5-14-2023

ARTIFICIAL INTELLIGENCE ASSISTED MOTOR LEARNING FOR SPEECH SOUND DISORDERS IMPACTING /ɹ/: THE PERCEPT PROJECT

Nina R. Benway
Syracuse University

ABSTRACT

Approximately 1-2% of the American population enters adulthood with a residual speech sound disorder (RSSD) that impacts the clear pronunciation of speech sounds, most frequently /ɹ/ in fully-rhotic dialects of American English. RSSD is commonly encountered by clinicians, but traditional treatment practices have not been consistently effective in improving an individual's clarity of /ɹ/. Unresolved RSSD may lead to a lifelong negative impact on quality of life.

Recent research has shown that motor-based intervention, which involves high intensity speech practice that adapts in difficulty, can improve production of /ɹ/ even for those who have not responded to traditional treatment. However, not all who might benefit from sufficiently-intense intervention are able to receive this level of clinical service in real-world scenarios, resulting in an *intervention intensity gap* between the traditional intensity of available practice and the practice demonstrated to be therapeutic in recent research.

Computerized therapy with automatic speech analysis might be one way to narrow the intensity gap, but no available system targets /ɹ/ and existing systems, broadly, are likely insufficient for clinical use. The three fundamental issues impacting the development of effective clinical speech technology systems have been the lack of clinically-relevant speech samples for system training, limited technical descriptions of system development, and few empirical assessments of therapeutic benefit for existing tools. Each of these issues are addressed in this dissertation, which describes the development of the PERCEPT speech analysis Engine.

Chapter 1 of this dissertation empirically assesses the benefit of a clinical speech technology system for RSSD in a phase II, multiple baseline single case clinical trial with five participants. Participants in this study received ten sessions of artificial intelligence (AI) assisted motor-based intervention. Practice during nine of these ten sessions was largely conducted by

computerized motor-based intervention software, Speech Motor Chaining, that was driven by perceptual predictions from the PERCEPT Engine. This combined tool is called ChainingAI. Study outcomes were derived from masked expert listener perceptual ratings of /ɹ/ produced by learners throughout no-treatment baseline, treatment, and post-treatment phases. Perceptual ratings of /ɹ/ in treated stimuli were rated as having significantly more rhoticity after ChainingAI than directly before, providing efficacy evidence for ChainingAI. Separately, perceptual ratings of /ɹ/ on untreated words showed significant nonoverlap with ratings from the no-treatment baseline phase indicating a response to the AI-assisted treatment package for three of the five participants. All five participants demonstrated statistically significant improvements in /ɹ/ from pre-treatment to post-treatment, with standardized effect sizes ranging from 0.36-1.6 and a group-average of 30% improvement over baseline accuracy. PERCEPT-Clinician agreement when rating the /ɹ/ in practice attempts (i.e., F1-score) was largely within the range of agreement seen between human clinicians for four of five participants. Exploration of survey data indicated that parents and participants largely felt that computerized intervention could positively impact service delivery for those with RSSD, most frequently mentioning hybrid models in which computerized systems facilitate at-home practice.

Chapter 2 presents a series of supervised machine learning experiments evidencing the technical development of the PERCEPT-R Classifier. The goal of these experiments was to determine the acoustic features that best distinguish clinically correct and incorrect /ɹ/ in word-level audio recordings from children with RSSD, as well as to train a neural network classifier to predict how a human clinician would have rated these sounds. All testing was done with speakers whom PERCEPT has never heard before, which is important for validly estimating accuracy for future use in therapy. To achieve these goals, formant features and Mel-frequency cepstral

coefficient features were extracted from the /ɹ/ within each recorded participant utterance. Shallow and deep neural network classifiers were trained to associate input feature patterns with PERCEPT-R Corpus labels indicating human perceptual judgment of /ɹ/ (i.e., correct/fully rhotic, incorrect/derhotic). Age-and-sex normalized formants outperformed other feature sets. In replicated experiments, the gated recurrent neural network trained on these features outperformed the participant-specific average F1-score from existing literature by 17 points ($\bar{x}$ = .81, σ_x = .10, med = .83, n = 48). An explainability analysis indicated that the age-and-sex normalized third formant was the most influential feature in classifier predictions, aligning with acoustic phonetic descriptions of /ɹ/. Exploration of model performance regarding age and sex of participants did not highlight model bias issues in the current set of participants regarding these demographic variables.

Chapter 3 details the curation of datasets that permitted the training of the PERCEPT-R Classifier. The open-access PERCEPT Corpora contain over 36 hours of 125,000 syllable, word, and phrase utterances. These data come from children, adolescents, and young adults aged 6-24 with speech sound disorder and age-matched peers, and have been published in PhonBank. Sample educational exercises are included with the chapter appendices to emphasize the educational utility of this corpus.

Together, the chapters of this dissertation directly confront three main hindrances to the development of clinical speech technology for RSSD impacting /ɹ/ in American English. This work accelerates the long-term development of paradigm shifting treatment for RSSD through clinician-supervised, AI-assisted precision-treatments that adapt to a child's specific speech patterns. Such tools may ultimately narrow the intervention intensity gap and improve therapeutic and quality of life outcomes for individuals with RSSD.

ARTIFICIAL INTELLIGENCE ASSISTED MOTOR LEARNING FOR SPEECH SOUND DISORDERS IMPACTING /ɹ/: THE PERCEPT PROJECT

by

Nina R Benway

B.A., Cornell University, 2008
M.S., The College of Saint Rose, 2011

Dissertation
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Speech-Language Pathology

Syracuse University
May 2023

ACKNOWLEDGEMENTS

I am indescribably indebted to my advisor, Dr. Jonathan Preston. It may be that Jon's tendency to eschew spotlight and recognition prevents many others from seeing the true depth of his spirit. He is brilliant, diligent, and patient. He is the absolute embodiment of integrity and trustworthiness. Charity and kindness inspire all that he does. May this dissertation serve as a testament to his mentorship and exemplary human character; any errors that remain in the chapters that follow reflect not on him but are, of course, my own doing.

I have the deepest appreciation for my family. This endeavor, like many others I have undertaken, would decidedly not have been possible without unending dedication and tolerance from my caring husband, Jason, supportive parents, James and Patricia, encouraging brother, Austin, and treasured friend, Chris. I have managed this journey by embodying the traits I most admire in each of these role models, and I am grateful for their guiding influence as well as for the support of many friends and family not mentioned.

I am thankful for the time and contributions of my dissertation committee, whose interdisciplinary influences have very tangibly improved this project. Dr. Asif Salekin has provided vital scaffolding regarding machine learning experimental design and interpretation. Dr. Tara McAllister's impeccable insights have strengthened this project at each and every turn. Dr. Soren Lowell has encouraged me to reach as a scholar and think toward future implications of this work. Dr. Victoria Tumanova's feedback has greatly improved the clinical implementation of these ideas. Dr. Carol Espy-Wilson has helped me refine my thinking in both the machine learning and acoustic phonetics spaces.

I could not have undertaken this journey without the Speech Production Lab and our close collaborators. I am indebted to Megan Leece, Nicolle Caballero, Benny Herbst, and Kerry

McNamara for their encouragement, intellectual support, and hospitality. Nathan Preston undertook the enormous task of developing Speech Motor Chaining from its original Excel encoding into a web app, and then interfacing the software with PERCEPT; even though these preparatory details do not take up much space in the chapters that follow, this dissertation would not have been possible without his tireless dedication to this task over several years. Special thanks are also due to frequent co-author Dr. Elaine Hitchcock for her leadership-by-example and encouragement. Yvan Rose and Greg Hedlund have provided persistent support regarding the encoding of PERCEPT Corpora into Phon, which I am thankful for. The written dissertation was greatly improved thanks to proofreading by Ethan Peterson August. I am likewise grateful for the support of current and previous department chairs, Dr. Kathy Vander Werff and Dr. Karen Doherty, and for the assistance of those mentioned in the individual chapters of this work.

My overarching clinical philosophy arises from the person-forward, strength-based teachings of The College of Saint Rose, a community I cherish to this day. I am extremely grateful to Dr. Jack Pickering, Dr. Dave DeBonis, Julie Hart, Dr. Bob Owens, Dr. Deirdre Muldoon, Melissa Spring, Katelynn Carroll, Dr. Jim Feeney, and countless others for their support over the past 14 years. I hope the continuing influences of Dr. Mark Ylvisaker and Dr. Charlene Bloom – Sister Char – are apparent to all who may read these pages.

My research journey began with Dr. Sue Hertz at Cornell University, and I thank her for instilling in me a strong foundation in acoustic phonetics, scientific thought, fine art, and the importance of laughter.

Lastly, this project would not have been possible without the families who have participated in this research, and I am grateful for their commitment to these projects.

nrg

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

LIST OF APPENDICES

PROLOGUE

The number of Americans that enter adulthood with chronic, residual speech sound disorder (RSSD), 1-2% of the American population (Flipsen, 2015), is estimated to be greater than the individual populations of 33 states. This high prevalence of RSSD, most frequently impacting clear articulation of the /ɹ/ sound in American English (Lewis et al., 2015), reflects a condition that is commonly encountered by speech-language pathologists, can be resistant to traditional treatment, and can have lifelong negative impact on quality of life (McCormack et al., 2009; Ruscello, 1995).

Some participants who have not improved their speech following traditional speech therapy have been shown to improve their speech following motor-based intervention (e.g., Benway et al., 2021), an evidence-based practice that is defined by high-intensity, adaptive delivery of therapeutic stimuli to help a learner update incorrect speech movements (Maas et al., 2008). However, access to sufficiently-intense motor-based intervention is hindered by clinician shortages internationally (e.g., Brandel & Frome Loeb, 2011), creating an *intervention intensity gap* between the intensity of traditional practice and the intensity of motor-based intervention demonstrated to be therapeutic in recent research.

Computerized therapy with automatic speech analysis may help narrow the intensity gap, but no available system targets /ɹ/ and existing systems, broadly, are likely sufficient for clinical use (McKechnie et al., 2018). The three fundamental issues impacting the development of effective clinical speech technology systems are the lack of clinical speech samples for system training, limited technical descriptions of system development, and few empirical assessments of therapeutic benefit for existing tools (Chen et al., 2016, Furlong et al., 2017, Furlong et al., 2018, McKechnie et al., 2018).

This dissertation introduces an automatic speech analysis tool, the PERCEPT Engine (*Perceptual Error Rating for the Clinical Evaluation of Phonetic Targets),* while directly addressing these fundamental hindrances to the development of clinical speech technology. Chapter 1 presents empirical evidence of therapeutic benefit following artificial intelligence (AI) assisted computerized motor-based intervention in which the clinician's perceptual judgment is simulated by the PERCEPT Engine. The web app for the motor-based intervention tool, Speech Motor Chaining, is in constant communication with the PERCEPT Server, from which the PERCEPT Engine and PERCEPT-R Classifier are run. The clinical tool that combines Speech Motor Chaining and PERCEPT is called ChainingAI. In ChainingAI, Speech Motor Chaining acts as the clinical "brain" and "voice", deciding what to do next in treatment and giving feedback to the learner, while the PERCEPT Engine acts as the clinical "ears" that listen to the learner's /ɹ/ sounds. This chapter draws upon clinical trial expertise to evaluate how /ɹ/ productions recorded directly after ChainingAI compare to those directly before ChainingAI, how the overall AI-assisted treatment package may lead to perceptual improvement in /ɹ/ productions compared a no-treatment baseline phase, the extent of agreement between the PERCEPT-R Classifier and human clinicians, and community stakeholder perspectives of the role of AI in speech therapy.

Chapter 2 of this dissertation draws upon machine learning experimental design to describe the technical development of the PERCEPT Engine and its neural network, the PERCEPT-R Classifier. The goal of these experiments was to determine the acoustic features that best distinguish clinically correct and incorrect /ɹ/ in word-level audio recordings from children with RSSD, and to train the PERCEPT-Classifier to predict how a human clinician would have rated the /ɹ/ in the audio. All testing is in these experiments is done with speakers

whom PERCEPT has never heard before, which is important for validly estimating accuracy during future use in therapy.

Chapter 3 draws upon data science techniques to describe the curation of the PERCEPT Corpora, which contain over 125,000 clinically-relevant child speech utterances. The PERCEPT Corpora have been published, open-access, to begin to offset the paucity of child speech audio for training of clinical speech technologies outside of the PERCEPT project.

Lastly, the epilogue describes the innovation inherent in this work, as well as ongoing and future research arising from the results described herein.

References

Benway, N. R., Hitchcock, E., McAllister, T., Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual /ɹ/ errors in American English: A single case randomization design. American Journal of Speech-Language Pathology.

Brandel, J., & Frome Loeb, D. (2011). Program intensity and service delivery models in the schools: SLP survey results. Language Speech and Hearing Services in Schools, 42(4), 461-490. https://doi.org/10.1044/0161-1461(2011/10-0019)

Chen, Y.-P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A., Doube, W., & Morris, M. E. (2016). Systematic review of virtual speech therapists for speech disorders. Computer Speech & Language, 37, 98-128. https://doi.org/https://doi.org/10.1016/j.csl.2015.08.005

Furlong, L., Erickson, S., & Morris, M. E. (2017). Computer-based speech therapy for childhood speech sound disorders. Journal of Communication Disorders, 68, 50-69. https://doi.org/https://doi.org/10.1016/j.jcomdis.2017.06.007

Furlong, L., Morris, M., Serry, T., & Erickson, S. (2018). Mobile apps for treatment of speech disorders in children: An evidence-based analysis of quality and efficacy. PLoS One, 13(8), e0201513. https://doi.org/10.1371/journal.pone.0201513

Maas, E., Robin, D. A., Austermann Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. American Journal of Speech-Language Pathology, 17(3), 277-298. https://doi.org/10.1044/1058-0360(2008/025)

McCormack, J., McLeod, S., McAllister, L., & Harrison, L. J. (2009). A systematic review of the association between childhood speech impairment and participation across the lifespan. International Journal of Speech-Language Pathology, 11(2), 155-170. https://doi.org/10.1080/17549500802676859

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for children's speech production: A systematic literature review. International Journal of Speech Language Pathology, 20(6), 583-598. https://doi.org/10.1080/17549507.2018.1477991

Ruscello, D. M. (1995). Visual feedback in treatment of residual phonological disorders. Journal of Communication Disorders, 28(4), 279-302.

CHAPTER 1 - ARTIFICIAL INTELLIGENCE ASSISTED SPEECH THERAPY FOR /ɹ/: A SINGLE CASE EXPERIMENTAL STUDY WITH PERCEPT AND CHAININGAI

Nina R. Benway, Department of Communication Sciences & Disorders, Syracuse University, Syracuse, NY, USA

Jonathan L. Preston, Department of Communication Sciences & Disorders, Syracuse University, Syracuse, NY, USA; Haskins Laboratories, New Haven, CT, USA

Corresponding author:

Nina R Benway, nrbenway@syr.edu, Ph: 1-315-443-3143, Dept of Communication Sciences & Disorders, 621 Skytop Road, Suite 1200, Syracuse, NY 13244

Funding:

This research was supported through an internal grant (CUSE II-14-2021; J. Preston, PI) and computational resources (NSF ACI-1341006; NSF ACI-1541396) provided by Syracuse University.

