Evaluating dictation-based ASR technologies for L2 pronunciation assessment

Carey Nelson

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

March 2026

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared
By: **Carey Nelson**
Entitled: **Evaluating dictation-based ASR technologies for L2 pronunciation assessment**

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:

_____ Chair
Dr. Giuliana Cucinelli

_____ External Examiner
Dr. Rosane Silveira

_____ Arm's Length Examiner
Dr. Aparajita Dey-Plissonneau

_____ Examiner
Dr. Julie Corrignan

_____ Examiner
Dr. Teresa Hernandez-Gonzalez

_____ Thesis supervisor
Dr. Heike Neumann

_____ Thesis Supervisor
Dr. Walcir Cardoso

Approved by _____
Dr. Casey Burkholder, Graduate Program Director

December 8, 2025
Date of Defense

_____
Dr. Pascale Sicotte, Dean of Arts and Science

ABSTRACT

Evaluating dictation-based ASR technologies for L2 pronunciation assessment

Carey Nelson, Ph.D.

Concordia University, 2026

Automatic Speech Recognition (ASR) refers to the process by which digital devices convert spoken utterances into text. Once limited to dictation software or call centers, ASR is now integrated into everyday tools such as smartphones, laptops, and smart speakers. For second language learners, this technology can provide immediate transcriptions that permit real time monitoring, repetition, and correction, offering an interactive approach to pronunciation practice. In language education, ASR has been embraced for its potential to support learning and assessment, particularly pronunciation, where intelligibility is closely linked to communicative competence and professional opportunity. Its availability on widely used platforms such as Google, Microsoft, and Apple makes it a promising resource that can align with broader educational shifts toward autonomy, technology integration, and access to feedback beyond the classroom. However, concerns remain regarding its accuracy, validity, and potential biases. Considering these issues, this dissertation investigated the potential of dictation-based ASR for valid pronunciation assessment through two empirical studies.

In Manuscript A, the study examined Apple Siri as a potential tool for pronunciation assessment, extending prior work on Google Voice Typing and Microsoft Transcribe. Fifty-six adult English learners at a Canadian university completed a five-sentence read aloud oral test designed to target increasing pronunciation difficulty. Recordings were scored both by experienced human raters and by Siri, using a rubric covering comprehensibility, segmental accuracy, connected speech, stress, and rhythm. Siri's output was analyzed for transcription

accuracy and compared with human ratings. Results demonstrated strong correlations between Siri and human raters in measures of intelligibility. These findings suggest that Siri, like other dictation-based ASR tools, can produce valid and cost-effective results for formative assessment contexts. The study provides evidence that off-the-shelf ASR systems can help reduce rater bias, lower costs, and expand access to pronunciation feedback.

In Manuscript B, the focus shifted to validity by investigating potential age-related bias in dictation ASR. A corpus of test responses from 1,000 university learners was analyzed, spanning five first language backgrounds and three age groups. Each recording was processed through Google Voice Typing, Microsoft Transcribe, and Siri to generate word accuracy scores, which were then compared across age groups using regression analyses. Results indicated that the three ASR systems systematically favored younger test takers, a bias that reached statistical significance. These findings highlight an underexplored limitation of dictation-based ASR in assessment: while the technology can provide efficiency and large-scale testing, it may also introduce validity concerns that disproportionately affect older learners.

Together, these studies highlight both the promise and the limitations of dictation ASR in L2 pronunciation assessment. On the one hand, tools such as Siri demonstrate potential for generating valid, accessible feedback that aligns with human ratings, making them useful in educational contexts. On the other hand, evidence of age-related bias across systems raises questions about validity in high stakes contexts. The findings contribute to ongoing debates about the reliability and validity of automated assessment tools, emphasizing the need for critical validation before their large-scale adoption. This dissertation ultimately suggests that while dictation ASR can support broader access to pronunciation feedback, responsible integration requires systematic evaluation of accuracy, validity, and bias.

Acknowledgements

For many, the completion of a PhD marks the beginning of their professional careers. For me, however, this dissertation has been an opportunity to deepen my knowledge in a subject I have long adored: second language acquisition, especially as it is enriched and transformed by technology. This trek has allowed me to immerse myself in questions of learning, assessment, and pedagogy in ways that have been both intellectually stimulating and personally rewarding.

I would like to express my deepest gratitude to my two thesis advisors. To Doctor Walcir Cardoso, whose expertise in technology-enhanced language learning, and especially in pronunciation, has shaped not only the scope of this dissertation but also my understanding of how learners can engage meaningfully with digital tools. Your guidance, encouragement, and insightful feedback consistently pushed me to refine my ideas and pursue questions I might not otherwise have considered. To Doctor Heike Neumann, I am equally indebted for your thoughtful mentorship in the areas of language testing and assessment. Your insistence on rigor, clarity, and careful attention to issues of validity and reliability has anchored this research within a strong methodological and theoretical framework. The complementary perspectives of both of you have been invaluable, and I feel privileged to have benefited from your combined wisdom.

I would also like to thank my two examiners. Professor Julie Corrigan, your expertise on writing and presenting research in an efficient and convincing way while raising issues of notions of fairness in assessment helped me to expand my thinking on equity and ethics in language testing. The questions you posed, and the generous spirit with which you posed them, continually reminded me of the broader social implications of this work. Doctor Teresa Hernandez-Gonzalez, your knowledge of pedagogy and gamification brought a creative and learner-centered dimension to my project. Your insights encouraged me to consider how research

can remain grounded in classroom practice while still contributing to broader theoretical debates. Your thoughtful perspectives challenged me to bridge innovation with practicality, ultimately strengthening both the rigor and relevance of my work.

I would also like to express my sincere gratitude to Doctor Aparajita Dey-Plissonneau, whose insightful and probing questions greatly enriched the discussion and deepened my reflections on speech recognition systems and to Doctor Rosane Silveira, your experience and judicious approach to integrating ASR into L2 pronunciation research offered a balanced and insightful perspective that enriched both the methodological and pedagogical dimensions of my work.

Finally, I would like to extend my appreciation to colleagues, peers, and friends who provided encouragement, feedback, and support at different stages of this journey, and to my family for their unwavering patience and belief in me. I am especially grateful to Carol Johnson, who helped blaze the trail of using ASR for L2 pronunciation assessment, allowing me to explore in depth the potential and risks of using different ASR systems. I am deeply grateful to my daughter, Emi, whose unwavering patience, boundless love, and readiness to embrace the many sacrifices of this journey have sustained and inspired me every step of the way. Finally, I dedicate this work to my parents, who never had the opportunity to finish high school but cherished the value of education and supported me in all my endeavours. This dissertation is as much a product of their generosity as it is of my own effort.

Contribution of Authors

This dissertation, presented in manuscript format, is organized into four chapters. The opening chapter introduces the study and provides the foundation for the manuscripts that follow in Chapters Two and Three. The concluding chapter synthesizes and discusses the findings. Manuscript A was submitted in 2025 to the *Computer-Assisted Language Instruction Consortium Journal* (CALICO) and is currently under review. Manuscript B has been accepted for publication in a 2025 special issue of the *Canadian Journal of Applied Linguistics*.

CRediT Author Statement (Manuscript A). Carey Nelson: conceptualization; methodology; validation; formal analysis; investigation; resources; data curation; writing – original draft; writing – review and editing; visualization; project administration. Walcir Cardoso and Heike Neumann: conceptualization (supporting); writing – review and editing (supporting); supervision.

Note: Beau Zuercher and Suzanne Springer were responsible for developing the pronunciation assessment rubric examined in this study and for collecting the human-rater scores, which was part of an earlier project. These scores served as secondary data for Manuscript A.

CRediT Author Statement (Manuscript B). Carey Nelson: conceptualization; methodology; validation; formal analysis; investigation; resources; data curation; writing – original draft; writing – review and editing; visualization; project administration. Walcir Cardoso and Heike Neumann: conceptualization (supporting); methodology (supporting); writing – review and editing; supervision.

Table of Contents

List of Figures

List of Tables

Chapter 1: General Introduction

Automatic Speech Recognition in Language Learning: Definition and Landscape

Automatic Speech Recognition (ASR) refers to the process by which digital devices, computers, or cloud services interpret and convert spoken utterances into written text (Li, 2022). Initially developed for specialized applications such as dictation software and automated customer service systems, ASR has since evolved into a widely accessible tool found in everyday technologies. Its integration into smartphones, smart speakers, laptops, and productivity software has made voice-enabled interactions commonplace (Levis & Suvorov, 2012). ASR offers learners the ability to speak directly into their devices and receive instantaneous transcriptions, enabling them to monitor their performance, identify recurring challenges, and make real-time adjustments to their speech production. This immediacy not only enhances learner engagement but also facilitates self-correction, repetition of problematic sounds, and development of strategic pronunciation habits over time (McCrocklin, 2019). In essence, ASR transforms what was once passive practice into a feedback-rich, interactive process that empowers learners to take greater control over their pronunciation development (Tejedor-García et al., 2021)

In the domain of language education, particularly second language (L2) acquisition, ASR has generated growing interest due to its potential to support both instruction and assessment (Chapelle & Voss, 2016). It is increasingly positioned as a resource that can offer real-time, personalized feedback on spoken language, allowing learners to improve pronunciation and overall intelligibility. This is especially important for those whose communicative competence may affect their access to academic opportunities, job prospects, or professional credibility (Cenoz & Lecumberri, 1999).

As pedagogical models shift toward greater learner autonomy and technological integration (e.g., Lan, 2018), ASR aligns with broader trends in education that emphasize personalized, data-driven, and self-directed learning (Isaacs, 2018). It enables access to pronunciation support beyond the classroom, circumventing barriers such as scheduling constraints, lack of instructional hours, or limited availability of trained pronunciation coaches (Litman et al., 2018). Within this context, ASR becomes a compelling option, as it can work with large numbers of students giving immediate feedback while being embedded in platforms that learners already use, for example, Apple, Google, Microsoft, etc. As artificial intelligence continues to shape the future of assessment and instruction, ASR tools represent a key point of intersection between human-centered learning goals and automated digital infrastructure (Li, 2022). However, the expanding use of ASR in evaluative contexts also raises important ethical and methodological concerns, particularly regarding validity and equity (Ngueajio & Washington, 2022). When employed for high-stakes purposes such as proficiency testing, ASR systems must be critically examined to ensure they operate transparently and without introducing unintended bias (Liu et al., 2021).

One of the primary advantages of contemporary ASR systems is their high degree of accessibility. Popular ASR systems such as Apple Siri, Google Voice Typing, and Microsoft Transcribe are pre-installed on most consumer devices and may not require additional hardware or a subscription. This widespread availability lowers the barrier to entry, allowing learners from varied socioeconomic and linguistic backgrounds to access pronunciation tools with minimal cost and effort. Their interfaces are often readily known by users, making them viable options even for those with low digital literacy (Miras et al., 2023). The surge in mobile learning and the normalization of remote education, which has been accelerated by global events such as the

COVID-19 pandemic, have only heightened the relevance of support such tools (Saikat et al., 2021).

Yet despite these advantages, the pedagogical potential of ASR can be affected by potential limitations in recognition accuracy, particularly when interpreting speech that is less intelligible (Cámara-Arenas et al., 2023). Most commercial ASR systems are trained on large language models that tap into corpora of first language speaker data, which means they are optimized for standard accents and fluency patterns. As a result, their ability to accurately transcribe speech from L2 learners, especially those with accents that exhibit atypical prosody or segmental deviations, may not perform as well (Aksënova et al., 2022). The consequences of ASR errors extend beyond momentary confusion: learners may receive inaccurate or misleading feedback that reinforces incorrect patterns or undermines their confidence. For example, a correctly articulated word may be flagged as incorrect due to system error, while mispronounced words might be accepted if the ASR model relies too heavily on contextual prediction (Feng et al., 2021; John, 2025). These inconsistencies raise important questions about the reliability and validity of ASR-based feedback. If learners cannot trust that the system's output accurately reflects their spoken input, the educational value of the tool can be compromised (Guskaroska, 2020). Educators and program designers must therefore critically assess how recognition errors affect learners' developmental trajectories and what safeguards can be implemented to mitigate those effects (Hinsvark et al., 2021).

Among some of the off-the-shelf ASR tools most frequently encountered by language learners, three systems warrant particular attention: Apple Siri, Google Voice Typing (GVT), and Microsoft Transcribe (MS-T). Each of these platforms is included in the major operating systems: Android, Windows, and iOS/macOS respectively. They are commonly used for

everyday dictation, device navigation, or note-taking. These tools are not designed specifically for language learning, yet their widespread availability, zero or low cost, and ease of use have made them attractive resources for pronunciation practice. Importantly, they differ in meaningful ways. GVT is known for its multilingual capabilities and adaptability. MS-T, available through the Office 365 suite, provides transcriptions through its online version of Word, though it may be less accessible for learners without an Office subscription. Finally, Siri is deeply integrated within Apple products. The ways in which these systems process varying levels of comprehensible speech and the degree to which they introduce or minimize bias both have significant implications for their pedagogical and evaluative use.

This dissertation examined the extent to which three ASR systems can support equitable pronunciation assessment in L2 contexts (Manuscripts A and B). It focused on evaluating Siri's performance in comparison to human raters (Manuscript A) and on investigating the potential influence of proficiency level (Manuscript A) and age (Manuscript B) across all three ASR systems.

Pedagogical Applications of ASR in L2 Pronunciation

Improving pronunciation is a central and often indispensable goal for many L2 learners, especially when their intelligibility directly influences academic or professional achievement (Dillon & Wells, 2021). For learners aspiring to enter fields that rely heavily on spoken interaction, comprehensible speech is not merely a linguistic objective, but a practical necessity. Intelligibility can shape how individuals are perceived in the workplace, influence hiring decisions, and affect opportunities for promotion and professional advancement. In academic environments, pronunciation can play a crucial role in group discussions, oral presentations, and classroom participation. Research has shown that pronunciation difficulties, even for learners

who demonstrate high grammatical and lexical proficiency, can still pose significant barriers to effective communication and may negatively impact learners' self-confidence and willingness to participate in spoken interactions (Derwing & Rossiter, 2002; Zielinski, 2012). In this context, ASR technologies present a practical and increasingly accessible solution for developing pronunciation awareness and accuracy (Dillon & Wells, 2021).

ASR tools function by transcribing spoken language into written text in real time, providing learners with a visual representation of their speech. This immediate feedback allows users to see which phonemes, syllables, words, or phrases are consistently misrecognized and infer which elements of their pronunciation may be unclear to their listeners. For instance, a learner who says "think" as /sɪŋk/ may see it transcribed as "sink," highlighting a common difficulty with the voiceless interdental fricative /θ/. These discrepancies serve as diagnostic cues, guiding learners toward specific areas for improvement (John et al., 2022). This process helps develop metacognitive awareness of speech production and fosters focused practice on problematic segmental and suprasegmental features, including vowel length distinctions, syllable stress, intonation contours, or final consonant clusters. Over time, repeated and targeted engagement with ASR feedback can support incremental improvement in intelligibility (Gutz et al., 2023). This can cultivate a sense of self-monitoring and correction that can carry over into natural, spontaneous conversation.

Beyond pronunciation practice, ASR systems are also gaining attention for their potential as objective pronunciation *assessment* tools. Traditional L2 pronunciation assessment has relied primarily on human raters, whose judgments, though expert, can be influenced by subjective factors such as rater bias, fatigue, expectations, or familiarity with a learner's first language (L1) accent (Saito et al., 2023). These elements can introduce variability and inconsistency into

scoring, which is problematic in high-stakes or large-scale testing contexts. In contrast, ASR-based assessment offers the potential for standardized and automated scoring of multiple samples (Isaacs, 2013). ASR systems can analyze a learner's spoken responses in terms of phonemic accuracy, word intelligibility, and assign scores based on how closely the input aligns with the target language or normative model (Van Moere & Suzuki, 2017). Tasks such as read-aloud passages, sentence repetition, or controlled pronunciation drills are particularly well-suited to automated scoring, and several studies have demonstrated moderate to strong correlations between ASR-generated scores and those provided by trained human raters (Tejedor-García et al., 2021). These findings suggest that ASR has considerable promise as a reliable tool in both formative classroom assessments and summative proficiency evaluations.

A key advantage of ASR technology in the pedagogical realm is its capacity to support individualized, self-directed pronunciation learning. While traditional classroom settings can be constrained by time, teacher availability, and group pacing (Celce-Murcia et al., 2010), ASR tools provide learners with the flexibility to practice at their own convenience. This autonomy is especially beneficial for adult learners juggling multiple responsibilities or for students in geographically remote areas with limited access to language instruction. Learners can engage with pronunciation practice in a low-stakes environment like placement testing where they can repeat words or sentences as many times as needed. Moreover, modern ASR-integrated platforms increasingly incorporate adaptive learning features such as performance tracking, voice journals, and personalized feedback logs (Cox & Davies, 2012). These elements allow learners to set milestones, revisit problematic items, and observe their progress over the long-term, supporting not only language development but also learner motivation and, perhaps more importantly, a sense of ownership over the learning process.

The immediacy of this feedback loop can be one of ASR's most pedagogically powerful features: It allows learners to adjust their pronunciation in real time, reducing the risk of fossilization, which is the inability to change incorrect forms. Unlike delayed or generalized teacher feedback, ASR tools give learners actionable input at the moment of speech, making each practice attempt more productive (Bashori et al., 2024). This active feedback loop reinforces the link between perception, production, and correction, contributing to deeper phonological awareness (John et al., 2022). In this way, ASR functions not merely as a reactive tool for flagging errors but as a dynamic and responsive tutor embedded in the learning process. This pedagogical potential has led to increasing interest in leveraging ASR for L2 pronunciation assessment, which can serve not only to evaluate learner output but also to inform and guide targeted pronunciation practice.

ASR in L2 Pronunciation Assessment

ASR technology has increasingly demonstrated its potential as a reliable and effective tool for evaluating L2 pronunciation, particularly in an English as a second language (ESL) context. Over the past decade, a growing body of research has examined the extent to which ASR-generated scores correlate with those produced by expert human raters (e.g., Johnson et al., 2024; Nelson & Cardoso, 2024). These studies typically analyze pronunciation through quantifiable measures such as word error rate, phoneme-level accuracy, speech rate, and fluency patterns (Bernstein et al., 2010). In many cases, especially in controlled testing environments involving structured tasks like word reading or sentence repetition, ASR systems have shown moderate to strong correlations with human evaluations. This alignment suggests that, under specific conditions, ASR has the capacity to produce valid, consistent assessments of pronunciation performance (Hollands et al., 2022).

ASR's ability to assess large numbers of recordings is one of its most significant advantages, particularly for institutions or programs seeking to administer pronunciation testing to large cohorts of learners. Traditional pronunciation assessment often requires substantial time and labor from trained evaluators, which can pose logistical and financial challenges in settings with limited human resources (Isaacs & Trofimovich, 2016). In contrast, ASR-enabled tools offer immediate scoring capabilities, uniform application of assessment criteria, and the infrastructure to support repeated testing without fatigue or drift in judgment. These features make ASR particularly appealing for formative classroom assessments, placement testing, and even low-stakes proficiency evaluations.

However, while ASR holds promise as a tool for pronunciation assessment, its reliability is not uniform across all contexts, linguistic features, or learner populations. The effectiveness of ASR systems depends significantly on how well they have been trained to recognize second or foreign language speakers' speech patterns (Hinsvark et al., 2021). Many commercial ASR models are primarily developed and fine-tuned using large corpora of L1 speaker data, which may not adequately reflect the pronunciation features of L2 learners (Inceoglu et al., 2023). As a result, recognition accuracy can vary substantially based on factors such as a speaker's first language (L1), age, proficiency level, and the presence of regional or foreign accents (Arora et al., 2018; Chan et al., 2022). For instance, learners whose L1 lacks certain English phonemes may produce substitutions that may be misinterpreted by the ASR system. This can lead to incorrect transcriptions and unfair scoring, even when the learner's speech would be intelligible to a human listener (O'Neill & Carson-Berndsen, 2023). Suprasegmental elements such as intonation, rhythm, and stress placement are especially difficult for ASR to evaluate reliably, given the nuanced and context-dependent nature of these features (Kang & Johnson, 2018).

Consequently, while ASR can provide detailed feedback on segmental pronunciation (e.g., individual consonants and vowels), its capacity to assess prosodic accuracy (e.g. rhythm and intonation) remains limited in most consumer-grade systems.

In sum, ASR-based pronunciation assessment offers substantial benefits in terms of efficiency, consistency, and accessibility, particularly when human raters are unavailable or impractical (Saito et al., 2016). As the field continues to develop, research should aim to evaluate these technologies with the objective of attaining equitable and meaningful assessment in L2 pronunciation learning.

