Assessing L2 Pronunciation with Automatic Speech Recognition Dictation Tools

Carol Johnson

A Thesis

In

the Department

of

Education

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Education) at

Concordia University

Montréal, Québec, Canada

August 2025

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Carol Johnson

Entitled: Assessing L2 Pronunciation with Automatic Speech Recognition Dictation Tools

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Education)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

Dr. Muhammad Ayaz Naseem	Chair
Dr. Kazuya Saito	External Examiner
Dr. Rosane Silveira	Arm's Length Examiner
Dr. Heike Neumann	Examiner
Dr. Teresa Hernandez-Gonzalez	Examiner
Dr. Walcir Cardoso	Thesis Supervisor

Approved by

Dr. Walcir Cardoso, Graduate Program Director

August 15, 2025
Date of Defense

Dr. Pascale Sicotte, Dean of Arts and Science

Assessing L2 Pronunciation with Automatic Speech Recognition Dictation Tools

Carol Johnson, Ph.D.

Concordia University, 2025

ABSTRACT

The use of automatic speech recognition (ASR) to score pronunciation placement tests offers language institutions an efficient alternative to human raters, addressing common challenges such as rater reliability and high labour costs. However, the customizable ASR systems used to create scoring models are costly, making them unaffordable for most institutions. As an alternative, this dissertation explored the feasibility of using transcripts from Google Voice Typing (GVT), a free and readily available dictation-based software, to provide automated scores for pronunciation assessments. Via two empirical studies, it addressed the following overarching research question "What are the affordances offered by dictation ASR in an L2 pronunciation assessment context?"

In the first study (Manuscript A), human-rated and GVT-rated scores of 56 pronunciation placement tests were compared, showing strong correlations. However, when the samples were divided by proficiency levels, there were weak correlations for high-proficiency users, raising concerns about the reliability of the test scores. To explain this finding, it was hypothesized that some high-proficiency test takers received low GVT scores due to problematic linguistic elements in some test items (e.g., highly infrequent words, unusual collocations). Overall, this study showed that scoring pronunciation tests with GVT is feasible, with the caveat that reliability issues need to be explored to ensure test validity and reliability.

As a follow-up, the second study (Manuscript B) explored the effect of word frequency, unusual collocations, and phonologically ambiguous items on GVT transcription accuracy, with

the aim of supporting the design of valid and reliable pronunciation tests. Four highly intelligible English speakers recorded 60 sentences targeting these three features. The recordings were transcribed by GVT and scored for accuracy, while eight human raters carried out an intelligibility transcription task which was also scored for accuracy. For GVT, the results suggest that lower-frequency vocabulary and phonologically ambiguous phrases were particularly challenging, while sentences containing names, proper nouns, or unusual collocations were almost always accurately transcribed. In contrast, transcriptions produced by human raters showed great variability and tended to be less accurate than those generated by GVT. These results indicate that certain features are difficult for both human raters and GVT to transcribe, even when produced by highly intelligible speakers. This highlights the importance of careful task design to avoid features that may compromise transcription accuracy. To ensure valid, reliable scores and fair decision-making, transcription accuracy must be verified through systematic test piloting.

The findings of this dissertation emphasize that GVT can be used to develop cost-effective scoring models for pronunciation placement tests, while also highlighting the need for careful task design to avoid the use of language that may compromise transcription accuracy and unfairly penalize test takers. These results also offer practical implications for classroom-based pronunciation assessments.

Acknowledgements

While it may feel like it at times, a journey through a PhD is not a solo trip. Many people have accompanied me throughout this process, and I owe a great deal to them.

I have said it many times and will continue to say it: I had the best supervisor anyone could wish for. Walcir Cardoso is generous with his time and knowledge, and he is truly gifted at teaching his students not only how to conduct research, but also how to present it to the larger community in manuscripts and conference presentations. Every door that has opened for me has been possible because of his guidance and mentorship. I am profoundly lucky to have had such a mentor, and it is difficult to imagine this journey without his constant presence and encouragement. I will especially miss the monthly Zooms, which so often kept me going through the challenges of this journey. Thank you from the bottom of my heart.

My thanks also go to Heike Neumann and Teresa Hernández-González for their service on my committee. Heike's guidance during our tutorial and her thoughtful and challenging feedback were especially helpful, and Teresa's comments throughout the program contributed to the development of this dissertation. I am grateful to both for their contributions to this work.

I am also grateful to Dr. Saito and Dr. Silveira for serving on my defence committee. Their thoughtful questions and constructive comments pushed me to reflect more deeply on my work and to consider new perspectives. I am grateful for the care they took in reading my dissertation and for the insights they shared during the defence. I also gratefully acknowledge the support of the Social Sciences and Humanities Research Council of Canada (SSHRC), whose Doctoral Fellowship made this research possible.

The best part of any journey are the people you meet along the way. I have made friends in the program whose presence over the last four years has been invaluable. We shared both

laughter and tears, and through it all, my friends got me through. For this, I want to especially thank Dr. Natasha MacDonald, Christine Kinsley, Anita Jandaly, and Carey Nelson. I am also grateful to Paul John, who, though not in the program, has been a wonderful research partner, mentor, and friend. They made the difficult days easier and the good days even better.

And finally, my family, who had no say in the matter but found themselves dragged along on this trip. My daughter, Anne-Marie, planted the idea all those years ago when she encouraged me to do my MA. Neither of us knew it would lead here. Her support, as my daughter and as my peer, has been invaluable and so special. So few people in academia get the chance to work with their daughter like I have. And my boys, Andrew and Ian, who put up with hastily made meals and my crazy schedule as I juggled teaching and studying but never complained and always understood and cheered me on. Hopefully, our lives get a bit more normal now! Without their love, support, and willingness to give me the space I needed to do this, I would not be here.

Though this particular journey has ended, I carry with me the people, lessons, and memories that will guide the many journeys still to come.

Contribution of Authors

This dissertation, presented in manuscript format, includes four chapters: an introductory chapter that lays the groundwork for the manuscripts in Chapters Two and Three, and a final chapter that synthesizes their findings. Manuscript A was presented at an international conference (EUROCALL) in 2022, and a preliminary version was published in the conference proceedings. The full manuscript was published *in The Journal of Second Language Pronunciation* in 2024. Manuscript B has been sent to *Computer Assisted Language Learning*, and is currently under review.

CRediT author statement (Manuscript A) – Carol Johnson: conceptualization; methodology; validation; formal analysis; investigation; resources; data curation; writing – original draft; writing – review & editing; visualization; project administration; funding acquisition. Walcir Cardoso: conceptualization (supporting); writing – review & editing (supporting); supervision; funding acquisition (supporting). Beau Zuercher: writing – review & editing. Kathleen Brenner: writing – review & editing. Suzanne Springer: writing – review & editing. N.B.: Zuercher, Brenner and Springer were responsible for the creation of the pronunciation assessment rubric analyzed in this study and for collecting the human-rater scores as part of a previous project. These scores were used as secondary data for Manuscript A.

CRediT author statement (Manuscript B) – Carol Johnson: conceptualization; methodology; validation; formal analysis; investigation; resources; data curation; writing – original draft; writing – review & editing; visualization; project administration; funding acquisition. Walcir Cardoso: conceptualization (supporting); methodology (supporting); writing – review & editing; supervision; funding acquisition (supporting).

Table of Contents

List of Figures

List of Tables

Chapter 1: General Introduction

As automatic speech recognition (ASR) technology became more advanced and was increasingly integrated into people's lives at the end of the 20th century, the field of applied linguistics quickly recognized its potential usefulness for teaching and assessing second language (L2) speakers (Levis & Suvorov, 2012). Indeed, two publications in *CALICO Journal* at the end of the century highlighted ASR's dual potential as a pedagogical and an assessment tool: Eskenazi (1999) described the theoretical benefits of computer-assisted pronunciation training (CAPT) using a customized ASR system, while Conium (1998) explored the possibility of using a non-customizable commercial dictation system to assess speaking proficiency. Since the publication of these two articles, research into using ASR for language learning has been conducted on all four language skills as well as vocabulary, grammar, and pronunciation. For pronunciation, the focus of this thesis, its effectiveness as a pedagogical tool was noted in Ngo's (2023) meta-analysis of studies using ASR, which found an overall medium effect size (g = 0.69). Concurrently, the technology has become accepted for use in speaking and pronunciation assessments and is even credited for an upsurge in L2 pronunciation assessment research (Isaacs, 2018a), which had become marginalized at the end of the twentieth century (Isaacs & Harding, 2017).

Use of this technology is not without its disadvantages. Commercial pronunciation teaching applications built using ASR have been decried for their lack of pedagogical theory and for prioritizing "native-like" output over intelligible speech (Rogerson-Revel, 2021). Automated scoring of pronunciation and speaking assessments using ASR was highly contentious at the beginning of the 21st century due to its reliance on constrained tasks (e.g., read aloud and repetition tasks), with a very public debate occurring in editorial pieces in *Language Assessment*

Quarterly (see Chun, 2006 and Downey et al., 2008). Since then, research has shown this scoring method to be reliable and comparable to human rating, and its use is now accepted and commonplace in large international language tests. However, the purchase, development, and maintenance costs for the technology are substantial (Isaacs, 2018a), placing its use unreachable for smaller testing organizations wishing to streamline the scoring process.

Free ASR-based dictation technology, embedded in popular word processing applications such as Google Voice Typing (GVT) in Google Docs, has been a solution to certain issues surrounding the teaching of pronunciation. While teachers may find it impossible to give immediate and individualized feedback to students, dictation technology can do so, anywhere, and anytime (Ngo et al., 2023). Research has shown GVT to be effective in improving learners' pronunciation, not only in terms of accurate segmental features (e.g., Dai & Wu, 2023; Evers & Chen, 2021) but also in terms of intelligibility (e.g., how understandable a person's speech is; see Mroz, 2020). As such, the use of dictation ASR aligns more closely with the current focus on intelligibility rather than accent reduction in pronunciation instruction (Derwing & Munro, 2015). Given the intertwining relationship between pedagogy and assessment, as well as the cost effectiveness of this technology, its use for assessing pronunciation would seem logical. Despite this, there has been little exploration into using dictation ASR as an assessment tool.

The goal of this chapter is to present my doctoral thesis, which examines the use of GVT to assess pronunciation. It begins with an overview of ASR technology, followed by reviews of literature into its use as a pedagogical and assessment tool for L2 pronunciation. This review will show a gap exists, one that motivates my dissertation: while research into both pedagogical and assessment uses have largely mirrored one another, dictation-based ASR remains underexamined

in assessment. The chapter concludes by outlining the structure of the dissertation and providing an overview of the two empirical studies.

This manuscript-based dissertation addresses this gap by proposing two related studies, with each addressing an aspect of using GVT to automatically score pronunciation assessments: (1) the feasibility of using GVT to assess pronunciation, and (2) the impact of problematic linguistic elements on GVT's ability to recognize a test taker's speech and effect this may have on the reliability of the score derived from the GVT output.

ASR and L2 Pronunciation

Automatic speech recognition (ASR)

ASR is a technology that automatically recognizes spoken language and converts speech to text. As such, it allows speakers the possibility of viewing what they said via a transcription on the screen of their device (Levis & Suvorov, 2012). Initially developed in the 1950s, this technology has become ubiquitous and is available for free in mobile phones, laptops, and cars. ASR systems are also used in personal digital assistants such as Alexa and Siri, which orally respond to the speaker rather than providing a written transcription (Moussalli & Cardoso, 2020). The conversion of speech to text by ASR systems is not based solely on the phonemic components of the utterance. The technology also uses language and lexical models to predict potential lexical and morpho-syntactic components of the utterance. The probability of various components appearing together in an utterance is statistically analyzed, and the output comprises elements with the highest likelihood of co-occurrence (Filippidou & Moussiades, 2020).

Developers wishing to integrate ASR into their product can choose from systems with varying levels of customization. They can purchase a fully customizable version of the technology and train the processor and models to recognize speech using a corpus appropriate for

their purpose (Filippidou & Moussiades, 2020). This provides them complete control over the manner in which language is processed in their system and the integration of the technology into their product. Other ASR systems are less customizable, such as Google's ASR application programming interface (API). In such a system, the processor and models are pre-trained, and the developer has no control over the way the system processes language. The API solely allows the developer to customize the integration of ASR into their product (Ashwell & Elam, 2017). The least customizable, dictation ASR, is already trained and integrated into products or devices frequently used such as Google Voice Typing (GVT) in Google Docs, or Apple's Siri (Inceoglu et al., 2020). Dictation ASR is free of charge, making it attractive to developers with financial limitations. However, there is no ability to customize the language processing, and the ability to integrate ASR into a product is limited. Studies on L2 pronunciation teaching and assessment have examined a wide range of customizable ASR tools, a topic that will be explored in the following sections.

ASR and the Teaching of L2 Pronunciation

The focus of research using ASR in second language acquisition (SLA) has been pronunciation (Levis & Suvorov, 2012). When ASR was becoming more commonplace in the 1990s, Eskenazi (1999) saw its potential and, in a seminal article, outlined its affordances specifically for pronunciation. He believed that it not only provides opportunity for increased language production, but also benefits learners due to the instant feedback received in the output provided on the device. Studies on utilizing ASR to improve L2 pronunciation have confirmed his perspective as they have consistently suggested that its use is beneficial. Ngo et al.'s (2023) meta-analysis of ASR-based pronunciation studies found a medium effect size ($g = 0.69$) in studies that examined learning gains, indicating its effectiveness as a pedagogical tool.

Dictation ASR was absent from L2 pronunciation pedagogy at the beginning of the new millennium, probably due to an oft-cited study by Derwing et al. (2000). They found low accuracy rates for L1 Spanish and L1 Chinese speakers compared to L1 English speakers when dictating English sentences into Dragon NaturallySpeaking Preferred. As a result, they concluded that dictation ASR was not yet ready to be used as a tool for improving L2 pronunciation.

Consequently, the field turned to the use of customizable ASR technology to develop pronunciation training tools. These tools, often commercial in nature, use customizable ASR, often alongside other technologies (e.g., videos showing articulations of sounds, PRAAT sound waves), allowing them to provide instant feedback about multiple aspects of pronunciation (Rogerson-Revel, 2021).

Research has been conducted into the use of pronunciation teaching tools designed using customizable ASR in a variety of contexts. Studies in which participants use ASR-based tools for supplemental practice outside class tend to show significantly greater levels of improvement in pronunciation of segmental features over participants in control groups who practise without the technology (e.g., Garcia et al., 2020; Pourhosein Gilakjani & Rahimy, 2019). Many studies have examined the use of ASR-based pronunciation tools to replace instruction rather than supplement it (e.g., Elimat & AbuSeileek, 2014; Saleh & Pourhosein Gilakjani, 2021). Again, significantly higher gains have been found in experimental ASR groups than in control groups receiving traditional explicit instruction in a classroom. Interestingly, Neri et al. (2008) found equal gains were realized by the experimental and the control groups. However, in their study, the learners using ASR received four 30-minute treatments while the control group received four 60-minute sessions of explicit pronunciation instruction. This would suggest that ASR is a more *efficient* method for improving pronunciation than traditional classroom methods.

Research has also examined the best practices for using ASR-based pronunciation training tools. Elimat and AbuSeileek (2014) examined the benefits of individual, paired, or group work when using the technology. They found individuals significantly outperformed pairs, which significantly outperformed groups. However, participants in this study were children, and the authors admit that the time spent in pairs and groups may not have always been on task. Tsai's (2015) found the opposite in a study of Taiwanese university students using MyET, an ASR-based tutorial system, either individually or collaboratively (i.e., sitting together around the computer and taking turns). Learning gains between the two groups were comparable even though participants working together collaboratively each had less time to practice since they shared a computer.

Despite the benefits associated with using customizable ASR to teach pronunciation, there are concerns regarding this technology. Rogerson-Revel (2021) highlights issues such as a lack of pedagogical theory behind the development of the tools and the focus on teaching students to produce "native-like" output rather than intelligible speech. This seeming adherence to the nativeness principle (Levis, 2005) can be seen in the names of some programs (e.g., Accent Master Software), and in the language in some research, such as Saleh and Pourhosein Gilakjani's (2021) frequent references to ASR-based pronunciation tools as "accent-reduction" software.

Recent advances in dictation ASR technology, which Derwing et al. (2000) found problematic over 20 years ago, may offer a solution to the concern about a lack of focus on intelligibility in pronunciation training tools designed with customizable ASR. To determine if the technology had improved its ability to recognize L2 speech over the period of two decades, McCrocklin and Edalathishams (2020) replicated Derwing et al. (2000). They asked highly

fluent ESL learners to dictate the same English sentences used in the original study into GVT, the dictation ASR system used in Google Docs. Accuracy had improved over the 20 years, with only a 5–7% difference in accuracy rates between L1 and L2 speakers. Other studies have also shown high accuracy rates in Google Translate, which is based on Google's ASR, for intermediate-level learners in an ESL context (i.e., up to 88% in John et al., 2024). Lower accuracy rates have been found in studies in English-as-a-foreign-language (EFL) contexts when using GVT (i.e., up to 75% in Inceoglu et al., 2023) as well as in the dictation technologies used by Microsoft and Apple (i.e., up to 71% in Hirai & Kovalyova, 2023). Both EFL studies, however, showed ASR accuracy rates strongly correlate with human listeners' judgments of intelligibility, indicating that the lower recognition rates reflect the participants' level of pronunciation rather than technological shortfalls. These studies suggest that dictation ASR has evolved into a tool that should be incorporated into L2 pronunciation classes.

Not only is dictation software free and readily accessible, which is not always the case with pronunciation training tools designed using customizable ASR, but it has the added benefit of seeming to be forgiving of accented speech, but less so for unintelligible speech. McCrocklin et al. (2019) found that while most accented words were usually transcribed correctly (e.g., incorrect first vowel and word stress in the word *spaghetti* did not result in an inaccurate transcription), this was not the case when a word was mispronounced as its corresponding minimal pair (e.g., desk/disk). In a study to determine if Google's API could be used in the development of a pronunciation training tool, Cámara-Arenas et al. (2023) also found Google's ASR would accurately recognize a word even when pronounced with an accent. This corroborates McCrocklin et al.'s (2019) findings that Google's ASR recognizes intelligible speech. With pronunciation research and teaching moving from accent reduction to intelligibility

(Derwing & Munro, 2015), dictation ASR technology, such as GVT, fits ideologically with current pedagogical practices.

Research into classroom use of dictation ASR has shown it to be a useful tool that benefits the students in various aspects of pronunciation. Studies using control groups as a comparison tend to show significant improvements in the dictation ASR groups in terms of pronunciation of segments in English (e.g., Dai & Wu, 2023; Evers & Chen, 2021; Guskaroska, 2020; Inceoglu et al., 2020) and in French (Liakin et al., 2014), accentedness and comprehensibility in English (e.g., Sun, 2023), and intelligibility in French (e.g., Mroz, 2020). McCrocklin (2018) found comparable rather than significant improvements in a study of two groups of university-level ESL students. However, the ASR group received explicit pronunciation instruction for only half of the time allotted to the non-ASR group.

Student perceptions of using dictation ASR as a tool for learning pronunciation tend to be positive despite some frustrations with misinterpretations of their speech (Dillon & Wells, 2021, Inceoglu et al., 2020; Liakin et al., 2017; McCrocklin, 2019). Mroz (2018) and Wallace (2015, 2016) both found students developed an awareness of their level of intelligibility and of the actions they needed to take to improve it, which resulted in an appreciation of the use of the tool. However, Chen et al. (2020) found only 22% of the participants would continue using dictation ASR on their phones as they were very frustrated with its inability to understand them. The authors hypothesize that this could be due to the lower-proficiency level of the students in their study as compared to other studies. This may indicate that more research is needed to determine the optimal level of proficiency at which dictation ASR should be used to minimize learner frustration.

Considering the ability of ASR to provide instant feedback on speech produced by language learners, it is not surprising that this technology has also greatly impacted language assessment, especially for speaking and pronunciation assessment. Specifically, the output from the ASR technology can allow for the generation of an automated score, removing the need for human raters who are expensive and are not always reliable. As a result, there has been and still is research into and development of automated scoring systems, which will be discussed in the following section.

