	is	a	division	of	white	light	into	many	beautiful	colours											
								Transcript	the	rainbow			isn't			it	it's a	mini	beautiful	colours											
AVA	French	85.5	331	30	14	4	48	Original	These	take	the	shape	of	a	long	round	arch,	with	its	path	high	above,	and	its	two	ends	apparently	beyond	the	horizon	
								Transcript	Describe		the	shape	of	a	long	round	arch,	with	his	Path	High	Above	and	his	two	ends	up going to	be on	the	horizon	

Program	Speaker	Accuracy	Words	S	O	A	Total	Sample											
Otter	Korean	88.56	411	41	3	3	47	Original	which	employs	rap	only	during	the	verses,	singing	choruses	in	a	pop	style
								Transcript	which	includes	work	only	during	the	vs	singing	courses	in	a	pub	style

Program	Speaker	Accuracy	Words	S	O	A	Total	Sample																														
AVA	British 2	75.76	330	57	23	0	80	Original	if	the	red	of	the	second	bow	falls	upon	the	green	of	the	first	the	result	is	to	give	a	bow	with	an	abnormally	wide	yellow	band			
								Transcript	if	the	red	of	the	second	both	fools	on	the	green	of	the	first	there was			YouTube			wouldn't hav		normally buy together		band				
AVA	German	78.72	343	38	35	0	73	Original	throughout	the	centuries	men	have	explained	the	rainbow	in	various	ways																			
								Transcript	throughout	the	centuries	men	have	explained	to	Rainbow	interviews		race																			
AVA	Chinese	70.23	346	81	19	3	103	Original	since	then,	physicists	have	found	that	it	is	not	reflection,	but	refraction	by	the	raindrops	with	which	caruses	the	rainbows										
								Transcript	sting	Stan	Physicists	-	-	-	It	is	not	reflections	by the	Glee fraction		The	Raindrops	with	which	caruses	the	rainbows										
AVA	Japanese	78.67	333	37	27	6	70	Original	These	take	the	shape	of	a	long	round	arch,	with	its	path	high	above,	and	its	two	ends	apparently	beyond	the	horizon							
								Transcript	This	takes	the	shape	of	a	long	round	Arch	-	Its	Path	High	Above	and	it's	too	and	upper entry	be on	the	rise							
AVA	Canadian F	73.64	331	47	40	0	87	Original	When	the	sunlight	strikes	raindrops	in	the	air,	they	act	as	a	prism	and	form	a	rainbow													
								Transcript	When	the	sunlight	strikes	raindrops	in	the	air		back to the			prism		Informer		rainbow												

Program	Speaker	Accuracy	words	S	O	A	Total	Sample																								
AVA	British 2	74.76	626	107	44	7	158	Original	with	very	significant	depopulation,	decline	in	material	standards	of	living	and	a	regression	to	much	simpler…								
								Transcript	with	very	sick	population	decline	and	material	standards	of	living	and	-	regression	too	much	simpler…								
AVA	Chinese	77.74	521*	73	38	5	116	Original	it's	actually	burried	within	some	old	Chinese	history																
								Transcript	it's	actually	fairly	Ravine	some	old	Chinese	history																
AVA	Canadian F	71.21	535	70	80	4	154	Original	So	the	idea	here	with	number	5	is,	"Why	would	we	make	mistakes	in	a	second	language,	[ ]*	that	one	is	really	referring	to	interference
								Transcript	So	the	idea	here	with	the respond		is	why	would	they	make	mistakes	and	second	language,	-	-	-	-	-	-	to	interference
AVA	Japanese	75.44	281	44	18	7	69	Original	So	sound	and	technology,	such	as	multiplane	camera																
								Transcript	So	some	and the	technology	such	as	a magical land	camera																

Program	Speaker	Accuracy	Words	S	O	A	Total	Sample																						
AVA	Korean	63.99	411	67	70	11	148	Original	which	employs	rap	only	during	the	verses,	singing	choruses	in	a	pop	style									
								Transcript	which	in.		Meet Ranger			versus	singing	courses	in	a	pop	style									
AVA	German	60.19	741	136	136	23	295	Original	So	I'm	talking	about	this	part	here,	the	part	behind	the	velum,	which	is	called	the	nasopharynx					
								Transcript	so	-	-	-		the party		-	-	-	-	-	which	is	called	the	nasal pharynx					
AVA	Canadian	51.76	510	80	166	0	246	Original	you	look	at	the	balls,	centered	at	zed,	radius	one	over	n,	you	take	their	closure,	take	their	compliment			
								Transcript	you	look	at	the	balls		and your dad		-	-	-	-	-	-	-	-	-	-	-			
AVA	British 1	46.9	435	101	125	5	231	Original	elegant	description	is	something	which	are [a]	very	important	part	of	science											
								Transcript	elegant	descriptions		something	we		should break up			the sun's											
AVA	American	50.54	370	54	113	16	183	Original	So	have	two	verbs	together	"to	write"	and	"to	lie",	produce	the	word	"book",	which	is	a	noun,	which	is	exocentric	
								Transcript		Where to purchase			together	to	write	in		July	British Summer			book	which	is	a	noun,	which	is	exercise	
AVA	French	61.67	347	75	53	4	132	Original	Dialogism	is	a	dialogical	life	and	every	life	is	dialogical												
								Transcript	Energy smoothies a diagnostic on tonight					and	every	night	is	a logical												