Bias and Validity: Evaluating ASR Performance Across Diverse Groups

Another important concern is the potential for bias in ASR-based pronunciation assessment (Feng et al., 2021; Liu et al., 2021). If the system regularly underperforms for particular groups of speakers, such as older adults, speakers with non-dominant accents, individuals from underrepresented linguistic backgrounds, or students of varying proficiency levels it can inadvertently reinforce inequities and produce misleading conclusions about learner competence (Aksënova et al., 2022; Ngueajio & Washington, 2022). This is particularly problematic in high stakes testing environments, where the outcomes of assessments may affect applications for immigration, certification, or access to employment opportunities. For ASR to be adopted responsibly in these contexts, it must be subjected to rigorous evaluation to ensure that its scoring mechanisms are fair, transparent, and representative of the diverse learner populations it is intended to serve (Kulkarni et al., 2024; Liu et al., 2022). As ASR technology becomes more deeply integrated into language assessment contexts, ensuring validity is not simply a desirable feature: it is a foundational requirement for responsible and ethical use.

In high-stakes applications where test results may influence access to academic programs, professional certifications, employment opportunities, and immigration pathways, the burden of proof lies squarely on developers and implementers to ensure that such tools uphold principles of equity and non-discrimination (Van Der Walt et al., 2008). These assessments are not casual learning tools, but formal mechanisms that can shape individuals' life trajectories. As such, the validity and reliability of any scoring system must be accompanied by a demonstrated commitment to validity across diverse populations (Bachman & Palmer, 1996). If ASR systems consistently yield biased results for specific groups, such as older adults, speakers of underrepresented L1s, or individuals with regional or second-language L2 accents, the validity of the entire evaluative process may be undermined.

Validity in ASR-based assessment must be understood as encompassing more than just technical accuracy or error rates (Szymański et al., 2020). It involves deeper responsibilities, including the obligation to recognize how technological systems can reproduce, amplify, or conceal structural biases already present in educational and social institutions (Chapelle & Lee, 2021). A truly fair language assessment system must ensure that all learners are evaluated based on the quality of their performance, not on how closely their speech resembles the data on which the ASR was trained (Babaeian, 2023). In this light, ensuring validity is not merely a computational goal; it is an issue of educational equity. Transparent reporting of system limitations, inclusive training datasets, and equitable benchmarking practices are critical elements in this process.

One of the most pressing concerns in this area is the potential for ASR systems to exhibit bias in how they process and evaluate speech across different demographic categories (Chan et al., 2022). Most commercial grade ASR models are trained on extensive datasets composed

predominantly of first language speaker speech. While such training data can support high levels of accuracy for standard speech inputs, it fails to represent the full range of linguistic and phonetic variation that characterizes L2 learner populations (Evers & Chen, 2021). This may lead to features that are misrecognized where learners are penalized for differences that are intelligible or incorrect in a communicative context.

While many studies have focused on the performance of ASR systems developed specifically for language learning, far less attention has been paid to general-purpose, dictation-based systems like Siri, GVT, and MS-T. These tools are widely accessible and frequently used by learners for everyday use or even pronunciation practice, yet their effectiveness and validity in assessing L2 speech remain largely unexamined. This area remains underexplored but deserves to be looked at given the increasing reliance on these systems for informal self-evaluation. Moreover, existing research has not always considered how such tools handle speech from L2 users across a range of age groups and proficiency levels. These demographic factors can significantly influence interaction with ASR technologies, potentially leading to inconsistent or biased feedback. Learners at different developmental stages may produce distinct pronunciation patterns, pause behaviors, and lexical choices, all of which could affect how accurately their speech is transcribed. Without comparative studies evaluating how these popular systems perform across diverse learner profiles, it is difficult to determine whether they can provide valid, equitable feedback or if they inadvertently privilege speech that aligns more closely with the systems' training data. This gap underscores the importance of evaluating widely used ASR tools not only in terms of transcription accuracy but also through the lens of validity across key learner characteristics.

Age-related differences in speech, which are one of the focuses of this dissertation, present an additional source of potential bias in ASR performance (Chen & Asgari, 2020). Since these systems are typically optimized for adult voices within a specific age range, their recognition accuracy may decline when processing the speech of older adults or younger children (Aman et al., 2013). Older learners may experience age-related physiological changes in vocal tract configuration, articulation patterns, or voice pitch, which can reduce recognition accuracy if the ASR model has not been exposed to similar speech during training (Chen & Asgari, 2020). This may lead to unfair outcomes where learners are evaluated less on their actual pronunciation skill and more on the system's unfamiliarity with age-related vocal characteristics. Moreover, the intersection of multiple identity markers can compound the potential for bias and remains to be studied.

One of these identity markers is accent diversity, which introduces yet another layer of complexity. Even among proficient English users, regional or non-standard accents can affect how ASR systems interpret spoken input. For example, speakers with Caribbean, West African, or South Asian English varieties may articulate English phonemes differently than North American or British English speakers, yet their speech may still be entirely intelligible to human listeners (Chan et al., 2022). When ASR systems fail to account for this diversity, they risk marginalizing legitimate linguistic variation and promoting a narrow view of what counts as acceptable pronunciation (Derwing & Munro, 2009). This is especially problematic in multicultural societies where linguistic pluralism is the norm, not the exception. If such systems are implemented in educational institutions, immigration offices, or professional testing bodies without proper bias evaluation, they may inadvertently penalize speakers who deviate from dominant speech norms through no fault of their own. This requires that learners and instructors

be educated about the limitations of ASR technology (Nickolai, 2024). Overreliance on ASR feedback without human interpretive support can lead to misconceptions about what constitutes accurate pronunciation or intelligible speech.

In conclusion, validity in ASR-based L2 pronunciation assessment is a multidimensional concern that intersects with issues of data representativeness, system design, linguistic diversity, and educational equity. As ASR continues to shape the future of language testing, ensuring that these systems function equitably across diverse user groups is not just a technical aspiration but a moral imperative. Any implementation of ASR in assessment must be accompanied by comprehensive evaluations of system bias, transparent documentation, and a commitment to inclusivity. Only then can ASR be trusted to play a meaningful role in promoting not only language proficiency, but validity and justice in language education.

This dissertation addresses this research gap by systematically analyzing the accuracy, reliability, and potential bias of Siri, GVT, and MS-T in the context of L2 pronunciation assessment, with the goal of informing future applications of ASR in language testing. Particular attention will be given to their recognition accuracy compared to human raters who are scoring using a rubric of different pronunciation targets. In addition, this study looks at the susceptibility to bias across diverse learner profiles with respect to age and proficiency levels. The findings can be used to inform best practices for integrating such tools into pronunciation assessment, contributing to a more equitable and effective use of ASR in second language education.

This Dissertation: Scope and Goals

ASR technology has evolved rapidly in recent years, becoming a significant tool in the instruction and evaluation of L2 pronunciation. Traditionally, in high-stakes language testing contexts, pronunciation teaching and pronunciation assessment have occupied distinct spheres of

research and practice, with instructional efforts focusing on awareness and intelligibility, and emphasizing accuracy, consistency, and standardization. However, the increasing sophistication and accessibility of dictation-based ASR platforms such as Siri, GVT, and MS-T, have opened up new possibilities for bridging these domains. These widely available tools not only provide learners with real-time transcription of their spoken input but also have the potential to function as low-cost, scalable assessment instruments. Their dual capacity to support both formative feedback and summative evaluation offers a promising yet largely untapped avenue for innovation in language learning and testing.

While there is growing interest in the pedagogical applications of ASR, not much is known about their effectiveness and limitations as tools for pronunciation assessment. Most existing research has focused on how ASR can support learners' pronunciation improvement by offering immediate feedback, improving metacognitive awareness, and allowing for self-paced practice. However, the use of dictation ASR for formal or semi-formal evaluation, especially within language testing frameworks, remains significantly underexplored. Key concerns persist regarding the consistency and validity of the scores these systems produce when evaluating L2 speech, especially given that most ASR technologies are trained predominantly on L1 speaker data. As a result, questions remain about how accurately these systems can assess speech that deviates from norms due to age-related variation.

This dissertation sought to address this critical gap in the literature through two empirical investigations (Manuscript A and Manuscript B) that examined the potential and limitations of dictation-based ASR platforms as tools for L2 pronunciation evaluation guided by one underlying question:

- Can Siri, GVT, and MS-T provide reliable, unbiased, and educationally viable pronunciation evaluation for L2 learners?

The first investigation (Manuscript A) examined the viability of Siri as a tool for pronunciation scoring by comparing its output to human rater judgments, as well as to the performance of Google Voice Typing (GVT) and Microsoft Transcribe (MS-T) when assessed against the same human-rated data. While GVT and MS-T have previously been examined in L2 contexts, no prior study had compared their performance alongside Siri and human raters within the same evaluation framework. As such, this study evaluates the degree to which Siri-generated scores align with human judgments. It also explored the implications for test usefulness, validity, and pedagogical application. An in-depth examination of potential age-related biases that may emerge when these same ASR systems (Siri, GVT, and MS-T) were used to evaluate the pronunciation of L2 learners across different age groups forms the basis of Manuscript B. A detailed description of the two manuscripts is presented below.

Manuscript A (Chapter 2) evaluated the feasibility of using Apple Siri to generate scores for a pronunciation placement test, applying Bachman and Palmer's (1996) test usefulness framework, which emphasizes not only the reliability and construct validity of a test but also its impact, practicality, and authenticity in applied settings. The placement test responses used in the study had previously been evaluated by trained human raters using standardized scoring rubrics focused on intelligibility and phonological accuracy. In this study, the same set of pronunciation samples of fifty-six university-level ESL students was processed through Siri's dictation engine, with word recognition accuracy serving as the basis for Siri-generated scores. Word accuracy was chosen as a proxy for intelligibility, under the assumption that higher transcription accuracy indicates more intelligible pronunciation (Lochland, 2020; Loukina et al., 2015). These

automatically generated scores were then compared to those of human raters as well as to scores generated by GVT (Johnson et al., 2024) and MS-T (Nelson & Cardoso, 2024) in previous studies. To investigate whether transcription accuracy varied meaningfully across different levels of language proficiency, students were systematically categorized based on their proficiency levels, allowing for a more nuanced analysis of performance patterns between groups. Following established research practices (e.g., Bernstein et al., 2010; Saito et al., 2023), correlation analyses were conducted between the ASR-derived and human scores, and the strength of these relationships were examined to determine the relative effectiveness of each ASR system for pronunciation assessment purposes. I am the first and main author, and the full-length article has been submitted to the *Computer Assisted Language Instruction Consortium* (CALICO) and is under review.

Manuscript B (Chapter 3) explored a potential age-related bias in ASR-based pronunciation assessment, an issue that remains underexplored in the field of applied linguistics. While ASR technology has demonstrated promising capabilities in evaluating L2 pronunciation, often showing strong alignment with human rater scores in prior studies, questions persist about its ability to fairly assess speakers whose vocal characteristics may differ from the training data, particularly due to physiological changes associated with aging. These may include shifts in pitch, articulation speed, and voice quality, all of which could affect ASR recognition accuracy. This study investigated whether Siri, GVT, and MS-T exhibit measurable age-related bias when used to evaluate English L2 pronunciation among 1,000 university-level candidates. Participants represented five L1 backgrounds (French, Spanish, Persian, Arabic, and Chinese) and were categorized into three age groups: under 30, 30–44, and over 44. The participants' speech samples were processed by all three ASR systems, and word recognition accuracy calculated as

the basis for comparison. Based on the results, the study proposes actionable recommendations to reduce the impact of age-related bias and improve the validity, transparency, and validity of ASR-based assessment tools. I am the first and main author, and the full-length article has been accepted for publication in a forthcoming special issue of the *Canadian Journal of Applied Linguistics*.

Taken together, this dissertation explored the validity and reliability of Siri, GVT, and MS-T in scoring English L2 pronunciation placement tests across different age groups and across different proficiency levels. It aims to contribute to best practices in ASR-based assessment by analyzing the impact of these systems on test usefulness, validity, and accessibility in both formative and summative contexts.

Table 1 summarizes the specific objectives, participant details, methodological considerations, and anticipated contributions of each study, offering a clear overview of how the two manuscripts work in tandem to address key challenges in automated pronunciation evaluation.

Table 1

Overview of Manuscripts

Manuscript	Goal	Participants	Potential Contributions
A*	Evaluate Siri's capacity to generate reliable and valid automated scores for L2 pronunciation assessment: • In comparison with other ASR systems • Accounting for variations in test-taker proficiency	University-level ESL students with various proficiency levels and L1s (n = 56)	Assess the feasibility of using Siri to automatically score L2 pronunciation assessments
B**	Determine if Siri, GVT, and MS-T exhibit age bias when used for L2 pronunciation assessment	University-level ESL students with various proficiency levels and L1s (n = 1000)	Highlight ASR-related age biases and underscore the need for validity in automated L2 pronunciation assessments

* Manuscript A has been submitted to the *Computer Assisted Language Instruction Consortium* (CALICO) and is under review. I am the first and main author.

** Manuscript B has already been accepted for a special issue of the *Canadian Journal of Applied Linguistics*. I am the first and main author.

Chapter 2: From Voice Assistant to Pronunciation Evaluator: Assessing Siri's Role in L2 Testing

Assessing language proficiency is essential for making informed decisions about second language (L2) competence in high-stakes contexts such as job applications, immigration procedures, and university admissions. These assessments may rely on internationally recognized standardized tests such as TOEIC, IELTS, or Versant, or they may be custom designed by individual institutions. Regardless of their format, most proficiency tests evaluate multiple linguistic skills to provide an overall picture of a candidate's abilities. When these assessments involve human raters applying scoring criteria to spoken responses, the grading process can be both time-consuming and resource-intensive (Coombe et al., 2020). In addition to these logistical demands, human scoring can introduce errors and variability stemming from differences in raters' interpretations of assessment criteria (Inbar-Lourie, 2017).

Automatic Speech Recognition (ASR) technology presents a promising tool for evaluating intelligibility in speaking tasks that would otherwise require multiple human raters (Bernstein et al., 2010). Human judgments, while indispensable, are susceptible to well-documented sources of bias and inconsistency, including rater severity, fatigue, and contextual effects, which can complicate the achievement of standardized evaluations (Boyd & Donnarumma, 2018).

ASR systems, by contrast, can offer a high degree of procedural consistency in scoring, thereby reducing variability attributable to human raters (Mroz, 2020). However, this consistency should not be conflated with fairness or impartiality. ASR systems are inherently shaped by their training data, model architecture, and input conditions, and their performance may vary systematically across learner populations, proficiency levels, and recording environments.

Consequently, while ASR may reduce certain sources of human-related variability, it introduces distinct forms of bias and instability, raising important questions about validity, reliability, and equity in L2 speech assessment.

To better understand this potential, this study explores how a specific dictation ASR system, Apple Siri, can be leveraged to assess second language pronunciation by examining the accuracy and consistency of its transcription outputs by calling upon Bachman and Palmer's (1996) test usefulness framework . As Siri is deeply embedded in the Apple ecosystem, investigating its performance offers valuable insights into the potential of everyday, readily accessible technology for language assessment, which can build on prior research focused on GVT and MS-T. The objective of this study looks to: (1) compare Siri's transcription accuracy with scores provided by human raters, (2) examine how test-taker proficiency level may affect ASR recognition performance, and (3) evaluate how its performance aligns with or diverges from that of GVT and MS-T when measured against the same set of human-rated benchmarks.

Literature Review

L2 Pronunciation Skills

Pronunciation skills are a fundamental part of learning oral production in a second or foreign language. They are not only essential for facilitating intelligible communication but also play a significant role in how language learners are perceived socially, academically, and professionally (Zielinski, 2011). In fact, pronunciation can shape first impressions, influence listener attitudes, and even affect learners' self-confidence and willingness to communicate (Derwing & Munro, 2009).

For over half a century, very little research addressed how L2 pronunciation was taught, and pronunciation was often marginalized within communicative language teaching approaches

(Isaacs, 2013). Historically, language programs placed greater emphasis on grammar and vocabulary, often leaving pronunciation to be informally corrected or implicitly acquired. However, more recently, research on second language pronunciation instruction has shifted, to prioritizing the more practical and learner-centered goal of intelligibility (Isaacs, 2018). In fact, Isaacs (2013) explains that the instructional goal of achieving "native-like" pronunciation, although traditionally regarded as the "gold standard," is problematic (see also Celce-Murcia et al., 2010, for similar claims). It is difficult to objectively delineate "native" pronunciation given the diversity of regional and social accents even among first language (L1) speakers, which can also be an unrealistic or unattainable goal for adult L2 learners (Celce-Murcia et al., 2010). More importantly, the pursuit of removing accentedness can carry sociolinguistic implications, as it often ignores learners' linguistic identities and reinforces L1 speaker norms as the ideal (Hansen et al., 2020). In contrast, intelligibility acknowledges that speakers can be highly effective communicators without fully conforming to L1 models. This reflects a broader movement in applied linguistics that emphasizes learner agency, multilingual realities, and functional outcomes over imitation of expected norms (Canagarajah, 2013).

As English continues to function as a global lingua franca, spoken among L2 speakers as often, or even more so, than among L1 speakers (Jenkins, 2000), the necessity of achieving highly fluent pronunciation, understood here as pronunciation approximating L1 accuracy and accent norms, as an instructional target becomes increasingly questionable. Instead, researchers advocate for a focus on functional intelligibility, defined as the extent to which listeners accurately understand a speaker's utterances (Isaacs & Trofimovich, 2016). This goal aligns more closely with real-world communication needs and provides a more inclusive and attainable

target for learners. Derwing and Munro (2015) support this view, noting that intelligibility is best assessed through listener comprehension rather than phonological accuracy relative to L1 norms.

To define what contributes to intelligibility, Derwing and Munro (2015) identify several core dimensions of spoken language. These include:

1. Segments – the production of individual consonants, vowels, and diphthongs in the target language, which form the building blocks of intelligible speech;
2. Prosody – the pattern of word and sentence stress, rhythm, and intonation that shapes how utterances are delivered and understood in context;
3. Accent – the speaker's overall sound system, including pitch patterns that can be especially crucial in tonal languages; and
4. Fluency – the rhythm, pacing, and smoothness of delivery, encompassing phenomena such as syllable timing, linking, elision, and the ability to produce continuous speech without excessive hesitation.

Isaacs (2013) reiterates this framework and emphasizes that intelligibility is not just a matter of accurate segmental production, but a multifaceted construct shaped by both linguistic (connected speech, syntax and grammar, or lexical choices) and social factors (power dynamics or bias, setting of interaction, or shared cultural references or not). Together, these pronunciation dimensions that make up a fundamental part of intelligibility offer a practical foundation for both instruction and assessment. They also serve as meaningful categories for learners to use when engaging in self-assessment and goal setting. By focusing on intelligibility, instructors and learners can shift toward a more realistic, inclusive, and effective construct that values clear communication over mimicry of expected L1 norms, and that recognizes the diversity of acceptable English pronunciations across global contexts (Isbell, 2019). One crucial aspect of

understanding the importance of learning L2 pronunciation lies in examining how it is evaluated, particularly through human rating.

Human Rating of L2 Pronunciation

In the field of L2 pronunciation assessment, human raters are important due to their ability to interpret speech with contextual sensitivity and nuanced understanding (Isaacs & Thomson, 2013). Their training enables them to recognize prosodic features, communicative intent, and subtle deviations from target pronunciation that automated speech recognition systems may overlook. Particularly in formative and instructional settings, human raters are well positioned to provide qualitative feedback that supports learner development. Analytic scoring rubrics further enhance this role by allowing raters to focus on specific features of pronunciation such as segmental accuracy (vowels and consonants), prosody (intonation, stress, rhythm), and broader measures like intelligibility and comprehensibility (Isaacs & Trofimovich, 2016). Although holistic rubrics offer faster scoring, they are often less effective for detailed diagnostic feedback (Saito, 2021). In high-stakes contexts such as academic placement, employment opportunities, or immigration applications, human raters are still considered a point of reference due to their ability to assess pronunciation within meaningful linguistic and pragmatic contexts.

However, despite their advantages, human ratings of L2 pronunciation face notable challenges, particularly the influence of subjective factors and biases that can compromise consistency and overall validity. As Isaacs (2013) observed, raters' judgments can be affected by extraneous factors such as the speaker's accent (e.g., familiarity or perceived ease of understanding) and the complexity of the speech task. Human raters' prior experiences with or expectations of L2 pronunciation may also shape their evaluations of pronunciation, introducing subjectivity into the process. In addition, raters can be influenced by various forms of bias that

can further affect the validity of assessment outcomes. When raters are more lenient toward accents with which they are familiar and less so toward unfamiliar ones, this tendency is referred to as L1 bias (Isaacs, 2013; Winke et al., 2013). Such bias can lead to systematically lower scores for learners whose first language is less commonly encountered or understood by raters.

Age-related bias has also been observed: younger speakers are sometimes perceived as more capable language learners and may receive more favorable ratings, whereas older speakers may not (Vipperla et al., 2010). Gender bias may influence ratings as well, with some studies suggesting that female speakers are sometimes rated more positively due to perceived politeness or clarity, while male speakers may be rated more harshly for assertiveness or pitch range (Kang & Rubin, 2009). Additionally, sociodemographic biases, including assumptions based on ethnicity, nationality, or perceived socioeconomic status, can affect raters' judgments unconsciously, particularly in high-stakes settings where raters may make implicit associations between accent, credibility, and competence (Derwing & Munro, 2009). These biases highlight the need for robust and frequent rater training, norming procedures, and awareness-raising to minimize subjective influences in pronunciation assessment.