ASR in Assessment of L2 Speaking and Pronunciation

The internationalization of education in the late 1980s led to a record number of students from non-English-speaking countries hoping to study at universities in English-speaking countries. To ensure these students had the linguistic abilities needed to succeed in English-medium universities, international testing organizations developed tests, such as IELTS and TOEFL iBT, to provide a portrait of a student's ability to university admissions departments (McNamara et al., 2019). For the speaking portion of these assessments, human raters often score the speech samples, which can be time-consuming and expensive in terms of human resources (Evanini & Zechner, 2019). There is ample evidence in the literature of other issues such as rater bias due to accent familiarity (e.g., Brown & Fulcher, 2016), shared languages with test takers (e.g., Winke & Gass, 2013; Winke et al., 2012), and experience with L2 speech (e.g., Kennedy & Trofimovich, 2008; Saito et al, 2016). Rater fatigue has also been shown to have an impact on the reliability of scores (e.g., Ling et al., 2014). Given the high-stakes nature of these assessments (e.g., acceptance into a university), reliability, or the consistency of measurement, is of the essence.

The period of rapid growth in international language testing coincided with significant advances in ASR technology, and automated scoring of assessments began to be seen as a possible solution to mitigate issues with human raters, as described above. Bernstein et al. (1990) published a study into the use of ASR to score a pronunciation assessment and found strong correlations with human raters. Since this time, large international commercial language assessment companies such as Pearson and Education Testing Systems (ETS) have researched and invested in the development of ASR-based systems for scoring speaking samples. Pearson's Versant, widely used by educational institutions and corporations, uses a highly customized ASR engine, along with other technologies, to measure different aspects of speech such as pronunciation, fluency, and pitch, to provide automated scores on the speaking portion of their language test (Bernstein et al., 2010). ETS's SpeechRater, likewise, employs ASR as one of the technologies used to generate automated scores for the TOEFL iBT test, which are combined with scores generated by human raters to provide the speaking score (Zechner & Evanini, 2020).

These tests have not been without controversy, as automated scoring relies on predictable speech to be accurate and reliable, resulting in the frequent use of constrained test tasks (e.g., read-aloud tasks). The use of constrained tasks has led to criticisms that these tests lack construct validity since they assess only a small portion of the speaking skill and neglect important communicative aspects such as interaction, content, and pragmatics (Isaacs, 2018b; Khabbazbashi et al., 2021). This can call into question the fairness of automated assessments of speaking as the psychometric quality of the tests may not fully represent the construct being measured (McNamara & Ryan, 2011). Indeed, a very public exchange of opinions on the matter of test fairness and validity appeared in *Language Assessment Quarterly.* Chun (2006) decried the use of constrained tasks in Pearson's PhonePass, the name initially given to Versant, stating

that the lack of interaction and authenticity made it an invalid speaking test. Downey et al. (2008) countered the criticisms with the argument that the tasks are designed to measure the construct of "facility in spoken English," and that the measures focus on the psycholinguistic aspects of language (e.g., language processing) rather than the sociolinguistic aspects (e.g., discourse).

ETS, the developer of the TOEFL test, takes a more sociolinguistic and communicative approach to the construct of speaking than Versant, and as such has worked toward having SpeechRater score spontaneous speech (Chen et al., 2018). This has proven to be difficult; accordingly, SpeechRater is used in conjunction with human scores for the TOEFL iBT test. When SpeechRater was first introduced in 2006, it was limited to measuring pronunciation and fluency; since this time, measures of different aspects of the speaking construct have been incorporated, such as vocabulary, grammar, rhythm, content, and discourse (Evanini & Zechner, 2019).

Given the controversy, the literature surrounding automated scoring of speaking assessments has often focused on research supporting the validity and reliability of these tests. Pearson and ETS have aimed to show the validity of their tests by showing concurrent validity with well-established human-scored speaking assessments (e.g., IELTS). For example, test-taker scores from Versant were found to have strong correlations (.75 - .77) with overall TOEFL iBT scores and IELTS speaking test scores (Bernstein et al., 2010). Conversely, SpeechRater was found to only have a correlation ranging from .57 to .68 in a study of over 7,000 samples of spontaneous speech (Zechner et al., 2009), showing the difficulty in using spontaneous speaking tasks rather than constrained tasks. Research into concurrent validity of smaller-scale tests using ASR to automate speaking scores have also shown strong correlations to other large-scale tests

(e.g., ACTFL speaking test in Graham et al., 2008; overall TOEIC scores in Luo et al., 2009) or with overall scores for the same test (e.g., Cox & Davies, 2012). This suggests that constrained tasks, while not representative of the full speaking construct, may allow for a valid assessment of the speaking skill. Smaller-scale tests have also been studied to ensure the reliability of scores by comparing the automated score to those of human raters scoring the same speech samples. Both van der Walt et al. (2008) and Saito et al. (2023) found strong correlations between the two types of scores; in fact, there were stronger correlations between the machine scores and those given by the human raters than the inter-rater reliability scores calculated for the human raters.

While the use of highly customized ASR has become the norm in pronunciation and speaking assessments (Isaacs, 2018a), the use of dictation ASR has been almost completely overlooked. Conium (1998) carried out an early study using Dragon NaturallySpeaking for an oral assessment. While he found it did discriminate between L1 speakers and higher-proficiency L2 speakers of English, he concluded that dictation ASR technology was not yet ready to assess lower-proficiency students. Perhaps his findings discouraged research, even as the technology improved. Whatever the case, the literature is scarce, as opposed to the growing research into its use as a pedagogical tool. However, since pedagogical and assessment practices are inherently intertwined, it is of the essence to explore dictation ASR's potential in L2 pronunciation assessment.

One of the few studies examining the use of dictation ASR for assessment rather than teaching purposes focused on student acceptance (Dillon and Wells, 2021). The authors found that after taking a formative test using a dictation ASR, students had positive perceptions despite some technical glitches. However, there is a dearth of research examining the feasibility of using dictation ASR as an assessment tool (e.g., the validity and reliability of the scores). I carried out

a preliminary study (Johnson et al., 2022 – expanded in Chapter 2 of this dissertation) that used speech samples from a database previously scored by human raters. The samples were run through GVT, and a score was produced based upon the transcription accuracy rates (referred to interchangeably in the dissertation as "accuracy" or "recognition rates," for stylistic reasons; both denote the percentage of words correctly rendered by GVT and can be interpreted as reflecting intelligibility). Nelson and Cardoso (2024) replicated the study using the same speech samples, but using Microsoft Transcribe (MS-T) as the dictation technology. Both studies found strong correlations between human-rater and automated scores for the overall score and for each criterion in the rubric used by the human raters. While Microsoft Transcribe (MS-T) had slightly stronger correlations with the overall scores provided by the human raters (.79 versus .78), correlations between the subscores of three of the five criteria on the rubric used by the human raters had stronger correlations with GVT than with MS-T (i.e., comprehensibility and production accuracy in the production of phonemes; and thought groups, sentence stress, and intonation).

The strong correlations between overall human-rater and ASR scores in both Johnson et al. (2024) and Nelson and Cardoso (2024) indicate the scores generated by dictation ASR are reliable. Of greater significance are the robust correlations also observed between each ASR system and every criterion evaluated by the human raters. This suggests that the test accurately reflects the pronunciation construct and, as such, does not compromise the test's validity (Loukina et al., 2018). However, when examining scores across proficiency levels, both studies found strong correlations for low-proficiency test takers, but not for high-proficiency test takers, suggesting potential bias in the scoring or in the test task itself. Johnson et al. (2024) hypothesize that this may be due to the high presence of low frequency vocabulary items in one of the test

versions. Since low frequency words are less probable, the algorithm may have opted for other lexical items, even for the speech of high proficiency test takers. However, the human raters knew the sentences and would have no issue with recognition of low-frequency words, leading to a divergence between the human-rater and the GVT scores.

Google's ASR's difficulty with recognizing low frequency words has been established by two studies that evaluated the accuracy of the ASR's transcriptions (Cámara-Arenas et al., 2023; Ashwell & Elam, 2017). However, these studies are limited in the range of vocabulary examined. Ashwell and Elam (2017) focused on people's names and unusual collocations and Cámara-Arenas et al. (2023) looked at highly infrequent vocabulary (i.e., glom). There is still much exploration needed into GVT's recognition of proper nouns other than people's names as well as unexpected collocations and a better understanding of GVT's recognition rate of lexical items from all levels of word frequency lists, not just the most infrequent words.

The findings of Cámara-Arenas et al. (2023) and Ashwell and Elam (2017) also raise questions about the effect of other linguistic elements on the ASR's output. For example, there is a lack of understanding of how ambiguous speech caused by natural speech phenomena such as syllabification (e.g., differentiating between "a nice man" and "an ice man") affects GVT's output. The possibility that natural speech phenomenon can negatively affect the transcription provided by GVT has a direct implication for the validity and reliability of test scores generated using dictation ASR. However, there does not appear to be any research addressing this issue. For this reason, Chapter 3 of this dissertation will examine the impact of infrequent vocabulary, unusual collocations, and ambiguous connected speech on GVT output.

As the pedagogical use of dictation ASR for pronunciation instruction increases, it is of the essence to explore its potential in L2 pronunciation assessment as the two uses are inherently

intertwined. This dissertation proposes to explore the potential of ASR-based dictation technology as a tool to assess L2 pronunciation. It aspires to be the starting point of a larger conversation about how to integrate dictation ASR into various assessment practices (e.g., placement tests, classroom assessment) to reflect its growing use as a pedagogical tool.

This Dissertation

ASR has become an essential technology for both the teaching and assessment of L2 pronunciation. While research into its use in both aspects has historically mirrored each other, this has not been the case for ASR-based dictation technology, such as GVT. Multiple studies indicate that this free and readily accessible technology is beneficial for teaching L2 pronunciation; despite this, there has been a dearth of research into its use for assessing pronunciation and the linguistic factors that may affect its accuracy and reliability in language testing.

To address this lack of research, the dissertation examined the feasibility of using GVT to assess pronunciation (Manuscript A) and the impact of problematic linguistic elements on GVT's ability to recognize speech (Manuscript B). Through a better understanding of such issues, best practices can be developed to minimize their impact on the validity and reliability of test scores. The overarching research question guiding these two studies was: What are the affordances offered by dictation ASR in an L2 pronunciation assessment context? To answer this question, two studies were undertaken. The following paragraphs summarize the two manuscripts.

Manuscript A (Chapter 2) explored the feasibility of using GVT to score a pronunciation placement test through the framework of Bachman and Palmer's (1996) Test Usefulness Model. In this study, human raters scored samples from a pronunciation placement test. The samples were then run through GVT, with the accuracy rate of the output functioning as the GVT score.

Following the literature (e.g., Bernstein et al., 2010; Saito et al., 2023), relationships between the human-rater scores and the GVT scores were determined by running correlations. A short summary of the study was published in *Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022* and the full-length article has been published by the *Journal of Second Language Pronunciation* (Johnson et al., 2024).

Manuscript B (Chapter 3) addressed the effect of potentially problematic linguistic elements affecting GVT output. L1 and high-proficiency L2 speakers of English provided speech samples that contained potentially problematic linguistic elements such as low frequency words, unusual collocations, and phonologically ambiguous phrases. Transcriptions of the speech samples were obtained through GVT and from eight human raters who carried out an intelligibility transcription task. All transcripts were scored for accuracy at the sentence level and at the linguistic feature level. The GVT and human transcripts were then compared and analyzed to determine if certain linguistic elements are more problematic for GVT. Based upon the findings, best practices were proposed to avoid or mitigate these issues in order to ensure test validity and reliability.

Overall, this dissertation examined the ability of GVT to provide valid and reliable scores for pronunciation placement tests by evaluating its impact on test usefulness and identifying best practices for its implementation. Table 1.1 provides an overview of the two studies in terms of goals, participants, and potential contributions.

Table 1.1

Overview of Manuscripts

Manuscript	Goal	Participants	Potential Contributions
A*	Assess GVT's ability to provide reliable and valid automated pronunciation scores	University-level ESL students with various proficiency levels and L1s (n = 56)	Determine the feasibility of automatically scoring pronunciation assessments with GVT
B	Explore the effect of problematic linguistic features on GVT output	L1 and high-proficiency L2 English speakers (n = 30) Human raters (n = 8)	Identify problematic linguistic features and develop best practices to mitigate their impact on test scores generated using GVT

* This peer-reviewed article has already been published. I am the first and main author:

Johnson, C., Cardoso, W., Zuercher, B., Brenner, K., & Springer, S. (2024). Assessing pronunciation using dictation tools: The use of Google Voice Typing to score a pronunciation placement test. *The Journal of Second Language Pronunciation*. 1–24. https://doi.org/10.1075/jslp.23033.joh

N.B.: The formatting of this article has been adapted from the journal's style to standard APA 7, as per dissertation guidelines.

A preliminary version of this paper appeared in:

Johnson, C., Cardoso, W., Zuercher, B., Brenner, K., & Springer, S. (2022). Using Google Voice Typing to automatically assess pronunciation. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds.), *Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022.* (pp. 203–207). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1459

Chapter 2: Assessing Pronunciation Using Dictation Tools:

The Use of Google Voice Typing to Score a Pronunciation Placement Test

Placement tests are often used by language schools to ensure students are placed in classes that match their language proficiency level. However, the development and administration of these tests are costly, requiring significant human and financial resources (Saito et al., 2023). The pronunciation component of these tests can involve multiple interviewers to elicit language from test takers and/or raters to assess their output, which is expensive. As well, this process can lack reliability, as it is prone to subjective scoring on the part of human raters' biases (Isaacs, 2018a). Studies have shown biases in terms of familiarity with a specific second language (L2) accent (Browne & Fulcher, 2016; Carey et al., 2011), in rater experience with L2 accents in general (e.g., English second language or ESL teachers are less strict in their ratings; Kennedy & Trofimovich, 2008; Saito et al., 2016), and in personal attitudes about certain accents (Yan & Ginther, 2018). The issue is compounded by the fact that raters develop fatigue and become less accurate and reliable during the long scoring sessions accompanying placement tests (Ling et al., 2014). Such issues carry significant consequences, as incorrect placement of students can create challenges for both students and instructors, leading to skepticism among students towards the validity of the test (van der Walt et al., 2008).

Issues with rater biases and fatigue could be mitigated with the use of automatic speech recognition (ASR) to automate the scoring of pronunciation tests, but development and maintenance costs of a customizable ASR scoring tool are high (Isaacs, 2018a; Saito et al., 2023), placing it beyond the reach of many language schools. Google Voice Typing (GVT), an ASR-based dictation technology, has been shown to have high accuracy rates with fluent L2 speakers (McCrocklin & Edalatishams, 2020), and may offer a solution to language schools

looking for cost-effective testing methods. Thus, the aim of this study is to examine the feasibility of using GVT to automate the pronunciation scoring for university English placement tests. Existing samples from a pronunciation placement test were scored by a team of experienced ESL teachers and through an analysis of the GVT written output. The scores were correlated across evaluation criteria and proficiency levels to determine the usability of ASR-based dictation tools for the purposes of automated scoring of pronunciation assessments.

Background

Human Rater Biases when Assessing Pronunciation

Until the advent of technology that allowed for automated assessment of complex linguistic tasks such as writing and speaking, human raters were the only option available for providing scores. Williamson et al. (2012) recognize human-rated scores to be the "gold-standard" (p. 7) in assessment, despite acknowledging the presence of inherent biases that can undermine the reliability, or "consistency across measures" (Bachman & Palmer, 1996, p. 19), of these scores. Research has shown that biases can influence human-rater scores of pronunciation in different ways. Much of this research has examined raters' perceptions of participants' comprehensibility (e.g., the degree to which an utterance is easy to understand) and intelligibility (e.g., how understandable a person's speech is). For example, raters' experiences with accented speech can influence the scores. When raters are familiar with a test taker's specific L2 accent, they find the speech more intelligible than when they are not familiar with the accent (Browne & Fulcher, 2016) and, as a result, they tend to give a higher score for the test taker's pronunciation (Carey et al., 2011). Furthermore, scores tend to be higher if the test taker's accent is from a language also spoken by the rater (Winke & Gass, 2013; Winke et al., 2012). Research has also indicated that experienced raters (e.g., trained ESL teachers) score L2 speakers' higher than

novice raters in terms of comprehensibility (Saito et al., 2016) and intelligibility (Kennedy & Trofimovich, 2008) since experienced language teachers are used to interpreting L2 speech (see Yan & Ginther, 2018 for a comprehensive survey of research into rater biases). The use of rating scales can guide raters and increase the level of reliability (Turner, 2014). However, raters need training to use the scales, which involves human resource expenses, and scales are not guaranteed to reduce a rater's idiosyncratic assessments (Yan & Ginther, 2018). Fatigue can also affect the quality and consistency of a well-trained rater's scores. In a study of 72 experienced TOEFL iBT raters, Ling et al. (2014) found that raters working three sessions during six-hour shifts had more accurate and reliable scores than raters working two sessions during an eight-hour shift.

Despite being considered the "gold standard," human-rater scores are prone to biases and can be affected by fatigue, negatively impacting the reliability of the assessment. One possible solution to mitigate the multi-faceted problem of rater subjectivity and reliability is the use of ASR to score the pronunciation of speech samples, as the machine-generated scores are not subject to the same biases as the human-generated scores (Isaacs, 2018a).

Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is a technology that transforms speech into text using algorithms to analyze and process spoken word (Levis & Suvorov, 2012). Using lexical and language model databases, the ASR system processes the spoken input, enabling it to generate hypotheses regarding the phonemes and/or words spoken. The phonemes/words with the highest probability are then transcribed as written text (Filippidou & Moussiades, 2020). Developers who want to integrate ASR into their own programs or websites have a multitude of systems to choose from, with varying degrees of customization required. Fully customizable

systems must be purchased and trained to recognize speech using a corpus of samples appropriate for the intended use (Filippidou & Moussiades, 2020). Less customizable options, such as Google's ASR application programming interface (API), are less expensive and come pre-trained. This is a more cost-effective solution; however, developers have no control over how speech is processed in the system (Ashwell & Elam, 2017). There are also free non-customizable ASR systems that allow people to dictate into computers (e.g., Google Voice Typing in Google Docs) or other personal devices (e.g., Amazon's echo) (Inceoglu et al., 2000). Like the less customized technology, developers have no control over the functioning of the speech recognition system.

Current use of Automated Assessment of Pronunciation

For the past two decades, large international language testing companies with substantial financial and human resources at their disposal have been developing and employing advanced, tailor-made automated assessment systems. Tests such as Pearson's Versant (Bernstein et al., 2010) and Educational Testing Service (ETS)'s SpeechRater, used in the TOEFL Practice Online (TPO) (Chen et al., 2018; Zechner et al., 2009), use highly sophisticated and customized ASR technologies and algorithms to assess pronunciation and speech. These commercialized tests still tend to rely heavily on the use of constrained tasks (e.g., read-aloud tasks) as predictable language ensures greater accuracy and reliability in scoring speech (Khabbazbashi et al., 2021). This has led to criticisms that the tests lack construct validity, the degree to which a score can be interpreted to reflect the construct being measured (Messick, 1989), since constrained tasks neglect important communicative aspects such as interaction, content, and pragmatics (Isaacs, 2018b).

One approach taken by testing organizations and researchers to counter the criticism that these tests lack validity is to show concurrent validity with other assessments scored using human raters. While it may seem counterintuitive to use human-rated scores as a benchmark, given the biases discussed above, they are the only measure against which ASR scores can be compared (Yan & Ginther, 2014). For example, Versant scores had a strong correlation of .77 with human-rated scores based on the IELTS speaking test rubric and a correlation of .75 with participants' official TOEFL iBT scores (Bernstein et al., 2010). Higher correlations have been found in the following language versions of Versant: Spanish (.92), Dutch (.81), and Arabic (.87) (Bernstein et al., 2010). Other studies have shown strong correlations with the American Council on the Teaching of Foreign Language (ACTFL) speaking test (Graham et al., 2008), overall TOEIC scores (Luo et al., 2009) and with overall placement test scores (Cox & Davis, 2012). Such findings indicate that using customized ASR to score oral proficiency tests can be beneficial as there is high concurrent validity and a "low cost when compared to human scoring" (Graham et al., 2008, p. 1608).