Conflict of Interest:

We are the developers of Speech Motor Chaining, PERCEPT Engine, PERCEPT Classifiers, and ChainingAI. The design of the PERCEPT Engine is patent pending: US Patent Application No. 63,450,762 (Benway, Preston, and Salekin).

Abstract

Purpose: This phase II clinical trial describes perceptual improvement in rhotic residual speech sound disorder following ten 40-minute sessions of artificial intelligence-assisted motor-based intervention.

Methods: Five participants who were stimulable for /ɹ/ participated in a multiple (no-treatment) baseline, A/B/A, single case experiment. Prepractice activities were led by a clinician and drill-based motor learning practice was automated by ChainingAI, a version of Speech Motor Chaining that receives perceptual ratings from a neural network, the PERCEPT-R Classifier. Study outcomes were derived from masked expert listener perceptual ratings of /ɹ/ from trained and untrained words recorded during every baseline, treatment, and post-treatment session.

Results: Perceptual ratings of /ɹ/ in treated prompts were perceived to have significantly more rhoticity after 30 minutes of ChainingAI than directly after human-clinician led prepractice. Three of five participants showed significant generalization to untreated words during the treatment phase (i.e., nonoverlap) compared to the no-treatment baseline. All five participants demonstrated statistically significant generalization of /ɹ/ to untreated words from pre-treatment to post-treatment. PERCEPT-clinician F1-score was largely within the range of expert clinician agreement, for four of five participants. Exploration of survey data indicated that parents and participants felt hybrid computerized-clinician service delivery could facilitate at-home practice.

Conclusions: This study provides the first evidence of participant improvement for /ɹ/ in untreated words in response to an AI-assisted treatment package. The continued development of this technology may someday mitigate barriers precluding access to sufficiently-intense speech therapy for individuals with speech sound disorders.

Introduction

Difficulty articulating speech sounds is estimated to impact 9% of American children at age nine and persist into adulthood as a chronic, residual speech sound disorder (RSSD) for at least 1-2% of the American population (Flipsen, 2015). The most common RSSD in rhotic dialects of American English, derhoticization of /ɹ/, can be resistant to traditional treatment. Treatment resistance has historically resulted in caseload dismissal without improvement, which can lead to lifelong educational, occupational, and/or social consequences (McCormack et al., 2009; Ruscello, 1995). Recent evidence, however, shows that some children, adolescents, and young adults who have not improved their speech with traditional treatment may respond to motor-based intervention (e.g., Benway et al., 2021; Preston, Leece, & Maas, 2017; Preston et al., 2014). Motor-based intervention contrasts with traditional treatment in that it is operationalized around frequently occurring, adaptive, high-dosage sessions. We believe that treatment resistance, and the relatively high prevalence of adulthood RSSD, can be addressed in part by reducing the *intensity gap* between the lower amount of practice traditionally available to a learner in real-life scenarios and the higher intervention intensity of interventions shown to facilitate treatment response in recent clinical trials.

Computerized motor-based intervention with automatic speech analysis may someday reduce the intensity gap by, for example, increasing access to clinical-grade practice trials and feedback in between sessions with a clinician. To this end, the present study reports a clinical trial of computerized evidence-based practice, in which Speech Motor Chaining (Preston et al., 2019) is controlled by an automated mispronunciation detection engine, the PERCEPT Engine (*Perceptual Error Rating for the Clinical Evaluation of Phonetic Targets)*. This synchronized tool is called ChainingAI. Within ChainingAI, Speech Motor Chaining coordinates the practice

trials and several principles thought to affect motor learning while the PERCEPT Engine processes recorded audio to predict clinician judgment of rhoticity in /ɹ/ (i.e., clinically "accurate" or "inaccurate" /ɹ/ according to an adult fully rhotic American English standard). In this sense, Speech Motor Chaining serves as the clinical "brain" and "voice" (i.e., grouping of practice trials, linguistic complexity of the next practice trial, frequency with which to provide feedback, type of feedback to deliver, and presentation of feedback to the learner), while the PERCEPT Engine serves as the clinical "ears" (i.e., predicting perceptual judgment of recorded practice attempts). The synergy between these two systems is discussed in detail later, including in Figure 3. Because ChainingAI automates all aspects of within-session decision-making in response to learner input, this tool simulates an artificial intelligence (AI) clinician. Note however, that clinician presence in assessment, treatment planning and delivery, and progress monitoring is a necessary ethical guardrail for clinical AI, broadly. The sections that follow detail the rationale for AI-assisted treatment, the ChainingAI treatment methodology, and a single-case experimental design examining perceptual improvement in /ɹ/ in a ChainingAI-assisted treatment program for five participants with RSSD.

Motor-Based Interventions Require Practice Adaptation That May Not Be Readily Available at a Sufficient Treatment Intensity

Motor-based interventions draw their theoretical foundation, the *principles of motor learning,* from schema theory (Maas et al., 2008). The principles of motor learning guide the clinician in adapting a therapy session, based on learner performance, to switch practice emphasis between *acquisition* of a speech sound motor plan (i.e., demonstration of the speech sound motor plan in practiced contexts) to generalized *learning*. Frequent adaptation is required

to maximize treatment outcomes because the principles that facilitate acquisition do not necessarily enable generalized *learning* of a movement (Maas et al., 2008). To use a sports analogy: if the learner spends all their practice time hitting softballs off a tee, they may never be able to to hit a curveball in a game.

As a learner progresses through motor-based intervention and the clinical goal changes from motor acquisition to learning, adaptations of motor learning principles might adjust the linguistic complexity of the stimulus, variability of practice, type of feedback, and frequency of feedback (Matthews et al., 2021; McAllister et al., 2021; Preston et al., 2020). Briefly, when the short-term goal is motor skill acquisition (e.g., establishing the sound at the syllable level), practice should prioritize consistent targets within linguistically homogenous treatment blocks and use frequent, immediate, and detailed feedback emphasizing articulatory positioning (i.e., knowledge of performance feedback, KP; Maas et al., 2008). After the target sound is established with some accuracy at the syllable level, practice may instead emphasize higher levels of linguistic complexity, random practice, and prosodic variation while reducing the amount of KP feedback in favor of summary feedback (i.e., knowledge of results feedback; KR) and no-feedback trials. These adaptations occur within the context of frequently occurring, high-dosage sessions.

Treatment Intensity

Approximately 5,000 effective trials are thought to be needed to generalize a new speech motor plan to continuous speech (Koegel et al., 1986; Koegel et al., 1988). The optimal intensity distribution of these trials is not known, but it is generally thought that more intense treatment improves speech treatment outcomes, particularly regarding *sessions per week*/*dose frequency* (Kaipa & Peterson, 2016; Shields & Hopf, 2023). A review by Hitchcock et al. (2019) similarly

concluded that cumulative intervention intensity, dose, and duration demonstrated a small but significant relationship with treatment outcomes in RSSD trials involving biofeedback. Empirically, Allen (2013) found that more intense dose frequency resulted in significantly higher adjusted mean percent consonants correct versus lower dose frequency in preschoolers with phonological speech sound disorders when cumulative intervention intensity was held constant. Likewise, in a randomized controlled trial of 48 older children with childhood apraxia of speech, Preston and colleagues (under review) found greater improvement in speech sound accuracy on untreated phrases for those randomized to an intensive schedule of Speech Motor Chaining versus a distributed schedule of twice per week (again, when treatment hours were held constant between groups). Overall, this body of evidence lends credence to the prevailing clinical thought that the intensity of treatment is an important consideration in addressing the prevalence of chronic RSSD.

Access to sufficiently intense intervention, however, is often limited. In the United States specifically, caseload size is concerning (Katz et al., 2010) and impacts the treatment frequency recommendations clinicians make for speech sound learners (Brandel & Frome Loeb, 2011). Provider shortages (ASHA, 2018), particularly in rural areas (MacDowell et al., 2010), may exacerbate caseload concerns. Similar stressors are seen in Australia and the United Kingdom (Health Workforce Australia, 2014; Pring et al., 2012; Sugden et al., 2018; Verdon et al., 2011).

Together, these issues point to an international system in which achieved speech-language treatment intensity generally falls short of evidence-based recommendations—creating the intervention intensity gap. While access to service is not well-studied for RSSD specifically, the above factors are likely compounded in the RSSD population, particularly by policy interpretation that wrongfully limits or denies services (Hitchcock et al., 2015) on the grounds

that RSSD does not constitute a sufficient educational impact (Silverman & Paulus, 1989). Insufficient access to service likely limits the amount of practice available for individuals with RSSD, which in turn likely contributes to the prevalence of RSSD.

Clinical Grade Feedback

The delivery of clinical feedback is also operationalized in motor-based intervention. Detailed KP feedback references articulatory movements, which, for /ɹ/, include a complex vocal tract configuration (see, for review: Boyce, 2015; Preston et al., 2020; Tiede et al., 2004). Specifically, to produce fully rhotic American English /ɹ/, clinicians may need to instruct clients to (1) elevate the tip/blade of the tongue, (2) brace the tongue laterally against the posterior molar teeth, (3) keep the posterior tongue dorsum low, (4) retract the tongue root into the pharynx, and (5) slightly round the lips (Preston et al., 2019). This tongue configuration imparts an acoustic speech signal that is perceived as “rhotic”; erred speech motor plans produce a more neutral tongue configuration that typically results in a sound with an insufficiently rhotic quality (i.e., “derhotic”, see Benway, Preston, Salekin, Xiao, et al., under review, for more detail).

It is less apparent in the literature, however, if effective feedback requires veridical judgments of "correct" and "incorrect" in response to a learner's speech sound productions. Veridical feedback is not explicitly codified as a principle of motor learning by Maas et al. (2008), perhaps because it is clinically intuitive that feedback should correctly reflect “accurate perception”. For the phoneme /ɹ/, however, providing veridical clinical feedback is perhaps not so forthright as to warrant this assumption. Previous ratings of clinical /ɹ/ tokens from children with RSSD have had only 85% agreement between expert listeners (Klein et al., 2013). Similarly, in sociophonetic corpus research, raters agree 80%-90% of the time when classifying the rhoticity of tokens from typical speakers of rhotic and non-rhotic dialects (Gupta &

DiPadova, 2019; Nagy & Irwin, 2010). Non-perfect agreement might be explained by evidence suggesting that perception of speech sounds, broadly, is not only grounded in acoustics; research suggests that speech perception is influenced by *expectation* of a sound (e.g., Grill-Spector et al., 2006; Miller, 2016), as well as phonotactic probabilities, phonetic, lexical, and prosodic context (Yi et al., 2019). These factors, particularly internal expectations of a sound's characteristics and perceptual restoration of ambiguous speech input (Leonard et al., 2016), likely contribute to the imperfect agreement in rating the perception of /ɹ/ sounds.

Complicating this issue in the context of the present investigation is that the children for whom independent practice is clinically indicated – those who can occasionally produce fully rhotic /ɹ/ – are those theorized to elicit the least reliable ground truths due to the production of ambiguous/intermediate tokens (Benway, Preston, Salekin, & McAllister, under review). This contrasts with productions from individuals with more salient errors (likely containing more robust derhotic-rhotic endpoints) who would likely not be candidates for independent practice. Indeed, recent ratings of children with speech sound disorder impacting /ɹ/ indicate that productions lacking salience as fully rhotic or fully derhotic may also elicit poor inter-rater reliability among trained listeners. In their study, Li et al. (2023) elicited listener perceptual ratings of the syllable /ɑɹ/ from speakers with RSSD and speakers with typical /ɹ/. Three listeners from a panel of 40 clinically trained listeners rated each of 567 productions using a 100-point visual analogue scale in which 0 was anchored as "incorrect" and 10 was anchored as "correct". Although Li et al., (2023) found the intraclass correlation coefficient (ICC) used to quantify interrater reliability was .90 (95% CI [.89-.91) for all rated tokens, including tokens from typical speakers, the authors report that ICC was poor for productions with an average rating between 2-8 (ICC = .39, 95% CI [.22-.48]). Productions with an average rating between the scale

endpoints—from the feature space between fully derhotic and fully rhotic /ɹ/—are conceptually analogous to the /ɹ/ productions expected from the population enrolled herein. Although these marginal or ambiguous productions are clinically valid and extant to RSSD treatment, it may be that reliability of listener ratings may be lower than in previous RSSD clinical trials in which non-stimulable speakers are enrolled (e.g., Benway et al., 2021).

Computerized Intervention with Automated Speech Analysis Can Address the Intensity Gap

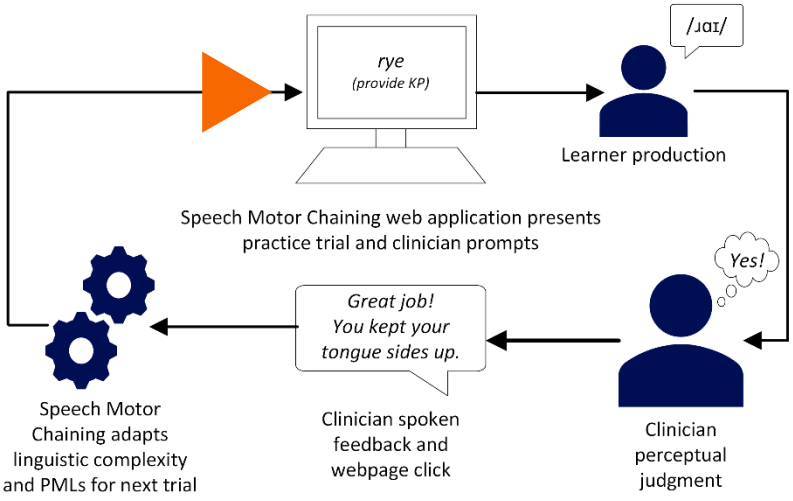
Computerized treatment could increase access to high-intensity, evidenced-based speech sound treatment (McLeod et al., 2020) with the potential to also alleviate the impact of poor audio connections during telepractice. There are no published data rating existing programs for the computerized treatment of /ɹ/ with automatic speech analysis, however, and systematic reviews broadly indicate that the existing literature for computerized speech-language therapy systems is heterogeneous and hard to interpret (Chen et al., 2016), the overall methodological quality of research is moderate to low (Furlong et al., 2017), and that many systems do not show potential for therapeutic benefit (Furlong et al., 2018). Together, these reviews point toward a paucity of high-quality research on computerized treatment that describes both the technical system and the empirical design of efficacy studies in adequate detail (Chen et al., 2016). The inadequacy of the existing computerized treatments contrasts sharply with the potential for technology to offset service delivery barriers for sufficiently intense motor-based intervention. Computerized intervention, with automatic speech analysis, could supplement clinician-led sessions by, for example, permitting those who might have only been able to receive treatment once per week to supplement a clinician-led session with a number of computerized sessions.