Bias may also manifest through interrater variability, especially when raters differ in their phonetic training, linguistic background, or exposure to L2 accents (Isaacs, 2008; Winke et al., 2013). Even with training and clear rubrics, raters may diverge in how they interpret pronunciation features, particularly in judgments of accentedness. A key distinction must be made between intelligibility (how much is understood), comprehensibility (how easy it is to understand), and accentedness (how different the speech sounds from L1 speaker norms; Munro & Derwing, 1999). While intelligibility and comprehensibility are more closely linked to communicative success and pedagogical relevance (Derwing & Munro, 2009), research shows

that raters may still be unduly influenced by perceived accentedness, even when intelligibility is not hindered. Such patterns reflect both inconsistency, stemming from interrater variability (a reliability issue), and bias, when raters' judgments of accentedness overshadow intelligibility or comprehensibility, ultimately threatening test validity in high-stakes contexts.

Human rating is also limited in terms of scalability and efficiency. Conducting pronunciation assessments with trained raters requires time-intensive scoring sessions, particularly when analytic rubrics are used. Rater fatigue can further compromise consistency over time, and the need for repeated calibration adds to administrative overhead (Isaacs & Thomson, 2013). These factors make large-scale implementation costly and logistically complex (Winke et al., 2013). As such, while human raters offer essential insight into L2 pronunciation, their potential bias and their use in high-volume assessment contexts presents practical challenges that continue to motivate the search for complementary or alternative approaches, including ASR-based scoring systems focused on intelligibility. Dictation-based ASR technology may come to the fore in intelligibility-oriented assessment, particularly because it can perform the traditional intelligibility task of orthographic transcription with a level of speed and scalability that is impractical for human raters. By automating transcription, ASR substantially reduces the time, cost, and logistical burden associated with intelligibility assessment. However, this efficiency advantage should not be interpreted as an absence of bias or inconsistency. ASR systems are shaped by training data, input conditions, and learner characteristics, and thus introduce distinct forms of bias and variability—an issue that the present study explicitly investigates.

ASR in the Evaluation of L2 Pronunciation: Assessment to Feedback

ASR technology is increasingly seen as a viable solution to persistent challenges in L2 pronunciation assessment, particularly in measuring intelligibility. In this context, it is also gaining traction as an assessment tool (Ngo et al., 2024). Language learning applications such as ELSA Speak, Duolingo, and Microsoft's Read Aloud utilize ASR to provide real-time, individualized feedback, enabling learners to compare their spoken output with transcriptions and thereby refine their pronunciation (Chapelle & Voss, 2016; Liakin et al., 2014). By drawing learners' attention to the discrepancies between their speech and the target forms, this type of feedback draws attention to problem areas of segmental accuracy for intelligibility (Guskaroska, 2020; McCrocklin, 2019). Because these tools are accessible and can be used autonomously, they are especially valuable in self-directed or out-of-class learning contexts. Beyond pedagogy, as previously mentioned, ASR systems offer potential for scalable, consistent, and efficient pronunciation assessment as it applies its dataset to fixed algorithmic criteria (Nickolai et al., 2024). This positions it as a useful tool for low-stakes diagnostic tasks and potentially as a component in larger-scale testing systems (Tejedor-García et al., 2021). Apple Siri, Google Voice Typing (GVT), and Microsoft Transcribe (MS-T) are readily available and allow learners to assess their speech without the need for formal testing environments.

ASR systems are particularly well suited to evaluating intelligibility, a central component of pronunciation proficiency (Derwing & Munro, 2009; Munro & Derwing, 1999). Successful transcription typically indicates that key phonemes were produced clearly enough to be understood. This makes transcription accuracy a practical, although indirect, measure of intelligibility. Moreover, ASR tools allow repeated, consistent feedback over time, which is harder to achieve with human raters due to time constraints and variability in scoring.

Nonetheless, important limitations remain. One concern is that many ASR systems do not effectively capture suprasegmental features such as stress, intonation, or rhythm, which are essential to comprehensibility (Derwing & Munro, 2015). Additionally, the predictive capabilities of some ASR tools may mask pronunciation errors by predicting likely words from context, rather than decoding based solely on phonetic input (Ashwell & Elam, 2017). This can lead to artificially high transcription accuracy and diminish the tool's diagnostic usefulness.

Proficiency levels can also play a further role in ASR performance. Lower-proficiency learners often produce more atypical and unintelligible segments, which may lead to higher word error rate in transcriptions (Knill et al., 2018). While this can reflect real intelligibility issues, it may also exaggerate perceived deficiencies due to the ASR system's limitations in handling non-native input. Misleading feedback may affect learners' confidence and progress and may also distort results when ASR data is used for tracking or placement decisions.

Despite these concerns, dictation ASR can prove to be a promising tool due to its consistency, speed, low cost, and accessibility. While proprietary systems such as ETS's SpeechRater and Pearson's Versant have been extensively studied, dictation ASR tools like Siri remain underexplored, despite being widely available by learners. It would be important to investigate how Siri's transcription output aligns with human judgments of intelligibility and whether proficiency level influences scoring reliability. These questions raise broader concerns about the validity and usefulness of ASR-based assessments, which are addressed in the following section.

Evaluating Siri's Usefulness in L2 Pronunciation Assessment

Previous research involving ASR systems Google Voice Typing and Microsoft Transcribe (e.g., Cox & Davies, 2012) has applied Bachman and Palmer's (1996) test usefulness

framework to systematically evaluate ASR-based pronunciation assessments. This framework outlines six interrelated qualities: reliability, construct validity, practicality, impact, interactiveness, and authenticity, which together define the overall usefulness of a language test. The present study applies this framework to examine Apple Siri's implementation as an automated scoring mechanism for L2 pronunciation, using secondary data drawn from an existing institutional placement test (see Method). Since the analysis is based on previously collected test responses, it is not possible to meaningfully assess the dimensions of authenticity (the degree to which test tasks reflect real-world language use), interactiveness (the extent to which test takers' characteristics engage with test tasks), or impact (the effects of the test on individuals and institutions), as doing so would pertain to the original test design rather than to Siri's function as an automated assessment tool. Therefore, the study focuses on the three dimensions most relevant to Siri's scoring role in this study: reliability, construct validity, and practicality, which are examined in detail below.

Reliability refers to the consistency and stability of test outcomes across different administrations and raters. In language assessment, this means producing scores that are replicable and unaffected by irrelevant factors (Bachman & Palmer, 1996). Siri's automated and fatigue-free scoring offers potential advantages in this regard by eliminating the inter- and intra-rater variability associated with human judgment (Saito et al., 2016). As such, Siri could enhance reliability by delivering consistent scores across test-takers and administrations. However, reliability alone does not guarantee the quality of measurement, and consistent scores may still misrepresent learners' actual abilities if the underlying measurement lacks validity.

Construct validity concerns the extent to which test scores accurately reflect the abilities they are intended to measure, without being influenced by irrelevant variables. In pronunciation

assessment, this typically involves evaluating a range of features, such as segmental accuracy, word stress, intonation, and rhythm, elements that are central to comprehensibility and oral proficiency (Loukina et al., 2015). Human raters often use analytic rubrics to assess these constructs explicitly. Siri, by contrast, can provide a single intelligibility-based score derived from word recognition accuracy, which limits construct representation. Research has shown that ASR systems can misinterpret mispronounced words correctly based on context (Ashwell & Elam, 2017), leading to false positives and inflated scores. Furthermore, ASR performance may vary across demographic groups, such as L1 background (Mehrabi et al., 2022), gender (Krishnan et al., 2024), and age (Ferland et al., 2019), introducing potential construct-irrelevant variance. Evidence also suggests that transcription accuracy may differ across proficiency levels (Inceoglu et al., 2023; McCrocklin & Edalatishams, 2020), possibly resulting in reduced sensitivity to subtle errors at higher levels. This potential ceiling effect among high-proficiency speakers raises further concerns about the system's ability to reflect fine-grained aspects of learner pronunciation.

Practicality refers to the logistical feasibility of an assessment in terms of time, cost, resources, and ease of implementation. Siri offers clear practical advantages: it is available on widely owned Apple devices, requires no specialized software, and generates immediate results. These features reduce both administrative overhead and financial costs, which is particularly beneficial for programs with limited resources. Siri's scoring system could support more frequent testing and reduce dependency on trained human raters. However, these gains must be balanced against potential compromises in score quality. If construct validity is undermined, especially through demographic bias or lack of sensitivity to prosodic features, then practicality alone cannot justify the use of Siri in high-stakes decision-making. Ultimately, the usefulness of

the test depends on balancing reliability, construct validity, and practicality in accordance with the testing context and goals. Although most existing research on ASR systems has focused on their pedagogical applications, its potential for use in automated pronunciation assessment remains underexplored. Considering Siri's accessibility and potential ability to differentiate among speakers at different proficiency levels, further empirical investigation is needed to determine how well Siri's outputs align with those of trained human raters and to what extent it can be used as a valid and useful assessment tool.

This study addresses this gap by comparing Siri-generated scores with human-rated pronunciation scores, with specific attention to intelligibility and the role of learner proficiency. Given the study's reliance on existing test data, the analysis is limited to three dimensions of Bachman and Palmer's (1996) framework: reliability, construct validity, and practicality, as the remaining criteria pertain to the original test design rather than the evaluation of Siri as an automated pronunciation assessment tool with the objective of determining its strengths and limitations.

The Current Study

Recent studies found a strong correlation between scores assessed by Google Voice Typing (GVT; Johnson et al., 2024) and Microsoft Typing (MS-T; Nelson and Cardoso, 2024), and those provided by human raters, demonstrating the reliability and validity of ASR system in evaluating L2 pronunciation based on a set of phonological criteria (e.g., phonemic accuracy, stress) and overall proficiency (e.g., comprehensibility). This raises the question: Can similar conclusions be drawn regarding Apple Siri? This study thus aims to build on these previous findings by examining whether Siri (readily found in the Apple ecosystem) has the potential to produce valid results and be used reliably for L2 pronunciation assessment, whether its scoring

aligns with ratings across different proficiency levels, and how its performance compares with that of GVT and MS-T.

This study called upon Bachman and Palmer's (1996) test usefulness framework by applying the concepts of reliability, construct validity, and practicality to determine whether Siri can perform on par with human scoring for pronunciation assessment. Demonstrating the usefulness of Siri for this purpose could pave the way for the adoption of more cost-effective and scalable assessment solutions. As such, this study aimed to answer the following research questions (RQs):

1. ASR vs. Humans: To what extent do Siri-generated pronunciation scores correspond with human-rated overall and subscores on an analytic rubric (RQ1)?
2. Proficiency Effects: Does the relationship between Siri-generated scores and human ratings vary according to participant proficiency levels (RQ2)?
3. Siri vs. other ASR systems: Does Siri's performance differ from that of GVT and MS-T when evaluated against the same human-rated scores (RQ3)?

Method

Context

This study presents a quantitative analysis of the feasibility of ASR dictation technology to assess pronunciation in a university English placement test. It is a conceptual replication of Johnson et al. (2024), which evaluated GVT, and Nelson and Cardoso (2024), who used the same data for MS-T. The data were originally collected and stored in a database by a Canadian university language department as part of an initiative to revise the pronunciation scoring rubric used in their institutional English placement test. These data were drawn from a computerized placement test, approximately 90 minutes in length, which assessed multiple language skills,

including listening, speaking, reading, writing, pronunciation, speech perception, syntax, vocabulary, and critical reading. Each skill was assessed independently based on skill-specific tasks and scored using skill-specific criteria. The present analysis focuses exclusively on the pronunciation component.

Participants

This study drew on an existing database of pronunciation task responses taken from over 10,000 placement tests administered between 2015 and early 2020 and used a subset of responses from 56 adult participants (N = 56; 21 males, 35 females; mean age = 28.09, SD = 7.78). All 56 participants were undergraduate students taking the placement test either to determine appropriate ESL course placement or to fulfill language requirements for their academic programs. According to their results from this placement test, their oral proficiency ranges from A1 to C2 on the Common European Framework of Reference for Languages (CEFR) scale (Council of Europe, 2001).

The 56 participants had diverse linguistic backgrounds, with French representing the predominant first language (n = 39; 68.4%), followed by Spanish (n = 4; 7.0%), Arabic (n = 3; 5.6%), and various other L1s (n = 5; 8.9%), encompassing L2 language groups overall. For this study, participants were categorized into two proficiency levels based on scores assigned by trained human raters using the institution's analytic rubric: the lower-proficiency group, which ranges from A1 to B1, scored below 72 (n = 20) and the higher-proficiency group, which goes from B2 to C2, scored 72 or higher (n = 36). To contextualize the scores, results between 72 and 86 qualified students for advanced pronunciation instruction, while 86 or above exempted students from further pronunciation training. For this study, the system of scoring was retained

for ecological validity as was also the case in the studies carried out by Johnson et al. (2024) and Nelson and Cardoso (2024).

Instruments

There were two main instruments for this study: the human rater's rubric to evaluate the participants' responses and the read aloud sentences used to elicit the responses. For the first tool, the rubric was composed of five key phonological dimensions: 1. *comprehensibility*, 2. *segmental accuracy* (individual phonemes), 3. *connected speech features*, 4. *word stress and rhythm*, and 5. suprasegmental elements such as *thought groups, prominence, and intonation*. Each dimension was scored on a scale from 1 (poor) to 5 (excellent), and sub-scores were totaled to give a raw score out of 25. Each criterion was scored independently to provide a comprehensive assessment of the speaker's pronunciation abilities. The raw score was then multiplied by 4 to yield a final score out of 100.

The second tool was a task with five read-aloud sentences drawn from a large bank of curated prompts designed to elicit the participants' responses. These sentences were developed by a team of L2 pronunciation experts overseeing the pronunciation curriculum and were designed to gradually increase in phonological and lexical complexity to differentiate students' L2 pronunciation abilities. The first sentence, known as the baseline sentence, was identical for all test-takers across all testing sessions. This standardization served as a quality control measure to help detect inconsistencies or potential cheating among repeat test-takers.

The five test sentences followed a structured progression: Sentence one featured monosyllabic, high-frequency words, universal consonants and vowels, and simple thought groups and linking. Sentence two introduced two- to three-syllable words, compound nouns, consonant clusters, and marked (universally infrequent) consonants. Sentence three focused on

intonation patterns and grammatical endings. Sentence four was written with increased difficulty with multisyllabic and lower-frequency words, while sentence five targeted infrequent academic vocabulary and words with idiosyncratic pronunciations. Due to confidentiality agreements, the exact content of the baseline sentence cannot be shared. Table 2 (adapted from Johnson et al., 2024 and Nelson & Cardoso, 2024) provides representative examples of the types of sentences read by participants during the pronunciation task.

Table 2

Sample Sentences

Level	Sentence
1	[Baseline 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 acetone and other toxic amalgams, either in the gaskets or in the valve assembly.*

Procedure

The data used in this study were originally collected as part of an initiative by a modern language department at a French-speaking Canadian university to revise the scoring rubrics used in their institutional ESL placement test. During the original assessment, participants first read aloud two practice sentences to ensure they understood the procedure and that the recording system was functioning correctly. This was followed by a task in which participants were

presented with five test sentences to read aloud. Each sentence appeared sequentially, with a 20-second time limit for participants to read and record their reading before the next sentence appeared.

The pronunciation responses were assessed using two methods: human scoring (which had already been carried out as part of the project for which the recordings were originally collected) and automatic scoring using Siri. With respect to human scoring, teams of three experienced ESL instructors, aged between 30 and 50 with extensive classroom and pronunciation evaluation experience reviewed each participant's set of responses and reached a consensus on all participants' responses. In instances where a pronunciation score appeared to be a statistical outlier relative to the participant's performance on other sections of the placement test, a dedicated quality control team conducted a secondary review to confirm the integrity of the rating and ensure validity in placement outcomes. This team evaluated the candidates' performance using the previously mentioned analytic rubric. To ensure interrater reliability and mitigate subjective bias, raters initially worked independently and then shared their scores with each other. In cases where discrepancies arose for a given criterion, consensus was reached through discussion. However, it is important to note that the interrater reliability (IRR) was not calculated by the university in question, and this study only had access to the scores that were reached by consensus by the human raters.

For the automated scoring, the same set of 56 pronunciation recordings was input into MacOS (15.5 Sequoia) making use of the offline version of Siri, which is the same that is available in iPhones and iPads. Word recognition accuracy was determined by calculating the percentage of words correctly identified in the transcriptions generated by Siri. Following the procedure outlined in Cox and Davies (2012), the number of correctly identified words were

divided by the total number of words across the five sentences, then multiplied by 100 to generate a percentage score for each participant. These results were hand scored by the first author. This approach allowed us to examine whether Siri-generated total pronunciation scores (intelligibility) corresponded with the human-rated overall scores, addressing the first part of RQ1. For the second part of RQ1, Siri's total scores were compared with the human-assigned subscores to assess the degree of correspondence between the two scoring methods. To address RQ2, correlations between human rater and Siri scores were examined across the two target proficiency levels. For RQ3, the correlation coefficients obtained for Siri were compared with those previously reported for GVT and MS-T in relation to human ratings, to evaluate how Siri's performance aligns with that of the other ASR systems.

To ensure that the Siri results were evaluated on the same footing as GVT (Johnson et al., 2024) and MS-T (Nelson & Cardoso, 2024), the following criteria was applied 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*).

Results

The analytical process involved entering Siri-generated scores, final human evaluations, and individual criterion sub-scores into SPSS (version 29) to investigate the relationship between Siri automated and human pronunciation assessments. Initial scatterplot examination was carried out to examine linear and monotonic relationships between variables, which revealed no significant outliers. However, histogram analysis indicated a non-normal data distribution, which the Shapiro-Wilkes test subsequently confirmed by showing significant deviation from normality across all variables ($p < .05$). As this was the case, a nonparametric Spearman's rank-order correlation was used to determine relationship strength and direction. To establish robust confidence intervals for both overall and sub score correlations, the analysis incorporated bootstrap resampling techniques, generating correlation coefficients with 95% bias-corrected and accelerated (BCa) confidence intervals based on 1,000 bootstrap samples. The MS-T and GVT results were taken from Nelson and Cardoso (2024) and Johnson et al., (2024) respectively.

Table 3

Descriptive Statistics

Variable	M	SD
Siri score (/100)	85.03	16.54
MS-T score (/100)	86.88	16.39
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: Siri score (/100)	68.32	17.85
Lower proficiency: Human-rated score (/100)	40.20	16.80
Higher proficiency: Siri score (/100)	94.34	5.14
Higher proficiency: Human-rated score (/100)	89.67	9.44

As shown in Table 3, Siri scores were consistently higher than human-rated scores across both proficiency groups. For the lower-proficiency group, the mean Siri score was 68.32 (*SD* = 17.85), while the human-rated mean was 40.20 (*SD* = 16.80). In the higher-proficiency group, Siri scores averaged 94.34 (*SD* = 5.14) compared to 89.67 (*SD* = 9.44) for human ratings.

Regarding RQ1 concerning the relationship between Siri and final human-rated scores, the results indicated a statistically significant strong correlation between the two variables, $r_s(54) = .789$, $p < .001$. For the second part of RQ1 pertaining to the relationships between the Siri scores and the human rater rubric criteria, the results indicated statistically significant strong correlations between the Siri score and each of the subscores (Table 4 presents a summary of the correlations). As noted above, these correlations are based on the overall Siri score, as Siri does not produce subscores for each criterion on the human rating scale.

With respect to the RQ2 addressing the relationship between Siri scores and test-taker proficiency, a significant strong correlation was found between the Siri and human-rated scores

for lower-proficiency test takers, $r_s(54) = .83$, $p < .001$. However, a non-significant medium correlation was found between the Siri scores and the human-rated scores for higher-proficiency test takers, $r_s(54) = .41$, $p = .100$ (see Table 5 for a summary of the correlations).

Table 4

Correlations Between Siri and Human-Rated Scores by Rubric Criteria

Rubric Criteria	rho	95% BCa Cis
Final score	.79**	.66, .87
Comprehensibility	.80**	.68, .88
Phonemes	.81**	.70, .89
Connected speech	.70**	.54, .82
Word stress and rhythm	.71**	.54, .82
Thought groups, sentence stress, and intonation	.81**	.68, .88

Note. Confidence intervals based on 1000 bootstrap samples.

**$p < .001$.

Table 5

Correlations Between Siri Score and Human-Rated Scores by Proficiency Level

Rubric Criteria	rho	95% BCa Cis
Lower-level proficiency	.83**	.61, .93
Higher-level proficiency	.41	.09, .66

Note. Confidence intervals based on 1000 bootstrap samples.