Other research has focused on the reliability of automated scores when compared to human-rated scores of the same speech samples. Such studies show strong correlations between the two scores, indicating that using ASR is a reliable method for scoring assessments. van der Walt et al.'s (2008) study regarding the automation of the scoring of a university's English placement test found that human-rater scores for speaking had a high correlation (.74) with ASR scores of the same samples. Similarly, Saito et al. (2023) observed strong correlations between human-rated scores measuring comprehensibility and automated scores over a series of three experiments. Indeed, in both studies, the correlation between automated and human-rated scores was stronger than the inter-rater reliability scores among the human raters. This was also the case

when Bernstein et al. (2010) carried out test-retest and split-test analyses on Versant test samples scored automatically and by human raters.

These research findings suggest that the automated scoring of speaking and pronunciation tests demonstrates concurrent validity with other assessment measures and exhibits reliability when compared to human-rated scores. This implies that ASR can be a cost-efficient method for assessing low-stake tests such as placement tests. Nevertheless, language schools still require sufficient funding and connections to experts to develop and maintain customizable ASR technology (Isaacs, 2018a). Recent advances in ASR-based dictation systems, such as GVT, may offer a cost-efficient option for scoring placement tests. Surprisingly, there is currently a lack of research investigating the effectiveness of such systems for scoring pronunciation placement tests.

ASR-based Dictation Tools and L2 Pronunciation

As opposed to the commercial and customizable ASR systems used in the studies discussed above, ASR-based dictation tools like GVT have primarily been designed to transcribe spoken language into text within an application. However, early versions of the technology were problematic for L2 speakers due to frequent incorrect transcriptions. This was borne out in Derwing et al.'s (2000) study, which found that proficient L2 speakers of English had accuracy rates approximately 20% lower than native (L1) speakers. Two decades after this initial research, the technology has advanced considerably, leading McCrocklin and Edalathishams (2020) to revisit this topic by examining the accuracy rates of GVT. Their replication study of Derwing et al. (2000) showed improved accuracy rates, with only a 3–5% difference between high-proficiency L2 English speakers and L1 English speakers when dictating sentences in English. Inceoglu et al. (2023) found much lower recognition rates for lower-proficiency

speakers, with GVT only accurately recognizing 40.81% of words. This would indicate that GVT may be an appropriate tool for differentiating between proficiency levels.

Research into classroom use of ASR-based dictation tools in pronunciation courses has shown that it is beneficial to students in various aspects of pronunciation. Studies using control groups as a comparison tend to show significant improvements in the ASR groups in terms of pronunciation of segments in English (e.g., Evers & Chen, 2021; Guskaroska, 2020; Inceoglu et. al., 2020) and in French (e.g., Liakin et al., 2014) as well as intelligibility in French (e.g., Mroz, 2020), demonstrating its pedagogical effectiveness. Furthermore, students have positive perceptions of its use as a language learning tool, despite some frustrations with misinterpretations of their speech in English (e.g., Dillon & Wells, 2021, Inceoglu et al., 2020; McCrocklin, 2016, 2018, 2019; Wallace, 2015, 2016) and in French (Liakin et al., 2017; Mroz, 2018).

While research has explored the efficacy of dictation technology as an L2 pedagogical tool, there remains a scarcity of studies investigating its application in assessing English pronunciation. A study by Dillon and Wells (2021) examined users' perceptions of using the technology in a testing context. We found no other studies exploring the utilization of ASR-based dictation tools for testing purposes, encompassing areas such as teacher and student perceptions, and the usability of such tests.

Google Voice Typing

Google Voice Typing (GVT) is a free ASR-based dictation technology available on Google Drive applications (e.g., Google Docs) when using Google Chrome. As noted earlier, McCrocklin and Edalathishams (2020) found GVT's accuracy rates for L2 speakers of English who had been living in the United States for approximately one year to be only 3–5% lower than

for native English speakers. In a study comparing GVT and Windows Speech Recognition (WSR), McCrocklin et al. (2019) found that: (1) WSR transcribed more words because GVT tended to stop mid-dictation, and (2) Google's system was significantly more accurate when transcribing L2 speakers' utterances. Similarly, Filippidou and Moussiades (2020) found that the Google's ASR interface provided more accurate transcriptions of the sentences produced by Greek L2 speakers of English than IBM's ASR or Wit, a free customizable ASR technology.

Interestingly, Google's dictation tool has been shown to be forgiving of accented speech, but less so of unintelligible speech (e.g., see Johnson & Cardoso, 2024). McCrocklin et al. (2019) found that while most accented words were usually transcribed correctly (e.g., incorrect first vowel and word stress in the word *spaghetti* did not result in an inaccurate transcription), this was not the case when a word was mispronounced as its corresponding minimal pair (e.g., desk/disk). Mroz (2020), in a study examining the use of the ASR in Gmail for pedagogical purposes, found strong correlations between human-rater scores assessing intelligibility and scores derived from the ASR output. While Mroz's (2020) study was pedagogical in nature and had French as the target L2, the findings support the idea that Google dictation technology focuses on intelligibility rather than accent. Cámara-Arenas et al. (2023) also found Google's ASR to be forgiving of accented speech. The authors asked participants to pronounce words correctly and incorrectly and found the technology frequently transcribed words that were intentionally pronounced incorrectly, indicating that the ASR recognizes intelligible speech. However, as the trend in pronunciation teaching has been shifting away from accentedness and toward intelligibility (Derwing & Munro, 2015), it appears that GVT could be an appropriate tool for assessing participants' intelligibility rather than the accentedness of their speech.

Given that accuracy rates vary depending on the proficiency level of the speaker and the fact that GVT tends to focus on intelligibility rather than accent, its use to score speech samples has the potential to improve the usefulness of pronunciation assessments. To explore this potential, our study will analyze the statistical relationship between GVT-based and human-rater scores of a pronunciation placement test. Based on the findings, we will discuss the impact of using GVT to automate the scoring of pronunciation placement tests via Bachman and Palmer's (1996) test usefulness model.

Test Usefulness

In their seminal work on language testing, Bachman and Palmer (1996) emphasize that "the most important quality of a test is its usefulness" (p. 17) and set out a model for evaluating this construct. Previous research has used this model to evaluate ASR-based speaking assessments (e.g., Cox & Davies, 2012). Test usefulness consists of six qualities: reliability (i.e., consistency of the results), construct validity (i.e., the extent to which the score can be interpreted to indicate the level of ability for construct that it purports to measure), authenticity (i.e., correspondence between the characteristics of the test task and actual language use), interactiveness (i.e., involvement of the test taker's language knowledge in accomplishing the test task), impact (i.e., effects on individuals and educational systems), and practicality (i.e., relationship between resources required and resources allotted). The importance of each of these qualities in relation to each other depends on the assessment context. As such, developers need to prioritize constructs that are most pertinent given the test's context and intended use, even if emphasizing those constructs negatively affects the evaluation of less relevant qualities. The debate over the use of constrained tasks in automated speaking assessments (e.g., reading aloud) is a perfect example of this balancing act: while constrained tasks may be limited in terms of

construct validity, such tasks are practical from the point of view of resources needed to score the test.

The implementation of GVT in the pronunciation test explored in this study has the potential to enhance its usefulness according to Bachman and Palmer's (1996) model. Automating the scoring process does not affect the authenticity, interactivity, and impact of the test since the task type and the effects of the results of this low-stakes placement test remain the same. However, automating the scoring process could affect the remaining three qualities, validity, reliability, and practicality, discussed below.

Construct validity of the assessment is high when human raters use a thoroughly tested rubric measuring multiple constructs of pronunciation (e.g., comprehensibility, phonemes, connected speech phenomena, word stress and rhythm, and intonation), as will be described in more detail in the methods section. The use of ASR to score pronunciation may negatively affect construct validity, as scores are based solely on one measure of pronunciation: intelligibility to the ASR system adopted. Concerns about construct validity are compounded by the fact that ASR does not measure important elements of intelligibility (e.g., stress and intonation), which could further reduce construct validity (Loukina et al., 2018). As well, GVT is predictive in nature, meaning that the output provided is not based solely on the phonemes that it recognizes. It also searches a lexical database to determine the collocations with the highest possibility of appearing given the context of the sentence (Ashwell & Elam, 2017). If GVT's predictive nature results in output containing false positives (e.g., the correct transcription of a word despite being produced incorrectly), then construct validity will be reduced.

Biases in GVT could impact the output, affecting the reliability of the scores. If a group is underrepresented in the corpus used for training, the system may have more difficulty

recognizing the speech of an individual from that group. McCrocklin and Edalathishams (2020) found that GVT recognized more words spoken by high-proficiency Spanish speakers than high-proficiency Chinese speakers, indicating potential L1 bias. Studies into gender bias in Google's ASR have been inconsistent, with some studies indicating greater accuracy for women (John et al., 2022), lower accuracy for women (Tatman, 2017), or no difference in accuracy between genders (John et al., 2024; Tatman & Kasten, 2017). While there do not appear to be any age-based bias studies for GVT, research into other ASR systems have shown biases against children's (Shahnawazuddin et al., 2020) and seniors' (Feng, 2021) voices. On the other hand, automated scoring could improve the reliability of the scores as machine scores are not subject to the biases inherent in the use of human raters, as discussed above (e.g., Ling et al., 2014, Saito et al., 2016).

Finally, reducing the expenses associated with hiring evaluators to score placement exams and improved efficiencies in the student placement process through quicker access to test results could improve the practicality of the test. Thus, although there might be a slightly negative impact on content validity, there is the potential for the test to be more useful overall as there could be improvements in reliability and practicality.

There is little research into using ASR-based dictation technology to automate pronunciation tests despite its improved ability to understand L2 speakers. Studies seem to focus on its pedagogical potential, yet its potential for creating a useful assessment tool that language departments looking for cost-effective solutions could adopt should be explored, especially given its apparent ability to differentiate between proficiency levels (e.g., Inceoglu et al., 2023; McCrocklin & Edalathishams, 2020).

The Study

The goal of this study was to examine the feasibility of using GVT to automate the pronunciation scoring for a university's English placement test by comparing scores from human raters. These relationships will be examined through the lens of Bachman and Palmer's (1996) test usefulness model to allow us to establish if using GVT to score a pronunciation assessment is useful when compared to using human scorers. Evaluating the usability of GVT to automatically score pronunciation tests would set the stage for language departments to develop low-cost testing solutions and facilitate stakeholders' decisions. The research questions guiding this study were:

1. What is the relationship between GVT-rated pronunciation scores and the final human-rated pronunciation scores?
2. What is the relationship between GVT-rated scores and human-rated scores across the set of pronunciation evaluation criteria used by the human raters (i.e., comprehensibility; phonemes; connected speech phenomena; word stress and rhythm; and thought groups, prominence, and intonation)?
3. What is the relationship between GVT-rated scores and human-rated scores of high- and low-proficiency levels?

Method

Overview

This study is a quantitative analysis of the ability of ASR-based dictation technology to reliably assess pronunciation for a university language department placement test. The stimuli were recordings of the pronunciation portion of 56 test-takers' placement that consisted of reading aloud five sentences in a timed task. Human raters used an analytic rubric to assess the

recordings. The same sentences were played into GVT in Google Docs, and the recognized words in the GVT output were computed. The scores from the human raters and GVT were correlated to determine if a relationship exists between them.

Research Context and Participants

A language department at a university in Canada originally collected the data for this study as part of the redesign of the rubric used to score the pronunciation portion of their English placement test. Test takers (*participants* henceforth) were undergraduate students taking a placement test for ESL classes or to fulfil linguistic requirements for specific programs. Their oral proficiency levels ranged from A1 to C2 (according to the Common European Framework of Reference for Languages—CEFR). The computerized placement test lasted 90 minutes and included tasks specific to listening, speaking, reading, writing, speech perception, pronunciation, critical reading, syntax, and vocabulary. An independent score was given for each language aspect being measured based on the task(s) for that aspect (e.g., a score for speaking was given based on the speaking task and a separate score for pronunciation was given based on the pronunciation task). For this study, only data about the pronunciation portion were analyzed. To ensure the sample contained varied levels of proficiency, seventy-five samples were randomly chosen from a collection of recordings of over 10,000 placement tests taken from 2015 to early 2020, based upon the pronunciation score given at the time of the test. Of these samples, only 56 were used due to issues with sound quality (e.g., the participant spoke so quietly that they were not comprehensible or intelligible to GVT). Thus, the final sample of participants were 56 adults (N = 56; 21 males, 35 females; average age: M = 28.09, SD = 7.78), with French (68.42%), Spanish (7.02%), and Arabic (5.62%) being the most frequently reported among 12 L1s.

The participants were divided into two proficiency levels. The lower-proficiency level group (n = 20) had obtained a score below 72 from the human raters on their scoring of this task using the analytic rubric, described below, with the higher-proficiency level group (n = 36) obtaining a score of 72 and above. This score was used as the cut-off point because a score of 72 to 86 indicated that the test taker was to be placed in an advanced pronunciation class, and a score of 86 and above resulted in an exemption from a pronunciation course.

Data Collection Materials

Pronunciation Samples

Participants read aloud five increasingly difficult sentences that appeared sequentially on a screen. The first sentence served as a baseline sentence and was identical for all students; the remaining sentences were randomly chosen from a bank of 12 groups of four sentences each, with sentences 2 to 5 increasing in difficulty. As the test items are currently in use and protected by confidentiality, it is impossible to disclose the baseline sentence used in the study. However, we are allowed to share one of the sentence groups, which will be used for all examples throughout the article (see Table 2.1).

Table 2.1

Sample Sentences

Level	Sentence
2	A trio sings to the audience as it streams onto the busy street in the cold rain.
3	These are more sophisticated pictures, aimed at a particular kind of filmgoer. Is she sure that this audience understands them?
4	After the stems are cut off the mushrooms, they are then going to be sautéed with a small onion, a clove of garlic, and an eighth of a cup of breadcrumbs.
5	Even though the trailer has been cleaned, there are still lingering traces of acetones and other toxic amalgams, either in the gaskets or in the valve assembly.

Rubric

The human raters used an analytic rubric consisting of five bands of proficiency (i.e., poor, fair, good, very good, excellent) to measure five criteria of pronunciation: (1) comprehensibility (how easy to understand is the speaker?), (2) phonemes (e.g., are the vowels and consonants in "sings" intelligibly produced?) (3) connected speech phenomena (e.g., are "t" and "o" in "cut off" resyllabified, as in "*cut off the mushrooms"?)*, (4) word stress (e.g., is "*assembly"* stressed on the penultimate syllable*?*) and rhythm (e.g., are alternating weak and strong stress patterns produced appropriately?), and (5) thought groups, prominence, and intonation (e.g., does "*Is she sure that this audience understands them?*" contain a prominent syllable with appropriate rising intonation to define a thought group?).

Procedure

Pronunciation Samples

All pronunciation samples were recorded at the language department testing centre at the university during actual placement tests. During the assessment, participants sat at computers and wore headphones with microphones. After completing the study consent screen, they read aloud two unevaluated practice sentences to ensure they understood what to do and to verify the functioning of the recording system. Participants then had 20 seconds to read aloud each sentence before the screen automatically advanced. Responses were recorded and saved in the university's database.

Human-rated Scores and Analysis

Three raters (1 male, 2 females) independently scored each participant's pronunciation sample using the analytic rubric. All raters were ESL instructors at the university, with five to twenty years of experience in teaching and evaluating pronunciation. Their ages ranged from 30

– 50 years, and they were L2 French speakers.[1] Raters gave a score of 1 to 5 for each of the five criteria described above. To ensure inter-rater reliability, the three raters shared their scores for each criterion for each participant. In case of disparity in a score for a criterion, the raters reached consensus through discussion. The sub-scores were added together for the final score, for a maximum of 25, then multiplied by 4 for a total out of 100.

GVT-rated Scores and Analysis

An analysis of the GVT output was used to score the same speaking samples. The recorded samples were played into GVT, resulting in an output of the words recognized by the ASR software. The first author analyzed the output for each participant to determine how many words GVT correctly identified in each sentence. Following the procedure established in Cox and Davies (2012), a proportion of correct words for each sentence was determined by dividing the number of correct words by the number of words in the group of 5 sentences, multiplied by 100 for a percentage score. It was not feasible to determine a GVT-score for the five pronunciation sub-scores from the analytic rubric described above for a reason that is evident: GVT cannot assess any of these measures. However, these sub-scores will be compared with the GVT scores for further analysis.

[1] As noted by an anonymous reviewer, there is a potential for bias in the human rater scores as all raters were L2 French speakers and 68.42% of the test takers were L1 French speakers. We tried to determine if scores for French speakers were higher than for any other language group. This proved difficult as there were sometimes only one or two test takers from a specific L1. For example, the L1 that received the highest average score was Romanian (M = 92), which is not a language spoken by any of the raters. However, there were only two test takers in this group, which does not allow for a meaningful comparison. We did compare the mean scores for all L1 French speakers (M = 77.37) to that of all other test takers combined (M = 61.41). While this indicates the scores were higher for L1 French speakers, it may not be a result of bias. It could be attributed to the fact that most of the L1 French speakers were from Quebec. In Quebec, ESL classes are mandatory throughout primary and secondary school. Students must pass a standardized ESL exam to graduate high school. They also take two more years of ESL courses at a post-secondary level before attending university. This, combined with Quebec's geographical location in North America, which allows for more interactions with L1 English speakers and English media than is typical for most L2 speakers, could have led to the higher scores. As such, it is entirely possible that the L1 French speakers were higher in proficiency that the other test takers. In fact, the corresponding means were almost identical for the GVT scores (M = 78.78 and M = 61.90 respectively; compare with the 77.37 and 61.41 illustrated earlier), indicating that L1 French speakers were more intelligible.

The following criteria were established for determining accuracy:

1) Repetitions of words or sentences were not analyzed to ensure reliability and fairness.
2) Correctly identified words were considered as accurate and were given one point.
3) Homophones such as *two* and *too* were considered correct.
4) Compound words (e.g., breadcrumbs, onto) written as two separate items were considered accurate. If only one of the elements was accurate, a half point was given.
5) Missing or added grammatical inflections (e.g., plural -s, past tense -ed) resulted in the loss of half a point.
6) Regular infinitive verbs (e.g., think) in place of irregular past tense verbs (e.g., thought) resulted in the loss of half a point.
7) Contracted forms of non-contracted words and non-contracted forms of contracted words were considered accurate (e.g., the pair *don't* and *do not*).

Data Analysis

The GVT score, the final score attributed by the human raters and the sub-scores they attributed for each criterion were entered into SPSS 28. A visual examination of the data using scatterplots seemed to indicate a linear relationship between the variables and did not indicate the presence of any outliers. An examination of the histograms indicated the data did not have a normal distribution. A Shapiro-Wilks test for normality verified that the data for all variables significantly deviated from normal ($p < .05$). Therefore, the decision was made to run the nonparametric Spearman's rank-order test. A visual examination of the data indicated a monotonic relationship between the variables, as required for a Spearman's rank-order test. To construct robust confidence intervals for the rubric criteria scores and proficiency level scores, a

bootstrap resampling procedure was utilized, generating correlation coefficients and associated 95% bias-corrected and accelerated (BCa) confidence intervals (CIs) based on 1,000 resamples.

Results

Regarding the first research question about the relationship between the GVT and final human-rated scores, the results indicated a statistically significant strong correlation between the two variables, $r_s\,(54) = .78$, $p < .001$.[2] For the second research question about the relationships between the GVT scores and the rubric criteria, the results indicated statistically significant strong correlations between the GVT score and each of the criteria (see Table 2.2 for all descriptive statistics and Table 2.3 for a summary of the correlations). As noted above, these correlations are between the GVT score and each criterion as the GVT output cannot assess the individual criterion. Regarding the third research question touching upon the relationship between GVT scores and test-taker proficiency, a significant strong correlation was found between the GVT and human-rated scores for lower-proficiency participants, $r_s\,(54) = .78$, $p < .001$. However, a non-significant weak correlation was found between the GVT scores and the human-rated scores for higher-proficiency participants, $r_s\,(54) = .28$, $p = .100$ (see Table 2.4 for a summary of the correlations).