Computerized Intervention with Speech Motor Chaining

Since the reviews by Chen et al. (2016) and Furlong et al. (2017; 2018), the Tabby Talks/Apraxia World app for the Nuffield Dyspraxia Program (Hair et al., 2021; McKechnie et al., 2020), the Challenge Point Program (McAllister et al., 2021), and our lab's Speech Motor Chaining (Preston et al., 2019) have begun accruing an evidence base. Generally, these programs are configured by the treating clinician to provide automatic reminders to deliver motor learning parameters that influence learning but are impractical to manipulate manually in real time, including *practice schedule*, *practice variability*, *frequency* and *type of feedback*, and *frequency of self-monitoring* (see: Maas et al., 2008 for a theoretical description and clinical examples of each principle). The linguistic complexity of practice prompts is also adapted in order to maintain functional difficulty around an individual's *challenge point* (Guadagnoli & Lee, 2004; Rvachew & Brosseau-Lapré, 2016). In Speech Motor Chaining, specifically, linguistic complexity is represented by levels of a Chain (i.e., monosyllabic words, multisyllabic words, and phrases built from a common core syllable, which is described in more detail in Methods). Speech Motor Chaining has recently been redeveloped as a web application that is freely available for clinician use (https://chaining.syr.edu), which logs all therapy activity and session data sheets for participant (pseudonyms) that are created by a clinician. General Speech Motor Chaining functionality is shown in Figure 1. Empirical evidence for the efficacy of Speech Motor Chaining has often been coupled with ultrasound biofeedback; however, Preston, Leece and Maas (2017) used a counterbalanced design to show average percent improvement in perceptually correct /ɹ/ was 30.2% [95 CI: 9.539 - 62.861] for six children in seven Speech Motor Chaining sessions with no biofeedback. The group-average treatment effect (Busk and Serlin's d_2) of 4.53 [95 CI: .26 - 8.79] was above the customary threshold for a clinically significant treatment response (Maas & Farinella, 2012). Overall, the several studies evidencing

efficacy of Speech Motor Chaining address previous concerns regarding technical description and clinical efficacy of computerized therapy tools.

Figure 1-1. Existing Speech Motor Chaining Web App Functionality



Note. PMLs = principles of motor learning

Automated Speech Analysis with the PERCEPT-R Classifier

Speech Motor Chaining's strong theoretical foundation, previous evidence of efficacy, and new availability as a web app provide a strong framework for the integration of an automated speech analysis tool. To this end, the recent development of the PERCEPT mispronunciation detection Engine and PERCEPT-R Classifier (Benway, Preston, Salekin, & McAllister, under review) has resulted in an algorithm that predicts clinician judgment of /ɹ/ as correct/incorrect relative to a fully rhotic American English standard. This prediction is possible because the PERCEPT-R Classifier has been trained to recognize the numerical patterns present in formant features from fully rhotic and derhotic /ɹ/. The two features found to most influence PERCEPT-R Classifier predictions were the (age-and-sex normalized) third formant (F3) and

second formant (F2), aligning with acoustic descriptions of fully rhotic American English /ɹ/ (low F3, high F2; Delattre & Freeman, 1968; Espy-Wilson et al., 2000). In lab testing, the PERCEPT-R Classifier achieved a participant-specific F1-score $\bar{x}$ = .81 (σ_x= .10; med = .83, n = 48) for novel speech from stimulable participants with RSSD, while the classifier's performance for the participants in this study is described in detail in a later section. Model cards (Kapoor & Narayanan, 2022; Mitchell et al., 2019) and datasheets (Gebru et al., 2018) for the PERCEPT project have been documented by Benway et al. (in press) and Benway, Preston, Salekin and McAllister (under review). This recent work represents one possible avenue for addressing the intervention intensity gap and increasing access to evidence-based motor-learning practice and clinical grade feedback for /ɹ/.

Purpose and Research Questions

The computerization of Speech Motor Chaining and the technical development of the PERCEPT-R Classifier provide the foundation for clinical assessment of computerized technology that could address the intervention intensity gap. The current investigation, therefore, is a phase II (Robey, 2004) feasibility trial testing the efficacy of a treatment program in which prepractice activities are led by a human clinician and drill-based practice is automated by a version of Speech Motor Chaining driven by clinician perceptual ratings from the PERCEPT Engine. We call this combined Speech Motor Chaining + PERCEPT tool *ChainingAI*, and we consider this to be an AI-assisted treatment package. The primary aim of the present investigation is to evaluate how /ɹ/ productions directly after ChainingAI compare to /ɹ/ productions directly before ChainingAI, and to evaluate how the overall AI-assisted treatment package may lead to perceptual improvement in /ɹ/ productions compared a no-treatment baseline phase.

Research Question 1 examines the ability of ChainingAI to facilitate the acquisition and short-term retention of /ɹ/: does ChainingAI result in near-immediate improvement in /ɹ/ in practiced Chains? We hypothesize that perceptual ratings of /ɹ/ will indicate more rhoticity in practiced Chains immediately after 30 minutes of ChainingAI than immediately before ChainingAI. **Research Question 2 examines the ability of the combined AI-assisted treatment package to facilitate motor plan generalization for fully rhotic /ɹ/ to untreated words: does the AI-assisted treatment package result in perceptual improvement in /ɹ/ on untreated words, compared to a no-treatment baseline?** Previous studies of Speech Motor Chaining alone have provided evidence of a large effect for untreated words after seven hours of Speech Motor Chaining (d_2 = 4.5) (Preston, Leece, & Maas, 2017). We hypothesize that we will similarly observe improvement in untreated words when comparing perceptual ratings of /ɹ/ from the treatment and post-treatment phases to perceptual ratings of /ɹ/ from the no-treatment baseline. **Research Question 3 examines the reliability of PERCEPT-R ratings compared to expert clinician judgment for within-treatment productions: what is the agreement between PERCEPT ratings and expert clinician ratings for /ɹ/ for in-treatment tokens?** The PERCEPT-R Classifier has been lab-validated, but lab validation may be overoptimistic in clinical speech technology (Berisha et al., 2022) and the stimulable participants we recruit to this study may likely have more ambiguous productions than prior lab-validation data for PERCEPT, which included participants who often had more canonically derhotic productions at the time of original study enrollment. We hypothesize that PERCEPT-Clinician agreement (i.e., F1-score) for the present participants will be within the range of agreement seen between human clinicians.

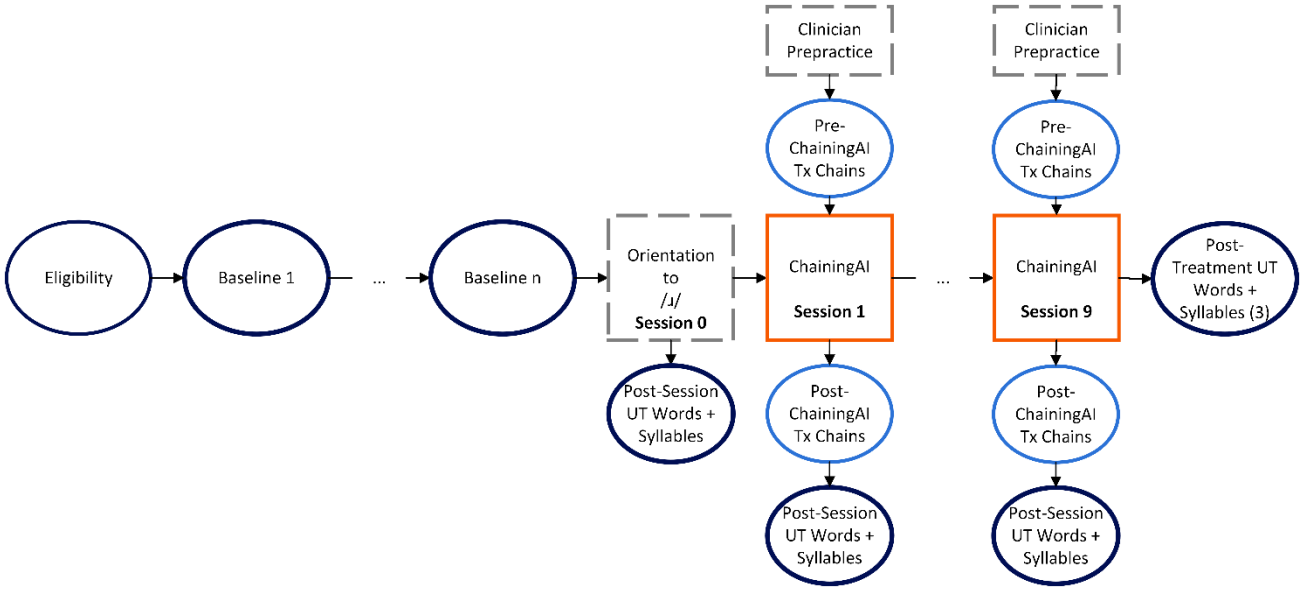
Lastly, **Research Question 4 is a survey-focused exploration** of the parent and participant end-user experience with AI-assisted intervention. We wished to explore what

participants and families thought about computerized speech therapy and its role in the treatment plans of individuals with RSSD.

Methods

This single-case A/B/A multiple baseline experiment with five replications was designed with reference to the What Works Clearinghouse standards (Kratochwill et al., 2010), and reported with reference to the SCRIBE guidelines (Single-Case Reporting Guidelines in Behavioral Interventions; Tate et al., 2016). See Appendix A for a study-specific summary of the SCRIBE methodological details. The participant research experience is summarized in Figure 2 and detailed in the sections that follow. Participants were assigned to between 5-10 possible baseline recordings using concealed start point randomization. Informed consent was obtained from parents/guardians and adult participants, while informed assent was obtained from child participants, in a manner approved by the Institutional Review Boards of Syracuse University and The College of Saint Rose. Treatment materials are available through the project's Open Science Framework page (https://osf.io/nqzd9/).

Figure 1-2. Clinical Trial Design



Note. Outcome measure recordings are represented by ovals and treatment events are represented by rectangles. The length of the multiple baseline phase varied between participants ($5 < n < 10$). *UT* = untreated, *Tx* = treated.

Participants and Recruitment

Recruitment occurred between October 2022 and January 2023, by advertising directly to speech-language pathologists and K-12 school personnel around Syracuse, NY and Albany, NY, as well as university clinics, previous research participants, and regional Speech-Language-Hearing Associations throughout the Northeastern and Mid-Atlantic United States. Advertisements indicated that the study was recruiting participants who could produce /ɹ/ in some syllables/words but not others.

A total of 21 families expressed interest in study participation using a web-based screening form. Three were excluded at the screening stage after self-reporting a history of neurodevelopmental diagnosis (n=2) or orthodontia blocking the roof of their mouth (n=1). 18

participant families were invited to individual consent/assent video conferences. All 15 participants who scheduled the first meeting with the researcher elected to provide consent/assent to continue determining eligibility for the study, and evaluation word reading and syllable repetition lists were recorded to determine /ɹ/ stimulability at baseline (stimulability criteria are discussed in detail later). Nine participants did not qualify for the full eligibility evaluation because they were below the floor criterion for syllable stimulability, while one participant exceeded the ceiling criterion for word accuracy. All five participants who qualified for the full eligibility evaluation were eligible for the treatment phase of the study. These participants were randomized to a baseline length and each completed all treatment sessions, as described later.

Eligibility Criteria

To pass initial screening, participating families reported that possible participants had difficulty producing the American English rhotic /ɹ/ and were within the study's age range as of the date of consent (which overlapped with the age range of participants represented in PERCEPT-R Classifier development: 9; 0–20; 11). Exclusionary criteria included characteristics that might confound therapeutic response, such as a neurodevelopmental disorder (e.g., Autism spectrum disorder, Oppositional Defiant Disorder), permanent hearing loss, and/or first exposure to American English after the age of 3/American English not (one of) the dominant language(s) (e.g., McAllister et al., 2020). Because this study involved independent practice that was operationalized to have minimal involvement from the clinician after the first 10 minutes of the study, a known diagnosis of ADHD/ADD at the time of the eligibility visit was exclusionary. Note, however, that one participant (1112) received a diagnosis of ADHD during the course of their participation in the study, which the family reported to the researcher post-treatment. Previous or concomitant speech, language, and learning difficulties, such as childhood apraxia of

speech, learning disability, dyslexia, history of otitis media and/or temporary hearing loss, were not inherently exclusionary provided that the participant passed the study's inclusionary speech, language, and hearing assessment tasks (described later).

Baseline stimulability for /ɹ/ was the most important eligibility requirement. Stimulable participants with difficulty using fully rhotic /ɹ/ in monosyllabic/multisyllabic words were theorized to be the most appropriate candidates for ChainingAI intervention. Stimulability was operationalized by the researchers as > 20% baseline accuracy on syllable repetition lists (either prevocalic /ɹ/ subsets, postvocalic /ɹ/ subsets, or both). Clinical indication of Speech Motor Chaining was operationalized as < 40% accuracy for /ɹ/ on word reading lists. All participants were initially assessed using the same 100-item evaluation word list, which was balanced according to syllable count, position of /ɹ/ in word, and frontness/backness of the adjacent vowel (as described in the following section). A custom Python script was used to examine word list accuracy ratings by each of these phonological contexts. If the participant's 100-item average accuracy was higher than the 40% inclusionary threshold for words, each combination of phonological contexts was examined to see if there was a permutation of syllable count, position of /ɹ/ in word, and frontness/backness of the /ɹ/ syllable nucleus that would be a candidate for treatment. In these cases, a second, custom, 100-item wordlist was made (e.g., word-final /ɹ/ in two syllable words) and word-level eligibility was re-evaluated. This occurred three times, with two participants meeting the accuracy criterion with more difficult word lists and one participant remaining above the threshold for eligibility. The monosyllable/bisyllable treatment targets and outcome measures selected for all eligible participants, including participants 1121 and 1130 who were eligible based on custom word lists, are described in detail in other sections that follow. All word/syllable list accuracy for eligibility tasks was rated by the first author. All

eligible participants had their eligibility corroborated by an expert research clinician, Megan Leece. There were no disagreements regarding inclusion/exclusion of participants based on these criteria.