**$p < .001$.

Table 6

Spearman Correlations Between Human Rating Criteria and ASR Scores

Rubric Criteria	Siri	MS-T	GVT
Comprehensibility	.80**	.83**	.85**
Phonemes	.81**	.76**	.78**
Connected speech	.70**	.78**	.72**
Word stress and rhythm	.71**	.73**	.71**
Thought groups, sentence stress, and intonation	.81**	.76**	.79**
Total	.79**	.79**	.78**

Note. Confidence intervals based on 1000 bootstrap samples.

**$p < .001$.

To address the RQ3, whether Siri's performance differs from that of Google Voice Typing (GVT) and Microsoft Transcribe (MS-T) when evaluated against the same human-rated scores, comparative correlations were examined across five different subscores. Overall, Siri performed on par with the other ASR systems, with a correlation of $r_s = .79$, identical to MS-T and slightly higher than GVT ($r_s = .78$; Johnson et al., 2024; Nelson & Cardoso, 2024;). Siri showed the strongest alignment with human ratings for the criterion *phonemes* ($r_s = .81$), outperforming both MS-T and GVT in this area. It also showed with the highest correlation for *thought groups, sentence stress, and intonation* ($r_s = .81$). However, Siri's performance was relatively weaker for *connected speech* ($r_s = .70$), where it lagged MS-T ($r_s = .78$) and GVT ($r_s = .72$). For *comprehensibility*, Siri scored slightly lower ($r_s = .80$) than GVT ($r_s = .85$) and MS-T ($r_s = .83$), though the difference was small. It is interesting to point out that all three ASR systems showed lower correlations with human ratings for *word stress and rhythm*, and the same pattern was also observed for Siri and GV-T in the case of *connected speech*. Across all categories, all three systems exhibited consistently significant correlations with human ratings ($p < .001$), suggesting

that while Siri performs as well as the other ASR systems overall, there are nuanced strengths and limitations in how each system captures specific aspects of L2 pronunciation.

The findings demonstrate a robust correlation between automated Siri-scored assessments and human evaluations of second language pronunciation quality, with this relationship remaining consistent across various human subscores. Notably, the correlation is stronger among lower-proficiency learners, whereas higher-proficiency speakers exhibited greater score variability between the two assessment approaches, suggesting reduced alignment at advanced levels. Finally, Siri holds its own when compared to both GVT and MS-T, demonstrating comparable correlations with human ratings and reinforcing its potential as a viable tool for automated pronunciation assessment.

Discussion

This study set out to examine the extent to which Siri-generated pronunciation scores align with human ratings, and how this relationship varies across evaluation criteria, participant proficiency levels, and in comparison, to other ASR systems. The results indicate statistically significant strong correlations between Siri-generated scores and human ratings (r_s = .79) for both the overall score and each sub score. When examined by proficiency level, scores from lower-proficiency participants showed statistically significant strong correlations between Siri and human ratings (r_s = .71–.81), whereas scores from higher-proficiency participants displayed a non-significant weak correlation, suggesting reduced alignment at more advanced levels. Correlation strength was classified according to the field-specific benchmarks proposed by Plonsky and Oswald (2014), where coefficients of .10 to .34 are considered small, .35 to .64 moderate, and .65 or higher large. Finally, Siri's performance was found to be comparable to that of both GVT and MS-T, with similarly strong correlations across most criteria.

These findings align with previous research demonstrating that ASR-generated scores with strong correlations with human evaluations (e.g., Graham et al., 2008; Johnson et al., 2024; Nelson & Cardoso, 2024; Nickolai, 2024). However, the correlation observed among higher-proficiency speakers contrasts with findings by McCrocklin and Edalatishams (2020), who reported high recognition accuracy for more advanced L2 speakers using GVT, although their study did not directly compare ASR output to human ratings. One possible explanation for this discrepancy might be that the quality of the audio recordings of more advanced participants negatively affected Siri's recognition, as they tended to speak faster and, at times, with lower volume or vocal intensity. Although human evaluators could readily comprehend their speech, Siri likely struggled with accurate transcription. Comparable audio difficulties emerged among certain lower-proficiency participants; however, because their speech presented challenges for human evaluators as well, both Siri and human assessments yielded consistently low scores. Another reason maybe be due to a ceiling effect (Taylor, 2010) which occurs when scores cluster near the top of a measurement scale, limiting the ability to detect meaningful differences among high-performing individuals. In this study, higher-proficiency learners may have received uniformly high scores from Siri, reducing variability and weakening the correlation with human ratings. This pattern is evident in the descriptive statistics: the mean Siri score was higher (M = 94.34) than the human-rated mean (M = 89.67), while the standard deviation was lower for Siri (SD = 5.14) than for human raters (SD = 9.44), suggesting restricted score dispersion. Despite this, the correlation between Siri and human ratings remained moderately positive (r_s = .41), reinforcing the interpretation that Siri can serve as a viable tool for L2 pronunciation assessment. Taken together, these findings highlight the importance of context when interpreting automated

scores and point to the need for further research to evaluate Siri's effectiveness across different proficiency levels.

After confirming that Siri can effectively evaluate L2 pronunciation, it is important to return to Bachman and Palmer's (1996) test usefulness framework to assess the quality and appropriateness of Siri's scoring as an automated evaluation tool, focusing on the dimensions most relevant to this context: reliability, construct validity, and practicality. Since this study draws on previously collected test responses, it is not possible to assess authenticity, interactiveness, or impact, as these constructs are not applicable to the current study, as discussed earlier.

Reliability

The strong correlation between Siri-generated scores and human ratings demonstrates the potential reliability of automated pronunciation assessment. Similar findings have been reported in research on other ASR systems. For instance, Bernstein et al. (2010) reported that automated scores from the Versant test, which employs more advanced technology than Siri, outperformed human ratings in terms of reliability across multiple datasets. Automated systems minimize variability linked to human factors such as accent familiarity (Browne & Fulcher, 2016), experience with L2 speech (Saito et al., 2016), personal bias (Yan & Ginther, 2017), and rater fatigue (Ling et al., 2014).

However, a consistent pattern observed in this study, which was echoed in previous research by Johnson et al. (2024), is that Siri produced weaker correlations with human ratings for higher-proficiency learners. Indeed, in this study, the correlation between Siri scores and human-rated scores was considerably stronger for lower-proficiency participants (r_s = .83, 95% CI [.61, .93]) than for higher-proficiency participants (r_s = .41, 95% CI [.09, .65]), slightly

outperforming GVT (r_s = .28, 95% CI [-.03, .55]). A key factor that may account for this discrepancy is that these higher-proficiency participants spoke more quietly and rapidly than others, which may have affected the audio input and, in turn, Siri's recognition accuracy. Although their speech remained highly intelligible to human raters, Siri may have struggled to accurately transcribe these samples. Similar audio challenges were present among some lower-level participants, but because their speech was also less comprehensible to human raters, both sets of scores remained comparably low. This study points to an underlying explanation: ASR systems may be vulnerable to the quality of the audio responses, which can undermine the reliability of the scores by introducing inconsistencies unrelated to the construct being measured.

Construct Validity

At the outset, there were concerns that using Siri to score pronunciation assessments might compromise construct validity. Human raters used a detailed rubric encompassing multiple pronunciation constructs, including comprehensibility, segmental accuracy (phonemes), and prosodic features. Siri, by contrast, provides a single intelligibility-based score (i.e., transcriptions), raising initial doubts about whether it could adequately reflect the full range of phonological abilities being evaluated. Surprisingly, Siri scores showed strong correlations with each individual criterion in the human rating rubric. This included robust relationships not only with segmental measures but also with prosodic features. Notably, the criterion *thought groups, sentence stress, and intonation* exhibited the second-highest correlation with Siri scores (rs = .81), a result that may appear unexpected given the prosodic nature of this criterion.

Importantly, neither human raters nor ASR systems were explicitly evaluating prosody in the transcription task. Rather, prosodic features may have influenced intelligibility indirectly. Previous research has shown that stress placement and rhythmic patterning facilitate lexical

segmentation and listener comprehension (Field, 2005; Hahn, 2004). While ASR systems such as Siri do not model prosodic features directly, stress and rhythm can affect transcription accuracy by shaping the acoustic realization of speech segments, including vowel reduction, segmental duration, and boundary cues. In this way, prosodic organization may contribute to both human and ASR intelligibility outcomes without being explicitly assessed. However, it is important to acknowledge that Siri's scoring is generated through a different process than human perception, and there is currently no direct evidence that it explicitly or systematically accounts for prosodic features beyond what can be inferred from correlation with human ratings.

Additionally, a potential ceiling effect introduced by high-proficiency speech suggests that these systems may lack the sensitivity needed to capture subtle pronunciation differences at advanced levels, which further raises concerns about the construct validity of the scores. From a test usefulness perspective, the findings suggest that Siri can be a suitable scoring mechanism for L2 pronunciation assessment. Although its approach is based on intelligibility rather than explicit evaluation of individual phonological constructs, its scores showed strong correlations with multiple features rated by humans, including segmental and prosodic aspects. These results indicate that while certain aspects of construct validity require careful consideration, Siri's intelligibility-focused scoring can still provide a meaningful reflection of overall pronunciation ability.

Practicality

Practicality, as defined by Bachman and Palmer's (1996) test usefulness framework, is one of the strongest advantages of using Siri for pronunciation scoring in placement tests. Although there are upfront costs for acquiring compatible equipment (such as a MacBook, iMac, or iPad) and for developing, integrating, and maintaining scoring applications, these expenses are

offset by significant long-term savings. Automating scoring not only lowers these costs but also frees instructors from time-intensive rating, allowing them to focus on teaching and supporting students. It further streamlines administration by delivering scores to students and program coordinators much faster, expediting placement and registration. Faster score delivery can also improve student satisfaction, as Isaacs (2018) noted with test taker feedback on Pearson's Versant platform, where quick results were highly valued. Together, these efficiencies allow language programs to save time, reduce costs, and reallocate resources more effectively, thereby enhancing the test's practicality and contributing to its overall usefulness in educational contexts.

Conclusion

Study objectives and findings

This research investigated whether Apple's Siri could effectively assess L2 pronunciation performance, as operationalized through intelligibility-oriented criteria, by comparing its outputs to human expert ratings, with the potential for automating pronunciation scoring in university language placement tests. The analysis revealed strong correlations between Siri-generated scores and human assessments, demonstrating that automated scoring maintains both reliability and validity. These results suggest that Siri could serve as an effective tool for language programs seeking to improve test efficiency while reducing costs through automation, without compromising assessment quality. While there are upfront expenses associated with acquiring compatible devices and developing integrated scoring tools, these can be outweighed by the long-term savings in efficiency and scalability, especially in programs assessing large numbers of students.

Study limitations

Several constraints affect the generalizability of these findings. The study involved only 56 participants completing a single read-aloud task, limiting the scope for broader application across diverse populations and assessment contexts. The audio quality of some recordings presented additional challenges, as some recordings contained background noise or reduced clarity, potentially affecting the accuracy of automated scoring. Another limitation of this study is that IRR was not calculated by the administering university, and only the final consensus scores assigned by human raters were available for analysis. Language institutions adopting Siri could implement it by using automated quality checks to flag questionable recordings for human review, thereby maintaining scoring reliability across varied audio conditions.

Future research directions

While this study demonstrates Siri's potential for low-stakes pronunciation assessment of intelligibility, several research areas warrant investigation. Previous research by Knill et al. (2018) found that automatic speech recognition systems performed differently across task types, with varying correlations between shadowing and read-aloud activities. Similar investigations should examine whether dictation-based technologies like Siri show comparable task-dependent performance patterns in tasks such as guided picture descriptions or spontaneous conversations.

Bias detection represents another critical research priority. Systematic analysis comparing Siri's performance across different first language backgrounds, gender groups, and age ranges is essential to ensure equitable assessment practices. Such research would identify potential systematic biases and inform the development of correction procedures or alternative approaches for affected populations. This line of inquiry has already been initiated in the second manuscript of this dissertation (in press in the *Canadian Journal of Applied Linguistics*).

Practical implications

The strong correlations demonstrated in this study between automated and human scoring provide evidence that institutions can integrate Siri-based assessment tools without compromising scoring standards. Implementing such technology has the potential to streamline placement procedures, reduce administrative burden, and enhance student experience through faster, more consistent assessment processes. Students can benefit from immediate feedback and standardized evaluation criteria, potentially increasing confidence in test outcomes and trust in institutional assessment practices.

Chapter 3: Automatic Speech Recognition for Second Language Pronunciation Assessment: Focus on Age-Related Bias

The use of technology is transforming language learning by introducing digital tools, apps, and online platforms that offer learners greater access to resources and opportunities for interactive practice. These innovations have redefined traditional pedagogical methods, making language learning more accessible, engaging, and adaptable to individual needs (Blake, 2016). This shift highlights the growing need to deepen our understanding of the theoretical foundations, practical implications, and challenges associated with these technological advances. The rapid development of artificial intelligence (AI) and machine learning has further revolutionized language education, particularly in the automated assessment of speaking and pronunciation (Kang & Johnson, 2018). This is beginning to reshape the landscape of language evaluation, opening new pathways for enhancing learning experiences and feedback mechanisms.

Automated Speech Recognition (ASR) technology has emerged as a promising tool in this transformation. Its integration into second language (L2) assessment represents a significant leap forward, offering opportunities for educators and learners alike to benefit from timely, consistent, and flexible feedback (Chapelle & Voss, 2016). However, the adoption of ASR has also raised critical questions about the reliability, validity, and equity of such systems in diverse educational contexts (Chapelle & Lee, 2021). The implementation of ASR technology in L2 learning and assessment carries tremendous potential but also poses significant challenges. Nonetheless, educational institutions striving to provide effective and equitable assessment tools may find that ASR systems offer a promising solution (Eskenazi, 1999). While ASR systems promise consistent and efficient evaluations, their successful integration into educational

contexts requires careful consideration of several factors. Among these, the issue of bias has emerged as a pressing concern. Studies have demonstrated that ASR systems can exhibit biases against certain user groups based on age (Bajorek, 2019; Vipperla et al., 2010), gender (Bajorek, 2019), or linguistic background (Koenecke et al., 2020). These biases can threaten the fairness and inclusivity of assessments, raising questions about whether ASR-based technologies equitably serve diverse learner populations. This is particularly concerning in contexts where language assessments play a critical role in shaping learners' academic and professional opportunities such as high stakes testing. While the technology can alleviate the logistical and subjective challenges associated with human raters, there is a need to examine these systems for their limitations with the objective of understanding how ASR systems perform when assessing users from different demographic groups. For instance, while some studies highlight the reliability of ASR in evaluating pronunciation performance, others reveal discrepancies in its accuracy (Chun et al., 2016; Liu et al., 2022).

This paper aims to explore the use of ASR technology in L2 pronunciation assessment, focusing on the potential manifestation of age-related bias in ASR systems. By critically examining these aspects, this study seeks to contribute to the discourse on how ASR technology can be effectively and equitably implemented in language education. The ultimate goal is to advance our understanding of ASR's potential to enhance L2 learning and pronunciation assessment while acknowledging and addressing its current limitations (Evanini & Wang, 2013).

Literature review

Research on ASR in L2 pronunciation assessment has highlighted its potential for providing objective, flexible, and immediate feedback. However, concerns about bias in ASR systems require deeper exploration. Given that speech characteristics naturally change with age

and that younger learners tend to achieve greater phonological accuracy in L2 acquisition (Lee, 2015), ASR may exhibit biases that affect its reliability across different age groups. Without accounting for age-related variation in speech input, there is a risk of disproportionate scoring discrepancies. This literature review examines the intersection of ASR technology, L2 pronunciation assessment, and age-related variability in speech, exploring how these factors may contribute to disparities in ASR-generated scores (Feng et al., 2021).

ASR in second language learning and assessment

ASR technology has impacted educational settings and L2 acquisition (Ngo et al., 2024) by employing advanced operations that rely on sophisticated algorithms and large datasets to convert spoken language into text, analyzing speech patterns and linguistic features (Levis & Suvorov, 2012). Specifically, ASR processes speech input by identifying phonemes and using extensive lexical databases and linguistic models to generate accurate textual output. This probabilistic approach to phonological matching has been refined through advances in deep learning and neural networks, making ASR potentially more reliable for language learning and assessment (Li, 2022).

The growing sophistication of ASR offers new opportunities for L2 teaching and evaluation by providing automated, objective assessments of spoken language proficiency (Inceoglu et al., 2023). Modern ASR tools such as Google Voice Typing (GVT), Microsoft Transcribe (MS-T), and Siri, can be used with confidence for pronunciation assessment (Johnson et al., 2024; Nelson & Cardoso, 2024). MS-T, for instance, demonstrates progress in this field, supporting ASR capabilities in over 125 language varieties, including Canadian English. MS-T's features include an extensive global vocabulary, real-time speech streaming, and adaptive speech processing that can recognize specialized and uncommon terminology (Urban, 2024). These

capabilities can allow learners to receive immediate, corrective feedback on pronunciation, potentially enabling more effective self-assessment and improvement.

ASR technology has the potential to improve language learning by making sophisticated assessment tools more widely available and accessible to educational institutions and individuals. This potential is supported by a meta-analysis by Ngo (2024) highlighting the significant role of ASR-based tools in pronunciation acquisition and self-assessment. By providing immediate textual feedback, these systems help learners identify discrepancies between their intended speech and the ASR's transcription, fostering greater pronunciation awareness (Evers & Chen, 2021; Inceoglu et al., 2023; Mroz, 2020). While some studies report learner frustration with transcription errors (Liakin et al., 2017), overall perceptions of ASR as a learning tool remain positive, with students recognizing its value in their language development (Dillon & Wells, 2021; Liakin et al., 2017; McCrocklin & Edalatishams, 2020).

Beyond pronunciation training, ASR facilitates learner autonomy and self-directed learning, as students can practice independently and receive instant feedback (Pérez Castillejo, 2021). This shift towards autonomous learning has profound implications for language education, particularly in contexts where traditional instructor-led instruction is limited or unavailable. Although ASR technology is not new, its application in L2 education has expanded rapidly, leveraging advanced algorithms to analyze phonemes, consult lexical databases, and apply linguistic models for more precise speech recognition (Filippidou & Moussiades, 2020; Levis & Suvorov, 2012).

ASR and human raters in assessing pronunciation

The comparison between ASR systems and human evaluators in scoring L2 pronunciation has gained considerable attention (Cámara-Arenas et al., 2023). While ASR offers

a consistent and objective approach to pronunciation assessment, human raters provide nuanced evaluations that account for context and other features that ASR cannot assess, such as fluency, and intonation. Both methods have strengths and limitations, and ongoing research explores how they can complement each other to deliver accurate and effective feedback (Bernstein et al., 2010).

Some ASR systems provide immediate, objective feedback by analyzing speech at the phonetic level and comparing it to extensive datasets of pronunciation samples (Tejedor-García et al., 2021; Evanini & Wang, 2013). These systems break speech into phonetic components, offering real-time feedback (Georgescu et al., 2021). For example, applications such as ELSA Speak can identify mispronunciations as they occur and deliver instant corrective suggestions, allowing learners to adjust and improve their pronunciation on the spot. This instant feedback loop benefits learners, enabling independent practice outside traditional classroom settings. Moreover, ASR reduces subjectivity in assessment, eliminating biases linked to raters' familiarity with certain accents (Kang & Rubin, 2009). Human assessments are not always consistent due to biases, fatigue, and individual differences in perception (Babaeian, 2023), with humans sometimes unconsciously rating familiar accents more leniently than unfamiliar ones, leading to inconsistent evaluations (Derwing & Munro, 2009). In addition, factors such as accent familiarity, speech rate, and linguistic diversity can also impact human raters' evaluations (Stolcke & Droppo, 2017). In contrast, ASR systems remain impartial, relying on vast datasets for evaluation. They are especially valuable in large language learning programs, where individualized feedback is difficult to provide due to time and resource constraints (McCrocklin, 2022). For example, while teachers may struggle to give detailed and timely feedback to every

student, ASR can assess thousands simultaneously, ensuring consistent pronunciation evaluation (Saito et al., 2016).

ASR technology also presents a cost-effective alternative in language programs, reducing the need for multiple human raters (Saito et al., 2016). Its large-scale automation enables integration into online learning platforms and language apps like Duolingo and Rosetta Stone, making pronunciation feedback accessible to a global audience (Nickolai, 2024). This expansion can foster more equity in language learning as it is cost-effective, particularly for learners in underfunded or underserved areas who may not have access to instructors or fluent speakers for feedback.

Despite its advantages, ASR is not without limitations. Background noise, rare dialects, and certain phonetic variations can challenge system accuracy (Hollands et al., 2022; Inceoglu et al., 2023). Also, unlike human raters, ASR struggles with suprasegmental features such as tone, intonation, and stress patterns, which are critical for natural speech and communicative competence (Kochem et al., 2022). For example, in English, an ASR application might fail to assess the stress patterns in words like "photography" versus "photograph", leading to incorrect feedback on the speaker's prosody. Finally, ASR training data lack sufficient representation of less commonly spoken languages and regional dialects, which can result in inaccurate feedback that does not truly reflect learners' pronunciation or communicative competence.