[2] Based on a reviewer's recommendation, we attempted to also run a two-way mixed-effect model intraclass correlation coefficient test (ICC) with absolute agreement to answer the first and the third research questions. For RQ 1, we verified that all assumptions were met. The results showed good to excellent reliability between the human-rater and GVT scores, ICC = .93, (95% CI, .88 to .96). We endeavored to run the same test for the low- and high-proficiency groups for RQ3; however, the data for the high-proficiency group did not meet the assumption of homoscedasticity, as required for ICC tests. As we do not wish to risk publishing possibly inaccurate and unreliable results, we decided to not run the ICC test for the proficiency levels. As such, we decided to keep the original plan of analysis and use the Spearman rank order test, as it is a nonparametric test.

Table 2.2

Descriptive Statistics

Variable	M	SD
GVT Score (/100)	73.09	22.49
Human-rated score (/100)	72.00	26.95
Comprehensibility (/5)	4.14	1.20
Phonemes (/5)	3.39	1.47
Connected speech (/5)	3.34	1.51
Word stress and Rhythm (/5)	3.63	1.34
Thought groups, sentence stress, intonation (/5)	3.50	1.51
Lower-proficiency: GVT score (/100)	47.74	13.78
Lower-proficiency: Human-rated score (/100)	40.20	16.80
Higher-proficiency: GVT score (/100)	87.18	10.96
Higher-proficiency: Human-rated score (/100)	89.67	9.44

Table 2.3

Correlations between GVT-score and Human-rated Scores by Rubric Criteria

Rubric Criteria	rho	95% BCa Cis
Final score	.78**	.64, .88
Comprehensibility	.85**	.74, .90
Phonemes	.78**	.63, .88
Connected speech	.72**	.53, .84
Word stress and rhythm	.71**	.52, .84
Thought groups, sentence stress, and intonation	.79**	.65, .88

Note. Confidence intervals based on 1000 bootstrap samples.

***p* <.001.

Table 2.4

Correlations between GVT-score and Human-rated Scores by Proficiency Level

Rubric Criteria	rho	95% BCa Cis
Lower-level proficiency	.78**	.56, .89
Higher-level proficiency	.28	-.03, .55

Note. Confidence intervals based on 1000 bootstrap samples.

***p* <.001.

To summarize, the results indicate a strong association between GVT and human-rated scores when assessing L2 pronunciation. This strong association is evident regardless of the

evaluation criteria adopted by the human raters. The findings also suggest a stronger association when assessing lower-proficiency participants, as an examination of the scores attributed to those with a higher-proficiency level showed that the scores were less comparable across the two scoring methods.

Discussion

This study set out to determine the relationship between GVT and human-rated pronunciation scores. The results for the first two research questions, which focused on the relationship between the GVT score and the human raters' final scores and scores for individual criterion of the rubric showed strong correlations and suggest a strong relationship. However, strong correlations were not found across proficiency levels, which was the focus of the third research question. The scores for the lower-level proficiency participants showed significantly strong correlations between GVT and human-rated scores while the correlations for the high-proficiency participants were weak. This suggests that GVT may score lower proficiency participants more accurately than those with a higher proficiency.

The finding that GVT and human-rated scores are highly correlated corroborates the literature examining the relationships between ASR and human-rated scores (e.g., Graham et al., 2008, Luo et al., 2009; Mroz, 2020). However, the weak correlation seen among the higher proficiency-level participants diverges from McCrocklin and Edalathishams (2020). Although their study did not compare GVT and human scores, they did find high accuracy rates among fluent L2 speakers when using GVT. It must be noted that the contexts of the two studies are very different. First, participants in McCrocklin and Edalathishams (2020) were speaking directly into GVT whereas this study used recordings. Second, the participants in the current study may have been under greater stress as they were taking a test with real-life consequences

(i.e., course placement). These differences may have had a negative impact on the quality of the speech samples in this study.

Cox and Davis (2012) and van der Walt et al. (2008) both concluded that using ASR to automate pronunciation placement tests is a cost-effective solution for language schools. This study partially corroborates their findings and contributes to the field as it found that free and readily accessible ASR-based dictation technology, such as GVT, is also effective at assessing pronunciation for placement tests and at a lower cost than purchasing and customizing ASR technology.

The positive impacts of using GVT to score pronunciation assessments are highlighted by examining the three qualities of Bachman and Palmer's (1996) test usefulness model affected by implementing this new technology: reliability, validity, and practicality (as discussed earlier, authenticity, interactivity, and impact are not relevant to the current study).

Reliability

The strong correlation between GVT and human-rated scores indicates reliability in scoring across the two methods. Due to the smaller-scale nature of this study and the fact that the raters reached a consensus rather than providing individual scores, we could not carry out meaningful tests on the reliability of the GVT scores versus the reliability of the scores of each rater. However, the literature has observed high levels of reliability in automated assessments for speaking and pronunciation in larger-scale studies (e.g., Bernstein et al., 2010). This is not surprising, since machine scoring adds objectivity to an inherently subjective task when completed by humans (Isaacs, 2018b) by eliminating the biases inherent to human raters, such as accent familiarity (Browne & Fulcher, 2016), experience with the target L2 speech (Kennedy & Trofimovich, 2008; Saito et al., 2016), personal biases (Yan & Ginther, 2018), and fatigue (Ling

et al., 2014). For this reason, the use of GVT to score pronunciation assessments should increase the usefulness of a test through an increase in reliability.

One area of concern, however, is the low correlation between GVT and human-rated scores for higher proficiency participants. We went back to the recordings and the GVT output for the six high-proficiency participants with the largest discrepancy between the two scores. The sound quality on the recordings was less than optimal for three of the six as they spoke more quietly and more quickly than the other participants. This seems to have made it more difficult for GVT to accurately transcribe these participants' speech, even though they were highly comprehensible to the human raters. There were also sound quality issues for some lower-proficiency participants who sometimes spoke more quietly and/or quickly than most participants, making it difficult for GVT to determine what they were saying. However, since these participants were also lower in comprehensibility to the human raters, both GVT and human-rated scores were low. GVT's difficulty in accurately transcribing speech from samples with sound quality issues, such as speech that is too quiet or too fast, may negatively impact the reliability of scores for high-proficiency more than lower-proficiency participants.

The reason behind the remaining high-proficiency test taker's low scores with GVT may be related to the test items (i.e., sentence group; see Table 2.1) they received. Overall, the average GVT score for all participants was slightly higher than the average human-rated score (73.09 vs. 72.00). However, the average GVT score for all participants, regardless of proficiency level, who read the same test items as these three participants was five percent lower than the average human-rated score for those sentences. A comparison of the frequency of the lexical items in the sentences read by these three participants to the other sentences used in the placement test was carried out in Lextutor's Vocabprofiler (Cobb, n.d.). The sentences in

question have a lower percentage of the most frequent 1,000 words (66.77 vs. 77.33), a higher percentage of the most frequent 2,000 words than average (9.28 vs. 4.39) and a higher percentage of off-list words than average (16.49 vs. 11.97). Of the 16 off-list tokens, 7 can be found in one sentence, possibly impacting the participants' ability to produce the utterance intelligibly as certain low-frequency words in the sentence (i.e., amalgams) may have been unknown to the participants. Mispronunciation of these tokens may have affected GVTs comprehension of the utterance and its ability to predict what was being said. Ashwell and Elam (2017) found that Google's ASR has difficulty with less frequent constructions such as unusual collocations (e.g., infrequent "first coat" versus "first time") and sentences with ambiguous constructions (e.g., tag questions) because it is predictive in nature. The human raters, on the other hand, may have been more forgiving of errors due to the high difficulty level of the sentences or may have found the speech more comprehensible because they knew the content of the sentences (Yan & Ginther, 2018). Another possible answer can be found in Cámara-Arenas et al.'s (2023) study of Google's API. They found the technology had difficulty recognizing low-frequency words. For example, the word "glom" was never recognized during the course of their study, regardless of whether it was pronounced correctly. Perhaps infrequent words, due to their nature, are less prevalent in GVT's lexicon and as such the technology has difficulty recognizing them, resulting in false negatives in the GVT output (e.g., an incorrect word is transcribed even though it was produced correctly). Therefore, when using GVT to assess pronunciation, test items should be verified in GVT to ensure that they present equal challenges to the technology. This would improve reliability and validity, as participants would not be penalized due to a decrease in intelligibility caused by the presence of linguistic elements in a certain item that are problematic for the system to understand.

Validity

Initially, there was a concern that using GVT to score the assessment could negatively affect construct validity. The human raters used a rubric measuring multiple constructs of pronunciation (e.g., comprehensibility, phonemes, prosodic features), but GVT scores were based solely on one measure of pronunciation (i.e., intelligibility). However, the GVT scores had strong correlations with each criterion in the rubric, indicating relationships between these measures. This was initially surprising, since two of the criteria, *word stress and rhythm* and *thought groups, prominence, and intonation*, are solely prosodic in nature. In fact, the criterion *thought groups, sentence stress and intonation* has the second-highest correlation ($r_s = .79$) with the GVT score. These correlations may be explained by past studies that have suggested that both lexical stress (Field, 2005) and primary sentence stress (Hahn, 2004) are factors affecting intelligibility. Perhaps despite GVT's difficulty in targeting prosody, prosodic features may affect it due to their impact on intelligibility. Whatever the reason, the strong correlations between GVT and the individual criterion scores suggest that using GVT to assess pronunciation did not reduce the breadth of the phonological aspects measured. Therefore, in terms of test usefulness, validity would seem to be unaffected by using GVT as the means to calculate scores.

Practicality

Practicality is perhaps the quality of test usefulness benefiting the most from the use of GVT to score pronunciation placement tests. Although there would be initial costs for developing a scoring application and integrating it with GVT and maintenance costs, these could be more than offset by reduced human resources expenses in terms of raters. In cases where teachers' contracts and workloads include scoring placement tests, this task could be replaced with other tasks that could further benefit students' language learning. As well, language departments could

realize efficiencies in the placement process and course registration as the scores would be available to students and administrators in a timelier manner. This could also have a positive impact on student satisfaction. As Isaacs (2018b) notes, the short wait time for results is a repeated theme in test taker testimonials appearing on Pearson's website for Versant. The increase in practicality could allow language schools and departments to save on time, money, and effort, which could be redistributed to benefit students in other aspects of their language learning, resulting in an increase in test usefulness.

Conclusion

The aim of this study was to determine if there was a relationship between GVT and human-rated pronunciation scores in view of using GVT to increase the test usefulness of the pronunciation portion of a university language department placement test. The robust correlations observed between the scores demonstrate a strong association, indicating that the GVT-based scores are both reliable and valid, especially for lower-proficiency speakers. This finding suggests that GVT should be further explored as an automated scoring tool for pronunciation, as it could offer language schools the potential to enhance the practicality of their placement tests and achieve cost savings through machine scoring.

It must be noted that there are certain limitations to this study that should be considered. With only 56 participants engaged in a constrained read-aloud task, the findings are not generalizable to other populations. As well, the sound quality of the recordings may have affected the output analyzed by the human raters and GVT. Recordings that were inaudible to GVT were not used; however, among the recordings that were retained, some had background noise or other minor issues with sound quality. This may have impeded GVT's ability to correctly process what was being said and negatively impacted the scores generated from the

output. However, there will always be participants who speak too quietly or other issues affecting the quality of the speech sample being assessed. When faced with similar sound quality issues, ETS developers ensured the reliability of the scores by establishing a system that flagged recordings with low sound quality and sent them to human raters (Isaacs, 2018a). To maintain the reliability of scores in the case of sound quality issues, language institutions could also consider establishing a system to test for and flag recordings with low sound quality and having human raters assess these samples.

This study indicates that GVT has the potential to be used to automatically score low stakes pronunciation assessments such as placement tests, but further research is needed. Luo et al. (2009) found different correlations between different task types (i.e., shadowing versus read aloud) when using ASR to score pronunciation. Research should investigate if the same holds true when using dictation technology. Research validating test items should also be carried out to ensure other linguistic features, such as vocabulary frequency, do not cause problems with reliability and validity. Finally, comparisons of scores for participants of different L1s, genders, and ages are needed to ensure biases do not exist within GVT.

Many university language departments and language schools have faced obstacles in implementing automated pronunciation assessment for placement tests due to the high expenses involved in acquiring, customizing, and maintaining the required ASR technology. ASR-based dictation tools like GVT offer a promising cost-effective solution. The results of this study demonstrate a strong correlation between GVT-generated scores and human-rated scores of the same speech samples. Utilizing this technology could assist institutions in streamlining the placement process and increase student satisfaction and confidence in the testing process.

Chapter 3: Real Eyes or Real Lies? Implications of Infrequent Vocabulary, Unusual Collocations, and Ambiguous Linguistic Elements for ASR-based Pronunciation Assessment

In recent literature surrounding second language (L2) pronunciation assessments, automated scoring has received increasing attention. In Kang and Ginther's (2018) edited volume examining current issues in pronunciation assessment, a quarter of the chapters are dedicated to machine scoring. This growing focus reflects its appeal to test developers: as Isaacs (2018a) notes, automated scoring is attractive because of potential long-term cost savings, despite the initial costs of developing the system. Additionally, automated scores are potentially more reliable compared to subjective human scores (Isaacs, 2018b). However, for university language departments and language schools wishing to develop automated systems, the costs associated with purchasing, training, and maintaining the automatic speech recognition (ASR) technology can be beyond their financial means.

Dictation ASR, the ASR technology found in word processing systems such as Google Docs, has been proposed as a low-cost alternative to customizable ASR. This technology is free or inexpensive, and training and maintenance are handled by the companies that developed it. In separate studies, Johnson et al. (2024) and Nelson and Cardoso (2024) used identical speech samples to compare human-rater scores with those generated using Google Voice Typing (GVT) and Microsoft Transcribe (MS-T), respectively. Both studies found strong correlations with human scores. However, when participants were divided by proficiency levels, correlations remained strong only for the low-proficiency group, which is counterintuitive, as recent research on GVT has shown high recognition rates for high-proficiency L2 speakers (McCrocklin & Edalatishams, 2020). Johnson et al. (2024) suggested the issue stemmed from lexical variation

across test versions, with some versions containing infrequent words that GVT failed to recognize. This aligns with Cámara-Arenas et al. (2023), who also found Google's ASR struggled with infrequent lexical items. High-proficiency participants may have been penalized due to limited awareness of GVT's lexicon.

Test developers working with customizable ASR understand fully how their system processes language as they trained and developed it. In contrast, dictation ASR comes with the drawback that test developers do not know this information since they did not train and customize the system. It is crucial that researchers and test writers explore how dictation ASR processes language to avoid the unwitting use of problematic linguistic features that negatively affect test scores. To address this concern, this study will examine the impact of potentially problematic linguistic elements (e.g., infrequent vocabulary, unusual collocations, and ambiguous connected speech) on GVT and human-rater transcripts to determine if these elements could impact pronunciation test scores generated using dictation ASR.

Literature Review

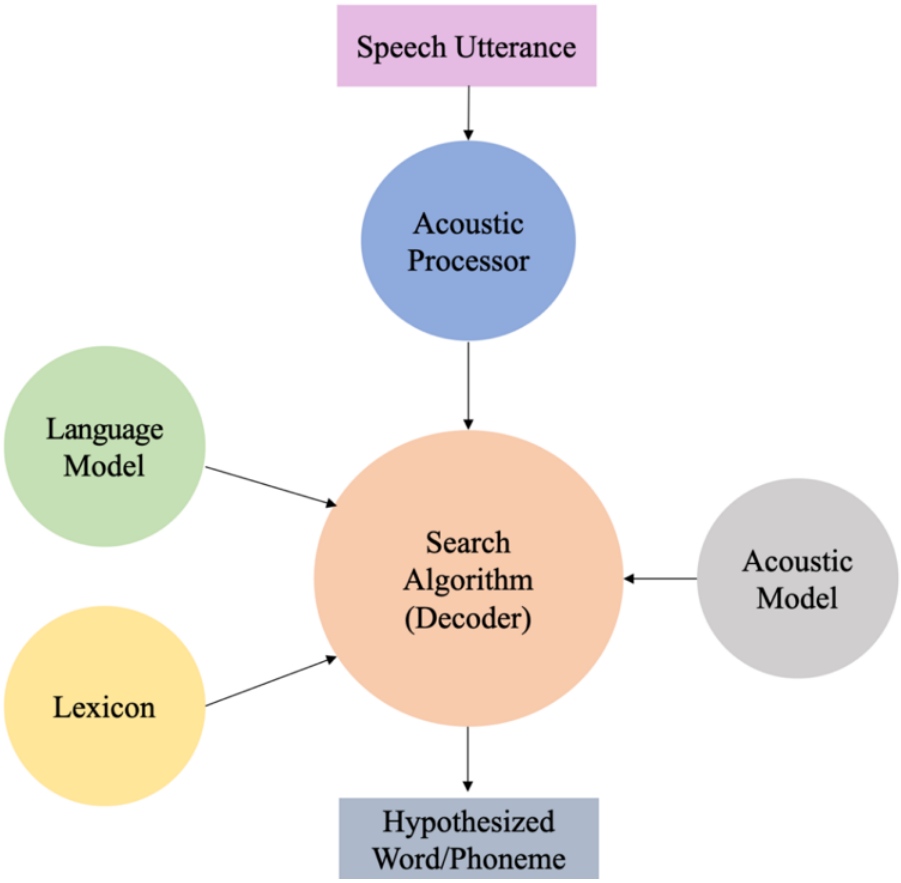
Automatic Speech Recognition (ASR)

ASR converts speech into text on the screen of a computer or electronic device. Speech is analyzed by an acoustic processor that determines the phonemic components of the utterance. It is then analyzed using a decoder algorithm to search lexical and language model databases, allowing it to hypothesize and predict the component items of the utterance through an analysis of potential lexical items and morphosyntactic options. In parallel, an acoustic model statistically analyzes the probability of various phonemes and/or lexical items appearing in the same utterance. The utterance with the highest probability is transcribed on the screen (Filippidou & Moussiades, 2020). Thus, as indicated in Figure 3.1 the pronunciation of a word's segmental

features is only one factor involved in determining the transcription; language and lexical features also play a large role.

Figure 3.1

Automatic Speech Recognition Architecture



Note. Adapted from Filippidou and Moussiades (2020, p. 75).

Developers wishing to integrate ASR into their product can choose from systems with varying levels of customization. Fully customizable ASR allows developers to train the models to recognize speech using a corpus appropriate for their purpose (Filippidou & Moussiades, 2020). However, purchasing, developing, and maintaining the system is expensive. Less customizable ASR systems, such as Google's ASR application programming interface (API) are pre-trained, leaving the developer little control over the way the system processes language but

offering a more cost-effective solution (Ashwell & Elam, 2017). The least customizable, dictation ASR, is already trained and integrated into products or devices frequently used such as GVT in Google Docs or Apple's Siri. Dictation ASR is usually free, making it attractive to developers with financial limitations. However, they cannot customize the language processing and have limited control over integration into their products.

Using Dictation ASR to Score Pronunciation Assessments

Fully customizable ASR has been used to score speaking and pronunciation proficiency and placement tests for the last two decades, and the validity and reliability of these systems have been established through multiple studies (e.g., Bernstein et al., 2010, Cox & Davies, 2012). Research into using APIs and dictation ASR for assessment purposes is relatively young, and to the best of my knowledge, only two studies comparing scores generated using dictation ASR and from human raters have been conducted: Johnson et al. (2024), which used GVT, and Nelson and Cardoso's (2024) replication study, which used MS-T. Using the same pre-recorded samples from a university language department's placement test, both studies found strong correlations between human-rater and dictation ASR scores. However, when comparing the scores across proficiency levels, those of the lower-proficiency test takers had strong correlations, while high-proficiency test-taker scores had weak correlations in both studies, suggesting that dictation ASR had more difficulties assessing higher proficiency test takers.

While these studies are encouraging about the feasibility of using dictation ASR to score pronunciation assessments, the weak correlation for the high-proficiency level is concerning and does not corroborate current research. Studies of L2 speakers using GVT have shown high transcription accuracy rates for high-proficiency users (e.g., McCrocklin & Edalatishams, 2020)

and correspondingly low rates for low-proficiency users (e.g., Inceoglu et al., 2023). It is confounding that the higher-proficiency test takers had lower accuracy rates lower than expected.