The five participants passing the Zoom screening completed the full eligibility evaluation to examine characteristics relative to inclusionary/exclusionary speech-language criteria in more detail. All eligibility visits were required to be in-person, either in the lab or, for participants living more than 75 miles from a study site, in the participant's home. All passed a pure tone hearing screening, bilaterally, at 20 dB HL for the frequencies of 500 Hz, 1000 Hz, 4000 Hz, and 8000 Hz at the in-person eligibility visit, except for one participant whose eligibility visit was conducted in the home who self-reported having recently passed a school-based hearing screening with no hearing changes since that time. A brief oral-mechanism screening confirmed that all evaluated participants could protrude their tongue past their lips and keep the tip of their tongue in contact with their alveolar ridge while lowering their jaw enough for the researcher to see inside the oral cavity, which was theorized to indicate tongue range of motion suitable for /ɹ/ in these stimulable participants. All participants scored within the study's inclusionary range on the Goldman Fristoe Test of Articulation - Third Edition (Goldman & Fristoe, 2015) (< 8th percentile) and the study's inclusionary range of the Clinical Evaluation of Language Fundamentals-Fifth Edition screening test (Wiig et al., 2013) (>= age-based criterion). Participants were required to pass one of two childhood apraxia of speech screenings (e.g., Preston et al., 2021) to rule out major sound sequencing difficulty in these stimulable participants. Participants who did not pass the first screening task with a maximum repetition rate greater than 4.4 syllables per second on the Maximum Performance Syllable Repetition task of the Max Performance Tasks (Thoonen et al., 1996) were required to demonstrate fewer than

three inconsistent productions in the Linguisystems Articulation Test-Normative Update Apraxia Screening (Bowers & Huisingh, 2018) along with fewer than four transcoding errors on the Syllable Repetition Task (Shriberg et al., 2009). All eligibility recordings were made with a shock-mounted Sennheiser MKE 600 super-cardioid microphone and Focusrite Scarlet audio interface. Descriptive information was also collected from participant families regarding speech sound disorder and previous intervention history, but this information was not exclusionary. After confirming study eligibility, participants were randomized to intervention start points, from among the predetermined number of possible baseline visits (5–10). The evaluation sessions between the first baseline visit and the intervention start point consisted of repeated measures of the baseline syllable and word list recordings, representing the no-treatment condition.

All five participants who met eligibility criteria completed the study, including: 5–10 baseline word list recording sessions, 10 treatment sessions, and three post-treatment word list recording sessions. All five participants, four male and one female, 10-19 years ($\bar{x}$ = 12.7, σ_x = 3.6), are reported herein. All self-reported as white, monolingual speakers of American English. Characteristics of enrolled participants are summarized in the Results section.

Probe Word List Stimuli and Treatment Targets

Probe stimuli were selected from a custom list of 2,361 single /ɹ/ words with rhotic phonemes, chosen from 21,315 candidate monosyllable and bisyllable words with rhotics grepped from the LIBRISPEECH speech recognition dictionary. From this custom list, a subset was randomly selected, with replacement, for each participant and each probe timepoint. The length of the word lists, 100 for pre–post word lists, 60 for repeated words lists, was motivated by the intention to phonologically balance the words lists and to create outcome measures long enough to mitigate against practice/learning effects of repeated trials while not being overly

fatiguing for the speaker. The Python script that sampled words from the custom list of 2,361 words was written such that words could be sampled by any of the following phonological properties: syllable length (only monosyllables, only bisyllables, include both); position in word (only word initial, only word final, include both, include only clusters); and characteristic of the /ɹ/-adjacent syllable nucleus (include only front vowels, include only back vowels, include only /ɹ/ nuclei, include all). This phonological information was generated using Phon (Hedlund & Rose, 2019). This sampling procedure permitted us to customize appropriately difficult, phonologically balanced word lists for the participants and mitigate the possibility of participants learning the individual word list items because of the frequency of measurement throughout the study.

Treatment targets were selected for each participant after evaluation but before the baseline sessions to allow for customization of the baseline word lists. Treatment targets were selected individually for each participant based on performance in evaluation word and syllable lists. Phonological characteristics of the participant's treatment targets were determined by examining accuracy in different combinations of the phonological properties described above (i.e., syllable length, position in word/cluster status, and/or adjacent nucleus). This was done to mitigate floor or ceiling effects increase the external validity of the repeated measures. Possible targets were drawn from non-accurate word-level contexts, prioritizing syllables that were stimulable if such contexts were available. The treatment targets selected for each participant, as well as example wordlist items, are discussed in Results.

Personalization of ChainingAI to the Participant's Speech Error Pattern

The probe syllable list stimuli administered during the evaluation, baseline, and "Orientation to /ɹ/" sessions (all described in more detail later) served the dual purpose of

evidencing the no-treatment baseline phase as well as providing examples of fully rhotic and derhotic /ɹ/ upon which to personalize the PERCEPT-R Classifier to an individual's unique pattern of /ɹ/ errors. Although unmasked first-author ratings were not used for reporting of any research outcomes, first-author ratings were used to provide participant-specific ground-truth labels for PERCEPT-R personalization. There was no data leakage between retraining, revalidation, and test personalization datasets (i.e., audio files were confirmed to not repeat across these datasets, which would otherwise represent a threat to validity). Because the experimental design dictated a different number of baselines for each participant, the size of the retraining set differed among participants ($\overline{x}$ = 497.2, $\sigma_{\overline{x}}$ = 242, min = 229, max = 888). The high number of retraining tokens (888) for one participant, 1121, reflects that there were not enough examples of fully derhotic /ɹ/ in his syllable lists, so word list exemplars were rated for his personalization datasets as well.

Personalization was completed in the following manner, one participant at a time. The employed method reflects a less automated version of the overall procedure by which the PERCEPT-R Classifier was initially trained (Benway, Preston, Salekin, & McAllister, under review). Briefly: tokens were extracted from session audio using boundaries manually set within Praat TextGrids and rated by the first author on a binary scale [0,1] to provide a derhotic/fully rhotic label for each production. Formant extraction parameters were set for each participant using the Praat Formant Ceiling values that visually optimized formant tracking through a manual grid search. The first, second, and third formant estimates for entire utterances were extracted from the syllable using custom Python scripts and the Praat "To Formant (Robust)" command with default settings, except for the participant-specific Formant Ceiling setting. Formant transforms were also calculated (F3-F2 distance, the Euclidian distance between the

third and second formants, and F3-F2 deltas, the first derivative of the F3-F2 trajectory). The timestamps of the rhotic-associated interval within the syllable were predicted by a custom implementation of the Montreal Forced Aligner (McAuliffe et al., 2017) embedded within the PERCEPT Engine, using the known syllable orthographic transcript and LIBRISPEECH adult American English acoustic models as adapted to the PERCEPT Corpus. These rhotic-associated interval timestamps were used to extract the formant and formant transforms associated temporally with the /ɹ/ phoneme in the utterance (and rhotic interval extraction occurred after formant estimation to avoid edge-effects issues in Praat). All formant estimates were z-normalized according to age-and-sex mean values for fully rhotic /ɹ/ from a published reference dataset, as shown to improve PERCEPT performance by Benway, Preston, Salekin and McAllister (under review). Because neural networks require all input to have the same number of samples in every dimension, and the number of formant estimates varied across tokens according to the temporal length of the spoken rhotic, formant estimates were standardized into 10 bins. Each bin represented the age-and-sex mean, median, min, max, standard deviation, variance, skew, and kurtosis of the formant estimates and transforms in each decile of the sample. This process resulted in, for each utterance, a three-dimensional feature matrix [5 age-and-sex normalized formants/formant transforms, 10 time windows, 8 aggregate feature statistics] that served as neural network retraining inputs.

The features representing a participant's baseline speech samples were then randomly separated into retraining (70% of utterances per participant), revalidation (15% of utterances), and test sets (15% of utterances). Membership in each of the sets was stratified by the first author's ground-truth rating, which ensured that a participant's derhotic and fully rhotic exemplars were constraining the model's learning at each step of retraining and evaluation.

Participant-specific models were created by fine-tuning the PERCEPT-R gated recurrent neural network within a hyperparameter tuning study facilitated by Optuna (Akiba et al., 2019). Notably, the gradients of the first several layers of the model were frozen and the gradients of the last layers of the model were updated based on the feature space for a given participant's retraining input. For participants 1107, 1111, and 1112, the number of updated layers was set heuristically as the last two fully connected linear layers and the output layer for the model, and the hyperparameters used in the model were fixed as the same hyperparameters from the participant-general PERCEPT-R Classifier. For participants 1121 and 1130, the personalization procedure was updated such that the number of layers with gradients allowed to freely vary was optimized as a hyperparameter through a search facilitated by the Optuna package. For these participants, other hyperparameters were permitted to vary as well. The fine-tuning process and the model accuracy for each participant is summarized for each participant in Table 1, with reported metrics for 1121 and 1130 reflecting the average of 5-fold cross validation strategy used as part of the updated personalization procedure. Model accuracy was rated through F1 score, the harmonic mean of precision and recall. F1-score is the metric used to tune the PERCEPT-R Classifier, which differed from the metric reported previously in the literature—80% agreement with human raters on incorrect sounds (McKechnie et al., 2018)—for two main reasons. First, percent agreement can be misleading in the case of imbalanced datasets, as we expect to have here. Reanalysis of 229,934 previous practice trials from Speech Motor Chaining indicates a 2:1 ratio of correct: incorrect ratings during practice (Preston et al., under review; Preston, Leece, & Maas, 2017). Second, for clinical reasons we wanted to tune the classifier on a metric that would minimize ground-truth rhotics being predicted as derhotic (i.e., false negatives), because (false)

negatives would keep these stimulable participants from advancing to higher levels of linguistic complexity in Speech Motor Chaining.

Table 1-1. PERCEPT Baseline F1-Score Performance

Participant	Out of Box	Personalized
1107	.708 [.72, .28 | .30, .70]	.792 [.82, .18 | .24, .76]
1111	.383 [.19, .81 | 0, 1]	.780 [.73, .27 | .14, .86]
1112	.520 [.24, .76 | .12, .88]	.735 [.71, .29 | .24, .76]
1121	.614 [.83, .17 | .61, .39]	.808 [.69, .31 | .08, .92]
1130	.458 [.03, .97 | 0, 1]	.842 [.71, .29 | .05, .95]

Note. Table entries represent F1-score [true derhotic, false rhotic | false derhotic, true rhotic], with contingency table values normalized by proportion of ground-truth label

Artificial Intelligence-Assisted Treatment Methodology

The treatment phase of the study was designed to contain ten 40-minute visits at a dose frequency of three times per week for ~3.5 weeks. Throughout the treatment, prepractice activities were led by the human clinician, while drill-based practice was led by ChainingAI. Treatment video examples and other treatment resources are available on the study's Open Science Framework page (https://osf.io/nqzd9/).

Orientation to /ɹ/ and Within-Session Clinician-Led Prepractice.

Prepractice refers to the period of motor-based treatment that ensures the participant understands the target speech movement and the articulatory/somatosensory criteria for success (Maas et al., 2008). The same (human) clinician-led therapeutic techniques appeared in each of the 10 treatment sessions, with the duration of prepractice differing between the introductory session and the remaining nine treatment sessions. Automated prepractice was not tested in the current study because we hoped to first informally observe which participants may or may not be

clinically suited for automated treatment as well as to use the study experience to inform the development of the automated prepractice modules.

During the introductory session, prepractice consisted of a standardized “Orientation to /ɹ/” as in previous and ongoing clinical trials (Benway et al., 2021; McAllister et al., 2020). The narrative script and sample images for this session are freely available to clinicians (Preston et al., 2020). In the present study, “Orientation to /ɹ/” prepractice was expanded to include an orientation to the independent use of the ChainingAI website as well as thematic elements related to ChainingAI (i.e., updated illustrations and Bitmoji cartoons of the computerized clinician). As in previous studies, prepractice included unrestricted real-time clinical cueing and clinical elaboration. This elaboration included unlimited, immediate KR and detailed KP feedback provided by the clinician. Visual aids included magnetic resonance images and illustrations representing the complex vocal tract configuration for correct /ɹ/ sounds, including articulatory GIFs. These animations highlight that correct /ɹ/ productions generally have tongue root retraction into the oropharynx while also having a constriction in the anterior oral cavity that is formed by a raised tongue tip or blade and are explained in more detail later. The “Orientation to /ɹ/” script ended with a period of articulatory cueing and shaping, during which the clinician noted facilitative cues that supported the production of a correct /ɹ/ in that participant. Prepractice ended following the elicitation of four correct rhotics for each of the four prepractice syllable targets, or until there was 5 minutes left in the session, at which time a short overview of how to use the website interface and audio recording was completed.

During the remaining nine treatment sessions, prepractice lasted for at most 10 minutes or until 16 correct syllable-level trials were elicited after less than one minute of /ɹ/ tongue shape review. Prepractice was also led by the clinician (e.g., Hair et al., 2021). Prepractice feedback

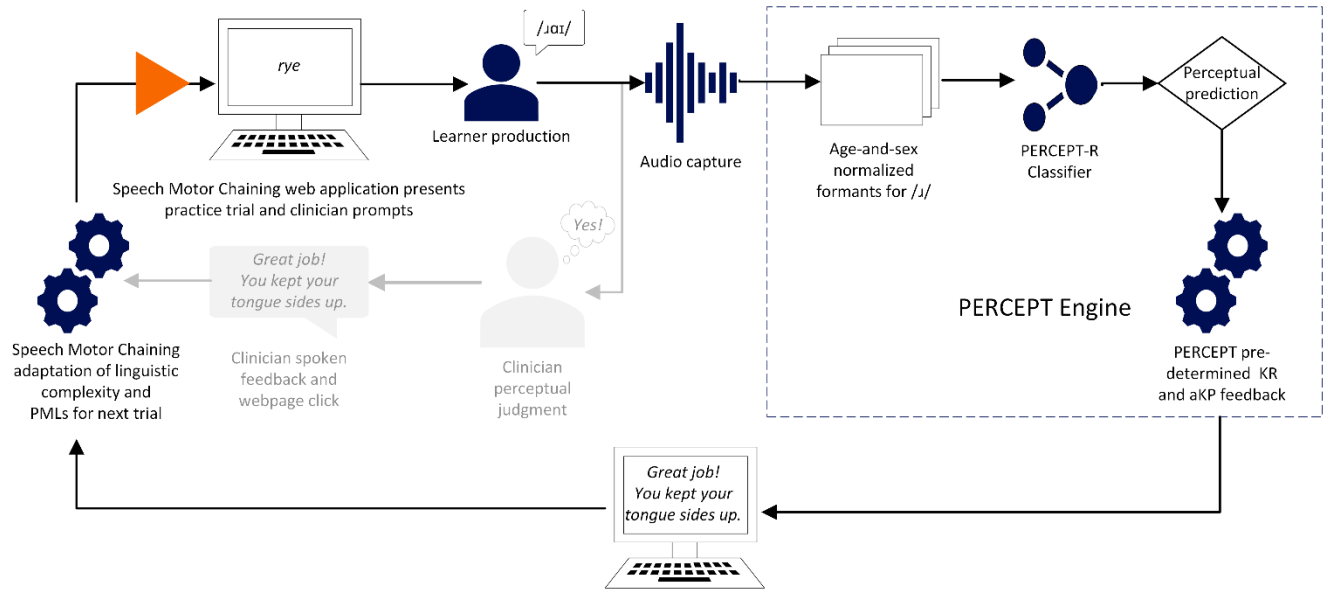
referenced the participant's previous ChainingAI practice (e.g., "The website was saying "not quite" in most of those words because you have an /ə/ sound at the end of your /ɹ/"). Items that were programmed for the non-syllable levels of ChainingAI (i.e., monosyllabic words, multisyllabic words, etc.) were never practiced with the human clinician; however, prosodic variation and nonsense syllables (e.g., /ɑɹθ/) were practiced if a coarticulatory context from a practice target was noted to be difficult. To be considered correct during prepractice, the entire syllable must have been rated as correct (e.g., entirely correct /ɑɹθ/, including preceding and following coarticulatory transitions, not just /ɹ/). On 7/45 occasions participants did not produce 16 correct trials before the allotted prepractice time expired, typically because the clinician was working to shape the coarticulatory transition between an /ɹ/ and consonant. Prepractice was the only human led treatment component, although the treatment sessions were monitored by the research clinician to provide real-time technical support to the participants using the website and record details about the participant's interaction with the website as fidelity monitoring.