Comparative studies show that automated ASR scores used as intelligibility measures often align closely with human ratings of pronunciation performance, particularly in distinguishing more and less precise segmental and suprasegmental realizations (Johnson et al., 2024; Nelson and Cardoso, 2024). In this context, phonetic precision refers to the degree to which speech sounds are produced with sufficient acoustic clarity and distinctiveness to support

reliable phonological categorization by a listener. Such precision is known to contribute to intelligibility, as deviations in segmental realization, stress placement, or temporal organization can reduce the likelihood that an utterance is accurately understood (Arora et al., 2018). While ASR systems do not assess phonetic precision directly, greater phonetic precision can increase transcription accuracy by improving the acoustic evidence available to the system. Importantly, ASR transcription outcomes reflect the combined influence of phonetic precision and other factors, including language model constraints and recording conditions, rather than phonetic precision alone. However, while ASR models can incorporate diverse accents and dialects to reduce bias, they may still lack the flexibility of human raters in adapting to individual speech variations. Consequently, ASR provides a strong, objective foundation for phonetic assessment but may introduce biases that impact fairness and inclusivity. This limitation can lead to lower scores for speakers from marginalized linguistic backgrounds, even when their speech is intelligible. Identifying and addressing these biases will improve ASR's accuracy, equity, and reliability, particularly in language assessment contexts where fairness and precision are critical.

Bias in automatic speech recognition

As previously noted, ASR systems can impact language assessment by offering unprecedented efficiency and flexibility in evaluating speaking tasks. However the complexity of human speech - influenced by factors such as individual vocal traits, gender, and age - presents significant challenges for these systems (Fuckner et al., 2023). In fact, the integration of ASR in automatic scoring mechanisms has raised significant concerns about potential biases that could unfairly disadvantage certain groups of test-takers (Feng et al., 2021). ASR systems are trained on extensive datasets, and if these corpora over- or under-represent certain groups—such as speakers of specific first languages (L1s), particular age demographics, or gender—biases can

arise that skew the system's scoring accuracy (Madnani et al., 2017). This can lead to ASR systems scoring more accurately for well-represented groups while potentially penalizing or misinterpreting those from underrepresented backgrounds.

Demographic and regional bias in automatic speech recognition

Bias in ASR systems can be persistent and complex, as it can reappear over time even with initial efforts to create balanced data and algorithms. Changes in the demographics of test-takers, for example, can introduce new biases as the system encounters speech patterns it was not originally designed to handle (Mehrabi et al., 2022). This means that addressing bias is an ongoing and iterative process, requiring constant monitoring and updates to ensure fairness. For smaller organizations, this poses a major challenge, as they often depend on external ASR providers and have limited influence over training data or control over update schedules. As a result, smaller organizations may need to implement regular bias checks on their own, which can increase costs and require additional resources to prevent unfair outcomes in scoring (Feng et al., 2021).

The biases inherent in ASR systems stem from various sources, each contributing to potential inaccuracies in speech recognition and, consequently, in scoring (Evanini, 2019). ASR can perform better with mainstream or widely recognized accents, often underrepresenting regional or indigenous speech patterns (Hinsvark et al., 2021). This limitation can lead to lower scores for speakers from marginalized linguistic backgrounds, even when their speech is comprehensible. This bias can lead to significantly lower recognition accuracy for language learners or those with accents that diverge from the majority. Koeneke et al. (2020) found, for instance, that ASR systems from major companies showed substantial disparities for races, with error rates for white speakers nearly half those for African American speakers.

Gender bias in automatic speech recognition

Gender-based differences in speech patterns and acoustic properties can also lead to biased outcomes in ASR systems. They perform differently for male and female voices, often with higher error rates for female speakers (Krishnan et al., 2024). These ASR performance differences are not just due to basic acoustic features like pitch or volume; they also reflect deeper differences in how men and women typically use intonation and articulation when speaking. Beyond imbalances in training data, this bias may also arise from variations in speech characteristics between male and female voices (Kathiresan, 2021), including differences in fundamental frequencies and spectral features, which can negatively impact speech recognition accuracy.

Age bias in automatic speech recognition

Research has consistently shown that speech characteristics evolve with age, leading to higher error rates in speech recognition for both children and elderly speakers (Ferland et al., 2019; Sobti et al., 2024; Dutta et al., 2022; Gao et al., 2024). Physiological changes, such as reduced lung capacity, decreased vocal cord elasticity, and shifts in muscle coordination, contribute to differences in voice quality, articulation, and speech fluency (Linville & Rens, 2001). Older speakers may experience increased hoarseness, breathiness, or slower speech rates, all of which can influence pronunciation and intelligibility. These natural variations in speech patterns can present challenges for ASR systems, which rely on standardized models trained primarily on younger adult voices.

The social norms surrounding age and language often lead to differences in communication style. For instance, an older speaker might say, “Would you mind helping me with this, please?” while a younger speaker might simply say, “Can you help me?” This

difference reflects generational norms about politeness, deference, and conversational structure (Coulmas, 2005); however, ASR systems, optimized for speech of younger users, may fail to accommodate these variations, potentially misinterpreting the nuances of older speakers. In a related vein, Badwan (2021) discusses how linguistic capital (shaped by social positioning, including age) affects how individuals express themselves. These sociolinguistic norms can influence both what is said and how it is pronounced, with implications for how ASR systems process speech from different age group.

The consequences of age bias in ASR systems can be wide-ranging and impactful, affecting both personal and professional aspects of people's lives. Age bias in ASR systems not only affects pronunciation assessment but also has broader implications for AI-based applications that rely on accurate speech recognition, such as voice assistants and automated transcription services. For example, flawed ASR outputs can affect AI-driven language learning platforms, accessibility tools, or virtual assistants, reducing their effectiveness for older adults and other underrepresented groups (Kulkarni et al., 2024). This cascading effect reinforces systemic bias at multiple levels, amplifying the disparities introduced by the initial ASR errors.

In professional settings, age bias in ASR can have serious implications for productivity and equity. In occupations or scenarios where ASR can be used, such as in business meetings, presentations, or legal proceedings, higher error rates for older speakers can lead to inaccuracies in communication and documentation (Vipperla et al., 2008). As automated transcription tools are increasingly adopted to improve efficiency, older professionals may find themselves at a disadvantage, facing higher transcription error rates that younger colleagues are less likely to experience, which may perpetuate age-based disparities in the workplace (Koenecke et al., 2020).

The Current Study

Age bias can be due in part to the large volumes of sound files used to train ASR systems, which may have integrated younger voices, largely because younger populations are more likely to adopt new technologies and contribute more frequently to digital corpora. Consequently, the acoustic models in ASR systems are typically better suited to speech patterns used by younger speakers, while older voices, characterized by distinct vocal qualities, often suffer from higher rates of recognition error (Werner et al., 2019).

For instance, the stiffening of vocal chords, loss of muscle tone in the larynx, and changes in lung capacity can alter an individual's pitch, resonance, and articulation, creating a voice that is acoustically different from that of a younger speaker (Linville & Rens, 2001). These physiological differences further complicate ASR's ability to accurately recognize speech from older adults, as the models may not be designed to interpret these shifts in acoustic patterns. ASR systems may produce inconsistent results across age groups due to two related factors: first, a lack of diverse age representation in the training data; and second, age-related physiological changes in speech production, such as slower articulation or reduced pitch range in older speakers. Additionally, age-related sociolinguistic norms, such as differences in politeness strategies, speech rate, or lexical choice, may further influence how speech is produced and interpreted by the system

These age-related differences pose significant implications for ASR applications in language learning and assessment. If ASR systems do not account for the variability introduced by age, they risk introducing bias in pronunciation evaluation, particularly in educational and testing contexts where fairness and accuracy are critical. Given these potential consequences, addressing these biases is crucial to ensuring that ASR and its dependent applications provide

equitable and accurate assessments for diverse user populations, especially for widely accessible applications such as GVT, MS-T, and Siri. Therefore, this study was guided by the following research question:

1. Does age moderate the relationship between human-rated pronunciation scores and ASR-generated scores by three ASR systems (GVT, MS-T, and Siri)?

Method

This quantitative study examines the potential age bias exhibited by the three ASR systems: GVT, MS-T, and Siri. The samples used for comparison consisted of recordings from the pronunciation portion of placement tests administered to 1,000 test-takers, in which participants read aloud five sentences under timed conditions. These recordings were first scored by human raters applying an analytic rubric and subsequently processed by the two target ASR systems to generate corresponding scores. The scores from the human raters, GVT, MS-T, and Siri were analyzed to determine whether the ASR systems exhibited age bias in their assessments.

Participants

The study involved 1,000 participants who partook in an English as a second or additional language proficiency test from a modern language department at a French-Canadian university. The gender distribution included 467 male and 533 female participants. The participants represented a diverse range of linguistic backgrounds: French (n = 208), Spanish (n = 207), Persian (n = 200), Arabic (n = 206), and Chinese (n = 179). To ensure a representative sample of learners, their proficiency levels covered the Common European Framework Reference for Languages (CEFR) spectrum (A1–C2). The average age of participants was 32.47

years, with 498 individuals aged between 19 and 29, 412 aged between 30 and 44, and 90 aged 45 or older.

Instruments

This study utilizes secondary data from a language proficiency test administered for placement in English courses, program admission, graduation, and international mobility purposes. The test has different sections such as reading, writing, syntax, speaking, and pronunciation. With respect to the pronunciation section, the target sentences were carefully selected to evaluate a range of phonetic and phonological competencies, with a focus on comprehensibility as the primary outcome. In addition, the primary construct assessed by human raters was comprehensibility, not intelligibility. Raters were instructed to evaluate how easily they could understand the speaker, which aligns with the standard definition of comprehensibility (Derwing & Munro, 1997). The participants read two practice sentences to prepare for the evaluation process, and then proceeded to read five sentences aloud, with a 20-second interval between each. These sentences progressively increased in pronunciation difficulty. The first sentence served as a baseline (identical for all students), while the remaining four were randomly selected from a pool categorized by proficiency level, roughly aligned with the CEFR of A1 to C2.

At the most basic level, participants encountered sentences with monosyllabic, high-frequency words and phrases that emphasized universal consonants and vowels, as well as simple thought grouping and linking patterns. As the complexity increased, sentences incorporated two- to three-syllable words, compound nouns, hard-to-pronounce segments, and complex consonant clusters (e.g., *The trio sings to the audience as it streams onto the busy street in the cold rain*).

More advanced tasks focused on intonation patterns and grammatical endings, requiring learners to produce English prosody in contextually appropriate ways (e.g., *These are more sophisticated pictures, aimed at a particular kind of filmgoer. Is she sure that this audience understands them?*). Participants were then challenged with multisyllabic words and lower-frequency vocabulary, testing their ability to maintain fluency while articulating longer phrases (e.g., *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*).

At the most advanced level, the task involved infrequent and academic vocabulary, as well as idiosyncratic pronunciations, requiring heightened attention to phonological subtleties (e.g., *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*). By progressing through these structured pronunciation challenges, participants engaged with a broad spectrum of phonetic and prosodic demands, allowing for a comprehensive assessment of their proficiency in spoken English.

Human-rater scoring

The participants' pronunciation was scored by trained human raters using a standardized rubric, which provided a structured and objective framework for the evaluation. The rubric assessed multiple criteria, each rated on a scale from 1 (poor performance) to 5 (advanced performance), resulting in a cumulative score out of 25. This score was multiplied by 4 to yield a maximum of 100, making it easier to interpret and compare results as a percentage. The criteria included:

1. **Comprehensibility**

- Assessment of overall comprehensibility, defined by Munro and Derwing (1995) as the listeners' perceptions of how easily they understand an utterance. In practice, this means that even if a speaker has a noticeable accent, their speech can still be considered highly comprehensible if the listener can easily understand the intended message.

2. **Phonemes**
    - Evaluation of the participants' articulation of individual segments in words and phrases. For instance, a relatively poor performance would be observed when a participant pronounces "heat" instead of "eat" in the phrase "I eat the cake," which could lead to a communication breakdown.
3. **Connected speech**
    - Evaluation of connected speech phenomena, including across-word resyllabification, segment deletion, and assimilation. In an advanced-level articulation that exhibits resyllabification and palatalization, "meet you" in "Nice to meet you" would be produced as /mi.ʧu/.
4. **Word stress**
    - Examination of appropriate emphasis on syllables within words (stress). A poor performance at the word level might occur when a participant misplaces the stress in "desserts" (sweet dish) in the sentence "I like desserts", mispronouncing is as /'dɛ.zərts/ (not /dɪ.ˈzɜrts/).
5. **Prosody: Thought groups, sentence stress, and intonation**

- This criterion encompasses all aspects of sentence-level pronunciation, including the segmentation of sentences into thought groups (e.g., via appropriate pauses and pitch contours), the assignment of sentence stress (rhythm), and intonation (e.g., the appropriate use of falling and rising pitch contours). For example, a poor performance in prosody can be observed if a participant fails to raise their pitch to a mid-level during phrases like "a small onion" and "a clove of garlic" (which are intended to signal suspense, or that the sentence is not yet complete) in the following utterance: "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."

Automated coding of ASR output

As is customary in research that evaluates ASR's output quality for pronunciation assessment (e.g., Cox & Davies, 2012; Cucchiarini et al., 1997; Saito et al., 2023; as well as our own research), the three adopted ASR systems analyzed the participants' speech samples to generate a textual output, which was then submitted to a script (created for this research) that calculated the proportion of accurately transcribed words. This metric provided a quantitative benchmark for evaluating the ASR system's alignment with human-rater judgments. Each correctly pronounced word earned one point, with homophones counted as correct. Errors involving missing or added grammatical inflections (e.g., plural markers, verb endings) were penalized with a half-point deduction, as these errors cannot be considered fully correct nor entirely incorrect. The ASR score was the same formula that was used in Johnson et al. (2024) and Nelson & Cardoso (2024):

$$\text{Accuracy Score} = \left(\frac{\text{Number of Accurate Words}}{\text{Total Number of Words}} \right) \times 100$$

Inter-rater reliability

To ensure the reliability of human ratings, inter-rater agreement was assessed through an evaluation of the scoring process. Each speech sample was independently evaluated by two experienced ESL instructors with substantial classroom and pronunciation assessment expertise. Final scores were determined through consensus while consulting the analytic rubric. In cases where a pronunciation rating appeared to be a statistical outlier (greater than fifteen points; roughly equivalent to one CEFR level, e.g., A2 to B2) compared to the participant's overall placement test performance, a quality control team conducted a secondary review to verify the rating's accuracy. This step helped mitigate potential subjective bias in human evaluations and maintain the reliability of scoring across participants.

In addition to human ratings, machine scoring was employed to analyze pronunciation accuracy using the three target ASR systems: GVT, MS-T, and Siri. To verify the reliability of machine scoring, 15% of the machine-transcribed sample was manually coded and compared against the script-generated ratings. Statistical validation of machine scoring was conducted using a two-way mixed-effects model intraclass correlation (ICC) for absolute agreement, yielding a near-perfect correlation of .99 (95% CI, CI, .980, .991), indicating a high level of alignment between human and machine evaluations. By incorporating both human consensus ratings and automated analysis, this study aimed to provide a strong and reliable assessment of L2 pronunciation proficiency.

Statistical analysis

Descriptive statistics were first computed to summarize participant characteristics, including gender, first language, and age. This provided an overview of the dataset and ensured a clear understanding of the sample distribution. To examine potential biases in machine scoring, a

linear regression analysis was conducted in SPSS 28. It was run using the PROCESS macro in SPSS (Hayes, 2022), using ASR-generated scores as predictors of human-rater scores. Test-taker age (29 and under, between 30 and 44 inclusively, and 45 and older) was included as a moderator variable to determine whether age influenced the relationship between machine and human scores. This approach allowed for the identification of any systematic discrepancies in ASR assessments across different age groups, helping to assess the fairness of automated scoring systems and consequently answer our research question. By incorporating this statistical model, the study aimed to quantify any age-related bias and evaluate whether machine scores aligned consistently with human evaluations across all age groups.

Results

To begin, we conducted a descriptive statistical analysis for each ASR, ensuring that outliers were removed before proceeding. The analysis of GVT and human scores (Table 7) across different age groups revealed notable trends. Participants aged 18 to 29 achieved the highest GVT scores, followed by those aged 30 to 44, with participants aged 45 and older scoring the lowest. For human scores, a similar pattern emerged. The 18 to 29 age group scored the highest, with scores decreasing in the 30 to 44 age group and in the 45 and older group. These findings suggest a decline in both GVT and human scores with increasing age, with the youngest age group consistently outperforming the older groups.

Table 7

GVT and Human Scores Across Age Groups

		GVT		Human	
Age	*n*	*M*	*SD*	*M*	*SD*
$18 \leq x \leq 29$	397	85.73	9.82	54.98	6.75
$30 \leq x \leq 44$	485	81.40	10.93	50.45	5.38
$x \geq 45$	88	79.37	12.59	49.63	6.11
Total	970	82.99	10.90	52.23	6.46

The analysis of MS-T and human scores across different age groups (Table 8) revealed a consistent pattern. Participants aged 18 to 29 achieved the highest MS-T scores, with a slight decline in scores for the 30 to 44 age group and the 45 and older group. For human scores, a similar trend was observed. The 18 to 29 age group again scored the highest, followed by the 30 to 44 group and the 45 and older group. These results indicate a moderate decrease in performance with age for both MS-T and human scores, with younger participants consistently outperforming older participants. This pattern suggests that age-related differences may influence performance outcomes, particularly in MS-T and human-evaluated scores.

Table 8

MS-T and Human Scores Across Age Groups

		MS-T		Human	
Age	*n*	*M*	*SD*	*M*	*SD*
$18 \leq x \leq 29$	402	91.33	8.31	55.06	6.90
$30 \leq x \leq 44$	493	88.40	9.55	50.49	5.39
$x \geq 45$	87	88.12	11.14	49.71	6.07
Total	982	89.58	9.33	52.30	6.53

Table 9 shows how performance on both Siri and human assessments varied across age groups. Among the 978 total participants, the youngest group performed best on Siri tasks. As with the other ASR systems, performance declined in older groups (i.e., those aged 30-44 and 45 and older). The same age-related pattern emerged in human scores. The 18-29 age group led, followed by the 30-44 group and finally the 45+ group. The results demonstrate that younger participants consistently outperformed older ones on both types of assessments.

Table 9

Siri and Human Scores Across Age Groups

		Siri		Human	
Age	*n*	*M*	*SD*	*M*	*SD*
18 ≤ x ≤ 29	397	84.90	9.75	54.98	6.77
30 ≤ x ≤ 44	492	82.28	10.53	50.51	5.38
x ≥ 45	89	80.76	12.11	49.65	6.08
Total	978	83.20	10.47	52.25	6.45

Multiple regression analyses were performed for each of the ASR systems. The first was conducted to examine the relationship between GVT scores and performance outcomes (human rater score out of 100), with age group interactions included as predictors (Table 10). The GVT score had a significant positive effect, demonstrating that higher GVT scores were associated with higher performance outcomes. The interaction between GVT scores and the 18 to 29 age group was also significant (p = .002), suggesting that younger participants may experience an additional positive impact from higher GVT scores compared to other age groups. However, the interaction between GVT scores and the 45 and older age group was not significant (p = .460), indicating no substantial difference in the effect of GVT scores on performance outcomes for older participants relative to the reference age group.