To explore this inconsistency, Johnson et al. (2024) examined the transcriptions of six high-proficiency test takers whose GVT scores were low. Three test-takers spoke very quietly, impacting the ASR's ability to recognize their speech. The other three had received the same version out of twelve possible test versions. A verification of the scores of all test takers who received that version revealed their GVT scores were five percent lower than the corresponding human-rater scores, despite higher overall GVT averages. Further analysis showed that the test version contained fewer high-frequency words and more low-frequency lexical items, including several off-list items (e.g., *amalgams*). GVT may not have recognized these words, even when correctly pronounced, resulting in low scores. This effect is less apparent among lower-proficiency test takers, who may struggle to pronounce infrequent words. As a result, high-proficiency test takers could be unfairly penalized.

There is a scarcity of studies into this issue. A review of the literature revealed only two studies examining the impact of linguistic elements on non-customizable ASR transcriptions: Ashwell and Elam (2017) and Càmara-Arenas et al. (2023). Both studies were undertaken during the development of computer-assisted pronunciation training programs using Google's API to score student attempts. No studies regarding the influence of linguistic items in a purely assessment-based context were found. This gap in research needs to be addressed considering that using certain linguistic features (e.g., low-frequency vocabulary) in test items may reduce the reliability of the scores, which compromises test fairness.

Reliability and Social Consequences of Score Inferences

The issue impacting high-proficiency test takers' scores indicates a lack of reliability, the consistency in the measurement of scores, when using GVT to assess pronunciation. This can occur when using non-uniform test versions (Bachman & Palmer, 1996), such as an unequal presence of linguistic features that GVT struggles to recognize, raising concerns about test fairness. While the test in Johnson et al. (2024) is a relatively low stake pronunciation placement test, inferences made from the results can impact students' academic paths, with denied access to courses aligned with their proficiency levels and language learning needs. The social consequences of inferences made from test scores cannot be ignored by test developers (Messick, 1989). Considering the potential negative consequences for test takers, further study of the impact of linguistic elements on GVT transcription accuracy is warranted. An examination of the GVT transcripts from Johnson et al. (2024) and a review of the literature on GVT accuracy uncovered three potentially problematic linguistic features: low-frequency vocabulary, unusual collocations, and ambiguous connected speech.

Potential Impact of Linguistic Features on GVT Transcriptions

Human raters have previous knowledge of test items, assisting in their understanding of the test takers (Yan & Ginther, 2018). GVT, however, does not have this prior knowledge and may struggle to recognize certain linguistic features, especially those lower in frequency or that do not follow expected patterns. The following section will examine three such features and their potential impact on GVT-generated scores: infrequent vocabulary, unusual collocations, and ambiguous connected speech.

Infrequent Vocabulary

If a word rarely appears in or is absent from the corpus used to train an ASR system, the algorithm may not consider it as a probable candidate for transcription. For example, the word *glom* was never recognized in Cámara-Arenas et al.'s (2023) study in which 13 native speakers (L1) of English read a list of 22 words into Google's ASR API. *Glom* falls within the lowest frequency band; that is, it appears within the last 1,000 of the 25,000 most frequent word families in the British National Corpus/Corpus of Contemporary American English frequency list (NBC/COCA; Nation, 2017). It is possible that the algorithm ignored it as a probable candidate. Interestingly, the word *wan*, despite being relatively low-frequency (i.e., within the first 4,000-word level), was almost always recognized in their study. Understanding which infrequent lexical items are not recognized by GVT is important since inadvertently writing a test item containing them would compromise reliability and test fairness.

Infrequent vocabulary also includes given names. Ashwell and Elam (2017), in a study of Google's ASR API, found that *Tom* was the word the least recognized when Japanese L2 speakers of English read sentences aloud, despite human raters consistently judging it as correct. However, some names, such as *Robert*, were always recognized. Given names are often not well represented in corpora as names are high-frequency in certain texts, but low-frequency or non-existent in others (Nation, 2013). As a result, they may play a minimal role in the training of ASR systems, raising questions about Google's ASR ability to recognize names and proper nouns in general. Yet, research into this issue is limited, highlighting the need to better understand how given names and other proper nouns impact the reliability and fairness of GVT-generated test scores.

Unusual Collocations

Collocations are words that commonly occur together (e.g., *last night*) or follow predicable patterns (e.g., *last* followed by a noun) (González Fernández & Schmitt, 2015). Ashwell and Elam (2017) found that Google's ASR is sensitive to collocation frequency. For example, *first coat* had a low accuracy rate, but the rate improved greatly when the authors changed the collocation to *first time*. In the COCA database (Davies, 2008), *time* is the most frequent collocate with the node word *first* with over 91,000 examples. However, *first coat* only appears twice, indicating that Google's ASR favours collocation frequency over acoustic information.

A problematic collocation from Johnson et al. (2024) is in the sentence *These are more sophisticated pictures aimed at a particular kind of filmgoer*. *Aimed* was incorrectly transcribed for five of the seven participants. In COCA, the pattern of *pictures* followed by a past participle (as in *pictures aimed*) is relatively uncommon. The most frequent collocate, *taken*, appeared only 84 times, while only one instance with *aimed* appeared. *Pictures aimed* is an unusual collocation because of lexical pairing and its syntactic structure (i.e., a past participle following *pictures*). Given that this collocation is unusual for both aspects, GVT's language model may have had difficulty processing it, underscoring the need to examine how such atypical collocations impact scoring and, ultimately, test fairness.

Ambiguous Connected Speech

Connected speech processes (CSP) occur when words are spoken in continuous speech, rather than in isolation. These phenomena can hamper segmentation and lead to ambiguous speech (Alameen & Levis, 2019). Consider, for example, the pronunciation of *eyes* and *lies*, which are pronounced differently when spoken in isolation ([aɪz] and [laɪz] respectively);

however, when the word *real* precedes them, their pronunciations become identical ([ɹiəlaɪz]), as coda /l/ re-syllabifies as the onset of the following vowel in *eyes*. CSPs are natural and encouraged in L2 speech. In fact, the human-rater rubrics used in Johnson et al. (2024), which compared a human- versus GVT-scored pronunciation assessment, rewarded their accurate production. However, if CSPs result in ambiguous speech that GVT cannot accurately recognize, a test taker would be unfairly penalized.

Other CSPs can also lead to ambiguous speech. One such process involves geminate consonants, which occur when the same consonant appears at the end of one word and the beginning of the next, blending across the word boundary. For example, *I watched TV* (past tense) and *I watch TV* (present) sound nearly identical, and even L1 English speakers struggle to differentiate them without contextual support (Strachan & Trofimovich, 2019). Another example is the difficulty in distinguishing *known ocean* ([noʊnoʊʃən]) from *known notion* ([noʊnːoʊʃən], which is rendered ambiguous by both linking and gemination (Mattys & Melhorn, 2007).

Similarly, assimilation, where a segment changes its place or manner of articulation (including voicing) due to the influence of a neighbouring sound, can also cause ambiguities in connected speech. For example, when /p/ precedes /b/, /p/ may assimilate to the voicing of /b/ and be realized as [b], resulting in *cap belonged* and *cat belonged* having identical pronunciations ([kæbːəlɑŋd]). Ambiguity due to assimilation is problematic for human listeners (Gaskell & Marslen-Wilson, 2001) and ASR systems (Pouw et al., 2024), particularly when context is limited.

Finally, another source of ambiguity in connected speech is homophones: words or word combinations that are phonetically identical but differ in meaning. For example, when spoken fluently, *car go* and *cargo* are pronounced identically ([kɑɹgoʊ]), making them difficult to

decipher without context clues (Kim et al., 2012). The presence of ambiguous speech in a test being scored by GVT could, ironically, benefit less fluent test takers while penalizing more fluent ones, undermining validity and test fairness.

A Cumulative Problem

The use of infrequent vocabulary, unusual collocations, and ambiguous connected speech may affect GVT transcription accuracy. The effect of these three features on a test taker's score is problematic; however, the issue is compounded by the predictive nature of GVT. The inability to recognize one word affects the analysis of the entire sentence in the lexical and language models, negatively impacting the system's capability to accurately predict or hypothesize the utterance and causing a ripple effect across the transcription. For example, when the problematic *Emma* was used after the word *would* in Ashwell and Elam's (2017) study, the system had trouble recognizing *would*. Replacing *Emma* with the more easily recognized *Robert* unwittingly improved the accuracy rate for *would*. *Would* is the 59th most frequent word in COCA (Davies, 2008); one can assume it is prevalent in GVT's lexicon. Regardless, its proximity to a problematic lexical item made it more difficult for GVT to recognize. Test takers may be exponentially penalized when scoring an assessment using GVT due to the presence of problematic linguistic features, highlighting the need for further research into this issue.

Current Study

Scores generated using dictation ASR have been shown to strongly correlate with human-rater scores. However, concerns have emerged that some linguistic features (infrequent vocabulary, unusual collocations, and ambiguous connected speech) are difficult for GVT to recognize even when pronounced accurately, disadvantaging high-proficiency test takers (Johnson et al., 2024). To ensure the reliability of GVT-generated scores, it is essential to

examine the potential impact of certain linguistic elements on transcriptions. Thus, this study was framed by the following research question (RQ): In what ways do transcripts generated by GVT and human raters differ for: (a) infrequent vocabulary, (b) unusual collocations, and (c) ambiguous connected speech?

Method

Design

This quantitative study addressed concerns that certain linguistic features affect GVT transcription accuracy, impacting the reliability of placement test scores generated using this technology. Four speaking participants were recorded reading 60 sentences containing the linguistic elements being examined (details provided below). Human raters completed an intelligibility transcription task for each recorded sentence, and the same recordings were input into GVT for automatic transcription.

The transcriptions from GVT and the raters were used to determine two accuracy scores for each sentence: one for the targeted linguistic feature and one for the entire sentence. Differences in the distribution of accurate and inaccurate transcriptions between GVT and human raters across features were examined using a chi-square test of independence. Additionally, GVT and human-rater transcripts were analyzed to explore the impact of infrequent vocabulary, unusual collocations, and ambiguous connected speech.

This study received ethics approval from Concordia University (Protocol: 30020735) and adhered to standard procedures, including obtaining informed consent, maintaining confidentiality, and ensuring voluntary participation.

Participants

The study included two categories of participants: speaking participants and expert raters. Speaking participants ($N = 4$), recruited using a convenience sampling method, were two adult L1 English speakers and two adult simultaneous bilinguals who identify French as their primary language. All four speak a North American variety of English. Expert raters ($N = 8$) were recruited through emails to language departments and to graduate students studying applied linguistics and through snowball sampling. All raters met the minimum requirements of: (1) a bachelor's degree specializing in Teaching English as a Second Language or a graduate degree in Applied Linguistics or a related field, and (2) at least three years' experience teaching L2 English. L1 backgrounds included English ($n = 1$), Arabic ($n = 1$), French ($n = 1$), Spanish ($n = 2$), and Brazilian Portuguese ($n = 3$). Teaching experience ranged from 3 to 31 years ($M = 14$), and all reported normal hearing.

Data Collection Materials

Sentences

Each speaking participant read 60 sentences, with 20 sentences for each linguistic feature being studied. Each sentence contained only one instance of the target feature, allowing for a controlled and targeted analysis. Following Munro and Derwing's (2020) transcription task procedure, the sentences had no more than 17 words. To avoid confounding variables, only lexical items among the 3,000 most frequent items in the NBC/COCA frequency list (Nation, 2017) were used, aside from the infrequent lexical items being studied. For ambiguous speech, some ambiguous items include words above the 3,000 level, allowing for an analysis of the impact of word frequency (e.g., *two lips*, both 1,000 level words versus *tulips*, which is in the 9,000 level). The sentences were free of context that could assist in understanding the sentence,

especially for infrequent vocabulary and ambiguous connected speech. For consistency, the targeted linguistic element was placed centrally in all sentences.

The 20 sentences with infrequent vocabulary were divided between infrequent lexical items, given names, and other proper nouns. To better understand at which point word frequency affects GVT transcriptions, ten sentences contained progressively less frequent lexical items as per the BNA/COCA frequency list (Nation, 2017). As well, five sentences containing given names were chosen from a list of 15 sentences created for Nicklin et al.'s (2023) study exploring the impact of names on reading fluency. The names for the current study were chosen to ensure that various sounds (e.g., stops, fricatives) occur in the word-initial position, thus controlling for the potential effects of manner of articulation on speech processing (e.g., stops are less salient than fricatives and potentially less likely to be processed by GVT). The remaining five sentences contained proper nouns belonging to different categories in the typology of proper nouns outlined in Atkins and Rundell (2008; e.g., famous places, organizations).

Infrequent collocations were operationalized as collocations using two words that are both in the most frequent 3,000 words on the BNA/COCA frequency list (Nation, 2017) but that appear together fewer than five times as collocations in COCA (Davies, 2008). Various types of collocations were used, such as adverb-adjective (e.g., *classically correct*) and noun-past participle (e.g., *act met*). Ambiguous sentences due to connected speech phenomena featured a range of CSPs such as assimilation (e.g., *cat belonged* vs. *cap belonged*) and geminates (e.g., *watch TV* versus *watched TV*). The sentences were adapted from those used in studies into human perceptions of ambiguous speech (e.g., Strachan & Trofimovich, 2019). Both versions of the sentence were read, allowing for an analysis of the impact of word frequency, as mentioned

above, and type of CSP. A sample sentence for each of the three linguistic features is found in Table 3.1; a complete list is provided in Appendix A.

Table 3.1

Sample Sentence by Linguistic Feature

Linguistic Feature	Sample Sentence
(1) Infrequent vocabulary:	
a. Lexical item	The couple enjoyed a **rustic** wedding, surrounded by all their friends.
b. Name	It was now getting very dark indeed and **Alice** thought there must be a storm coming.
c. Proper noun	They spent hours studying the **Eiffel Tower**, noting its elaborate details.
(2) Unusual collocation	The plant spent days in a **dry glass** and could not take root.
(3) Ambiguous speech	There were reports about **a tax** on buses in the paper yesterday. There were reports about **attacks** on buses in the paper yesterday.

Procedure

Speech Samples

Speaking participants recorded the speech samples independently using the Poodll plugin in Moodle. The sentences were randomized to avoid an order effect. Moodle quizzes used to collect the rater transcriptions were created using the recordings. To avoid an order effect in human-rater transcriptions, four quizzes were compiled, each comprised of 25% of the sentences appearing in a random order. Only one version of the four recordings appeared per quiz to mitigate a familiarity effect.

Transcriptions

Raters transcribed the sentences in a Moodle quiz. Procedures followed those in Kennedy and Trofimovich (2008): raters were asked to listen to each sentence only once; they carried out an orthographic transcription of what they heard; if they were not sure, they were to guess; if they could not guess, they were to type an underscore. In a session with the researcher, the

procedure was explained, and raters were given ten practice sentences to familiarize themselves with the task.

To avoid an order effect for the quizzes, raters were divided into four groups that completed the quizzes in different orders. To mitigate the familiarity effect and possible fatigue, raters were asked to complete the quizzes on four separate days. Each quiz had five practice sentences to be completed before starting the task. The same speech samples were played into GVT to obtain the GVT transcriptions in Google Docs.

Accuracy Scores

Two accuracy scores were generated for all transcriptions. Sentence-level accuracy scores were determined using a script specially designed to score dictation ASR accuracy, allowing for an efficient and consistent scoring of the large number of transcriptions collected in the study. The script was shown to have a .99 intraclass correlation coefficient with human raters in previous research into dictation ASR accuracy (Nelson & Cardoso, in press). Before running the script, spelling and typing errors in human-rater transcriptions were corrected. The script read the transcriptions and provided an automated score for each sentence, following the scoring procedures established in Johnson et al. (2024). A point was awarded for every correct word. Homophones, compound items that appear separately (e.g., *free masons* instead of *Freemasons*), and non-contracted verbs that were contracted in the prompt and vice versa also counted for one point. Half points were given to words with incorrect grammatical inflections (e.g., *walk* instead of *walked*). Repetitions of words or parts of sentences were not considered. The number of points was divided by the number of words in the sentence then multiplied by 100. The second accuracy score was at the item-level for each of the three targeted linguistic features. The item-level

accuracy codes were manually generated by the researcher, using a binary system: 0 for inaccurate and 1 for accurate.

Analysis

Accuracy rates for the targeted linguistic features for each rater were entered into SPSS version 30.0. A chi-square test of independence was run to examine whether the distribution of accurate and inaccurate transcriptions differed among raters within each feature. All assumptions for chi-square tests were met (e.g., expected cell counts). Additionally, transcriptions were visually inspected to explore whether patterns of inaccuracy emerged at the targeted feature or sentence level. Deletions and substitutions of the targeted linguistic features were noted for all raters. GVT transcription errors beyond the targeted features were also noted to identify any trends in potential systematic errors.

Results

To answer the research question, which pertained to the differences in human-rater and GVT transcriptions, both sentence-level and feature-specific accuracy rates were examined. Table 3.2 presents the descriptive statistics for sentence-level accuracy scores organized by rater and linguistic feature. These results indicate that GVT tended to provide higher accuracy scores for sentences with smaller standard deviations than human raters (R), for all targeted linguistic features.

Table 3.2

Descriptive Statistics for Sentence Accuracy by Linguistic Feature and Rater

Rater	Infrequent Vocabulary		Unusual Collocations		Ambiguous Speech	
	M	SD	M	SD	M	SD
R1	97.67*	4.11**	97.23	5.04	96.40*	10.22
R2	98.11*	4.14**	98.20	5.32	96.43*	10.07
R3	91.32	12.63	87.80	17.44	94.51	10.49
R4	84.48	16.71	81.31	18.99	93.10	12.49
R5	96.00	7.10	96.95	4.96	95.79	10.54
R6	93.79	5.77	93.50	9.74	94.36	10.26
R7	90.25	7.30	90.17	9.65	92.72	12.08
R8	86.65	11.00	85.74	11.57	91.67	12.03
GVT	97.55	4.36	98.96	2.84	95.96	5.45

Note. * Higher mean than GVT.
** Smaller standard deviations than GVT.

Descriptive statistics for the accuracy percentage of the targeted linguistic features are found in Table 3.3. For infrequent vocabulary, GVT had higher accuracy rates than six raters. Only one rater exceeded GVT in accuracy for unusual collocations. However, five raters outperformed GVT for ambiguous speech.

Table 3.3

Accuracy Percentage for Targeted Features

Linguistic Feature	R1	R2	R3	R4	R5	R6	R7	R8	GVT
Infrequent Vocabulary	88.6**	91.5**	75.9	59.5	81.0	55.7	57.0	53.2	83.5
Unusual Collocations	89.9	97.5**	67.1	53.2	88.6	79.7	54.4	44.4	93.7
Ambiguous Speech	81.0**	78.5**	75.9**	30.4	77.2**	74.7**	62.0	63.3	72.2

Note. **Accuracy greater than GVT

A chi-square test showed a significant difference in transcription accuracy of targeted features across raters for infrequent vocabulary, X^2, (8, N = 711) = 73.04, p <.001 and unusual collocations, X^2, (8, N = 711) = 131.52, p <.001. However, no significant difference in transcription accuracy was found among raters for ambiguous speech, X^2, (8, N = 711) = 13.92, p <.084. This indicates that accuracy rates differed significantly among raters for the first two

linguistic features but not for ambiguous speech. The following examination of the transcriptions by linguistic feature provides greater detail about the similarities and differences in GVT and human-rater transcriptions.

Infrequent Vocabulary

Lexical Items

As the targeted lexical items decreased in frequency, GVT and human-rater transcriptions became increasingly inaccurate. Table 3.4 presents the accuracy rates for all infrequent lexical items used in the study, illustrating this trend. Three interesting examples are *furor, meniscus,* and *fane*. Despite being a higher-frequency word than *meniscus* and *fane*, *furor* was never accurately recognized by GVT, while *meniscus* was recognized with 100% accuracy and *fane* with 50%. Accuracy rates for human raters tended to be lower than those of GVT, with the most extreme difference being for *gelding* (38% for human raters versus 100% for GVT). For both types of raters, sentence accuracy decreases as the targeted lexical item becomes less frequent. However, the decrease is due to nonrecognition of the targeted lexical item. For example, 90% accuracy in GVT for the sentence containing *orache* reflects a single-word mistranscription in an otherwise correctly transcribed ten-word sentence, where each word contributes equally to the score. This indicates that transcription errors were *item-specific* rather than diffused throughout the sentence.