ChainingAI

In each of the nine sessions following "Orientation to /ɹ/", the 30 minutes of the session following prepractice was facilitated by the Speech Motor Chaining web app (Preston et al., 2022). This web app is full-stack custom software built upon a C# and SQL Server Database backend and a JavaScript/HTML frontend, accessed through a browser window (https://chaining.syr.edu). The intervention in this study used a prealpha release of ChainingAI. ChainingAI is powered by the PERCEPT Engine and PERCEPT-Classifier, meaning that the role of the clinician was automated using the participant's personalized mispronunciation detection algorithm. Figure 3 illustrates how ChainingAI contrasts with typical Speech Motor Chaining. As discussed, the clinician typically enters their judgment of a learner's individual

practice attempts into the website and responds to website reminders for when to give feedback, what type of feedback to give, when to collect the learner's self-monitoring response, and the linguistic complexity of the next trial. Instead, the PERCEPT Engine uses a participant's personalized mispronunciation detection algorithm to predict the clinician's binary rating of the attempt (i.e., "derhotic" or "fully rhotic") and sends that prediction to Speech Motor Chaining as if it were a clinician mouse click. Speech Motor Chaining then delivers

Figure 1-3. Schematic Showing Clinician Involvement in Speech Motor Chaining (Gray) Replaced with the PERCEPT Engine and PERCEPT-R Classifier

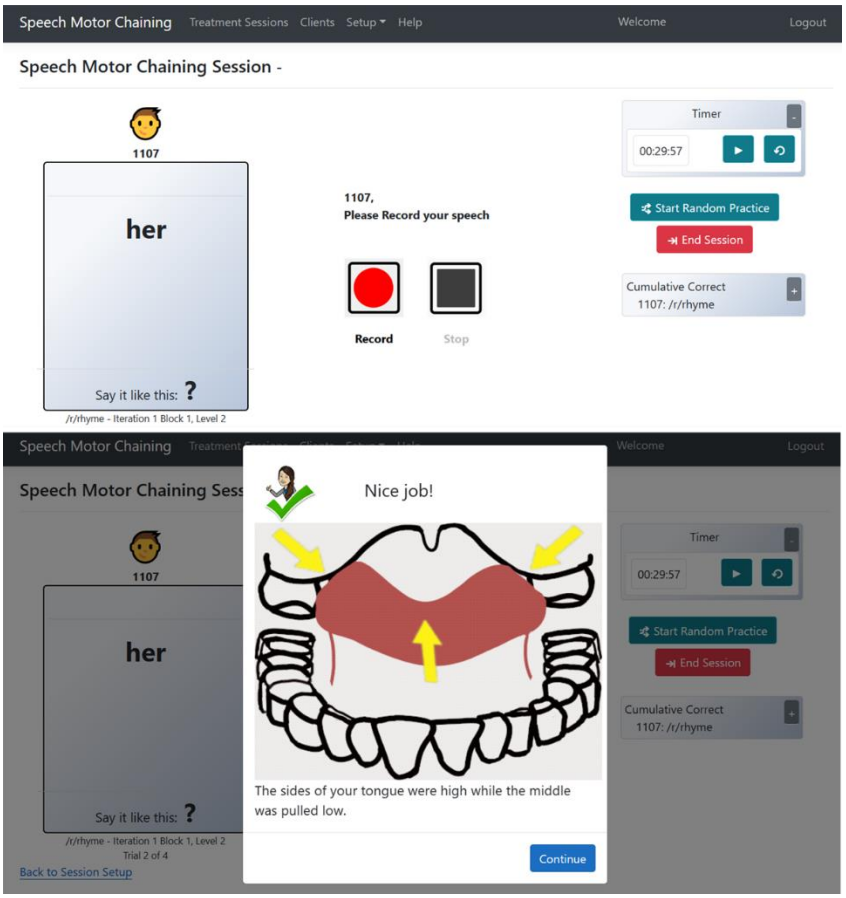


The PERCEPT Engine is packaged with feedback prompts that are, based on PERCEPT's prediction of clinician judgment. With this level of automation, KP is considered *approximate knowledge of performance feedback* because PERCEPT does not (yet) provide utterance-informed articulatory feedback. Instead, aKP represents random selections from a range of general articulatory cues for /ɹ/, much like a clinician may deliver when detailed visualization of a production is not available through e.g., ultrasound biofeedback. KR and aKP

feedback are delivered to the participant through a Bitmoji clinician representing the first author, a written prompt, a GoogleSpeech text-to-speech rendition of the written prompt, and an animation showing the articulatory gestures needing to transition a neutral tongue shape to the rhotic tongue shape determined to be most facilitative for that participant during "Orientation to /ɹ/". The treatment interface design (e.g., Jacko, 2012; Figure 4) was informed by human-computer interaction literature. The animated articulatory gestures incorporate points by Preston et al. (2020) and are freely available on the PERCEPT Project Open Science Framework webpage referenced prior.

Figure 4

Figure 1-4. Examples of ChainingAI User Interface



Note. The top panel shows the prompt and recording buttons while the bottom panel shows approximated knowledge of performance feedback (note, however, the tongue shape movement is animated, and the feedback read to the learner through text-to-speech synthesis).

Chaining Practice

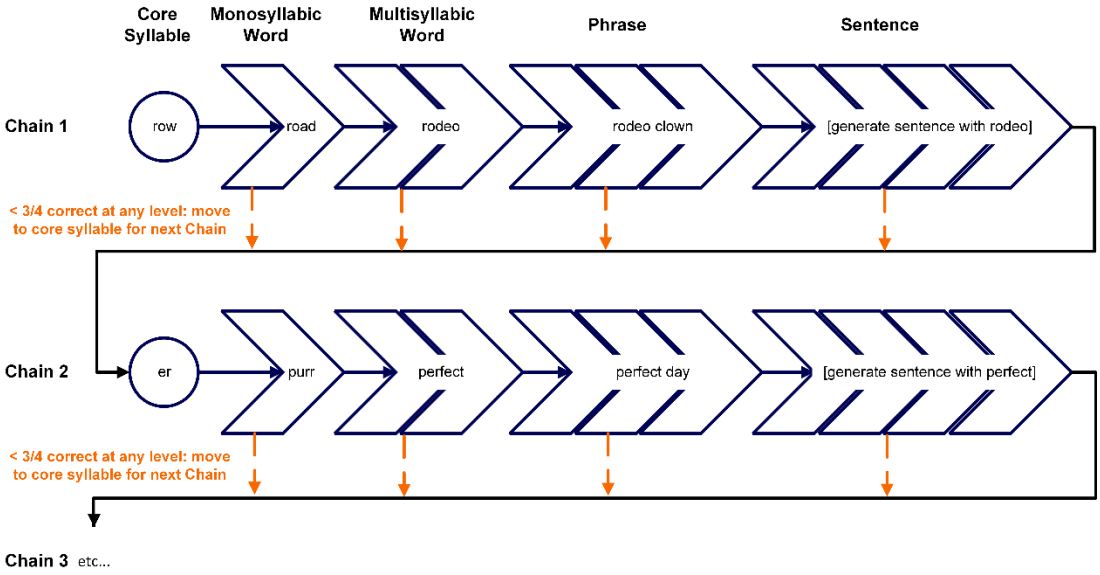
In Speech Motor Chaining, treatment targets are organized into chains that are comprised of blocks representing syllables, monosyllabic words, multisyllabic words, phrases, and sentences, presented in order of increasing difficulty. Common to each chain is the foundational syllable target, such that a sample chain for the target /ɹo/ might include /ɹo/, *road*, *rodeo*, and *rodeo clown*. In the present investigation, chains could represent nucleic /ɝ/, prevocalic /ɹ/, and postvocalic /ɹ/. As in our previous studies (Preston et al., under review), when the participant met the accuracy criterion for a block of sentences, that chain was replaced by a new chain for the same foundational syllable. When two chains for the same sound variant were replaced in this manner, a new /ɹ/ sound target was selected.

After meeting the prepractice accuracy criterion described prior, participants advanced to Chaining Practice following procedures adapted from Preston et al. (2019) with each trial entirely facilitated by ChainingAI. ChainingAI Practice configurations mainly differed from the Speech Motor Chaining Practice default in two related ways: blocks were reduced from six to four trials (under the assumption that participants who could say a good /ɹ/ in some words would benefit from practice that adapted more quickly to higher levels of linguistic complexity), and the criteria to advance to higher complexity was lowered from 5/6 to 3/4.

The chaining hierarchy common to Speech Motor Chaining and ChainingAI is shown in Figure 5. In ChainingAI, the prompt for the current trial is presented on the screen for the

participant to read, with the first prompt per block spoken aloud with GoogleSpeech text-to-speech audio. A heuristically determined proportion of trials at the monosyllabic word level and above were accompanied by a prosodic cue (i.e., loud, fast, question, exclamation, declarative), the prosody of which was also reflected in the GoogleSpeech text-to-speech audio. In this version of ChainingAI, all sentence-level prompts were produced as the same phrase: "I'm saying a sentence with [target]."

Figure 1-5. Adaptive Chaining Algorithm



Note. Productions from lower levels of linguistic complexity are provided with relatively more feedback than productions from higher levels. The relative frequency of aKP reduces in higher levels of linguistic complexity as well.

Following the presentation of the prompt, participants recorded their practice attempts using start/stop recording buttons in the ChainingAI user interface and a study-provided Shure MV5 tabletop cardioid USB microphone. Except when meeting with the researcher for

"Orientation to /ɹ/", participants used their own Windows or Mac computer for the duration of the study and were free to use their browser of choice[1] to access ChainingAI. For each practice trial, the participant's browser packaged the captured audio into a .wav container and uploaded the file to the PERCEPT Engine server at Syracuse University. This audio capture method was designed to circumvent quality and latency issues when streaming audio across a network connection. The PERCEPT Engine processed the audio through the Montreal Forced Aligner and the PERCEPT-R Classifier, then returned a prediction of clinician judgment of rhotic accuracy to ChainingAI. Once the PERCEPT prediction was received, ChainingAI prompted the participant for their self-evaluation (i.e., "correct", "not quite") for selected trials.

Based on the participant's self-evaluation, PERCEPT's predicted accuracy rating, and the feedback type that ChainingAI randomized to this trial (i.e., KR, aKP, or no feedback), ChainingAI delivered clinical feedback to the client. Feedback to the participant appeared as a Bitmoji clinician, a written prompt, a text-to-speech rendition of the written prompt, and, in the case of aKP feedback, an animation showing the articulatory gestures needed to transition a neutral tongue shape to a rhotic tongue shape in either the "bunched" or "retroflexed" configuration. Rotating through aKP that are thought to be effective has been recommended in the context of /ɹ/ treatment without biofeedback, as it is difficult to visually see the aspects of the tongue involved in articulation for the /ɹ/ vocal tract configuration (Preston et al., 2020). The wording of each KR/aKP prompt was customized to reflect the level of agreement between the child's self-evaluation and PERCEPT's prediction (e.g., "I agree, not quite…." or "Actually, I

[1] ChainingAI was observed to work as expected in Mozilla Firefox as well as Chromium-based browsers (i.e., Brave, Chrome, and Edge).

thought that sounded good!"; Guadagnoli & Lee, 2004). As the role of the research clinician was to provide only real-time technical support, no feedback on production accuracy nor commentary on PERCEPT predictions was provided by the clinician to the learner during the sessions.

Random Practice

In the present study, Random Practice began when there were five minutes remaining in each ChainingAI session. Random practice follows a similar prompt, production, feedback framework as Chaining Practice except stimuli in Random practice are presented in random order and, instead of presenting stimuli from all levels of the chains, Random Practice only includes the stimuli from the highest accurate levels achieved during Chaining Practice. Because Random Practice is meant to support generalization, aKP is provided only for chains that did not advance past the syllable level; all other feedback in Random Practice was KR.

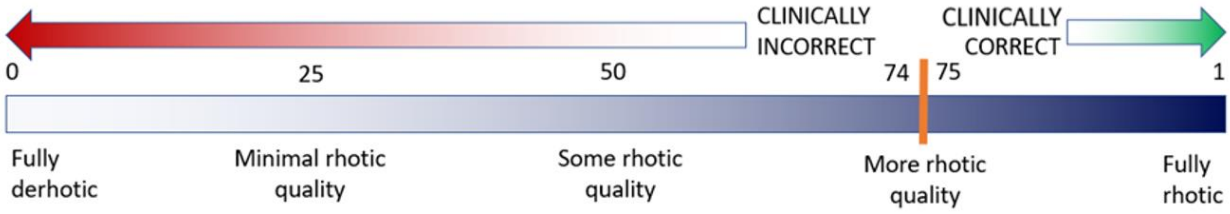
Outcome Measure: Perceptual Judgment of /ɹ/

Each user-captured ChainingAI recording and each production from the repeated word list probes (described later) was judged by three expert raters from a listening panel of licensed speech-language pathologists with expertise in rhotic speech sound disorder intervention. All raters completed training modules and exceeded 90% agreement with expert perceptual judgment for a category goodness task containing words from individuals with speech sound disorders saying fully rhotic and derhotic /ɹ/ (Ayala et al., 2023). All seven listeners met this threshold on the first attempt. The research-clinician was not involved in these ratings.

All ratings were completed using a 100-point visual analogue scale in which raters moved a scroll bar to indicate the amount of rhoticity in the target sound relative to anchor points: "fully derhotic", "minimal rhotic quality", "some rhotic quality", "more rhotic quality", and "fully rhotic" (Figure 6). Scale training included verbal descriptions and audio examples for

each of these anchor points. The scale was also anchored such that 75% delineated “clinically incorrect” and “clinically correct” relative to fully rhotic dialects. This anchor point was known to the raters and was used to collapse continuous scale ratings to binary categories for research questions examining generalization. During the ratings, listeners were masked to participant identity and timepoint of utterance collection. The order of file presentation was randomized. Listeners rated tokens according to a strict “diagnostic” standard to quantify change in /ɹ/ for Research Questions 1 and 2, and a “treatment standard” for agreement with PERCEPT predictions for Research Question 3. The treatment standard was anchored around the question “would you have told the participant that this was correct, within a session?”, and was meant to account for common clinical practice of accepting more marginal productions from participants for reasons such as rapport and not letting a quest for perfect productions stand in the way of “very good”.

Figure 1-6: Outcome Rating Scale



Intra- and inter-rater reliability for ratings are discussed alongside results. A random 5% of tokens were selected for intra-rater reliability. When we analyzed the continuous ratings from the visual analogue scale, we quantified reliability with ICC using the psych package (Revelle, 2019). When we analyzed the derived binary ratings we quantified reliability with Gwet’s chance-corrected agreement coefficient (γ), which has been found to be more robust than kappa

in cases of high or low binary class prevalence (Gwet, 2014; Wongpakaran et al., 2013) using the irrCAC package (Gwet, 2019). Benchmarking of the strength of Gwet's coefficient was completed in the same package, and describes the agreement coefficient and its standard error relative to the coefficient benchmarks of Altman (1990; i.e., poor, fair, moderate, good, very good). All reliability calculations were computed in R (R Core Team, 2013).