Table 10

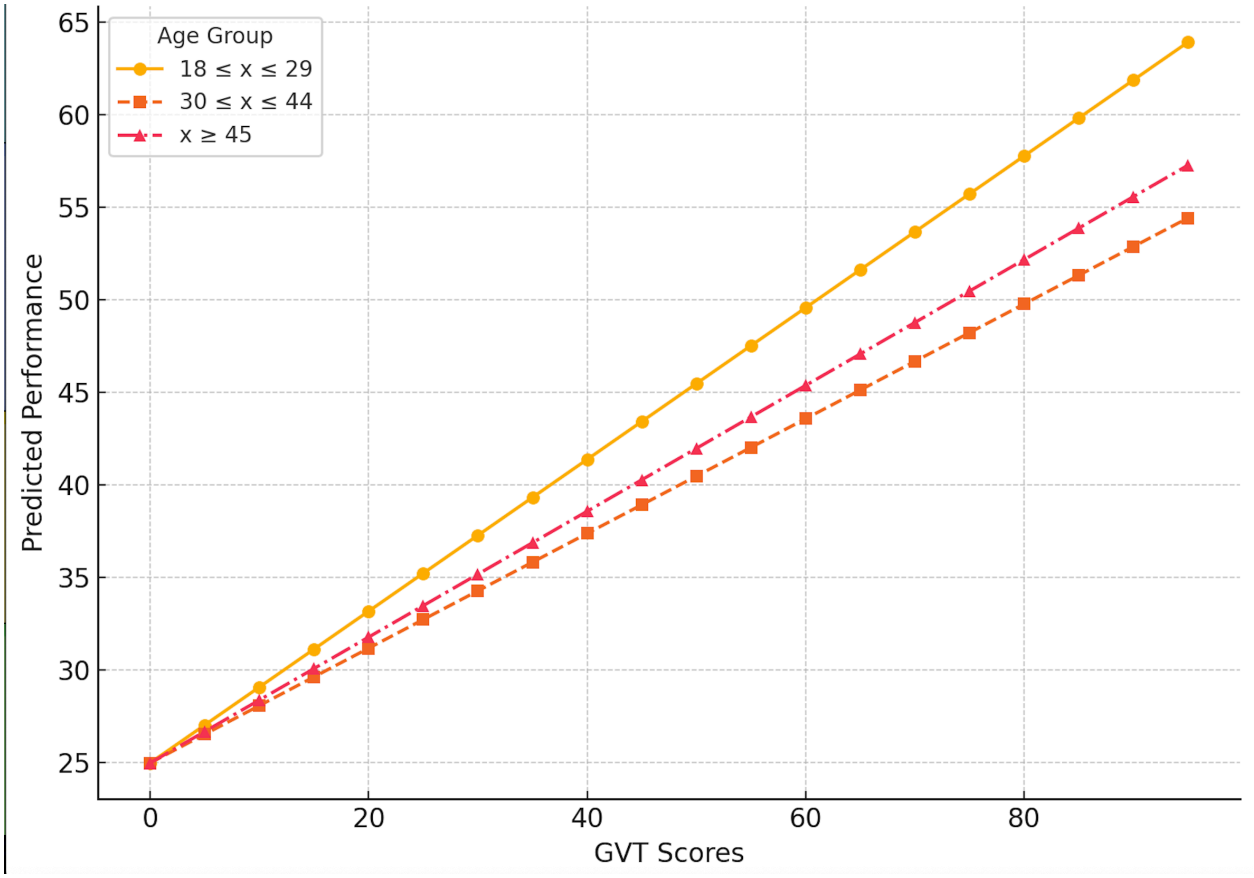
GVT and Performance Outcomes with Age Group Interactions

	b	*SE*	*t*	*p*
Constant	24.96 [21.59, 27.27]	1.61	15.47	<.001
GVT	.31 [.27, .35]	.02	15.94	<.001
(18 ≤ x ≤ 29)*GVT	.10 [.04, .16]	.03	3.18	.002
(x ≥ 45)*GVT	.03 [-.05, .12]	.04	.74	.460

Overall, the results suggest that GVT scores positively predict performance, as shown in Figure 1, with a notable enhancement for the youngest age group, while the older age group does not demonstrate a significant interaction effect.

Figure 1

Predicted Performance by GVT Scores and Age Group



The second multiple regression analysis (Table 11) examined how MS-T scores related to performance outcomes across different age groups. MS-T scores showed a robust positive relationship with performance, with age influencing the extent of this effect. Participants aged 18-29 showed an enhanced benefit from higher MS-T scores ($p = .006$), performing better than other age groups with equivalent MS-T scores. In contrast, participants 45 and older showed no

significant difference in how MS-T scores affected their performance (p = .292) compared to the reference group.

Table 11

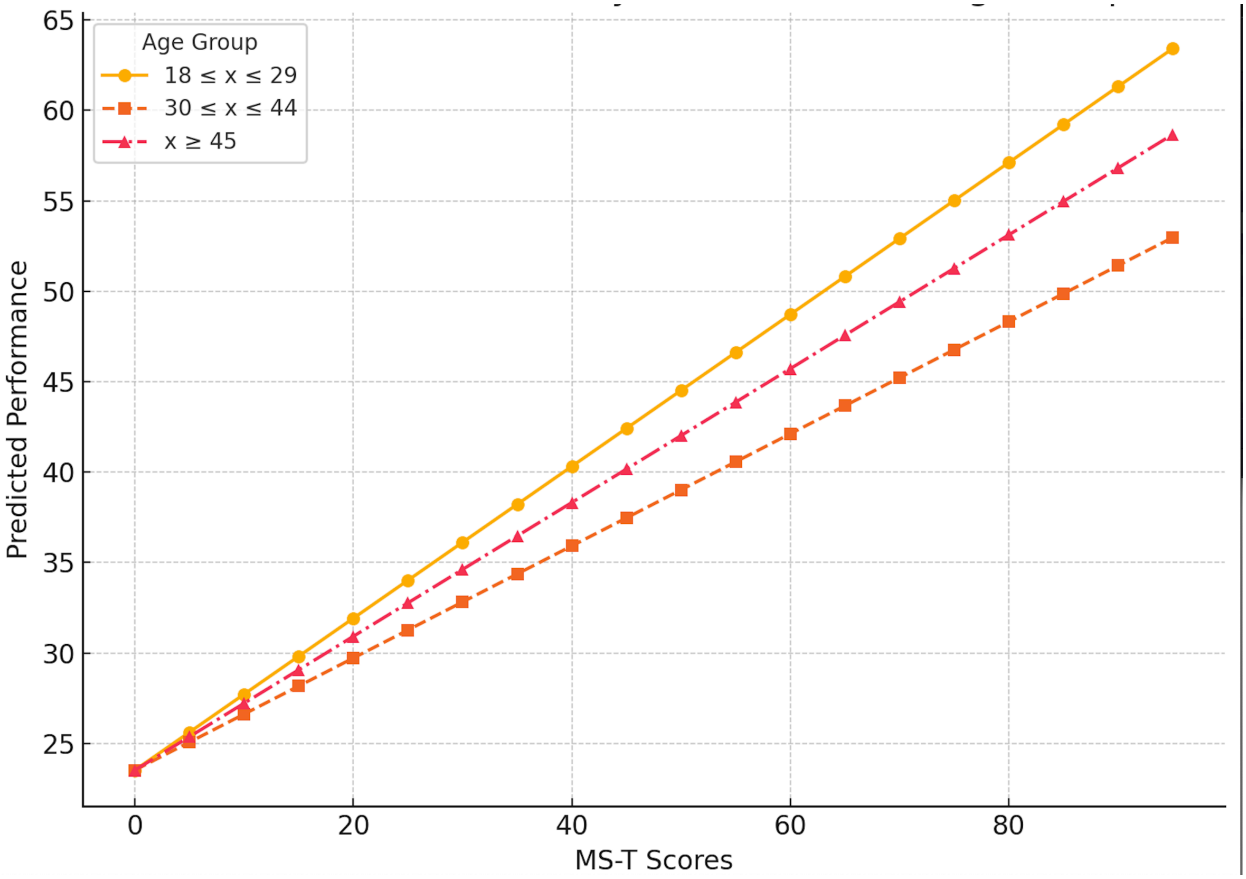
MS-T and Performance Outcomes with Age Group Interactions

	b	*SE*	*t*	*p*
Constant	23.52 [19.25, 27.79]	2.18	10.81	<.001
MS-T	.31 [.26, .35]	.02	12.47	<.001
(18 ≤ x ≤ 29)*MS-T	.11 [.03, .19]	.04	2.78	.006
(x ≥ 45)*MS-T	.06 [-.05, .17]	.06	1.05	.292

As shown in Figure 2, these findings demonstrate that while higher MS-T scores generally predict better performance across all ages, this relationship is particularly strong among young adults aged 18-29.

Figure 2

Predicted Performance by MS-T Scores and Age Group



Finally, a multiple regression analysis examined how Siri scores influenced performance outcomes across age groups (Table 12). Siri scores demonstrated a strong positive relationship with performance, revealing age-related differences in their impact. The youngest group (18-29 years) showed an additional benefit from higher Siri scores ($p = .022$) compared to other age groups. However, participants aged 45 and older showed no significant difference in how Siri scores affected their performance ($p = .461$) compared to the reference group.

Table 12

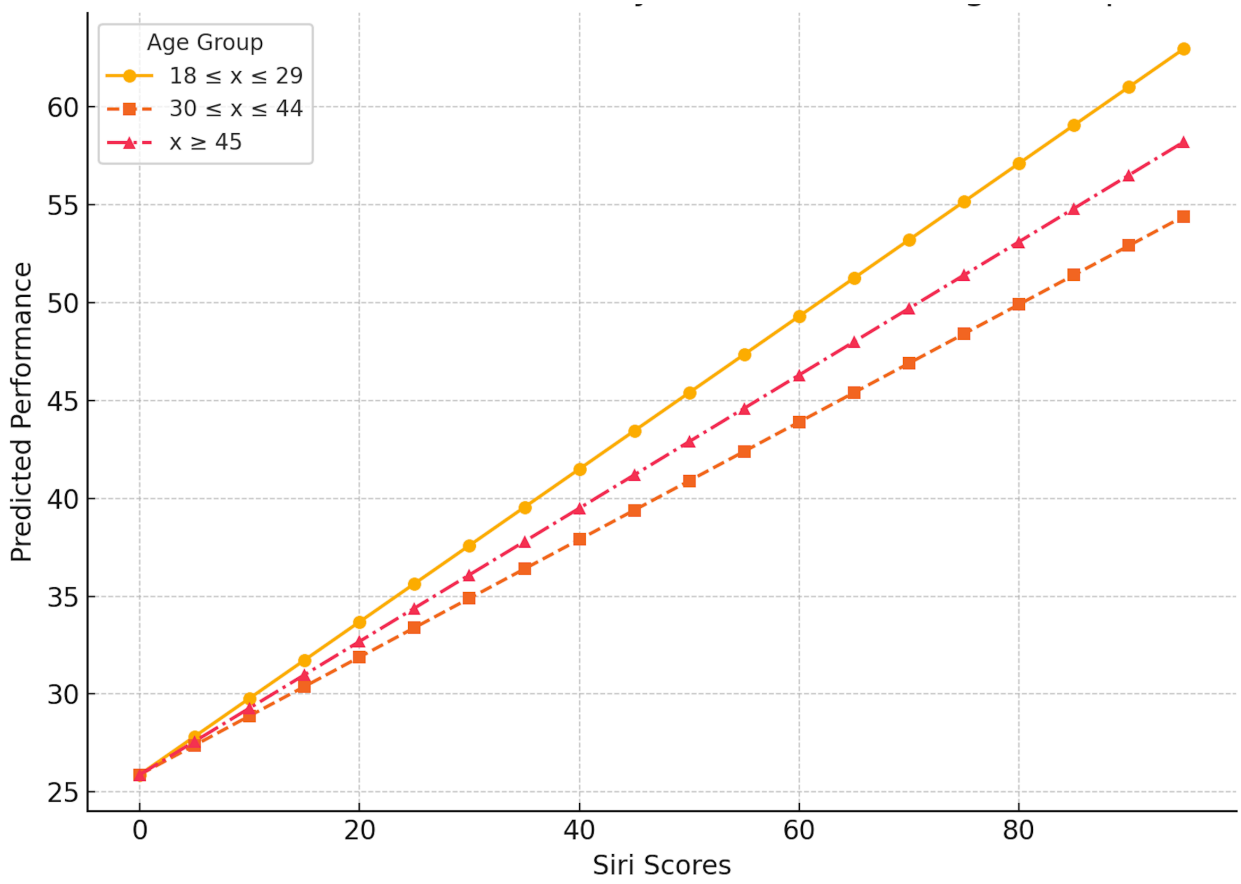
Siri and Performance Outcomes with Age Group Interactions

	b	*SE*	*t*	*p*
Constant	25.89 [22.44, 29.33]	1.75	14.75	<.001
Siri	.30 [.26, .34]	.02	14.15	<.001
(18 ≤ x ≤ 29)*Siri	.09 [.02, .15]	.03	2.64	.022
(x ≥ 45)*Siri	.04 [-.06, .13]	.05	.74	.461

These results indicate, as seen in Figure 3, that while higher Siri scores generally predict better performance across all ages, this relationship is particularly pronounced among young adults aged 18-29.

Figure 3

Predicted Performance by Siri Scores and Age Group



Discussion and Concluding Remarks

The present study examined the relationships between GVT, MS-T, and Siri scores and performance outcomes across different age groups, highlighting the influence of age-related interactions. The findings consistently demonstrated a positive association between higher scores on GVT, MS-T, and Siri assessments and improved performance, with notable variations among age groups.

Across all analyses, the youngest age group (18 to 29 years) consistently outperformed older groups, both in terms of raw scores and interaction effects. Specifically, GVT scores were significantly associated with performance outcomes, with an additional positive effect observed

in the 18 to 29 age group. A similar trend was noted for MS-T scores with a significant interaction for younger participants. Siri scores also showed a robust positive effect on performance with the youngest age group demonstrating a unique advantage. In contrast, the interactions between scores and the oldest age group (45 years and older) were not significant across all ASR systems, indicating that the performance outcomes for this group did not differ significantly from the reference age group when considering GVT, MS-T, or Siri scores. This finding aligns with Ngueajio and Washington (2022), reinforcing the notion that age-related differences in ASR performance may plateau in older adults. Although these findings cannot be directly compared to existing literature due to the absence of similar comparative studies, they align with research suggesting that ASR systems tend to be less accurate in recognizing the speech of older people (e.g. Ferland et al., 2019; Sobti et al., 2024, Ngueajio & Washington, 2022), likely due to age-related changes in voice quality, articulation, and fluency, which are not well-represented in ASR training datasets.

The consistent trend of younger participants benefiting more from higher GVT, MS-T, and Siri scores may reflect age-related cognitive and technological adaptability during the computerized placement test (Hartshorne et al., 2018). Younger individuals may possess greater familiarity and comfort with the digital and analytical skills required by these assessments, contributing to their enhanced performance (Bennett et al., 2008). Furthermore, cognitive processing speed and learning adaptability, which tend to decrease with age (Bialystok et al., 2008), might partly explain why older participants did not exhibit significant interaction effects. In the context of L2 acquisition, age plays a crucial role in pronunciation development, with research indicating that younger learners tend to achieve greater phonological accuracy compared to older learners (Birdsong, 2006). This phenomenon is often linked to the critical

period hypothesis (Hartshorne et al., 2018), which suggests that early exposure to a second language facilitates its acquisition. Older learners, in contrast, may struggle with certain phonemes that do not exist in their first language, retain a stronger accent, or exhibit greater difficulty in modifying established speech habits (Caldwell-Harris & MacWhinney, 2023).

For the oldest age group, the lack of significant interaction effects might indicate a ceiling effect or a reduced influence of these assessments on performance outcomes. While older adults can achieve high scores (Johnson et al., 2024; Nelson & Cardoso, 2024), the incremental gains in performance associated with these scores appear less pronounced than in younger groups. This pattern could be attributed to critical period phenomena (Hartshorne et al., 2018), as discussed earlier, and may have implications for tailored interventions or support strategies that focus on maximizing performance across different age demographics.

These findings have practical implications, particularly in educational, professional, and assessment settings. The robust positive effects of GVT, MS-T, and Siri scores suggest that enhancing skills related to these assessment settings of the computerized placement test could lead to improved performance outcomes, especially for younger individuals. Along the lines of Knowles (2015), training programs and learning environments might consider incorporating adaptive strategies to support older participants, such as increased familiarization with ASR-mediated tasks, clearer pacing and instructions, or targeted preparatory activities designed to enhance speech clarity in technologically mediated contexts.

In addition to pedagogical adjustments, adaptive strategies may also involve task design and assessment-use considerations, including the provision of consistent recording conditions, allowance for processing-time differences, and cautious interpretation of ASR-based scores for older test takers. Such measures acknowledge the physiological and cognitive changes that may

accompany aging, such as reductions in speech clarity, vocal strength, or processing speed—without framing these changes as deficits and help ensure that ASR-mediated pronunciation assessment is used in a fair and valid manner across age groups. While the study provides valuable insights, several limitations must be acknowledged. The cross-sectional design limits causal interpretations, and future research could explore longitudinal approaches to understand how these relationships evolve over time. Additionally, future research should examine whether variables that often covary with age, such as prior exposure to speech technologies, educational background, or cognitive processing demands, function as sources of construct irrelevant variance in ASR based pronunciation assessment. Investigating these factors would help clarify whether observed age related differences reflect genuine variation in intelligibility or systematic biases in ASR performance across age groups, thereby contributing to a more robust evaluation of validity and reliability. Future research could explore how the design of read-aloud sentence tasks might be optimized to improve accessibility and performance for older L2 learners when evaluated using GVT, MS-T, and Siri. By identifying and addressing potential barriers to performance gains in this demographic, researchers and practitioners could foster more equitable outcomes across all age groups.

Overall, the present study highlights the significant impact of GVT, MS-T, and Siri scores on pronunciation performance, with younger individuals showing greater benefits. These insights contribute to a broader understanding of how age and assessment performance interact and offer a foundation for developing tailored interventions to support individuals across the lifespan.

Chapter 4: Discussion and Conclusion

Purpose and Scope

The overarching purpose of this dissertation was to examine the role of dictation-based ASR systems, specifically GVT, MS-T, and Siri in L2 pronunciation assessment. At the heart of this project lies the question of whether these widely available technologies can generate scores that align with human ratings and whether they can do so without introducing validity threats that might compromise their usefulness in both educational and testing contexts. The possibility that ASR tools could offer reliable and accessible alternatives is an issue of both practical and theoretical significance, as pronunciation assessment by human raters can be often costly and time consuming. By systematically comparing the performance of these systems to trained human evaluators, this dissertation sought to identify the extent to which dictation-based ASR can serve as a trustworthy resource, while also acknowledging its limitations. The two manuscripts that form the core of this dissertation are presented as complementary studies that approach this central question from different but connected perspectives.

Manuscript A investigated the extent to which Siri's transcriptions aligned with human ratings of L2 pronunciation. Using a detailed rubric that considered comprehensibility, segmental accuracy, prosody, and rhythm, human raters evaluated learner samples that were also processed by Siri. The study then situated Siri's performance within the broader landscape by drawing on prior findings related to GVT and MS-T. This allowed for a comparative view of how these three dictation systems perform in relation to human benchmarks, highlighting both areas of convergence and points of divergence.

Manuscript B extended this inquiry by addressing a different dimension of validity, namely the potential for demographic variation to affect ASR recognition accuracy. Specifically,

it examined whether the systems demonstrated systematic differences when scoring speakers from different age groups. While Manuscript A demonstrated that Siri could, under many conditions, approximate human judgments as effectively as GVT and MS-T, Manuscript B raised a more critical issue by showing that younger speakers tended to receive higher scores from all three ASR systems. Regression analyses of 1,000 learner responses across three age groups and five L1 backgrounds demonstrated that this pattern was consistent and statistically significant. Importantly, similar age-related trends have been reported in human comprehensibility research, suggesting that higher scores among younger participants may partially reflect genuine differences in intelligibility rather than ASR specific bias alone. At the same time, the replication of this pattern across multiple ASR systems raises concerns about how age-related variation is operationalized and weighted in dictation-based scoring. From a validity perspective, this pattern constitutes a potential threat not because it conclusively demonstrates ASR bias, but because it underscores the need to examine whether ASR mediated scores systematically interact with age in ways that may affect score interpretation, reliability, and fairness in evaluative contexts.

Together, these studies underscore the central theme of the dissertation: the exploration of validity in ASR-based scoring of L2 pronunciation. Manuscript A demonstrated that dictation-based ASR can yield scores that align closely with human ratings, thereby supporting its potential validity as an assessment tool, while also highlighting its advantages as a low-cost and scalable option. Manuscript B, in turn, revealed an important limitation by identifying age related variation as a source of potential bias. Viewed side by side, the two manuscripts do not simply provide separate findings but instead form a more comprehensive picture of the affordances and constraints of ASR in L2 assessment. The combination of positive evidence for correlation with human ratings and cautionary evidence about age bias contributes to a balanced and nuanced

understanding of the potential of these systems. Ultimately, the findings of this dissertation emphasize the need for careful validation before dictation-based ASR tools are adopted for widespread use in language testing, ensuring that their benefits can be realized without compromising validity or disadvantaging different groups of learners.

Synthesis of Key Findings

Across both manuscripts, the three dictation-based ASR systems, Siri, MS-T, and GVT, produced scores that correlated strongly with human evaluations of L2 pronunciation. In Manuscript A, Siri demonstrated robust overall alignment with human raters (r_s = .79, p < .001), showing significant convergence across all rubric criteria, including phonemes (.81), comprehensibility (r_s = .80), and prosody (r_s = .81). Comparable results were reported for MS-T and GVT, with each producing correlations in the range of r_s = .78 to r_s = .83 across the different phonological criteria. These findings reinforce that ASR systems can approximate human judgments in meaningful ways. Notably, Siri's correlations were particularly strong for lower proficiency learners (r_s = .83, p < .001), suggesting that ASR systems can reliably capture broad differences in intelligibility when learner speech is less advanced and errors are more frequent. For this reason, these tools appear especially promising in formative contexts where many learners are working toward basic intelligibility and need immediate, accessible feedback.