Table 3.4

Accuracy Rates (%) by Word Frequency Levels

Lexical Item	Frequency Level	GVT		Human Raters	
		Feature Accuracy	Sentence Accuracy	Feature Accuracy	Sentence Accuracy
Compel	4k	100	98	97	96
Rustic	7k	100	100	88	93
Implore	9k	100	100	81	97
Gelding	11k	100	97	38	90
Salacious	13k	100	91	69	90
Furor	16k	0	92	63	90
Meniscus	18k	100	100	84	90
Plebe	20k	25	89	38	89
Fane	23k	50	96	41	88
Orache	25k	0	90	4	89

Given Names

The accuracy rate for given names was high for GVT and human raters. GVT showed higher accuracy, with 100% for all given names, while human-rater scores ranged from 88% *for Laura* and 100% for *Susan*. Errors were substitutions for similar names (e.g., *Norah* instead of *Laura*) or phonologically similar words (e.g., *tiny* for *Tony*). Sentence accuracy remained high for GVT (98%-100%) and human raters (94-98%), suggesting again that errors are not diffused. These results indicate that the given names chosen for this study are not problematic for GVT.

Proper Nouns

The accuracy rates for proper nouns varied greatly. GVT had a 100% accuracy rate for all proper nouns compared to ranges of 41% (for *Hades*) to 91% (for *Eiffel Tower*) accuracy for human raters. The targeted noun was often deleted by human raters, indicating they struggled to understand the word. As well, *Freemasons* and *Matterhorn* were often written as two words. Sentence accuracy was higher, ranging from 77% to 99% but the lower ranges are below those of the other elements of infrequent vocabulary, suggesting that difficulties experienced with proper

nouns could, sometimes, affect raters' ability to understand the entire sentence. GVT sentence accuracy was high, at 99%, indicating that proper nouns are not problematic for GVT transcriptions.

Unusual Collocations

GVT did not struggle with unusual collocations, with an accuracy rate of 94% for the targeted structure and 99% for the sentences. The only mistranscribed collocations were *worn trunk* (transcribed as *warm trunk*), *an act met* (rendered as *enact met*), and *honours* being transcribed in the singular. The rest of the sentences were transcribed correctly, except for eight errors, half of which were morphological inflections unrelated to the collocations (e.g., *use* instead of *used*). Conversely, most human raters showed greater difficulty with this feature, with accuracy rates ranging between 44%-98%. Sentence accuracy rates were higher, ranging between 81% and 98%, substantially lower than GVT, indicating that unusual collocations may have affected sentence transcriptions.

Human raters often deleted elements of the collocation or substituted other words. They seemingly focused on the main words and struggled with modifier collocates. Six adjectives were deleted 18 times, while their corresponding nouns were only deleted twice. The same trend emerged for adverbs, with eight adverbs deleted 19 times, while their corresponding adjectives or verbs were deleted only three times. Substitutions were even more frequent: adjectives had 32 substitutions, while their associated nouns were substituted only twice. Additionally, adverbs were substituted 33 times compared to five substitutions among the words they modified. The results suggest that unusual collocations posed minimal difficulties to GVT but were challenging for some human raters.

Ambiguous Speech

Unlike other features, ambiguous speech presented challenges for both GVT and human raters, highlighting the persistent difficulty of phonological ambiguity in connected speech. Ambiguous speech yielded the lowest accuracy rates for GVT (72%) and for four human raters. The range between human raters was large, spanning from 30% to 81%. Sentence accuracy was high, with 92%—96% for human raters and 95% for GVT, indicating that the ambiguity resulted in localized errors rather than affecting the entire sentence.

Regarding homophone pairs, human raters and GVT appeared to prefer the two-word option (e.g., *two cans* versus *toucans*). However, for human raters, *attacks*/*a tax* was an extreme outlier: *attacks* was substituted for *a tax* 29 times, while *a tax* was substituted for *attacks* only once. GVT, however, did not exhibit the same strong preference. Human raters made errors in sentences containing geminate consonants/linking (e.g., *known ocean/known notion*), with no clear directional pattern, as bidirectional substitutions occurred. GVT did not make any errors with these pairs, indicating that GVT may be able to detect syllable boundaries better than humans. Human raters had difficulties with geminate consonants arising from the past tense -ed inflection as they substituted *watched* with *watch* seven times but making the reverse substitution only one time. Unlike human raters, GVT made no errors when transcribing *watched* and *watch.* In sentences targeting assimilation, GVT and human raters frequently substituted *cap* for *cat* and *neck* for *net.* GVT and human raters struggled with transcribing ambiguous speech, with error patterns frequently, though not consistently, different in nature.

Discussion

This study set out to explore the impact of potentially problematic linguistic features on GVT transcription accuracy by asking: In what ways do transcripts generated by GVT and

human raters differ for infrequent vocabulary, unusual collocations, and ambiguous connected speech? Both GVT and human raters showed significant differences in their transcriptions of infrequent vocabulary (RQ-a) and unusual collocations (RQ-b), but not for ambiguous speech (RQ-c). The following sections explore each linguistic feature in detail to illustrate the differences between human-rater and GVT transcriptions.

Infrequent Vocabulary

GVT was generally accurate in transcribing names and proper nouns, outperforming human raters on these items, but both GVT and human raters struggled with infrequent lexical items. These lower-frequency words posed challenges for all raters; as lexical frequency decreased, so did transcription accuracy for GVT and human raters, corroborating Cámara-Arenas et al. (2023), who found Google's ASR had difficulties in transcribing infrequent words. It also offers an empirically grounded explanation for the weak correlation between GVT and human-rater scores for high-proficiency test takers in Johnson et al. (2024). The high-proficiency test takers who had divergent GVT and human-rater scores had read from a test version containing multiple low-frequency and off-list words such as *acetone* and *amalgams*. Other test versions did not contain a comparable level of low-frequency words. Given the finding that GVT has difficulty transcribing lower-frequency words, test takers may have been unintentionally penalized by the unequal use of low-frequency lexical items in test versions.

Some notable exceptions shed light on GVT's transcription of low-frequency words. For example, while human raters displayed difficulties transcribing *gelding,* with an accuracy rate of 38%, GVT had a 100% accuracy rate, suggesting that this specialized term was perhaps unfamiliar to some raters. GVT may be able to transcribe less frequent specialized vocabulary, provided they are not off-list words. Two other words also offer insight into the words GVT has

difficulty recognizing: *furor* and *meniscus*. While GVT consistently substituted *furor* with more frequent, phonologically similar alternatives like *fewer*, it never mistranscribed the less frequent *meniscus*, possibly because fewer or no plausible higher-frequency phonological alternatives exist. This suggests that transcription accuracy may depend not solely on word frequency, but also on the availability of more frequent *phonological* alternatives. A further insight is offered by *fane*, the second least frequent word used. It was often transcribed by both GVT and human raters as its more frequent homophones *feign* and *fain*. This pattern suggests that both GVT and humans privileged acoustic similarity and word frequency over contextual clues when interpreting the input. These examples demonstrate that frequency alone does not determine GVT transcription accuracy: factors such a familiarity with specialized vocabulary, phonological similarity, and the existence of homophonous alternatives also appear to influence recognition.

Given that names did not pose problems for most human raters (93% accuracy), occasional substitutions of similar-sounding names (e.g., *Norah* for *Laura*) suggest that phonological similarity may have influenced listeners' perception. In contrast, GVT transcribed names with 100% accuracy, counter to Ashwell and Elam's (2017) finding that Google's API struggled with given names, an outcome that may reflect advancements in ASR technology in the nearly ten years since their study. A second possibility is that *Tom*, problematic for Ashwell and Elam (2017), has many high-frequency phonologically similar alternatives (e.g., *bomb, calm*). The given names chosen for this study happen to not have high-frequency phonetically similar alternatives. This explanation may support the patterns noticed with infrequent lexical items: transcription accuracy may be affected by the availability of other phonologically similar options.

Proper nouns was the only infrequent vocabulary category in which GVT vastly outperformed human raters. GVT had 100% accuracy, but human raters struggled, with *Hades* being recognized only 41% of the time. Only the iconic *Eiffel Tower* had a high transcription rate of 91%. Compound proper nouns (e.g., *Freemasons*) were often transcribed as two separate words, and coded as correct, indicating that the raters may have been relying on auditory perception rather than lexical knowledge. While it was hypothesized that GVT may have difficulty with proper nouns due to their underrepresentation in many corpora, this hypothesis is not supported by the findings. Conversely, human raters struggled to retrieve these nouns. Neuropsychological research suggests that different parts of the brain are active when retrieving proper nouns and common nouns, indicating different storage methods in long-term memory (Proverbio et al., 2001). While common nouns are stored with broad networks that include visual and sensory associations, proper nouns tend to be stored as labels, making them vulnerable to retrieval difficulties. The findings of the current study seem to align with this theory, as human raters often had difficulty retrieving proper nouns for which they did not have visual or semantic associations, relying instead on phonological clues.

Unusual Collocations

Unusual collocations posed almost no issues for GVT, a finding that neither corroborates Ashwell and Elam's (2017) analysis nor aligns with the instances reported in Johnson et al. (2024). As with given names, this may reflect improvements in ASR technology that now enables more accurate transcriptions of unusual collocations. Human raters, however, faced a much greater difficulty, often substituting or deleting modifiers while maintaining the node words (nouns and verbs) essential for conveying meaning. These substitutions often involved higher-frequency alternatives. For example, *firmer denial,* which appears only once as a

collocate in COCA and is ranked as the 301st most frequent collocate of *denial*, was replaced seven times by *formal denial.* While *formal denial* is still relatively infrequent, appearing only five times in COCA, it is ranked much higher, with *formal* being the 130th most frequent collocate of *denial*. The phonological similarity and higher frequency of the replacement *formal* may explain the raters' tendency to select the more frequent alternative over the original. Deleting and substituting modifiers aligns with linguistic research into how collocations are stored as representations in our grammar (Bybee, 2006). Nouns and verbs occur more frequently, and as such are stored as the core of a collocation. Modifier collocates, however, are stored more peripherally, and are subject to deletion or modification when confronted with an unusual collocation. GVT appears able to process the modifier collocate and node word with greater consistency than human raters.

Ambiguous Speech

Ambiguous speech was the linguistic feature that many raters found the most difficult to transcribe and, notably, it was the one feature with no significant difference between raters. However, the transcripts show that the reasons for the lower accuracy rates differed between human raters and GVT. Human raters struggled to segment words when geminate consonants and linking created ambiguity (e.g., *known notion*/*known ocean*), a finding consistent with Mattys and Melhorn (2007). They also had difficulties in instances where gemination occurs because of past tense inflections (e.g., *watch TV*/*watched TV*), corroborating the findings of Strachan and Trofimovich (2019). In contrast, GVT did not have difficulty with these two types of ambiguity.

While the findings regarding *watch* versus *watched* make it tempting to claim that GVT does not have issues with transcribing past tense -ed, three instances in the dataset fail to support

this conclusion. The first appeared in an ambiguous sentence pair (*cat*/*cap belonged to*). Of the eight GVT transcriptions of *belonged to*, six were missing the past tense -*ed* inflection. As well, in the sentence *I closed the door behind me so that Laura would not see or hear him*, GVT incorrectly transcribed the verb *closed* as the present form *close* in 50% of the cases. Finally, GVT transcribed *used to* as *use to* 75% of the time in *To them, the words we used to intend honour seemed unimportant*. None of these cases involve geminate consonants: the transcription errors are likely due to assimilation or deletion. This suggests that GVT does not struggle with /tt/ geminates for past tense -ed forms but does have difficulty transcribing the inflection when assimilation or deletion occurs.

The error in transcribing -ed inflections that had undergone assimilation corresponds with the difficulty GVT and human raters had in accurately transcribing the sentences specifically designed to target assimilation. For instance, *cap* was often substituted for *cat,* and similarly, *neck* was substituted for *net*, echoing the findings of Gaskell and Marslen-Wilson (2001) with human listeners, and Pouw et al. (2024) with ASR transcriptions. The findings of Pouw et al. (2024) also explain the one difficulty GVT had with unusual collocations: the substitution of *warm* for *worn* in the collocation *worn trunk*. This would not usually be considered due to assimilation, since it is atypical for /n/ to assimilate to /m/ before /t/. However, Pouw et al. (2024) found that ASR can overgeneralize assimilation rules and apply them in contexts that do not typically apply when lacking lexical or semantic cues.

Both humans and GVT also showed difficulty with accurately transcribing homophones (e.g., *two lips* versus *tulips*), consistent with the findings of Kim et al.'s (2012) study on listener perception. The one-word option was often substituted by the two-word alternative, likely due to the higher frequency of the individual words in the latter. One notable exception in which human

raters and GVT differed was *a tax on buses* versus *attacks on buses.* Human raters tended to break the pattern, transcribing *attacks on buses,* while GVT favoured *a tax*. While *a tax on buses* seems an unlikely scenario in contemporary contexts, *attacks on buses* is, unfortunately, more plausible. This real-world knowledge may have influenced human raters, whereas GVT may lack the awareness to prioritize plausibility over statistical frequency. GVT appears to be relatively strong at segmenting word boundaries when geminate consonants and linking occur. However, like human raters, GVT exhibited difficulties with homophones and ambiguities caused by assimilation, even in sentences that had not been intended to target these features.

A Contained Problem

Ashwell and Elam (2017) found that transcription errors often lead to additional errors in adjacent words, raising concerns that a single mistake could exponentially penalize test takers. In the current study, this was not found to be an issue. As noted in the results, when a targeted feature was mistranscribed, the error was localized and did not affect the transcriptions of the surrounding words. GVT transcriptions were verified for errors in the words directly before or after the targeted feature. The sole instance of an adjacent error was the word *plebe*, which was transcribed as *plea but.* Sentence-level accuracy rates were high, even when the targeted feature was mistranscribed. This suggests that Google's ASR has advanced since Ashwell and Elam was published, mitigating the previous concerns about reliability.

Assessment Implications

The findings of this study have implications for test developers wishing to use GVT to score pronunciation assessments while maintaining the construct validity (accurately measuring pronunciation ability as intended) and reliability (consistency of scoring; e.g., Messick, 1989). One potential threat to construct validity is that the presence of an infrequent lexical item in a

speaking prompt may inadvertently be a test of GVT's ability to recognize the word, rather than the test taker's ability to pronounce it. Construct-irrelevant variance will also occur with test items that lead GVT to over-rely on phonological cues and inaccurately transcribe ambiguous speech such as in the *cat/cap belonged* contrast. As seen in Johnson et al. (2024), these transcription errors penalize higher-proficiency speakers, reducing reliability and test fairness.

These limitations should not discourage the use of GVT to generate scores, as it offers important advantages in terms of scalability and cost efficiency (Johnson et al., 2024). However, test items need to be carefully designed to ensure validity and reliability. Corpora should be consulted to ensure test versions are comparable in their use of lexical items from the various frequency bands. At the same time, test designers must be sensitive to phonological ambiguity inadvertently creeping into test items. For example, although *worn trunk* was originally included as an unusual collocation, transcription errors occurred due to an assimilatory process, highlighting how easily an unanticipated linguistic feature can compromise transcriptions. To ensure scoring accuracy and alignment with the construct being assessed, all test items should be piloted with highly intelligible speakers of the targeted language.

Additionally, this study revealed the potential to leverage GVT's ability to accurately transcribe some lower-frequency lexical items and unusual collocations in the strategic design of test items. For example, choosing less frequent words recognizable by GVT allows for the creation of progressively more difficult test items. Low-frequency words like *meniscus*, which have few or no higher-frequency alternatives, could prove to be invaluable in distinguishing proficiency levels. GVT's ability to accurately transcribe unusual collocations could likewise inform test design as the frequency of formulaic sequences can impact processing by L2 speakers (Ellis et al., 2008). This pattern may be related to the cognitive mechanisms described in

Skehan's (2018) Limited Attention Capacity Hypothesis, which posits that working memory and attention are limited: accuracy declines as task complexity increases. As a result, lower-proficiency learners are more likely to rely on familiar, predictable chunks of language. Higher-proficiency learners, on the other hand, are better equipped to produce low-frequency collocations and are therefore more likely to perform better on a test that includes such items. These findings help temper the concerns raised at the outset of this study regarding test fairness and the inclusion of infrequent lexical items and unusual collocations in test items. Indeed, while some infrequent lexical items are problematic, the strategic and well-piloted inclusion of certain infrequent lexical items and unusual collocations could be a strength of GVT-scored assessments, offering opportunities to differentiate proficiency levels via controlled lexical variation.

Pedagogical Implications

The findings of this study also have pedagogical implications. Research has shown that dictation software helps students improve pronunciation skills and that students appreciate using it (e.g., McCrocklin, 2019). One affordance of dictation ASR is the confirmative feedback provided to students, as it indicates when pronunciation is accurate (John et al., 2025). To fully capitalize on this affordance, teachers must ensure students are assigned speaking prompts that do not inadvertently lead to inaccurate transcriptions that deprive students of the confirmative feedback they need. Students have reported frustration when speech they perceive as intelligible is not recognized by dictation ASR (Inceoglu et al., 2023; McCrocklin, 2019). Sometimes, this may be due to student pronunciation proficiency levels (i.e., their level of intelligibility may be lower than their self-assessment of it). However, it may sometimes be due to errors in transcriptions caused by connected speech processes. Ensuring that students are not using speaking prompts destined to result in inaccurate transcriptions will increase student confidence

in dictation ASR and facilitate their learning. It also ensures that classroom tests, be they formative, summative or self-assessments, are valid, reliable, and fair.

Conclusion

This study explored the differences between human-rater and GVT transcriptions of infrequent vocabulary, unusual collocations, and ambiguous connected speech. Unusual collocations were not problematic for GVT, as opposed to human raters who struggled with them. In contrast, both ambiguous items resulting from connected speech processes and infrequent vocabulary proved challenging for both ASR and human raters. However, GVT demonstrated greater accuracy with low-frequency words that lacked higher frequency phonologically similar alternatives. These findings indicate that a pronunciation test scored using GVT can be valid, reliable, and fair if test items are carefully designed and piloted with highly intelligible speakers. The findings also raised the possibility of strategically creating test items capitalizing on GVT's potential affordances for distinguishing proficiency levels.

While these results support using dictation software for scoring pronunciation assessments, an important limitation should be acknowledged: The voice samples used were drawn from only four speakers, all representing a single dialect of North American English. As a result, these findings are not generalizable to other English varieties. Future research should examine how these linguistic features affect GVT transcriptions across other English varieties, to safeguard the fairness of pronunciation tests used in diverse regional contexts. The hypothesis that strategically designed test items leveraging GVT's ability to transcribe certain lower-frequency features could discriminate between proficiency needs further investigation. Designing test items to capitalize on its strengths could elevate the use of dictation ASR to score pronunciation tests.

University language departments and language schools can use dictation ASR as a cost-effective method to score valid, reliable and fair pronunciation assessments. While customizable ASR has been used to assess speaking and pronunciation for two decades, the use of dictation ASR is now emerging as a promising tool in the automated assessment of pronunciation.

Chapter 4: Concluding Remarks

Advances in automatic speech recognition (ASR) technology have made the automated assessment of speaking and pronunciation possible, with its use gaining widespread acceptance over the last two decades. It has been adopted by many large testing organizations, such as ETS and Pearson, and it is widely accepted by governments, employers, and universities that require language testing to inform their decision-making process. However, the technology behind the assessments is expensive to purchase, develop, and maintain, requiring the expertise of specialists and making it prohibitively expensive for language schools and departments seeking to create their own automatically scored assessments.

Free and accessible dictation ASR, such as Google Voice Typing (GVT) in Google Docs, have been increasingly used in classrooms and researched for their affordances in supporting L2 learners who wish to improve their pronunciation. Studies show that it is an effective tool in improving both segmental (e.g., McCrocklin, 2019) and suprasegmental aspects of pronunciation (e.g., Mroz, 2020). However, there is limited research into its use to *assess* pronunciation, with only one study exploring student perceptions of its use (i.e., Dillon & Wells, 2021). This dissertation addressed this gap, by answering the overarching research question "What are the affordances offered by dictation ASR in an L2 pronunciation assessment context?" The two manuscripts explored the feasibility of using GVT to assess pronunciation and the impact of certain linguistic features on transcription accuracy, with the aim of ensuring test fairness.