The following sections review the timepoints at which the rated recordings were originally collected.

Baseline Phase: Untreated Words and Syllables

Each baseline session consisted of an untreated word list reading probe and syllable stimulability repetition probes. The word lists were between 54-60 items each, with the exact length impacted by the number of phonological contexts selected during the resampling/balancing script described prior. The syllable repetition list did not change from session to session and was 45 items long (each of 15 syllables repeated in order three times; Miccio et al., 1999).

Treatment Phase: Treatment Productions, Treated Chain Retention, Untreated Words, and Syllables

The exact utterances from the participant's list of treated Chains on a given day were recorded after prepractice (directly before ChainingAI) and again following ChainingAI in order to quantify immediate retention of practiced targets. This was meant to isolate any effect of ChainingAI versus the combined human-pre-practice + ChainingAI treatment package. Following each session, untreated word probes and syllable stimulability repetition probes were administered to quantify /ɹ/ acquisition and generalization that occurred up to that point of treatment. The untreated word probes were sampled in the same way as previously described.

Post-Treatment Phase: Untreated Words and Syllables

The probes administered in the three visits in the post-treatment phase of the study were identical duplicates to those obtained during the first three no-treatment baseline probes (i.e., the exact same prompts in the same order; no resampling).

Treatment Fidelity, Achieved Treatment Intensity, and Safety Monitoring

The aspects of the intervention that are thought to be therapeutic are typically quantified during fidelity measures, to ensure that the treatment was delivered as set in the study protocol. In motor-based intervention this would reflect if the principles of motor learning were dosed at the correct frequencies. Because ChainingAI is computerized, however, we expect such fidelity to be 100% (see: Preston et al., 2019 for a review of clinician fidelity with Speech Motor Chaining). Therefore, in the present study, we examined the time it takes the PERCEPT-R Classifier to make a prediction, the number of clinician redirections to the website required by participants, and the frequency of technical errors. Redirection was defined as a verbal cue to attend to the task, for example, in the context of inattentiveness or avoidance. Fidelity results are summarized in Table 2. The average time for PERCEPT to complete a prediction was 3.6 seconds. 99.3% of files returned a PERCEPT prediction in less than 10 seconds; the first file of every session was observed to take longer (~15 seconds) and on two occasions predictions took over one minute (after which the PERCEPT Engine was restarted, and normal operation was subsequently observed). These estimates, however, only include the time the PERCEPT Engine was processing. Extrapolating from the average file size (462,788 Bytes), the average transfer time for recorded utterances would likely fall in the range of 1.12 seconds (56 Kbps dial-up/modem), .02 seconds (DSL/phone line), to < .01 seconds (ethernet and above). Extrapolating further, the average time of the round trip of the data from the user's computer to the PERCEPT

Engine and back would fall between 4.72 seconds on a dial-up connection and 3.61 seconds on an ethernet connection. Lastly, as this was a feasibility study, ChainingAI was programmed to reproduce the entire debugging traceback for the research clinician to examine. Based on these tracebacks, participants were most frequently guided to re-record the word with more pause time between using the recording buttons and speaking. These instances were counted for fidelity under "technical support".

Table 1-2. Treatment Fidelity and Achieved Cumulative Intervention Intensity

Fidelity Item	**1107**	**1111**	**1112**	**1121**	**1130**
Frequency of redirection	0.47%	0.21%	1.00%	0%	0.16%
Frequency of technical support	1.7%	3.8%	1.8%	2.5%	0.80%
Total prepractice productions	126	165	173	152	162
Cumulative intervention intensity	827	576	735	1220	939
Average minutes: seconds spent in practice	24:30	23:22	23:32	26:37	23:21
ChainingAI productions per minute	3.75	2.73	3.47	5.09	4.47

We examined the average dose and achieved cumulative intervention intensity (Warren et al., 2007) for each participant. The average achieved cumulative intervention intensity for the 9 half-hour sessions of ChainingAI in the present study was $\overline{x} = 859.4$ ($\sigma_{\overline{x}} = 241$, min = 576, max = 1220). We also tracked the time spent interacting with the website, because the actual time each participant spent in ChainingAI within the allotted 30 minutes was influenced by the amount of time needed to collect the before ChainingAI word list, log the participant on to the website, and check the participant's microphone levels would not cause clipping. Participant 1121 spent the longest time in ChainingAI likely due to his overall stimulability and the efficiency with which he was able to meet the prepractice accuracy criteria. This stimulability and his dose rate are the two factors likely contributing most to him also having the highest cumulative intervention

intensity. Lastly, we asked parents and participants at the treatment midpoint and during the post-treatment phase if participants had experienced any possible side effects from study participation. No side effects were reported.

Analysis Methodology for Research Outcomes

In this report we focus on outcomes of clinical interest. Readers who are interested in additional technical details for the prospective validation of ChainingAI are referred to Benway and Preston (under review).

Research Question 1: Does ChainingAI result in near-immediate improvement in /ɹ/ on practiced Chains?

We fit (generalized) linear mixed-effects models to examine if ChainingAI resulted in near-immediate improvement in the average clinician perceptual rating of /ɹ/ on practiced Chains. These models were fit with the lmer and gmler functions in the R package lme4 (Bates et al., 2014) following the modeling strategy explained by Harel and McAllister (2019). Fixed effect terms modeled rhoticity across treatment sessions 1-9 (*session*) and rhoticity before/after ChainingAI (*time*), as well as a *session-by-time* interaction. We compared overall model fit for different, generalized linear model families and link functions (e.g., gaussian, binomial, identity, log, probit) by monitoring fit/convergence warnings and distribution of model residuals. We also fit several random effects structures, with the maximal structure including random intercepts for participants and random slopes for *participant-sessions*. As in Harel and McAllister (2019), the random intercepts account for the nested data structure and the random slope accounts for participant-specific trajectories in treatment. Parameters were fit with maximum likelihood estimation which allowed us to evaluate candidate random effects structures through the Akaike information criterion in addition to fit warnings. The lme4 package uses variance components

covariance structure for the random effects model matrix, so the effect of different covariance structures on model fit was not compared.

Research Question 2: Does the AI-assisted treatment package result in perceptual improvement in /ɹ/ on untreated words in post-session probes, compared to a no-treatment baseline?

We primarily used hallmark single-case visual analysis methods to determine if the total AI-assisted treatment package resulted in perceptual improvement in /ɹ/ on untreated words in post-session probes, compared to a no-treatment baseline. We examined perceptual ratings for /ɹ/ with regard to trend, level, stability, and overlap, focusing on the between-condition changes from the no-treatment baseline to the intervention phase (Lane & Gast, 2014). The baseline stability envelope was operationalized as 80% of values falling within ± 25% of the median baseline value (Ledford et al., 2018). Overlap was quantified using the nonoverlap of all pairs statistic (NAP, i.e., Wilcox Signed Rank test), which analyzes pairwise comparisons between baseline and treatment data points and has been found to outperform traditional overlap indices such as percent of nonoverlapping data (Parker & Vannest, 2009). All visual analysis methods were facilitated with the SCAN package (Wilbert & Lüke, 2023) in R.

We quantified pre-treatment to post-treatment effects individually for each participant according to thresholds of statistical and clinical significance. The statistical significance of pre-treatment to post-treatment change was quantified with a linear mixed model containing no fixed effects, random intercepts for participants, and random slopes for time (pre to post). Determining change through this method, also used by Benway et al. (2021), allows for significance testing of individual pre–post change without requiring adjustment for multiple comparisons (Gelman et al., 2012). These linear mixed models were fit with SAS PROC MIXED to quantify participant-

specific slopes and intercepts. Clinical significance was quantified by effect size (Busk and Serlin's d_2; Beeson & Robey, 2006), using a mean-level increase of one standard deviation from pre-to-post as the customary threshold for clinically significant improvement in motor-based intervention research (Maas & Farinella, 2012).

Research Question 3: What is the agreement between PERCEPT ratings and expert clinician ratings for /ɹ/ for in-treatment tokens?

We evaluated the agreement of PERCEPT predictions with human clinician judgment for each participant. The main outcome measure was F1-score, the harmonic mean of precision and recall (i.e., positive predictive value and sensitivity). The F1-score is a measure of overall accuracy in a way that values predicting positives. In addition to being a commonly used evaluation metric in the machine learning literature (and, the evaluation metric used in PERCEPT-R development), we believe emphasis of positive hits is appropriate in the context of ChainingAI, where correct ratings result in (stimulable) participants advancing to more challenging practice contexts. In order to calculate a single F1-score we used the mode of visual analogue scale ratings for a (thrice-rated) utterance, after those ratings had been collapsed to binary ratings of "fully rhotic"/"derhotic" based on that anchor point on the scale. It quickly became clear, however, that this sample of participants elicited lower reliability of ratings between human clinicians than in our previous work, so we contextualized the overall, mode-referenced F1-score with pairwise F1-sore PERCEPT-human and human-human comparisons. These pairwise comparisons illustrated the range of human-human perceptual agreement and how PERCEPT performed relative to that range. For human-human comparisons, the F1-score is reported as an average of two calculations, allowing both humans to serve as ground truth. For

human-PERCEPT comparisons, only human perceptual judgment served as ground truth for F1-score.

Research Question 4: Exploring parent and participant end-user experience

We explored parent and participant end-user experience with this study as part of clinical trial safety monitoring halfway through treatment and at the first post-treatment visit, and explored end-user perspectives on AI-assisted intervention, broadly, using research-generated surveys collected at the first post-treatment visit. Note that we asked our adult participant to complete both the parent and participant surveys. In the present reporting, we focus on stakeholder perspectives and overall opinion of ChainingAI. In line with our previous and ongoing work (e.g., McAllister et al., 2020), we asked: is there anything you would like us to know about how the study may be impacting [the participant/you], positively or negatively? We also asked parents three stakeholder questions: (1) what do you think would be the right balance of clinician-led sessions and computer-led sessions for children with speech sound disorders; (2) how do you think the use of artificial intelligence in speech therapy, generally, would impact daily life for children and young adults with speech sound disorders; and (3) is there anything else we should know about your thoughts on computerized speech therapy? Item 1 was presented as a multiple-choice item (*Person*, *Computer*, *Sometimes a Person/Sometimes a Computer*). Item 2 was presented as a visual analogue scale (0 = *make daily life worse*, 50 = *neutral*, 100 = *make daily life better*). We asked participants one stakeholder multiple choice question: if they would rather have speech lessons from a person or a computer (*Person*, *Computer*, *Sometimes a Person/Sometimes a Computer*). For summary impressions of ChainingAI, we asked participants to tell us the three best/worst things about the website and two related Likert scale questions: how often would you have agreed that the speech app was (1) awesome and (2) terrible?

Results

General Description of Participants at Baseline

Participants (Table 3) demonstrated a variety of rhotic error patterns at baseline. We provide detail on these error patterns under the assumption that it may be relevant in exploring participant profiles that are best served by AI-assisted intervention. Participant 1107 had a context-consistent pattern of /ɻ̩/ derhoticity. Acoustic review of his speech during classifier personalization indicated that this derhoticity was accompanied by a relatively low F2/high F3. Formant ceiling fit was informally judged to be good (5000 Hz). Participant 1111 demonstrated derhoticity marked by slight velarization in prevocalic contexts and underarticulation with some lowering of F3 followed by a schwa-like offglide in postvocalic contexts. Formant ceiling fit was good (5200 Hz). Participant 1112 had good rhotic quality with an appropriately low steady-state F3, but transitions included notable derhotic onglides and offglides. Formant ceiling was set to 5000 Hz, but formant estimates did not consistently track the spectrogram at this nor other formant ceiling values. Participant 1121 had a textbook, fully rhotic /ɹ/ in most monosyllables (whereas monosyllable errors were due to minimal onglide/offglide intrusion), but unstressed /ɹ/ in bisyllables was marked by underarticulation and incomplete lowering of F3, particularly in iambs. Formant ceiling fit was good (5500 Hz). Participant 1130 had a textbook, fully rhotic /ɹ/ in prevocalic contexts but demonstrated atypical derhoticity in nucleic and postvocalic contexts. These productions were characterized by low-back vowelization similar to /ɔˠ/ with a strangled vocal quality and notable facial strain. Formant ceiling fit was good (4500 Hz) and the baseline productions that were reviewed acoustically indicated minimal F3 lowering in the rhotic-associated interval. Overall, these impressions are reflected in the treatment targets and untreated word list items shown in Table 4.

Table 1-3. Participant Characteristics

Participant	Age (Y;M)	Sex	Base-line n	GFTA-3 Standard Score	CAS Screening	Baseline Syllable Accuracy	Previous /ɹ/ treatment	Concurrent Speech Goals
1107	10;7	M	8	53	4.39 syl/s, 1/12 LAT, 1 SRT-a	15/45 (0, .8, 0)	5-6 years	None
1111	11;10	F	10	40	7.29 syl/s	19/45 (.6, 0, .8)	< 6 months	None
1112	11;5	M	5	40	4.07 syl/s, 0/12 LAT, 0 SRT-a	17/45 (.7, .1, .5)	8 years	/θ/, /s/, /l/
1121	10;9	M	6	42	5.44 syl/s	21/45 (.5, .3, .6)	7 years	None
1130	19;3	M	7	40	6.36 syl/s	25/45 (0, .8, .6)	Never	None

Note. Age is reported as years; months. M = male, F = female. n = number. GFTA-3 = Goldman-Fristoe Test of Articulation, Third Edition (Goldman & Fristoe, 2015), LAT = LinguiSystems Articulation Test–Normative Update Apraxia Screening (Bowers & Huisingh, 2018), SRT-a = Syllable Repetition Task number of Additions(Shriberg et al., 2009). Baseline syllable accuracy is reported as n/45, and then percent correct for nucleic ɚ, prevocalic /ɹ/, postvocalic /ɹ/. CAS = childhood apraxia of speech; screening results in syls/s represent multisyllable repetition rate (Thoonen et al., 1996), and participants under 4.4 syl/s had to pass the LAT (inconsistency < 3) and SRT-a (additions < 4). Concurrent speech goals refer to (school-based) speech treatment occurring in parallel with the present study.