System level differences also emerged in nuanced ways. Although GVT, MS-T, and Siri showed nearly identical total correlations with human ratings (GVT with r_s = .78, MS-T with r_s = .79, and Siri with r_s = .79), each displayed distinct strengths and weaknesses. While GVT demonstrated the strongest link with human judgments of comprehensibility (r_s = .85), MS-T was comparatively more effective at capturing connected speech, perhaps reflecting its design for continuous dictation tasks. Finally, Siri aligned most strongly with human raters on phoneme

accuracy and intonation patterns, suggesting particular sensitivity to segmental features and prosodic cues. These results indicate that these tools' transcription engines may privilege global intelligibility over finer phonetic distinctions. These patterns show that while each system broadly converges with human raters, they do so in slightly different ways, emphasizing the importance of considering system choice in light of specific assessment priorities.

A recurring limitation highlighted in Manuscript A was that ASR struggled to capture subtle distinctions among higher proficiency speakers. For learners rated highly by human evaluators (mean score = 89.67/100), Siri produced very high average scores (94.34/100), yet correlations with human judgments dropped to a $r_s = .41$. This ceiling effect suggests that when pronunciation approaches near native levels, ASR systems are less effective at differentiating between strong and very strong performances. In practical terms, this means that while ASR tools can reliably flag learners with lower proficiency, they are less capable of distinguishing advanced learners whose speech already meets high standards of intelligibility. Such findings point to a boundary condition on the validity of dictation ASR, particularly in contexts where fine grained judgments are required, such as advanced placement testing or high stakes certification.

Manuscript B added another layer to this picture by showing that age is a significant factor shaping ASR performance. Across all three systems, younger test takers (18–29 years old) consistently received higher ASR scores than older participants, even when human ratings revealed only modest differences between age groups. Regression analyses confirmed that the interaction between ASR scores and younger age was significant for Siri (b = .09, p = .022), MS T (b = .11, p = .006), and GVT (b = .10, p = .002). In contrast, older learners (≥45 years old) showed no significant interaction effects. This pattern indicates that younger participants

benefited from systematically higher recognition accuracy across all systems, raising validity concerns for older test takers whose scores did not align as closely with their actual performance. These findings suggest that dictation-based ASR systems may unintentionally favor younger speakers, possibly due to the underrepresentation of age-related vocal and articulatory characteristics of older speakers in the training data used to build commercial engines.

Taken together, the findings from both manuscripts demonstrate that ASR systems can provide reliable and efficient assessments of L2 pronunciation when compared to human raters, particularly at lower proficiency levels, where errors are frequent and intelligibility is variable. However, the validity of these systems depends on learner profile (e.g., proficiency, age) and assessment context. Importantly, these findings are based on read-aloud tasks, and further research is needed to examine how well ASR systems perform on more open-ended, spontaneous speech tasks, which pose different challenges for both human and automated scoring. Additionally, the systems diverge in their sensitivity to specific pronunciation features, show reduced alignment with human judgments at higher proficiency levels, and exhibit age-related differences that privilege younger speakers. Table 13 and Figure 4 summarize the correlations between ASR scores and human ratings, as well as age-related effects across the three systems. All three ASR systems showed strong correlations with human raters for total scores, indicating a high level of criterion-related validity. Age effects were small but more pronounced among younger participants (18–29), with correlation coefficients between age and ASR–human score differences ranging from .09 to .11 across systems. In contrast, age effects for older participants (≥ 45) were close to zero and not statistically significant, with confidence intervals overlapping zero. This pattern suggests that while all systems align closely with human judgments overall, younger speakers may receive slightly higher ASR scores, whereas age does not appear to

systematically affect scores for older speakers. Taken together, these findings indicate that teachers and evaluators can rely on any of the three ASR systems to provide valid and consistent pronunciation scores across age groups.

Table 13

ASR Validity: Human Correlations and Age Effects

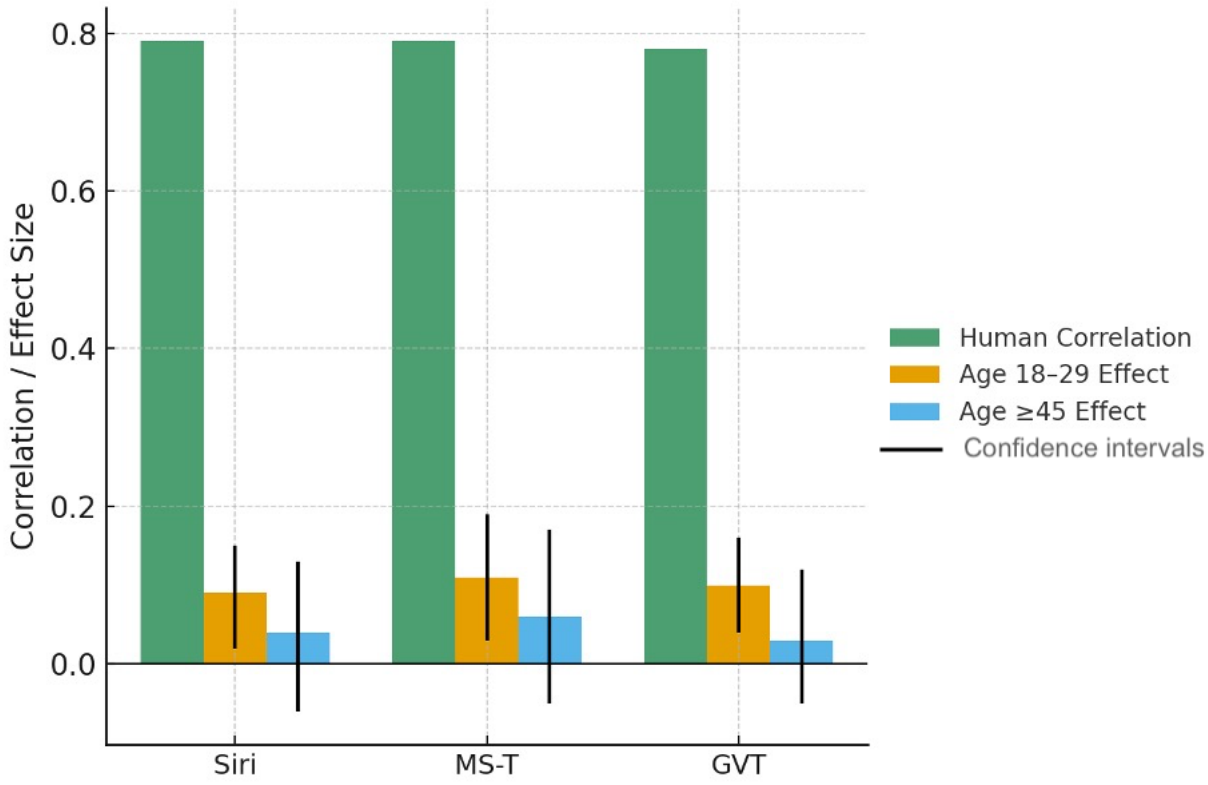
Results	Siri	MS-T	GVT
Correlation with human raters (total score)	.79**	.79**	.78**
Age effects ($18 \leq x \leq 29$)	.09 [.02, .15]	.11 [.03, .19]	.10 [.04, .16]
Age effects ($x \geq 45$)	.04 [-.06, .13]	.06 [-.05, .17]	.03 [-.05, .12]

Note. Confidence intervals based on 1000 bootstrap samples.

**$p < .001$.

Figure 4

ASR Validity: Human Correlations and Age Effects



These results demonstrate both the potential and the limitations of dictation ASR for pronunciation assessment. The promise of these systems lies in their ability to offer immediate, low cost, and standardized evaluations, but their limitations highlight the ongoing need for critical validation, refinement of training datasets, and safeguards against demographic bias if they are to be responsibly integrated into language testing and pedagogy.

Both manuscripts also converge on the conclusion that Siri, GVT, and MS-T exhibit very similar and significant correlations with human raters, confirming their utility as automated evaluators of L2 pronunciation. At the same time, the results across age groups reveal a consistent concern: While the systems generally scored well and aligned with human

evaluations, all three exhibited biases against older participants. This suggests that although dictation-based ASR can be trusted to capture broad patterns of pronunciation accuracy, further refinement is necessary to ensure validity across diverse learner demographics.

Takeaways

This dissertation demonstrates that dictation-based ASR systems, specifically GVT, MS-T, and Siri, can provide reliable assessments of L2 pronunciation that converge strongly with human ratings. Manuscript A established that Siri's scores correlated significantly with human evaluations across multiple rubric criteria, ranging from segmental accuracy to prosodic features such as rhythm, stress, and intonation. Rather than simply confirming that ASR can capture a broad set of pronunciation characteristics, these findings illustrate that mainstream dictation systems are sensitive to both segmental and suprasegmental features that directly affect intelligibility. In doing so, they build on previous research demonstrating that intelligibility-based transcription scores can serve as valid indicators of pronunciation quality (for similar findings see Derwing & Munro, 1997; Neri et al., 2006; Witt, 2012). Just as importantly, Siri's performance was broadly comparable to that of GVT and MS-T, with all three systems producing almost identical overall correlations with human raters. This convergence suggests that learners and educators who use different platforms are likely to experience comparable outcomes, an important consideration given the wide availability of these tools across devices and operating systems. At lower proficiency levels, correlations were especially strong, indicating that ASR is particularly effective at distinguishing broad differences in intelligibility among learners who are still developing foundational pronunciation skills. At higher proficiency levels, however, the correlations weakened, and a possible ceiling effect was observed, suggesting that ASR may not yet have the sensitivity to detect the finer distinctions in speech

quality that human listeners can perceive. Taken together, these findings highlight both the potential and the current limits of ASR in L2 assessment, supporting its use for formative assessment and placement while cautioning against its use in high stakes contexts where subtle distinctions are critical.

Manuscript B extended this inquiry by focusing on demographic factors, particularly age, as a source of variability in ASR performance. Across GVT, MS-T, and Siri, younger test takers consistently benefited from higher recognition accuracy, with statistically significant interaction effects for the 18 to 29 group across all three systems. Their speech was consistently transcribed with greater precision, producing stronger alignment between ASR scores and human ratings. In contrast, participants aged 45 and older did not show significant interaction effects, meaning that while their human rated performance remained strong, the ASR systems did not align as closely with their scores. These results suggest that dictation-based ASR inadvertently privileges younger speakers, possibly due to age related changes in voice quality, articulation patterns, or speaking rate that are underrepresented in the training data used to build commercial ASR engines. The findings resonate with broader discussions in applied linguistics about the influence of age on L2 pronunciation learning (Feng et al., 2024; Serditova et al., 2025), and they underscore that demographic variation must be carefully considered when evaluating the validity of automated systems. For assessment purposes, this raises an important challenge: if older learners are systematically disadvantaged, then the validity of ASR based testing is compromised, particularly in contexts where outcomes may influence educational or professional opportunities.

Taken together, the two manuscripts provide complementary perspectives that lead to several important takeaways. GVT, MS-T, and Siri converge strongly with human ratings,

confirming their potential as scalable and practical tools for pronunciation assessment. Their strengths lie in providing immediate, low cost, and relatively accurate feedback, particularly for learners at earlier stages of development. At the same time, their reduced sensitivity at advanced levels and their demonstrated bias toward younger speakers highlight that validity is not guaranteed across all learner populations or contexts. The central contribution of this dissertation is to show that dictation-based ASR can play a meaningful role in language testing, but only if its limitations are explicitly acknowledged and addressed.

Theoretical Contributions

This dissertation makes several theoretical contributions to the study of L2 pronunciation assessment and the role of ASR in language testing. First, both manuscripts contribute to ongoing discussions about validity in assessment. Manuscript A demonstrated that a general-purpose dictation-based ASR system such as Siri can produce scores that converge strongly with human ratings. This suggests that intelligibility-oriented transcription accuracy, while not designed to capture constructs like stress or intonation directly, can nonetheless serve as a meaningful indicator for broader aspects of pronunciation competence. The implication is that validity in pronunciation assessment may not require the explicit scoring of each construct individually but can instead be achieved through global measures of intelligibility (i.e., speech transcription) that reflect how speech is processed in real time. This insight contributes to theoretical models of assessment by showing that construct validity may be achieved through alignment of practical technology-based measures with human judgments of communicative competence.

Second, the results refine our understanding of test usefulness and score meaning across proficiency levels. The strong correlations between ASR and human ratings for lower

proficiency learners support theoretical claims that automated systems are most effective at identifying large salient differences in speech quality. However, the reduced correlations at higher proficiency levels highlight a ceiling effect that limits sensitivity to subtle proficiency-related variation. While automated systems can validly measure global intelligibility, they may fall short in contexts where fine grained distinctions are necessary. This finding contributes to the theoretical debate about the scope of ASR applicability in L2 testing, clarifying that its usefulness is conditional on learner profile and test purpose.

Finally, by comparing GVT, MS-T, and Siri alongside human raters, this dissertation provides theoretical insights into the interchangeability of ASR systems in L2 assessment. The finding that all three systems converged to a similar degree with human ratings suggests that the construct being measured is not unique to any single platform but rather a function of how dictation-based ASR interprets intelligibility (Knill et al., 2018). These contributions extend the literature on technology mediated assessment and provide a foundation for future theoretical and empirical work on the role of ASR in evaluating L2 pronunciation.

Practical Implications

The findings from these manuscripts also carry practical implications for testing programs, developers, and educators seeking to integrate dictation-based ASR systems into L2 pronunciation assessment.

For testing programs, one outcome is that ASR can act as a valuable complement to human scoring, particularly in contexts where efficiency, large-scale testing, and rapid turnaround are essential. Placement tests and classroom-based assessments are strong examples of such contexts, as they often involve large cohorts of learners and require immediate results to inform instructional decisions. The evidence shows that ASR systems like GVT, MS-T, and Siri

produce scores that align strongly with human evaluations, especially for lower proficiency learners whose speech patterns tend to generate more consistent recognition outcomes. This reliability makes ASR well suited for providing quick feedback, reducing administrative burdens, and lowering costs in repeated or large-scale testing scenarios. However, results must be interpreted with caution for advanced learners, where ASR systems struggle to capture subtle features of pronunciation such as rhythm, stress, and intonation. A similar caution applies to older learners, for whom ASR scores showed weaker alignment with human ratings. For both groups, overreliance on automated scores could lead to misleading judgments. Thus, while ASR can strengthen the efficiency of testing programs, human evaluation remains indispensable for nuanced decisions in high stakes contexts.

For developers, the results highlight the need to improve the foundations on which ASR systems are built. At present, training datasets largely reflect younger speakers and more standard varieties of speech. This narrow representation can limit the validity of ASR outputs for diverse L2 users, those at advanced proficiency levels, and older learners who may exhibit age related changes in pitch, articulation, or speech rate. Expanding training datasets to include a broader range of accents, proficiency levels, and age groups could reduce bias and improve the overall validity of automated scoring. This step is crucial if dictation-based ASR is to move beyond its current role as a supportive tool and into wider use in formal testing contexts. Developers also need to be transparent about system limitations and provide clearer documentation about the populations and speech varieties that are best represented in their models. Such improvements would not only raise the technical quality of ASR but also foster greater trust among testing organizations, educators, and learners.

For educators, the findings suggest that automated scores are most effective when positioned as supportive. In the classroom, ASR can empower learners by providing immediate feedback, highlighting recurring pronunciation difficulties, and encouraging self-monitoring. Learners benefit from the chance to see how their spoken output is transcribed and can adjust their pronunciation in real time. Yet, automated scores alone cannot capture the full complexity of pronunciation performance, particularly at higher levels where subtle segmental and suprasegmental features come into play. Educators could therefore use ASR feedback as a diagnostic aid or as a springboard for targeted practice, while continuing to rely on their judgment for nuanced evaluation. Program administrators can also benefit from ASR by using it to support formative learning and low stakes placement, while reserving high stakes assessment for combined or human driven approaches.

Taken together, these practical implications suggest that dictation-based ASR systems have an important role to play in the future of L2 pronunciation assessment. Their strengths lie in scalability, cost efficiency, and accessibility, especially when supporting learners outside traditional classroom settings. At the same time, realizing this potential depends on careful application by testing programs, sustained development by technology providers, and thoughtful pedagogical use by educators. When employed responsibly, ASR can extend opportunities for feedback and assessment to a much wider population of learners. However, this requires acknowledging and addressing the limits of current systems, particularly their reduced sensitivity with advanced and older learners. By combining the speed and consistency of automation with the nuance and judgment of human evaluation, institutions can achieve both efficiency and validity in pronunciation assessment.

Limitations

Several limitations should be noted when interpreting the findings of this dissertation. First, the analyses relied on secondary data drawn from an existing placement test, which limited the degree to which the tasks could be adapted to address the specific research questions. All speaking tasks consisted of read-aloud items. While these tasks are useful for eliciting controlled speech and capturing a range of learner performance, they do not reflect other task types, such as spontaneous or interactive speech. Consequently, the findings cannot be generalized to how ASR systems might perform on more open-ended tasks, which may pose different challenges for both human and automated scoring. In addition, the use of secondary data meant that test taker perceptions could not be gathered, leaving an important dimension of assessment validity unexplored. Learner attitudes toward ASR scoring, as well as their sense of trust or mistrust in automated feedback, are essential to understanding the broader educational impact of these systems.

A second limitation concerns the scope of demographic and individual variables that were examined. The two manuscripts focused primarily on proficiency level and age as sources of variability, both of which produced significant and interpretable findings. Other potentially influential factors, such as L1 background, gender, and educational experience, were not systematically included in the analyses. Data on L1 and gender were available but were not analyzed in depth, as these variables were not central to the research questions and incorporating them would have required a broader and more complex design than was feasible within the scope of the two manuscripts. In contrast, information on participants' educational experience was not collected in sufficient detail to support meaningful analysis. These variables could interact with ASR recognition accuracy in important ways, and future studies should expand the range of

demographic characteristics considered to provide a fuller picture of how ASR systems function across diverse populations.

In more open-ended speaking tasks, such as picture description or extended response prompts, speakers produce less predictable and more variable language, which may place different demands on ASR systems. It remains unclear whether the strong correlations between ASR and human ratings observed in this study would extend to these types of tasks, which more closely reflect authentic communicative use than read-aloud formats.

Taken together, these limitations suggest that while the dissertation demonstrates promising evidence for the use of dictation-based ASR systems in L2 pronunciation assessment, caution is needed in extending the findings to broader contexts. Addressing these constraints in future research, through the collection of primary data, the inclusion of a wider range of learner variables, and the use of more authentic speech tasks, will be essential for building a more comprehensive and robust understanding of the role of ASR in language assessment.

Future Research Directions

Future research should extend the scope of investigation beyond controlled read-aloud tasks to include spontaneous L2 oral assessment tasks. Spontaneous production introduces greater variability in pronunciation, rhythm, and lexical choice, all of which may challenge ASR systems in ways that are not captured in scripted tasks. Examining how GVT, MS-T, and Siri perform on tasks that elicit language more representative of authentic communicative contexts would provide a clearer understanding of their practical potential for both classroom and testing settings.

Another promising direction involves exploring intersectional factors that were beyond the scope of this dissertation. For instance, examining how age interacts with first language

background or how proficiency interacts with speech rate may reveal nuanced performance patterns that are not visible when variables are considered in isolation. Such analyses could help to identify which learner profiles are most likely to benefit from ASR based scoring and which may be at greater risk of misrecognition.

Task design also deserves closer attention, particularly with respect to improving ASR sensitivity for advanced learners. The present findings suggest that dictation-based ASR systems may struggle to differentiate fine grained features at higher proficiency levels, sometimes leading to ceiling effects or reduced correlations with human raters. Designing tasks that include more phonologically challenging material, or prosodic variation may help ensure that ASR systems can more effectively capture the subtleties of advanced learner speech.

Longitudinal approaches would also provide valuable insights into how ASR validity evolves over time. By following learners across multiple testing or learning cycles, researchers could evaluate whether the systems consistently track improvement and whether alignment with human ratings changes as learners' proficiency develops. Such studies would also allow for the investigation of potential adaptation effects, where learners modify their pronunciation in response to ASR feedback.

Finally, future research could examine the washback and broader consequences of using ASR in placement testing and classroom feedback. Understanding how learners and teachers respond to automated scores, how these scores shape study strategies, and whether reliance on ASR influences classroom dynamics are essential questions for evaluating the impact of these tools in educational practice. Such work would help ensure that ASR integration into assessment does not merely provide efficiency but also supports positive learning outcomes.

Conclusion

This dissertation has shown that widely available ASR tools such as GVT, MS-T, and Siri hold strong potential as valid and practical tools for pronunciation assessment. Across both empirical studies, these systems demonstrated significant correlations with human rater judgments and offered a scalable and low cost means of generating pronunciation scores in educational contexts. Their accessibility and integration into everyday technologies further underscore their appeal for both testing programs and classroom applications.

Overall, the findings highlight that validity is not an inherent property of ASR-based assessments, but rather of the interpretations and uses of the resulting scores, which are contingent on factors such as learner characteristics and task type. The evidence suggests that ASR systems are most effective when used with low to mid proficiency learners and in formative settings, where immediacy and efficiency are prioritized over fine grained diagnostic accuracy at higher proficiency levels. In contrast, validity appears weaker for advanced learners, whose nuanced phonological performance often exceeds the sensitivity of dictation-based ASR, and for older learners, whose age-related vocal characteristics may not be well represented in training datasets. These limitations remind us that while ASR can approximate human judgment in many contexts, it cannot fully serve as a comprehensive substitute for expert evaluation.