This chapter will summarize the dissertation, with particular focus on the two manuscripts, highlighting their key findings, and their important implications for the employment of dictation ASR in proficiency/placement assessments and classroom-based evaluations. It will also consider the needs and perspectives of key stakeholders such as test

designers, teachers, and students. Finally, the chapter will address the limitations of the research and suggest further directions for research into using dictation ASR in pronunciation assessment.

Overview and Key Findings

Contributions of Thesis

The first study in this dissertation addressed a gap in the literature, which has thus far largely overlooked the potential of using dictation ASR for assessment, despite its widespread pedagogical use. The second study, motivated by the findings of the first, addressed the gap in knowledge about the best practices of using dictation ASR, such as the degree to which problematic linguistic elements could have unintended negative consequences for test takers, potentially undermining the validity and reliability of scores generated by GVT and the fairness of decisions informed by these scores.

Together, these manuscripts addressed the overarching RQ of "What are the affordances offered by dictation ASR in an L2 pronunciation assessment context?" The evidence from both studies showed that dictation ASR offers the affordance of using a low-cost solution to create an automated scoring system. Potential affordances may extend beyond usability, as the strategic inclusion of certain linguistic elements could enhance the test's ability to differentiate test takers' proficiency levels. The findings have implications for multiple stakeholders, including test developers, educators, and learners.

Manuscript A. Assessing Pronunciation Using Dictation Tools: The Use of Google Voice Typing to Score a Pronunciation Placement Test

Manuscript A (Chapter 2) addressed the overarching research question by exploring the feasibility of using GVT to score a pronunciation placement test through the lens of Bachman

and Palmer's (1996) framework of test usefulness. The study sought to answer the three following research questions:

1. What is the relationship between GVT-rated pronunciation scores and the final human-rated pronunciation scores?
2. What is the relationship between GVT-rated scores and human-rated scores across the set of pronunciation evaluation criteria used by the human raters (i.e., comprehensibility; phonemes; connected speech phenomena; word stress and rhythm; and thought groups, prominence, and intonation)?
3. What is the relationship between GVT-rated scores and human-rated scores of high- and low-proficiency levels?

Fifty-six recordings from a university language department's placement test were given an accuracy score based upon the GVT transcript, as well as a human-rater score determined by three raters using a rubric to assess multiple aspects of pronunciation.

The results indicated that the scores derived from the GVT transcripts correlated strongly with human-rater scores. The correlation was not limited to the overall score, but also for each criterion on the rubric used by the human raters, indicating that GVT can generate scores that are valid and reliable. However, it was noted that while human-rater and GVT scores for lower-proficiency test-takers were strongly correlated, this was not the case for higher-proficiency test takers. An examination of the test version taken by a large proportion of the higher-proficiency test takers with diverging scores revealed that it included fewer high-frequency lexical items and a greater number of low-frequency and even off-list items. It appeared that GVT had difficulty transcribing these items, calling into question the fairness of the test.

These findings address the overarching research questions by showing that GVT offers the affordance of using a cost-effective dictation ASR system to score pronunciation assessments. However, concerns were raised about the validity and reliability of the scores, as well as the fairness of any decisions based on them, due to the weak correlation observed for high-proficiency test takers, a concern that was explored in the subsequent study (Manuscript B).

Manuscript B. Real Eyes or Real Lies? Implications of Infrequent Vocabulary, Unusual Collocations,

and Ambiguous Linguistic Elements for ASR-based Pronunciation Assessment

Manuscript B (Chapter 3) strengthened the claim that it is feasible to use GVT to score pronunciation assessments, underscoring the affordances of this technology outlined in the first study. This second manuscript explored the effect of infrequent vocabulary, unusual collocations, and ambiguous speech on GVT transcription accuracy, highlighting the need to carefully design test items that: are valid assessments of pronunciation and provide reliable scores upon which fair inferences can be made. This concern was addressed by answering the following research question: In what ways do transcripts generated by GVT and human raters differ for: (a) infrequent vocabulary, (b) unusual collocations, and (c) ambiguous connected speech? Sixty sentences containing the three target linguistic elements were recorded by four highly intelligible speakers of English and transcribed by eight expert raters. For comparison, the same recordings were also played into GVT.

A chi square test of independence showed that there were significant differences in the distribution of transcription accuracy across all raters, including GVT, for infrequent vocabulary and unusual collocations, but not for ambiguous speech. An examination of the transcripts showed that GVT struggled to transcribe infrequent lexical items that have phonologically

similar, high-frequency alternatives. Unlike human raters, GVT had no difficulty transcribing unusual collocations. Both human raters and GVT, however, showed low accuracy rates for ambiguous speech, with assimilation being the greatest challenge for GVT. These findings highlight that test items to be scored using GVT need to be piloted with highly intelligible speakers to avoid unwittingly penalizing high-proficiency test takers. They also suggest the potential for strategically using GVT's ability to accurately transcribe certain less frequent structures to discriminate test taker proficiency levels, an interesting emerging affordance of the technology.

Takeaways

This dissertation highlights multiple affordances in three categories: operational affordances, assessment affordances, and pedagogical affordances. These affordances have practical implications for different stakeholder groups involved in designing, administering, and using pronunciation assessments.

Using GVT to Score Pronunciation Assessments: Assessment Affordances

Perhaps the most important takeaway from this dissertation is that scores from pronunciation tests scored using GVT are valid and reliable, and fair inferences can be made using those scores. Test designers do need to verify that the linguistic elements used in test items are accurately transcribed by GVT when spoken by highly intelligible speakers of the targeted language. Steps to ensure that test items that lead to valid and reliable scores should already have been in place; using GVT simply adds an additional step to the process. Testing items should also be an iterative process, as transcription accuracy may change with updates and modifications made by Google.

For language departments using third-party assessments that are automatically scored, using dictation ASR to create their own placement test would give them autonomy and control over their placement testing process. A test could be designed that is more in line with their language learning reality and philosophy, one that better reflects their language learning culture. Their own test would be more meaningful and inferences from the scores would be fairer since they would be based on the realities of the department.

An emerging affordance surfaced during the second study. The discovery that GVT accurately transcribes certain infrequent lexical items and unusual collocations raises the possibility of using these features to effectively and efficiently differentiate between proficiency levels. Since less proficient test takers should have more difficulty producing these items, their resulting score should be lower than scores for higher proficiency test takers. Further empirical testing is needed to validate this hypothesis.

Using GVT to Score Pronunciation Assessments: Operational Affordances

University language departments and language schools that want to develop their own automated scoring systems could use GVT as the core technology. While there would initially be development and continuing maintenance costs, they would be significantly lower than those associated with using customizable ASR. Not only is it cost efficient in terms of technology, but it also reduces human resource expenses by reducing the need for human raters.

Automated assessments also provide scores more quickly than human raters, speeding up the process of course registration. Test takers would know their scores sooner, allowing them to plan and register for their semester in a timelier manner. As a result, departments would have almost instantaneous data indicating which courses students should be registering in. This would increase student satisfaction and make the enrollment process more efficient.

Using GVT to Score Pronunciation Assessments: Pedagogical Affordances

Feedback about pronunciation is essential, yet difficult to provide in a classroom. Speaking prompts based upon the findings of the second study could give students access to the corrective and confirmative feedback afforded by GVT. Additionally, because GVT is freely accessible anytime anywhere, students can work autonomously on individualized progression paths outside the classroom. The findings of the second study raise the possibility that GVT could be used to help students improve their pronunciation of certain low-frequency words. This is especially beneficial for students learning English for a specific purpose. Additionally, Skehan's (2018) Limited Capacity Model can be interpreted to suggest that students may improve oral fluency when practicing constructions that include an unusual collocation. As indicated in Manuscript B, this model posits that the production of complex language leads to the proceduralization of formulation skills. As a result, a student who practices a sentence containing an unusual collocation such as *horribly healthy* will become more fluent, not through memorization of the sequence, which is unlikely to occur in a real-world conversation, but due to an increased ability to process language.

Despite the affordances set out in the previous paragraph, students have been found to resist using GVT due to unexplained errors in its transcriptions. The results of the second study suggest that such inaccuracies may be due to the presence of infrequent lexical items or to the technology's difficulty handling ambiguous connected speech. Ensuring that speaking prompts do not contain problematic features that affect transcription accuracy may increase student

acceptance of this beneficial tool, reducing frustration and increasing confidence in the feedback it provides.

The use of GVT for scoring in-class pronunciation assessments that are valid and reliable is an important affordance for educators. These tests could be used for multiple purposes: diagnosis, self-assessment, summative feedback, or formative grades. In this context, students would see the transcriptions and receive real-time feedback, enhancing their awareness of specific pronunciation challenges that could potentially lead to improvement. In standard (non GVT-enhanced) classrooms, teachers must listen (often more than once) to recorded speech samples and assign a grade based upon a rubric designed to assess various pronunciation features, making the process time-consuming and labour-intensive. The integration of GVT would simplify the process, especially if paired with a script that automates accuracy scoring and spares teachers from manually counting words.

Language learning is a lifelong process that extends well beyond the classroom. Instilling in students the habit of using GVT in class and guiding them to see the benefits of its corrective and confirmative feedback sets them up as long-term users of the technology even when they have finished taking classes. Teaching in the 21st century includes providing students with tools for autonomous use in real life situations and is essential to support their ongoing progress after they leave the classroom. GVT will support their learning, especially if teachers help students build trust in it by using speaking prompts crafted to ensure trustworthy feedback.

Takeaways: Summary

Table 4.1 presents synthesized takeaways organized by stakeholders to support practical application of the research.

Table 4.1

Takeaways by Stakeholder

Stakeholder	Key Takeaways
Test Designers	– GVT can yield valid, reliable scores if items are piloted with intelligible speakers. – Transcription accuracy must be checked regularly. – Certain lexical items and unusual collocations may help distinguish proficiency levels (requires further research).
Language Departments	– Automated scoring with GVT is cost effective. – GVT allows departments to build in-house tests aligned with their pedagogy. – Faster scoring helps streamline course registration. – Results reflect local teaching goals and values.
Test Takers	– Placement decisions based on GVT can be fair if tests are well designed. – Faster results support earlier registration and planning.
Teachers	– GVT can simplify pronunciation assessment if items are pre-verified. – Pairing with scoring scripts reduces marking time. – Useful for formative, diagnostic, or summative feedback.
Students	– Free, reliable tool for independent pronunciation practice. – Accurate transcription boosts motivation and trust. – Supports learning of infrequent items and unusual collocations (requires further research). – Promotes long-term, autonomous skill development.

Limitations And Future Research Directions

Limitations

This dissertation addressed a gap in the research into automated scoring of pronunciation assessments and has found that the use of GVT to score pronunciation placement tests does not negatively impact validity and reliability if test items are verified in advance. However, several

limitations mut be acknowledged. First, since GVT does not offer an option of uploading audio files, all recordings in both manuscripts were played into the system in real time. Any background noise that inadvertently occurred during this process may have reduced the quality of the transcriptions. Second, the study reported in Manuscript A was carried out using secondary data that contained only 56 audio samples, making it difficult to generalize the results. In addition, the samples were not equally distributed between proficiency levels, with most test takers having a lower-proficiency level. Finally, the speaker participants in the study reported in Manuscript B all spoke a dialectical variety of North American English, limiting the generalizability of the findings to other varieties of English. What follows is a more detailed discussion of these limitations.

The first limitation, that of the necessity to play the recordings into GVT for both studies rather than uploading them, is an inherent constraint of using GVT. Nelson and Cardoso (2024) used the same recordings as those in the first study in Microsoft Transcribe, which allows for uploads of files. Their correlations were similar to the ones observed in Manuscript A (e.g., r = .79 for MS-T for the overall human-rater score versus r = .78 for GVT) and followed the same trend (e.g., strong correlations for low-proficiency test takers but weak correlations for high-proficiency test takers), suggesting that playing the recording may *not* have substantially influenced the results. However, without an opportunity to upload files, it is impossible to determine what impact playing the recordings into GVT may have had on the results.

The generalizability of the results of Manuscript A (Chapter 2) is limited due to the small sample size (N = 56). This small number also constrained other potential comparisons, such as correlations by test version or the exploration of potential L1 bias, as the resulting sub-groups were too small for meaningful statistics. Additionally, the sample was skewed in terms of

proficiency, with only 20 low-proficiency test takers compared to 36 high-proficiency test takers, a disparity that may have influenced the observed correlation patterns. A larger and more evenly distributed sample in terms of proficiency would have allowed for more robust analyses of the validity and reliability of the scores.

The final limitation, related to Manuscript B (Chapter 3), is that the design focused exclusively on North American English. Since all speaker participants spoke a variety of this dialect, the findings may not be generalizable to other English varieties. While GVT has settings for different English dialects, this study did not examine their effects and did not include speakers of other varieties. Additionally, all unusual collocations were drawn from COCA (Davies, 2008), a corpus representing North American English. The collocations may not be as unusual (or may not occur at all) in other English varieties. If the study is replicated in another English-speaking context, the target collocations should be verified and modified accordingly.

Future Directions

By answering the overarching research question "What are the affordances offered by dictation ASR in an L2 pronunciation assessment context?", this dissertation has laid the foundations for the use of dictation ASR, such as GVT, to automatically score pronunciation tests. However, further research is needed to optimize its use by ensuring scoring quality and to explore emerging affordances. This section presents future directions for research into strategic item design, bias, connected speech processes, multilingual use, and system comparisons to advance fair, accurate, and pedagogically useful machine-scored assessments.

The findings of Manuscript B (Chapter 3) raise the possibility of strategically using certain infrequent lexical items and unusual collocations as a means of differentiating between proficiency levels when using GVT to score pronunciation tests. Further research is needed to

determine if creating test items including these elements can effectively distinguish between levels, as they may be more difficult for lower-proficiency test takers to produce. Studies targeting this emerging affordance could expand the usefulness of GVT, positioning it as more than just a scoring tool.

ASR systems are prone to bias when certain groups are underrepresented in training data. When this occurs, ASR can have difficulty accurately transcribing speech depending on the speaker's gender, age, L1, and language variety. In the context of assessment, this can affect the validity and reliability of scores, rendering any decisions made using them unfair to the test taker. Research on bias in GVT should be carried out with the goal of identifying biases, allowing test developers to establish mathematical formulas that will reestablish equity and reliability in the scoring model. Given that GVT is regularly updated by Google, bias testing needs to be iterative to ensure the scoring model accurately represents the pronunciation abilities of test takers. However, such testing can be time consuming and expensive. To address this, researchers and language learning institutions that want to use GVT for scoring pronunciation could collaborate to promote a better understanding of machine-based scoring with dictation ASR and to reduce the burden of iterative validation.

The findings of Manuscript B (Chapter 3) highlighted GVT's difficulties when transcribing ambiguous connected speech. As there are multiple connected speech processes (CSP) that were not used in the sentences in Manuscript B, it is important to determine the degree to which the other CSPs affect transcription accuracy. For instance, it is possible that processes such as segmental insertion (e.g., some/p/thing) or deletion (e.g., did ~~h~~e) could cause ambiguous speech and prove challenging for GVT to transcribe. Such insights would facilitate the writing of test items, helping test designers ensure that their tests are valid and reliable when

using GVT-generated scores to inform decisions. As well, better understanding the impact of connected speech processes on transcription accuracy is important for teachers and students to understand when using the tool for pedagogical purposes. Knowing that certain constructions can be difficult for GVT to transcribe could guide teachers in avoiding them when creating speaking prompts and could lead to reduced frustration among learners.

This research was premised on using GVT to assess *English* pronunciation tests. However, other dictation ASR systems exist (MS-T, Apple Dictation), and each system supports multiple languages. For instance, GVT offers 40 languages, with many of those having multiple varieties available for use. Research is needed into the feasibility of using the different dictation systems to assess pronunciation, as has begun with Nelson and Cardoso's (2024) study of MS-T, as well as into using them to assess other languages. Being able to assess multiple languages would allow language departments to maximize their financial investment in automated scoring. It would also allow practitioners to compare across systems and determine the dictation ASR that is best suited to their specific needs and contexts. Systematic studies would be needed for each system and in each language to determine linguistic features that impact transcription accuracy, to verify whether there are transcription issues with different varieties of languages, and to find biases. Such studies would provide a wealth of information that could inform test design, ensuring that automated scoring of pronunciation across languages and systems is valid and reliable.

Final Remarks

The aim of this dissertation was to determine the affordances of using GVT to assess pronunciation placement tests. This journey began seven years ago, in the first semester of my MA in Applied Linguistics, when I proposed a study using dictation ASR to increase learners'

willingness to communicate. Although the trajectory to investigating pronunciation assessment was not a direct one, my strong belief in the multi-faceted affordances of this free and accessible technology has consistently guided my research, both within and beyond the scope of my PhD, with the aspiration that test designers and language teachers will better understand its potential. Manuscript A (Chapter 2) shows the feasibility of developing a lower-cost automated scoring model using GVT. Building on this, my aim is that language departments and instructors embrace this tool for assessing pronunciation, while also taking into account the impact of linguistic items when designing speaking prompts, as demonstrated in Manuscript B (Chapter 3). By sharing my findings, I hope other researchers will join me in advancing this technology as the basis for automated scoring systems that yield valid and reliable scores and ensure fairness in the decisions those scores inform.

References

Alameen, G. & Levis, J. (2019). Connected speech. In M. Reed & J. Levis (Eds.), *The handbook of English pronunciation*. Wiley-Blackwell.

Ashwell, T., & Elam, J. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *The JALT CALL Journal, 13*(1), 59–76. http://dx.doi.org/10.29140/jaltcall.v13n1.212

Atkins, B., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. *Proceeding of the First International Conference on Spoken Language Processing (ICSLP 1990)*, 1185–1188. https://doi.org/10.21437/ICSLP.1990-313

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377. https://doi.org/10.1177%2F0265532210364404

Browne, K., & Fulcher, G. (2016). Pronunciation and intelligibility in assessing spoken fluency. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 37–53). Multilingual Matters. https://doi.org/10.21832/ISAACS6848

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711—733. https://doi.org/10.1353/lan.2006.0186

Cámara-Arenas, E., Tejedor-García, C., Tomas-Vázquez, C., & Escudero-Mancebo, D. (2023). Automatic pronunciation assessment vs. automatic speech recognition: A study of

conflicting conditions for L2-English. *Language Learning and Technology, 27*(1), 1–19. https://hdl.handle.net/10125/73512

Carey, M., Mannell, R., & Dunn, P. (2010). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201—219. http://dx.doi.org/10.1177/0265532210393704

Chen, L., Zechner, K., Yoon, S-Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, M., Mundkowsky, R., Lu, C., Leong, C., & Gyawali, B. (2018). Automated scoring of nonnative speech using the SpeechRaterSM v. 5.0 engine. *ETS Research Report Series*, 1, 1–31. https://doi.org/10.1002/ets2.12198

Chen., W., Inceoglu, S., & Lim, H. (2020) Using ASR to improve Taiwanese EFL learners' pronunciation: Learning outcomes and learners' perceptions. In O. Kang, S. Staples, K. Yaw., & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation and Second Language Learning and Teaching Conference* (pp. 37 – 48). Iowa State University.

Chun, C. W. (2006). An Analysis of a Language Test for Employment: The Authenticity of the PhonePass Test. *Language Assessment Quarterly*, *3*(3), 295–306. https://doi.org/10.1207/s15434311laq0303_4

Cobb, T. (n.d.). Web Vocabprofile. *Lextutor*. https://www.lextutor.ca/vp/eng/

Coniam, D. (1998). The use of speech recognition software as an English language oral assessment instrument: An exploratory study. *CALICO Journal, 15*(4), 7–23. http://www.jstor.org/stable/24147601

Cox, T., & Davies, R. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal, 29*(4), 601–618. http://dx.doi.org/10.11139/cj.v29i4.54

Cucchiarini, C., Strik, H., & Boves, L. (2000). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication, 30*(2–3), 109–119. https://doi.org/10.1016/S0167-6393(99)00040-0

Dai, Y., & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: a mixed-methods study. *Computer Assisted Language Learning*, *36*(5–6), 861–884. https://doi.org/10.1080/09588221.2021.1952272

Davies, M. (2008). *The corpus of contemporary American English (COCA).* https://www.english-corpora.org/coca/.

Derwing, T., & Munro, M. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.