Table 1-4. Treatment Targets

Participant	Syllables practiced with clinician	Example ChainingAI Chain	Untreated word list composition	Baseline accuracy in this context	Example untreated word list item	Average pMLU of word list items
1107	/ɛɹ/, /ɑɹ/, /ɪɹ/, /ɝ/	purr, perfect, perfect day	nucleic/postvocalic monosyllables	2.7%	burned	7.45

1111	/ɑɹ/, /ɛɹ/, /ɑɹ/, /ɝ/, /ɔɹ/, /ɪɹ/	fire, firehouse, visit the firehouse	monosyllables	16.9%	spire	7.75
1112	/ɝd/, /ɹm/, /θɹ/, /ɹd/	throw, throwing, throwing the ball	monosyllables	24.6%	frog	7.76
1121	/ɑɹ/, /ɝ/, /ɔɹ/, /ɹk/	mark, marquee, names on the marquee	unstressed /ə/, unstressed /ɹ/ in iambs, monosyllable onsets	33.4%	forbode	9.22
1130	/ɑɪɹz/, /ɝz/, /ɑɹ/, /ɝ/, /ɛɹ/, /ɑɪt/	hers, mother's, mother's day	nucleic/postvocalic (monosyllable and bisyllables)	30.7%	vampires	8.7

Note. Participants who have more than four treatment targets met the proficiency criterion for certain syllables during treatment. Baseline accuracy in this table refers to eligibility determination. Phonological mean length of utterance (pMLU; Ingram, 2002) is provided to contextualize the relative difficulty of outcome measures for a given participant. Syllables listed without a vowel (e.g., /ɹm/) indicate an emphasis on rhotic-consonant coarticulatory transitions and combined multiple vowel contexts, typically of the same frontness/backness.

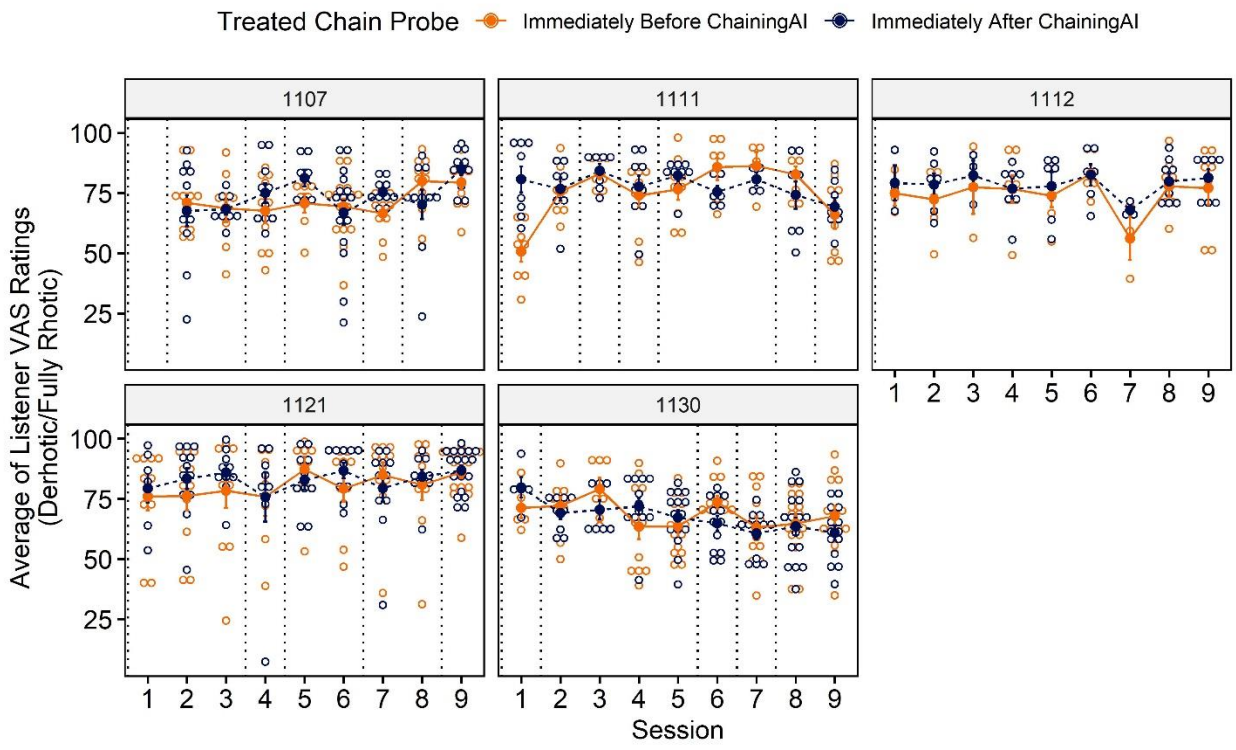
Research Question 1: Does ChainingAI result in near-immediate improvement in /ɹ/ on practiced Chains?

We analyzed 2,388 100-point continuous visual scale ratings for 796 Chaining prompts spoken during probe word list reading. These recordings were collected between human clinician-led prepractice and the start of ChainingAI, and again after ChainingAI to isolate any immediate effect of ChainingAI from human clinician-led prepractice. Inter-rater ICC for this sample was 0.61 (95% CI [.58, .64]; two-way mixed effects, absolute agreement, multiple raters/measurements). Intra-rater ICC averaged .81 ($\sigma_{\bar{x}}$ = .10, min = .67 max = .95; single rater, absolute agreement).

Participant-specific time series visualization are shown in Figure 6 (note: there was data loss for participant 1107 treatment session 1). In these visualizations, instances where the blue

dotted line is higher than the orange dotted line indicate that participant's /ɹ/ productions were (on average) more rhotic following ChainingAI, and instances where the blue dotted line is lower than the orange dotted line indicate that the participant's /ɹ/ productions were (on average) more rhotic immediately after prepractice. The plotted dots illustrate data distribution while the vertical dotted lines delineate when different, treated word lists were introduced after a participant achieved the sentence-level proficiency criterion for a chain in the session prior.

Figure 1-7. Change in Perceptual Rating of /ɹ/ Immediately After ChainingAI



Note. Vertical dotted lines indicate breakpoints for individual word lists, due to participants having met mastery criteria for a chain.

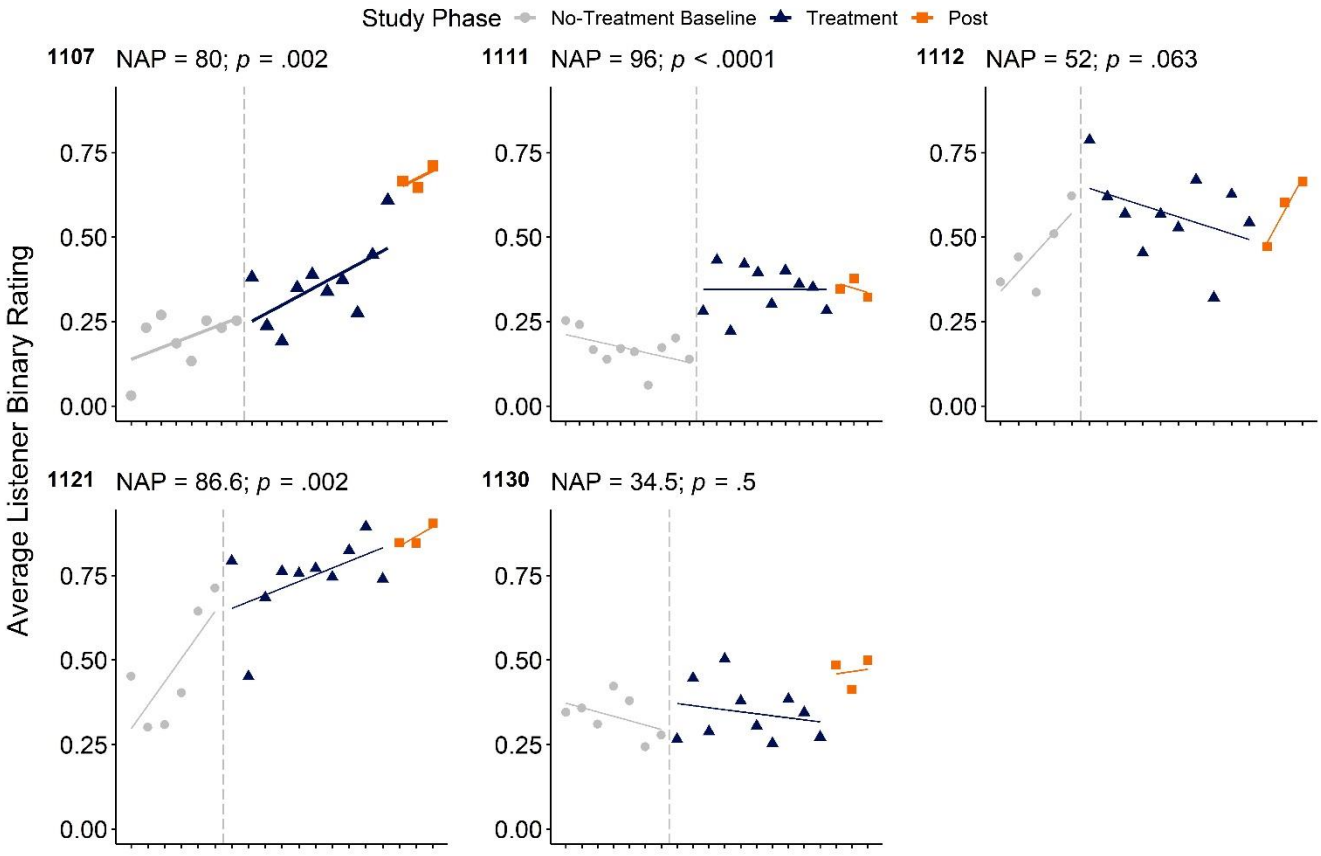
The best-fitting linear mixed model ran to completion with no warnings related to convergence or singular fit. This model was fit with lmer, with fixed effects of *session*, *time*, and a *session-by-time* interaction, and the maximal random effects structure (random intercepts for *participants* and *linguistic complexity*, and random slopes for *participant-sessions*). The fixed effect of session was not significant ($\hat{\beta}$ = .720, *SE* = .49, *df* = 5.87, *t* = 1.46 *p* = .19). The fixed effect of time (before/after ChainingAI) was significant ($\hat{\beta}$ = 6.56, *SE* = 2.35, *df* = 782, *t* = 2.79, *p* = .005), as was the session-by-time interaction ($\hat{\beta}$ = -.95, SE = .39, *df* = 782, *t* = -2.5, *p* = .02). The positive coefficient for the fixed effect of time indicates that tokens recorded after ChainingAI were rated, on average, 6.55 points more fully rhotic on the visual analogue scale than tokens recorded before ChainingAI. The negative coefficient for session-by-time interaction indicates that before/after ChainingAI differences were larger in the early treatment sessions than the later sessions. This may reflect that ChainingAI did not adapt its rating strictness in response to participant progress (although ChainingAI itself does adapt linguistic difficulty, which might explain the non-significant effect of *session*). Overall, we interpret these results as evidence for a (small) therapeutic effect for ChainingAI that was stronger in earlier treatment sessions that were temporally closer to the baseline elicitation of the productions on which PERCEPT-R was personalized.

Research Question 2: Does the AI-assisted treatment package result in perceptual improvement in /ɹ/ on untreated words in post-session probes, compared to a no-treatment baseline?

We analyzed 4,315 rated productions recorded during no-treatment baseline sessions and post-treatment sessions according to their binary "fully rhotic/derhotic" categories on the visual analogue scale. Recall that benchmarking for Gwet's chance-corrected agreement coefficient

was relative to the standards set by Altman (1990). Intra-rater reliability was "good" for binary ratings ($\bar{\gamma}$ = .78, $\sigma_{\bar{\gamma}}$ = .08, .66 ≤ $\bar{\gamma}$ ≤ .88) while omnibus inter-rater reliability was "moderate" ($\bar{\gamma}$ = .37, SE = .006, 95%CI [.36, .39]). Timeseries line graphs showing participant performance on untreated word probes in the no-treatment baseline phase, AI-assisted treatment phase, and post-treatment phase are shown in Figure 7. Quantification of level, trend, and overlap is shown in Supplemental Table 1. In these figures and tables, higher values reflect increased agreement regarding perceptual improvement in /ɹ/.

Figure 1-8. Single-Case Timeseries Data Showing Perceptual Improvement in Untreated Words



Note. Derhotic = 0, Fully rhotic = 1. NAP = rescaled nonoverlap of all pairs between the no-treatment baseline and treatment phases. X-axis ticks represent the sequential order of sessions

within that phase. Note that the first datapoint from the treatment phase was elicited immediately after clinician-led "Orientation to /ɹ/". Table 4 illustrates the linguistic complexity of the words represented by these data points.

Participant 1107 had a 5–6-year history of speech therapy prior to this study. Increases in mean level occurred between all phases, with the highest level in the post-treatment phase. The trend was positive and similar in magnitude in all phases. Two of eight baseline points fell outside of the stability envelope, indicating an unstable baseline. All trends appeared reasonably linear except for perhaps a nonlinear acceleration at the end of the treatment phase. Nonoverlap between treatment and baseline productions was statistically significant ($NAP_{rescaled} = 80$, $p = .002$). Overall change from pre-treatment to post-treatment (Figure 8) was significantly different from zero, and the effect size exceeded the threshold for clinical significance ($\hat{\gamma} = .47$, $SE = .021$, $t = 22.59$, $p < .0001$, $d_2 = 1.6$). Clinically, the participant resolved /ɻ/ derhoticity into a natural-sounding /ɹ/ in the monosyllables tested. Overall, these patterns suggest perceptual improvement in untreated words associated with the introduction of the AI-assisted treatment package, with continued generalization in the week following the culmination of treatment.

Participant 1111 had six noncontinuous months of speech therapy prior to this study. Three of ten baseline points fell outside of the stability envelope, indicating an unstable baseline. The data, however, demonstrate an immediate, abrupt increase in mean level from baseline to treatment. Post-treatment mean level is similar (but slightly lower) than treatment mean level. The trend is linear and negative in baseline and post-treatment phases, and flat in treatment. Nonoverlap between treatment and baseline productions was statistically significant ($NAP_{rescaled} = 96$, $p < .0001$). Overall change from pre-treatment to post-treatment (Figure 8) was

significantly different from zero, but the effect size did not reach our threshold for clinically significant change after the prescribed number of sessions ($\hat{\gamma}$ = .18, *SE* = .020, *t* =8.59, *p* <.0001, d_2 = .69). This participant's productions elicited about-chance agreement between PERCEPT-humans and the lowest agreement between humans ("fair"; described in detail in the next section). Clinically, however, the participant resolved underarticulation in some words. Taken together, these patterns provide the strongest evidence herein for immediate, abrupt perceptual improvement in untreated words associated with the introduction of the AI-assisted treatment package, with gains largely sustained in the post-treatment phase. The flat trend in the treatment phase, however, suggests a clinical plateau. This participant, notably, achieved the lowest cumulative intervention intensity. Her engagement with the website was steady but slow; furthermore, her family's computer was more than 11 years old, and its processing capabilities/intermittent deactivation of the track pad may likely have slowed down the overall pace of the sessions.