Looking ahead, ASR should be embraced as a promising innovation in language testing, but its adoption must be guided by a comprehensive view of validity. This includes not only the degree of alignment with human raters, but also considerations of construct coverage, reliability, practicality, and the broader consequences of use for different learner populations. If integrated thoughtfully, ASR systems can play a valuable role in expanding access to pronunciation

assessment and feedback, complementing human expertise while helping shape more equitable and efficient language testing practices.

References

Aksënova, A., Chen, Z., Chiu, C.-C., van Esch, D., Golik, P., Han, W., King, L., Ramabhadran, B., Rosenberg, A., Schwartz, S., & Wang, G. (2022). Accented speech recognition: Benchmarking, pre-training, and diverse data (arXiv:2205.08014). *arXiv*. https://doi.org/10.48550/arXiv.2205.08014

Aman, F., Vacher, M., Rossato, S., & Portet, F. (2013). Speech recognition of aged voice in the AAL context: Detection of distress sentences. *2013 7th Conference on Speech Technology and Human–Computer Dialogue (SpeD)*, 1–8. https://doi.org/10.1109/SpeD.2013.6682669

Arora, V., Lahiri, A., & Reetz, H. (2018). Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *The Journal of the Acoustical Society of America, 143*(1), 98–108. https://doi.org/10.1121/1.5017834

Ashwell, T., & Elam, J. R. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *JALT CALL Journal*, 13(1), 59–76. https://doi.org/10.29140/jaltcall.v13n1.j212

Babaeian, A. (2023). Pronunciation assessment: Traditional vs modern modes. *Journal of Education for Sustainable Innovation, 1*(1), 61–68. https://doi.org/10.56916/jesi.v1i1.530

Bachman, L. F., & Palmer, A. S. (1996). Test usefulness: Qualities of language tests. In L. F. Bachman & A. S. Palmer (Eds.), Language testing in practice: Designing and developing useful language tests (pp. 17–38). Oxford University Press.

Badwan, K. (2021). Language and the sociolinguistic market. In M. Martin-Jones & D. Block (Eds.), *Language in a globalised world: Social justice perspectives on mobility and*

contact (pp. 23–42). Springer International Publishing. https://doi.org/10.1007/978-3-030-77087-7

Bajorek, J. P. (2019, May 10). Voice recognition still has significant race and gender biases. *Harvard Business Review.* https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). I can speak: Improving English pronunciation through automatic speech recognition–based language learning systems. *Innovation in Language Learning and Teaching, 18*(5), 443–461. https://doi.org/10.1080/17501229.2024.2315101

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9), 1–39.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355–377. https://doi.org/10.1177/0265532210364404

Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 859–873. https://doi.org/10.1037/0278-7393.34.4.859

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning, 56*(S1), 9–49. https://doi.org/10.1111/j.1467-9922.2006.00353.x

Blake, R. (2016). Technology and the four skills. *Language Learning & Technology, 20*(2), 129–142. https://doi.org/10.64152/10125/44465

Boyd, E., & Donnarumma, D. (2018). Assessment literacy for teachers: A pilot study investigating the challenges, benefits, and impact of assessment literacy training. In D.

Xerri & P. V. Briffa (Eds.), *Teacher involvement in high-stakes language testing* (pp. 131–147). Springer. https://doi.org/10.1007/978-3-319-77177-9_7

Browne, K., & Fulcher, G. (2017). 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

Caldwell-Harris, C. L., & MacWhinney, B. (2023). Age effects in second language acquisition: Expanding the emergentist account. *Brain and Language*, *241*, Article 105269. https://doi.org/10.1016/j.bandl.2023.105269

Cámara-Arenas, E., Tejedor-García, C., Tomas-Vásquez, C. J., & Escudero-Mancebo, D. (2023). Automatic pronunciation assessment vs. automatic speech recognition: A study of conflicting conditions for L2-English. *Language Learning & Technology, 27*(1), 1–19. https://doi.org/10.64152/10125/73512

Canagarajah, S. (2013). *Translingual practice: Global Englishes and cosmopolitan relations.* Routledge.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201–219. https://doi.org/10.1177/0265532210393704

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge University Press.

Cenoz, J., & García Lecumberri, M. L. (1999). The acquisition of English pronunciation: Learners' views. *International Journal of Applied Linguistics, 9*(1), 3–18. https://doi.org/10.1111/j.1473-4192.1999.tb00157.x

Chan, M. P. Y., Choe, J., Li, A., Chen, Y., Gao, X., & Holliday, N. (2022). Training and typological bias in ASR performance for world Englishes. In *Proceedings of Interspeech 2022*, (pp. 1273–1277). ISCA. https://doi.org/10.21437/Interspeech.2022-10869

Chapelle, C. A., & Lee, H. (2021). Conceptions of validity. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 17–31). Routledge. https://doi.org/10.4324/9781003220756-3

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology, 20*(2), 116–128. https://doi.org/10.64152/10125/44464

Chen, L., & Asgari, M. (2020). *Refining automatic speech recognition system for older adults* (arXiv:2011.08346). *arXiv*. https://doi.org/10.48550/arXiv.2011.08346

Chen, T., & Sun, S. (2025). Evaluating automated evaluation systems for spoken English proficiency: An exploratory comparative study with human raters. *PLOS ONE, 20*(3), Article e0320811. https://doi.org/10.1371/journal.pone.0320811

Chun, D., Kern, R., & Smith, B. (2016). Technology in language use, language teaching, and language learning. *The Modern Language Journal*, 100(S1), 64–80. https://doi.org/10.1111/modl.12302

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia, 10*, Article 3. https://doi.org/10.1186/s40468-020-00101-6

Coulmas, F. (2005). Communicating across generations: Age as a factor of linguistic choice. In *Sociolinguistics: The study of speakers' choices* (pp. 52–67). Cambridge University Press. https://doi.org/10.1017/CBO9780511815522

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press.

Cox, T. L., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal, 29*(4), 601–618. https://doi.org/10.11139/cj.29.4.601-618

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System, 30*(2), 155–166. https://doi.org/10.1016/S0346-251X(02)00012-X

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(4), 476–490. https://doi.org/10.1017/S026144480800551X

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research* (Vol. 42). John Benjamins. https://doi.org/10.1075/lllt.42

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

Dutta, S., Tao, S. A., Reyna, J. C., Hacker, R. E., Irvin, D. W., Buzhardt, J. F., & Hansen, J. H. L. (2022). Challenges remain in building ASR for spontaneous preschool children speech in naturalistic educational environments. In *Proceedings of Interspeech 2022*, (pp. 4322–4326). ISCA. https://doi.org/10.21437/Interspeech.2022-555

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology, 2*(2), 62–76. *https://doi.org/10.64152/10125/25043*

Evanini, K., & Wang, X. (2013). Automated speech scoring for non-native middle school students with multiple task types. In *Proceedings of Interspeech 2013*, (pp. 2435–2439). ISCA. https://doi.org/10.21437/Interspeech.2013-566

Evanini, K. (2019). Overview of automated scoring. In K. Evanini & K. Zechner (Eds.), *Using language technologies to score spontaneous speech* (pp. 3–20). Routledge.

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. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. In *Proceedings of Interspeech 2021*, (pp. 1664–1668). ISCA. https://doi.org/10.48550/arXiv.2103.15122

Feng, S., Raj, B., & Watanabe, S. (2024). Towards inclusive automatic speech recognition. *Computer Speech & Language, 87*, Article 101566. https://doi.org/10.1016/j.csl.2023.101567

Ferland, L., Huffstutler, T., Rice, J., Zheng, J., Ni, S., & Gini, M. (2019). Evaluating older users' experiences with commercial dialogue systems: Implications for future design and development. *arXiv.* https://doi.org/10.48550/arXiv.1902.04393

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. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial intelligence applications and innovations* (Vol. 583, pp. 73–82). Springer. https://doi.org/10.1007/978-3-030-49161-1_7

Fuckner, M., Horsman, S., Wiggers, P., & Janssen, I. (2023). Uncovering bias in ASR systems: Evaluating Wav2vec2 and Whisper for Dutch speakers. In I. Janssen (Ed.), Proceedings of the 2023 International Conference on Speech Technology and Human-Computer Dialogue – SpeD (pp. 146–151). IEEE. https://doi.org/10.1109/SpeD59241.2023.10314895

Gao, L., Tejedor-Garcia, C., Strik, H., Cucchiarini, C. (2024) Reading miscue detection in primary school through automatic speech recognition. In *Proceedings of Interspeech 2024*, (pp. 5153-5157). ISCA. https://doi.org/10.21437/Interspeech.2024-1180

Georgescu, A.-L., Pappalardo, A., Cucu, H., & Blott, M. (2021). Performance vs. hardware requirements in state-of-the-art automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing, 2021*(28), 1–30. https://doi.org/10.1186/s13636-021-00217-4

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

Gutz, S. E., Maffei, M. F., & Green, J. R. (2023). Feedback from automatic speech recognition to elicit clear speech in healthy speakers. *American Journal of Speech-Language Pathology, 32*(6), 2940–2959. https://doi.org/10.1044/2023_AJSLP-23-00030

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

Hansen Edwards, J., Chan, K. L. R., Lam, T., & Wang, Q. (2020). Social factors and the teaching of pronunciation: What the research tells us. *RELC Journal, 52*(1), 35–47. https://doi.org/10.1177/0033688220960897

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Hinsvark, A., Delworth, N., Rio, M. D., McNamara, Q., Dong, J., Westerman, R., Huang, M., Palakapilly, J., Drexler, J., Pirkin, I., Bhandari, N., & Jette, M. (2021). *Accented speech recognition: A survey* (arXiv:2104.10747). arXiv. https://doi.org/10.48550/arXiv.2104.10747

Hollands, S., Blackburn, D., & Christensen, H. (2022). Evaluating the performance of state-of-the-art ASR systems on non-native English using corpora with extensive language background variation. In *Proceedings of Interspeech 2022*, (pp. 3958–3962). ISCA. https://doi.org/10.21437/Interspeech.2022-10433

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 257–270). Springer. https://doi.org/10.1007/978-3-319-02261-1_19

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

Isaacs, T. (2013). Assessing pronunciation. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., pp. 140–155). Wiley. https://doi.org/10.1002/9781118411360.wbcla012

Isaacs, T. (2018). 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

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Trofimovich, P. (2016). Key themes, constructs, and interdisciplinary perspectives in second language pronunciation assessment. In T. Isaacs & P. Trofimovich (Eds.), *Second language pronunciation assessment* (pp. 1–22). Multilingual Matters. https://doi.org/10.21832/ISAACS6848

Isbell, D. R. (2019). Diagnostic language assessment for L2 pronunciation: A worked example. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching Conference* (pp. 127–140). Northern Arizona University.

Jenkins, J. (2000). *The phonology of English as an international language: New models, new norms, new goals*. Oxford University Press.

John, P., Cardoso, W., & Johnson, C. (2022). Evaluating automatic speech recognition for L2 pronunciation feedback: A focus on Google Translate. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds.), *Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022* (pp. 197–202). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1458

Johnson, C., Cardoso, W., Zuercher, B., Brannen, 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), 10–34. https://doi.org/10.1075/jslp.23033.joh

Kang, O., & Rubin, D. L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology, 28*(4), 441–456. https://doi.org/10.1177/0261927X09341950

Kang, O., & Johnson, D. (2018). The roles of suprasegmental features in predicting English oral proficiency with an automated system. *Language Assessment Quarterly, 15*(2), 150–168. https://doi.org/10.1080/15434303.2018.1451531

Kathiresan, T. (2021). Gender bias in voice recognition: An i- and x-vector-based gender-specific automatic speaker recognition study. In C. Bernardasci, D. Dipino, D. Garassino, S. Negrinelli, E. Pellegrino, & S. Schmid (Eds.), *L'individualità del parlante nelle scienze fonetiche: Applicazioni tecnologiche e forensi* (Vol. 8, pp. 113–122). Officinaventuno. https://doi.org/10.17469/O2108AISV000006

Kerimbayev, N., Umirzakova, Z., Shadiev, R., & Kiv, A. (2023). A student-centered approach using modern technologies in distance learning: A systematic review of the literature. *Smart Learning Environments, 10*(61). https://doi.org/10.1186/s40561-023-00280-8

Knill, K., Gales, M., Kyriakopoulos, K., Malinin, A., Ragni, A., Wang, Y., & Caines, A. (2018). Impact of ASR performance on free speaking language assessment. In *Proceedings of Interspeech 2018*, (pp. 1641–1645). ISCA. https://doi.org/10.21437/Interspeech.2018-1312

Kochem, T., Beck, J., & Goodale, E. (2022). Use of ASR-equipped software in the teaching of suprasegmental features of pronunciation: A critical review. *CALICO Journal, 39*(3). https://doi.org/10.1558/cj.19033

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

Krishnan, A., Abdullah, B. M., & Klakow, D. (2024). On the encoding of gender in transformer-based ASR representations. In *Proceedings of Interspeech 2024*, (pp. 3090–3094). ISCA. https://doi.org/10.21437/Interspeech.2024-2209

Kulkarni, A., Couceiro, M., & Trancoso, I. (2024). Unveiling biases while embracing sustainability: Assessing the dual challenges of automatic speech recognition systems. In *Proceedings of Interspeech 2024*, (pp. 4628–4632). ISCA. https://doi.org/10.21437/Interspeech.2024-2494

Lan, Y. J. (2018). Technology enhanced learner ownership and learner autonomy through creation. *Educational Technology Research and Development, 66*, 859–862. https://doi.org/10.1007/s11423-018-9608-8

Lee, J. Y. (2015). Aging and speech understanding. *Journal of Audiology and Otology, 19*(1), 7–13. https://doi.org/10.7874/jao.2015.19.1.7

Levis, J., & Suvorov, R. (2012). Automatic speech recognition (ASR). In C. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 214–220). John Wiley & Sons.

Li, J. (2022). Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing, 11*(1), 1–63. https://doi.org/10.1561/116.00000050

Liakin, D., Cardoso, W., & Liakina, N. (2017). Mobilizing instruction in a second-language context: Learners' perceptions of two speech technologies. *Languages, 2*(3), 11. https://doi.org/10.3390/languages2030011

Ling, G., Mollaun, P., & Xi, 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

Linville, S. E., & Rens, J. (2001). Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of Voice, 15*(3), 323–330. https://doi.org/10.1016/S0892-1997(01)00034-0

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly, 15*(3), 294–309. https://doi.org/10.1080/15434303.2018.1472265

Liu, C., Picheny, M., Sarı, L., Chitkara, P., Xiao, A., Zhang, X., Chou, M., Alvarado, A., Hazirbas, C., & Saraf, Y. (2021). *Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions* (arXiv:2111.09983). arXiv. https://doi.org/10.48550/arXiv.2111.09983

Liu, Z., Veliche, I.-E., & Peng, F. (2022). Model-based approach for measuring the fairness in ASR. In *Proceedings of ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6532–6536). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9747654

Lochland, P. (2020). Intelligibility of L2 speech in ELF. *Australian Journal of Applied Linguistics, 3*(3), 196–212. https://doi.org/10.29140/ajal.v3n3.281

Loukina, A., Lopez, M., Evanini, K., Suendermann-Oeft, D., Ivanov, A. V., & Zechner, K. (2015). Pronunciation accuracy and intelligibility of non-native speech. In *Proceedings of Interspeech 2015*, (pp. 1917–1921). ISCA. https://doi.org/10.21437/Interspeech.2015-423

Madnani, N., Loukina, A., Von Davier, A., Burstein, J., & Cahill, A. (2017). Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 41–52). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1605

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. (2022). Exploring technologies available for teaching and learning second language pronunciation. In J. Levis (Ed.), *Technological resources for second language pronunciation learning and teaching* (Vol. 3, pp. 1–20). John Benjamins.

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

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Miras, S., Ruiz-Bañuls, M., Gómez-Trigueros, I. M., & Mateo-Guillen, C. (2023). Implications of the digital divide: A systematic review of its impact in the educational field. *Journal of Technology and Science Education, 13*(3), 936-950. https://doi.org/10.3926/jotse.2249

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

Nelson, C., & Cardoso, W. (2024). Evaluating the effectiveness of Microsoft Transcribe for automating the assessment of pronunciation in language proficiency tests. In B. Bédi, Y. Choubsaz, K. Friðriksdóttir, A. Gimeno-Sanz, S. Björg Vilhjálmsdóttir & S. Zahova (Eds.), *CALL for all Languages - EUROCALL 2023 Short Papers*. https://doi.org/10.4995/EuroCALL2023.2023.17007

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy–technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning, 15*(5), 441–467. https://doi.org/10.1076/call.15.5.441.13473

Ngo, T.-N., Chen, H.-J., & Lai, K.-W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. ReCALL, 36(1), 4–21. https://doi.org/10.1017/S0958344023000113

Ngueajio, M. K., & Washington, G. (2022). Hey ASR system! Why aren't you more inclusive? In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *HCI International 2022 – Late breaking papers: Interacting with eXtended reality and artificial intelligence* (Lecture Notes in Computer Science, Vol. 13518, pp. 392–407). Springer. https://doi.org/10.1007/978-3-031-21707-4_30

Nickolai, D. (2024). Quantifying the impact of ASR-based instruction: What does the iSpraak platform learner data show? *The EuroCALL Review, 31*(1), 16–23. https://doi.org/10.4995/eurocall.2024.20221

O'Neill, E., & Carson-Berndsen, J. (2023). *Investigating the sensitivity of automatic speech recognition systems to phonetic variation in L2 Englishes* (arXiv:2305.07389). arXiv. https://doi.org/10.48550/arXiv.2305.07389

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Pérez Castillejo, S. (2021). Automatic speech recognition: Can you understand me? In T. Beaven & F. Rosell-Aguilar (Eds.), *Innovative language pedagogy report* (pp. 122–127). Research-publishing.net. https://doi.org/10.14705/rpnet.2021.50.9782490057863

Saikat, S., Dhillon, J. S., Wan Ahmad, W. F., & Jamaluddin, R. A. (2021). A systematic review of the benefits and challenges of mobile learning during the COVID-19 pandemic. *Education Sciences, 11*(9), 459. https://doi.org/10.3390/educsci11090459

Saito, K. (2021). What characterizes comprehensible and nativelike pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly, 55*(3), 866–900. https://doi.org/10.1002/tesq.3027

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

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

Serditova, D., Tang, K., & Steffens, J. (2025). Automatic speech recognition biases in Newcastle English: An error analysis. In *Proceedings of Interspeech 2025* (pp. 3204–3208). ISCA. https://www.isca-archive.org/interspeech_2025/serditova25_interspeech.html

Sobti, R., Guleria, K., & Kadyan, V. (2024). Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. *Multimedia Tools and Applications, 83*(35), 81933–81995. https://doi.org/10.1007/s11042-024-18753-4

Stolcke, A., & Droppo, J. (2017). Comparing human and machine errors in conversational speech transcription. In *Proceedings of Interspeech 2017* (pp. 137–141). ISCA. https://doi.org/10.21437/Interspeech.2017-1544

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J., & Carmiel, Y. (2020). *WER we are and WER we think we are* (arXiv:2010.03432). arXiv. https://doi.org/10.48550/arXiv.2010.03432

Taylor, T. (2010). Ceiling effect. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 133–134). SAGE Publications. https://doi.org/10.4135/9781412961288.n44

Tejedor-García, C., Cardeñoso-Payo, V., & Escudero-Mancebo, D. (2021). Automatic speech recognition (ASR) systems applied to pronunciation assessment of L2 Spanish for Japanese speakers. Applied Sciences, 11(15), Article 6695. https://doi.org/10.3390/app11156695

Urban, E. (2024, October 25). Language support—speech service—Azure AI services. Microsoft. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/language-support

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. (2017). Using speech processing technology in assessing pronunciation. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 137–152). Routledge.

Vipperla, R., Renals, S., & Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices. In *Proceedings of Interspeech 2008*, (pp. 2550–2553). ISCA. https://doi.org/10.21437/Interspeech.2008-632

Vipperla, R., Renals, S., & Frankel, J. (2010). Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing, 2010*, 1–10. https://doi.org/10.1155/2010/525783

Wang, Y., & Young, S. S.-C. (2014). Effectiveness of feedback for enhancing English pronunciation in an ASR-based CALL system. *Journal of Computer Assisted Learning, 31*(4), 450–465. https://doi.org/10.1111/jcal.12079

Werner, L., Huang, G., & Pitts, B. J. (2019). Automated speech recognition systems and older adults: A literature review and synthesis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63*(1), 42–46. https://doi.org/10.1177/1071181319631121

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. In O. Engwall (Ed.), Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT) (pp. 1–8). KTH Royal Institute of Technology.

Yan, X., & Ginther, A. (2017). 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.

Zielinski, B. (2012). The social impact of pronunciation difficulties: Confidence and willingness to speak. *Pronunciation in Second Language Learning and Teaching Proceedings, 3*(1), 18–26.