Derwing, T., Munro, M., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly, 34*(3), 592–603. https://doi.org/10.2307/3587748

Dillon, T., & Wells, D. (2021). Student perceptions of mobile automated speech recognition for pronunciation study and testing. *English Teaching, 76*(4), 101–122. https://doi.org/10.15858/engtea.76.4.202112.101

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M. & Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly, 5*(2), 160–167. https://doi.org/10.1080/15434300801934744

Elimat, A., & AbuSeileek, A. (2014). Automatic speech recognition technology as an effective means for teaching pronunciation. *The JALT CALL Journal, 10*(1), 21–47.

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375—396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Eskenazi, M. 1999. Using a computer in foreign language pronunciation training: What advantages? *CALICO Journal, 16*(3), 447–69. https://doi.org/10.1558/cj.v16i3.447-469

Evanini, K., & Zechner, K. (2019). Overview of automated scoring. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment: Using language technologies to score spontaneous speech* (pp. 3–20). Taylor & Francis Group. https://doi.org/10.4324/9781315165103

Evers, K., & Chen, S. (2021) Effects of automatic speech recognition software on pronunciation for adults with different learning styles. *Journal of Educational Computing Research, 59*(4), 669–685. https://doi.org/10.1177/0735633120972011

Feng, S., Kudina, O., Halpern, B., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. In *22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021)*. International Speech Communication Association. https://doi.org/10.48550/arXiv.2103.15122

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399—423. https://doi.org/10.2307/3588487

Filippidou, F., & Moussiades, L. (2020). A benchmarking of IBM, Google and Wit automatic speech recognition systems. *Artificial Intelligence Applications and Innovations*, 583, 73–82. https://doi.org/10.1007/978-3-030-49161-1_7

Garcia, C., Nickolai, D., & Jones, L. (2020). Traditional versus ASR-based pronunciation instruction: An empirical study. *CALICO Journal, 37*(3), 213–232. https://doi.org/10.1558/cj.40379

Gaskell, G., & Marslen-Wilson, W. (2001). Lexical ambiguity resolution and spoken word recognition: Bridging the gap. Journal of Memory and Language, 44(3), 325–349. https://doi.org/10.1006/jmla.2000.2741

González Fernández, B. & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure. *International Journal of Applied Linguistics*, 166 (1), 94–126. https://doi.org/10.1075/itl.166.1.03fer

Gow, D., Jr., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: evidence from priming. *Journal of Experimental Psychology. Human Perception and Performance, 21*(2), 344–359. https://doi.org/10.1037//0096-1523.21.2.344

Graham, C., Lonsdale, D., Kennington, C., Johnson, A., & Mcghee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D.Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (pp. 1604–1610). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2008/pdf/409_paper.pdf

Guskaroska, A. (2020). ASR-dictation on smartphones for vowel pronunciation practice. *Journal of Contemporary Philology, 3*(2), 45–61. https://doi.org/10.37834/JCP2020045g

Hahn, L. (2004) Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, (38)*2, 201–223. https://doi.org/10.2307/3588378

Hirai, A., Kovalyova, A. (2023). Using speech-to-text applications for assessing English language learners' pronunciation: A comparison with human raters. In MdM Suárez & El-W.M. Henawy (Eds.), *Optimizing online English language learning and teaching*. Springer. https://doi.org/10.1007/978-3-031-27825-9_17

Inceoglu, S., Chen, W-H., & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and automatic speech recognition. *ReCALL, 35*(1), 89–104. https://doi.org/10.1017/S0958344022000192

Inceoglu, S., Lim, H., & Chen, W-H. (2020). ASR for EFL pronunciation practice: Segmental development and learners' beliefs. *The Journal of Asia TEFL, 17*(3), 824–840. http://dx.doi.org/10.18823/asiatefl.2020.17.3.5.824

Isaacs, T. (2018a). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. Thomson, and J. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 570–584). Routledge.

Isaacs, T. (2018b). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly, 15*(3), 273–293. https://doi.org/10.1080/15434303.2018.1472264

Issacs, T., & Harding, L. (2017). Pronunciation assessment. *Language Teaching, 50*(3), 347-366. https://doi.org/10.1017/S0261444417000118

John, P., Cardoso, W., & Johnson, C. (2022). Evaluating automatic speech recognition for L2 pronunciation feedback: A focus on Google Translate. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds.*), Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022*. http://dx.doi.org/10.14705/rpnet.2022.61.1458

John, P., Johnson, C., & Cardoso, W. (2024). Assessing Google Translate ASR for feedback on L2 pronunciation errors in unpredictable sentence contexts. In *CALL for all Languages - EUROCALL 2023 Short Papers*. https://doi.org/10.4995/eurocall2023.2023.16987

John, P., Johnson, C., & Cardoso, W. (2025). Exploring automatic speech recognition for corrective and confirmative pronunciation feedback. *Journal of Second Language Pronunciation.* https://doi.org/10.1075/jslp.24035.joh

Johnson, C. & Cardoso, W. (2024). Hey Google, let's write: Examining L2 learners' acceptance of automatic speech recognition as a writing tool. *CALICO Journal, 41*(2), 122-145. https://doi.org/10.1558/cj.22431

Johnson, C., Cardoso, W., Zuercher, B., Brenner, K., & Springer, S. (2024). Assessing pronunciation using dictation tools: The use of Google Voice Typing to score a pronunciation placement test. *Journal of Second Language Pronunciation, 10*(1), 122–145. https://doi.org/10.1075/jslp.23033.joh

Johnson, C., Cardoso, W., Zuercher, B., Brenner, K., & Springer, Suzanne. (2022). Using Google Voice Typing to automatically assess pronunciation. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds.), *Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022*. (pp. 203–207). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1459

Kang, O, & Ginther, A. (2018). *Assessment in Second Language Pronunciation*. Routledge.

Kennedy, S., & Trofimovich, P. (2008) Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Modern Language Review, 64*(3), 459–489. http://dx.doi.org/10.3138/cmlr.64.3.459

Khabbazbashi, N., Xu, J., & Galaczi, E. (2021). Opening the black box: Exploring Automated speaking evaluation. In B. Lanteigne, C. Coombe, & J. D. Brown (Eds.), *Challenges in Language Testing Around the World* (pp. 333–343). Springer. https://doi.org/10.1007/978-981-33-4232-3_25

Kim, D., Stephens, J., & Pitt, M. (2012). How does context play a part in splitting words apart? Production and perception of word boundaries in casual speech. *Journal of Memory and Language, 66*(4), 509–529. https://doi.org/10.1016/j.jml.2011.12.007

Lenhard, W. & Lenhard, A. (2014). Hypothesis tests for comparing correlations. *Psychometrica*. https://www.psychometrica.de/correlation.html

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–377. https://doi.org/10.2307/3588485

Levis, J., & Suvorov, R. (2012). Automatic speech recognition. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* John Wiley & Sons.

Liakin D, Cardoso W, Liakina N. (2017). Mobilizing instruction in a second-language context: learners' perceptions of two speech technologies. *Languages, 2*(3), 1–21. https://doi.org/10.3390/languages2030011

Liakin, D., Cardoso, W., & Liakina, N. (2014). Learning L2 pronunciation with a mobile speech recognizer: French/y/. *CALICO Journal, 32*(1), 1–25. https://doi.org/10.1558/cj.v32i1.25962

Ling, G., Mollaun, P., & Xiaoming, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479–499. https://doi.org/10.1177/0265532214530699

Loukina, A., Davis, L., & Xi, X. (2018). Automated assessment of pronunciation in spontaneous speech. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 153–171). Routledge.

Luo, D., Minematsu, N., Yamauchi, Y., & Hiroshi, K. (2009). Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences. In *Proceedings of SLaTE*. ISCA. http://www.eee.bham.ac.uk/SLaTE2009/papers/SLaTE2009-20-v2.pdf

Mattys, S., & Melhorn, J. (2007). Sentential, lexical, and acoustic effects on the perception of word boundaries. *The Journal of the Acoustical Society of America*, 122 (1), 554–567. https://doi.org/10.1121/1.2735105

McCrocklin, S. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System, 57*, 25–42. https://doi.org/10.1016/j.system.2015.12.013

McCrocklin, S. (2018). Learners' feedback regarding ASR-based dictation practice for pronunciation learning. *CALICO Journal, 36*(2), 119–137. https://doi.org/10.1558/cj.34738

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98–118. https://doi.org/10.1075/jslp.16034.mcc

McCrocklin, S., & Edalatishams, I. (2020). Revisiting popular speech recognition software for ESL speech. *TESOL Quarterly, 54*(4), 1086–1097. https://doi.org/10.1002/tesq.3006

McCrocklin, S., Humaidan, A., & Edalatishams, E. (2019). ASR dictation program accuracy: Have current programs improved? In J. Levis, C. Nagle, & E. Todey (Eds.),

Proceedings of the 10th Pronunciation in Second language learning and Teaching Conference, (pp. 191–200). Iowa State University.

McNamara, T., & Ryan, K. (2011) Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test, *Language Assessment Quarterly, 8*(2), 161–178. https://doi.org/10.1080/15434303.2011.565438

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment*. Oxford University Press.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8. https://doi.org/10.1111/j.1745-3992.1995.tb00881.x

Moussalli, S., & Cardoso, W. (2020). Intelligent personal assistants: Can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning, 33*(8), 865–890. https://doi.org/10.1080/09588221.2019.1595664

Mroz, A. (2018). Seeing how people hear you: French learners experiencing intelligibility through automatic speech recognition. *Foreign Language Annals, 51*(3), 617–637. https://doi.org/10.1111/flan.12348

Mroz, A. (2020). Aiming for advanced intelligibility and proficiency using mobile ASR. *Journal of Second Language Pronunciation, 6*(1), 12–38. https://doi.org/10.1075/jslp.18030.mro

Munro, M., & Derwing, T. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation, 6*(3), 283–309. https://doi.org/10.1075/jslp.20038.mun

Nakatani, L., & Dukes, K. (1977). Locus of segmental cues for word juncture. *The Journal of the Acoustical Society of America, 62*(3), 714–719. https://doi.org/10.1121/1.381583

Nation, I. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. (2017). *The BNC/COCA word lists* (Version 1.0.0) [Data file]. Victoria University of Wellington. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nelson, C., & Cardoso, W. (2024). Evaluating the effectiveness of Microsoft Transcribe for automating the assessment of pronunciation in language proficiency tests. In Y. Choubsaz et al. (Eds.), *EUROCALL 2023 Short Papers.* (pp. 117–122). Editorial Universitat Politècnica de València. https://doi.org/10.4995.EuroCALL2023.2023.17007

Nelson, C., & Cardoso, W. (in press). Dictation automatic speech recognition for second language pronunciation assessment: Focus on age-related bias. *Canadian Journal of Applied Linguistics.*

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer-assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, *21*(5), 393–408. https://doi.org/10.1080/09588220802447651

Ngo, T., Hao-Jan Chen, H., & Kuo-Wei Lai, K. (2023). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL*, 1–18. https://doi.org/10.1017/S0958344023000113

Nicklin C., Patterson A., & McLean, S. (2023). Quantifying proper nouns' influence on L2 English learners' reading fluency. *Studies in Second Language Acquisition, 45*(4), 906–929. http://doi.org/10.1017/S027226312200050X

Pourhosein Gilakjani, A., & Rahimy, R. (2020). Using computer-assisted pronunciation teaching (CAPT) in English pronunciation instruction: A study on the impact and the teacher's role. *Education and Information Technologies, 25*, 1129—1159. https://doi.org/10.1007/s10639-019-10009-1

Pouw, C., de Heer Kloots, M., Alishahi, A., & Zuidema, W. (2024) Perception of phonological assimilation by neural speech recognition models. *Computational Linguistics, 50*(4), 1557-1585. https://doi.org/10.1162/coli_a_00526

Proverbio, A., Lilli, S., Semenza, C., & Zani, A. (2001). ERP indexes of functional differences in brain activation during proper and common names retrieval. *Neuropsychologia*, 39(8), 815–827. https://doi.org/10.1016/S0028-3932 (01) 00003-3

Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (CAPT): Current issues and future directions. *RELC Journal, 52*(1), 189–205. https://doi.org/10.1177/0033688220977406

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T., & Minematsu, N. (2023). Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, *45*(1), 234 - 263. https://doi.org/10.1017/S0272263122000080

Saito, K., Trofimovich, P., Isaacs, T., & Webb, S. (2016). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs, & P. Trofimovich (Eds.), *Second language pronunciation assessment: Interdisciplinary perspectives* (pp. 141–156). Multilingual Matters. https://doi.org/10.21832/ISAACS6848

Saleh, A. & Pourhosein Gilakjani, A. (2021) Investigating the impact of computer-assisted pronunciation teaching (CAPT) on improving intermediate EFL learners' pronunciation ability. *Education and Information Technologies*, 26, 489–515. https://doi.org/10.1007/s10639-020-10275-4

Shahnawazuddin, S., Adiga, N., Kumar, K., Poddar, A., & Ahmad, W. (2020). Voice conversion based data augmentation to improve children's speech recognition in limited data scenario. In *Proceedings from INTERSPEECH 2020* (pp. 4382–4386). https://doi.org/10.21437/Interspeech.2020-1112

Skehan, P. (2018). *Second language task-based performance: Theory, research, assessment.* Routledge. https://doi.org/10.4324/9781315629766

Strachan, L., & Trofimovich, P. (2019). Now you hear it, now you don't: Perception of English regular past -ed in naturalistic input. *The Canadian Modern Language Review, 75*(1), 84–103. http://dx.doi.org/10.3138/cmlr.2017-0082

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: A mixed methods investigation. *Frontiers in Psychology*, 14, 1–14. https://doi.org/10.3389/fpsyg.2023.1210187

Tatman, R. (2017). Gender and dialect bias in Youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 53–59). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W17-1606

Tatman, R., & Kasten, C. (2017). Effects of talker dialect, gender, and race on accuracy of Bing speech and Youtube automatic captions. In *Proceedings from INTERSPEECH, 2017* (pp. 934–938). International Speech Communication Association. http://dx.doi.org/10.21437/Interspeech.2017-1746

Turner, C. (2014). Ratings Scales for Language Tests. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. John Wiley & Sons.

van der Walt, C., de Wet, F., & Niesler, T. (2008). Oral proficiency assessment: The use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies, 26(*1), 135–146. https://doi.org/10.2989/SALALS.2008.26.1.11.426

Van Moere, A., & Suzuki, M. (2018). Using speech processing technology in assessing pronunciation. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 137–152). Routledge.

Wallace, L. (2015). Reflexive photography, attitudes, behavior, and CALL: ITAs improving spoken English intelligibility. *CALICO Journal, 32*(3), 449–479. https://doi.org/10.1558/cj.v32i3.26384

Wallace, L. (2016). Using Google web speech as a springboard for identifying personal pronunciation problems. In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds.). *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference* (pp. 180–186). Iowa State University. https://apling.engl.iastate.edu/wp-content/uploads/sites/221/2016/06/PSLLT-7-Proceedings.pdf

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762–789. https://doi.org/10.1002/tesq.73

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing. 30*(2) 231–252. https://doi.org/10.1177%2F0265532212456968

Yan, X., & Ginther, A. (2018). Listeners and raters: Similarities and differences in evaluation of accented speech. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 67–88). Routledge.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883–895. https://doi.org/10.1016/j.specom.2009.04.009

Appendix A

Sentences for Speech Samples

Infrequent Lexical Items

A. Lexical Items

1. She hopes the facts will **compel** people to change their minds. (4k level)
2. The couple enjoyed a **rustic** wedding, surrounded by all their friends. (7k level)
3. We asked him to **implore** her to stay with the group. (9k level)
4. They admired the strong **gelding** before deciding approach it. (11k level)
5. The website shared **salacious** details about the famous woman's private life. (13k level)
6. The politician's remarks caused a **furor** among the people who were listening. (16k level)
7. She carefully studied the **meniscus** before noting its features. (18k level)
8. The older students gave the **plebe** a challenging job to complete. (20k level)
9. The villagers gathered at the old stone **fane** to offer prayers and seek blessings. (23k level)
10. The salad was made with fresh **orache** and tasted delightful. (25k level)

B. Proper Nouns

Note: Proper nouns cannot be assigned a frequency level because they are not included in word frequency lists.

Generic Names

1. It was now getting very dark indeed and **Alice** thought there must be a storm coming.
2. The woman was sure that **Tony** had something to do with the murder.
3. Her presence in the house made it possible for **Becky** to receive gentleman visitors.
4. I closed the door behind me so that **Laura** would not see or hear him.
5. That morning, he received a letter from **Susan** in which she asked him to come home.

Place and Organization Names

1. They spent hours studying the **Eiffel Tower**, noting its elaborate details.
2. The new mission has **Houston** organizing how the groups will work.
3. Many people feared **Hades** because of its unknown nature.
4. General rules that are followed by **Freemasons** are unknown to most people.
5. It is easy to recognize **the Matterhorn** as it has been photographed so often.

Unusual Collocations

1. The plant spent days in a **dry glass** and could not take root. (4 times in COCA)
2. His books were in a **worn trunk** in the corner of the room. (2 times in COCA)

3. He decided to **paint loosely** for this piece, breaking away from his usual style. (4 times in COCA)
4. It was nice to watch them **playfully recreate** the events rather than fight. (1 time in COCA)
5. He refused to offer a **firmer denial** even though he felt that he had done nothing wrong. (1 time in COCA)
6. She watched him **drink expertly** from the tiny teacup at the party. (1 time in COCA)
7. Thery had wanted a **stylish island** that combined form and practicality. (2 times in COCA)
8. They were using advanced **tool changers** to reduce downtime. (2 times in COCA)
9. When they got their **four-legged daughter**, they did not realize how much she would shed. (1 time in COCA)
10. The school lunches became **horribly healthy**, much to the students' disappointment. (1 time in COCA)
11. The new policy was an **act met** with mixed reactions from the voters. (1 time in COCA)
12. My favourite pet, **cat that he is**, often brings dead mice home. (does not appear in COCA)
13. The actor showed a **strangely pure** emotion that moved everyone to tears. (2 times in COCA)
14. They found that it was **hugely handy** to have the information in one place. (1 time in COCA)
15. It is challenging to **meaningfully measure** the effect of art on a community's culture. (2 times in COCA)
16. To them, the words we used to **intend honor** seemed unimportant. (1 time in COCA)
17. She had to select the **round square** tool option to draw that shape. (4 times in COCA)
18. The ghost story centered around a **lonely cottage** with a mysterious past. (1 time in COCA)
19. The way she played the piece was **classically correct**, but it lacked emotion. (3 times in COCA).
20. The children should try to lift their **uphill knees** to keep proper form. (3 times in COCA).

Ambiguous Speech

The ambiguous elements *in bold* the following 20 sentences (e.g., **a tax** versus **attacks**) were found in the sources cited at the end of each sentence pair. I co-created all sentences with the generative artificial intelligence tool, Perplexity (www.perplexity.ai).

1. There were reports about **a tax** on buses in the paper yesterday.
2. There were reports about **attacks** on buses in the paper yesterday. (Kim et al., 2012)
3. The people that asked for **more rice** were rude with the server.
4. The people that asked for **more ice** were rude with the server. (Mattys & Melhorn, 2013)
5. Is there a **known ocean** that matches the description that they provided?
6. Is there a **known notion** that matches the description that they provided? (Natikani & Dukes, 1997)
7. After this, **we'll own** the car for the weekend trips.
8. After this, **we'll loan** the car for the weekend trips. (Natikani & Dukes, 1997)

9. The picture of the **two lips** was beautiful against the black wall.
10. The picture of the **tulips** was beautiful against the black wall. (Gow & Gordon, 1995)
11. He turned and saw the **car go** around the corner.
12. He turned and saw the **cargo** around the corner. (Gow & Gordon, 1995)
13. She was looking for the **two cans** but couldn't find them.
14. She was looking for the **toucans** but couldn't find them. (Gow & Gordon, 1995)
15. They do not like that I **watched TV** in the bedroom.
16. They do not like that I **watch TV** in the bedroom. (Strachan & Trofimovich, 2019)
17. He knew who the **cat belonged** to but would not tell us.
18. He knew who the **cap belonged** to but would not tell us. (Gaskell & Marslen-Wilson, 2001)
19. During my explorations, my **net got** cut, which caused some problems.
20. During my explorations, my **neck got** cut, which caused some problems. (Gaskill & Wilson, 2001)