Participant 1112 had an 8-year history of speech therapy prior to this study. Increases in mean level occurred from baseline to treatment to post-treatment, with post-treatment levels being largely similar to treatment levels. A positive trend was seen during the baseline phase but only one of five baseline points fell outside of the stability envelope, which met the operationalization for a stable baseline. A negative, possibly nonlinear, trend was seen during the treatment phase. Nonoverlap between treatment and baseline productions was not statistically different from chance ($NAP_{rescaled}$ = 52, *p* = .063). Overall mean level change from pre-treatment to post-treatment (Figure 8) was significantly different from zero, but the effect size was not clinically significant after the prescribed number of sessions ($\hat{\gamma}$ = 0.13, *SE* = .03, *t* = 5.14, *p* = <.0001, d_2 = .36). This participant's positive baseline trend, negative treatment trend, and the

session-to-session variability in the treatment phase complicate interpretation of treatment response. The jump in level between the baseline and the first, human-led, treatment session and the negative trend during the treatment phase may likely indicate that this participant would have made more progress in clinician-led sessions. Note however, that Figure 5 prior shows his speech was consistently rated more rhotic after ChainingAI than after human-led prepractice, so he may have still experienced some benefit from the AI-assisted treatment package that had not yet generalized to untreated words. Informal observations of this participant with the website indicate that his word-initial practice attempts were notably lengthened (i.e., (“thrrrrrrrrrow”), with a notable derhotic onglide followed by good rhoticity. Post hoc exploration indicated the PERCEPT Engine was not considering the entire rhotic attempt when rating, perhaps due to the length of the interval; more information on the technical performance of the PERCEPT Engine in these participants is reported by Benway and Preston (under review). This systematic error resulted in ChainingAI limiting his ability to practice more complex targets, particularly for word-initial rhotics. Of note, the parent of 1112 disclosed a diagnosis of ADHD after treatment concluded; this was also the participant who expressed the most frustration with the website and distraction with Zoom filters. Lastly, observation suggests he may have benefitted from recording hotkeys rather than a trackpad, as he often required several clicks per practice trial to activate the recording device.

Participant 1121 demonstrated an unstable, rising baseline that appears to accelerate toward the end of the phase, with four of six baseline points falling outside the stability envelope. Positive trends were also seen in treatment and post treatment phases. Mean levels increase from baseline to treatment to post-treatment. Nonoverlap between treatment and baseline productions was statistically significant ($NAP_{rescaled}$ = 86.6, p = .002). Overall change from pre-treatment to

post-treatment (Figure 8) was significantly different from zero, and the effect size exceeded the threshold for clinical significance ($\hat{\gamma}$ = 0.39, *SE* = .02, *t* = 17.35, *p* = <.0001, d_2 = 1.30). Clinically, this participant generalized his fully rhotic /ɹ/ in stressed syllables to unstressed syllables. This participant also achieved the highest cumulative intervention intensity in the study, and his in-treatment productions elicited the most reliable responses from PERCEPT and human clinicians. This overall pattern (particularly the rising baseline) suggests this participant was equipped with some self-monitoring abilities at baseline and benefitted from structured times to practice. Interpretation of the full, isolated effect of the AI-assisted treatment package is likely complicated by the rising, unstable baseline and ceiling effects in the untreated word probes.

Participant 1130 never had speech therapy prior to this study and was the only participant with an atypical pattern of derhoticity. The trend was negative in baseline and treatment phases but positive in the post-treatment phase. One of seven baseline points fell outside of the stability envelope, which met the operationalization for a stable baseline. Mean level was nearly identical before and after the start of the AI-assisted treatment package, corroborated by nonoverlap between phases that was not statistically different from chance ($NAP_{rescaled}$ = 34.5, *p* = .5). Overall mean level change from pre-treatment to post-treatment (Figure 8) was significantly different from zero, but the effect size was not clinically significant after the proscribed number of sessions ($\hat{\gamma}$ = .13, *SE* = .02, *t* = 6.13, *p* = < .0001, d_2 = .40). This participant's productions elicited less-than-chance agreement between PERCEPT-humans and only "fair" agreement between humans (described in detail in the next section). Clinically, this participant improved his /ɹ/ from atypical derhoticity to derhoticity marked by onglides and offglides. Informally, we noted that this participant likely advanced out of the feature space on which the classifier was

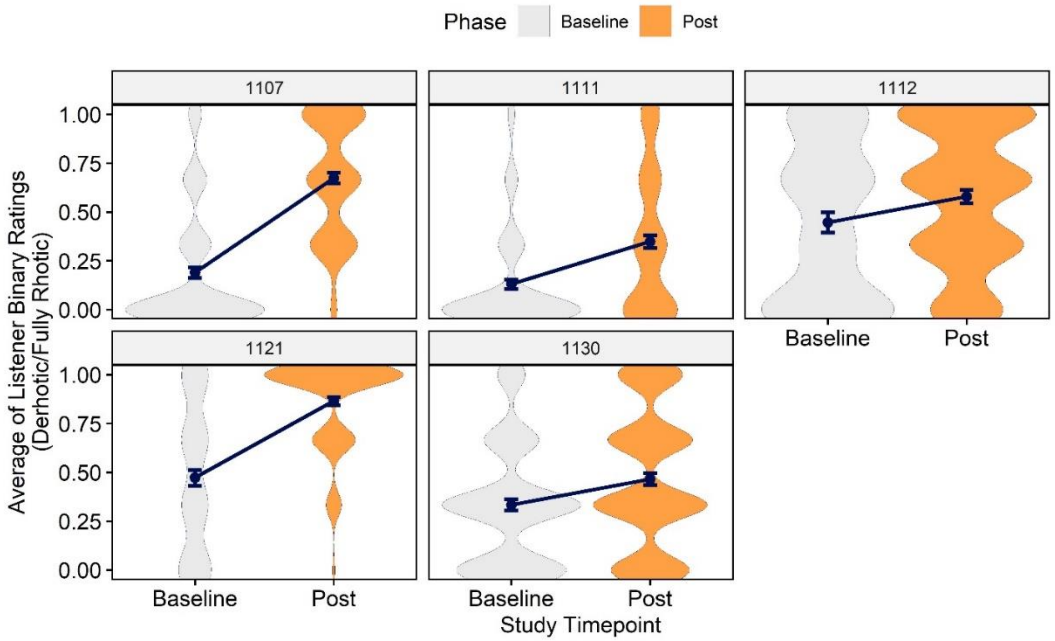
initially personalized, around session 4. In other words, this participant improved such that his baseline "incorrect" tokens became more similar to the exemplars the research clinician labeled as "correct" for PERCEPT retraining. It would be clinically intuitive to adopt a stricter KR threshold to reflect this improvement, but methods for the current study did not specify the re-personalization of PERCEPT-R.

Supplemental Table 1. Single Case Analysis Metrics

ID	Trend			Mean (SD) Level			Nonoverlap
	BL	TX	Post	BL	TX	Post	TX vs BL
1107	.017	.024	.022	.20 (.08)	.36 (.12)	.68 (.03)	72/80, NAP_r = 80, W=8, p=.002
1111	-.007	.000	-.013	.17 (.05)	.35 (.03)	.35 (.04)	98/100, NAP_r =96, W = 2, p < .0001
1112	.058	-.017	.096	.46 (.13)	.57 (.10)	.58 (.11)	38/50, NAP_r =52, W = 12; p = .063
1121	.069	.020	.028	.47 (.17)	.74 (.12)	.87 (.03)	56/60, NAP_r =86.6, W = 4; p = .002
1130	-.013	-.006	.007	.33 (.08)	.34 (.05)	.47 (.05)	35/70, NAP_r =34.5, W = 34.5; p = .5

Note. BL = baseline, TX = treatment, Post = post-treatment. Non-overlap reported as number of BL-TX pairs where treatment point is higher versus all possible BL-TX pairs, rescaled nonoverlap of all pairs value ($NAP_{rescaled}$, ranging 0–100), Wilcoxon signed rank test statistic (W), p value.

Figure 1-9. Amount and Distribution of Pre–Post Change for Mean Listener Rating of Untreated Words



Note. This figure illustrates the distribution of data within the no-treatment baseline and post phases shown in Figure 7. All slopes were significantly different than zero; mean level change exceeded the clinically significant threshold for two participants, 1107 (d_2 = 1.6) and 1121 (d_2 = 1.3). Error bars represent the 95% confidence interval around the mean. Table 4 illustrates the linguistic complexity of the words represented by these data points.

Lastly, we asked parents/participants if there was anything the

Research Question 3: What is the agreement between PERCEPT ratings and expert clinician ratings for /ɹ/ for in-treatment tokens?

We reviewed 3,776 recorded productions from ChainingAI treatment sessions. Recall that these productions were rated by three expert listeners according to a treatment standard illustrated by the question, “would you have rated the /ɹ/ in this utterance as correct during a

therapy session?". Reliability is shown in Table 5. Gwet's chance-corrected agreement coefficient (γ), averaged for inter-human reliability at the level of the participant, was "moderate" ($\bar{\gamma}$ = .41, $\sigma_{\bar{\gamma}}$ = .17). The average participant-specific F1-score compared to the mode of listener binary judgments in the present investigation was .61 ($\sigma_{\bar{x}}$ = .16).

Table 1-5. PERCEPT-R Performance During ChainingAI

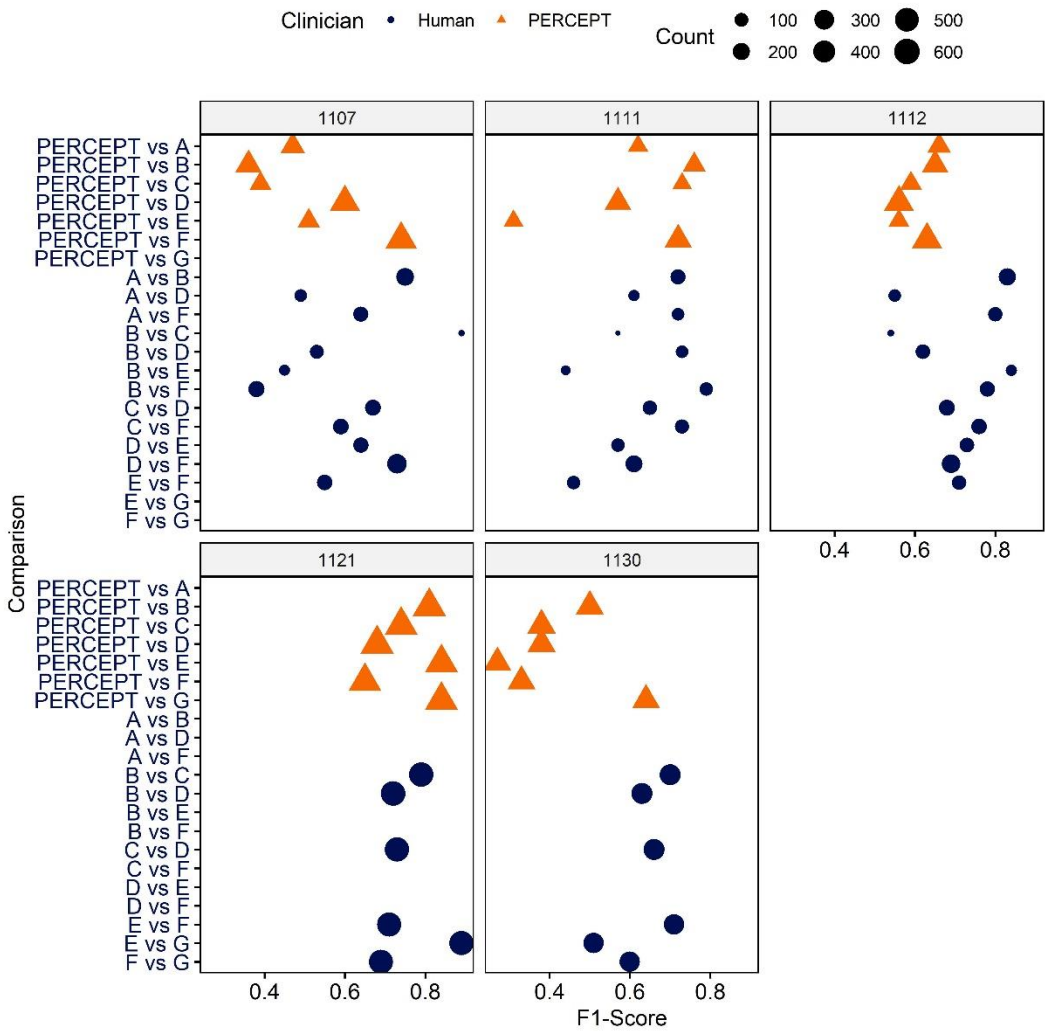
ID	F1-Score and Confusion Matrix	Interhuman Reliability (Gwet's γ)
1107	.55 [.29, .71 | .18, .82]	.23 [.18, .28], Fair
1111	.70 [.27, .73 | .08, .92]	.36 [.30, .43], Fair
1112	.62 [.46, .54 | .33, .67]	.50 [.45, .55], Moderate
1121	.81 [.54, .46 | .15, .84]	.65 [.69, .67], Good
1130	.39 [.18, .82 | .07, .93]	.31 [.26, .36], Fair

Note. F1-score relative to clinician mode; [true derhotic, false rhotic | false derhotic, true rhotic], normalized by ground-truth) and inter-rater reliability of ground-truth measures (γ, 95% CI, benchmark).

Because these results suggest that it made be hard to reliability define a singular "ground truth" among the clinician panel, we also examined pairwise PERCEPT-human and human-human comparisons. All such comparisons are included in Figure 9. Mean F1-score of human-human comparisons was .66 ($\sigma_{\bar{x}}$ = 0.12, range = [.38-.89]) versus .57 ($\sigma_{\bar{x}}$ = 0.16 range = [.327-.84]) for PERCEPT-human comparisons. Average intra-rater reliability for these data (across all participants) was "good" ($\bar{\gamma}$ = .75, $\sigma_{\bar{\gamma}}$ = .08). PERCEPT-human performance fell entirely within the range of human-human performance for participant 1112. PERCEPT-human performance was within the human-human performance range for all but one PERCEPT-human comparison for participants 1107, 1111, and 1121. For one participant, 1130, PERCEPT-human performance was lower than the range of human-human performance for all but one PERCEPT-human

comparison. The number of false positives for this participant indicate that PERCEPT was too permissive.

Figure 1-10. Pairwise Performance (F1-Score) of Raters for In-Treatment Productions



Note. A-G represent human expert listeners.

Research Question 4: Exploring parent and participant end-user experience

Parents and participants: Is there anything you would like us to know about how the study may be impacting [the participant/you], positively or negatively?

Three parents volunteered that they perceived functional improvement in their child's /ɹ/ sound during study, citing increased confidence, self-monitoring, clarity of speech, and noticeable carry over; one parent did not answer this question. Two participants volunteered that they perceived improvement in their /ɹ/ production, citing increased intelligibility when speaking with their parents, and that the study helped "a lot". Three participants indicated there was nothing they wished to share in response to this question.

Parents: What do you think would be the right balance of clinician-led sessions and computer-led sessions for children with speech sound disorders?

Three of the four parents (plus the adult participant himself) indicated that computer-led sessions had some place in treatment for children with speech sound disorders (multiple choice selection: *Sometimes a Person/Sometimes a Computer*). These individuals identified that computerized components would be especially useful for practice between sessions with a clinician. The adult participant elaborated: *if the technology improved[,] I would much prefer speech lessons from a computer because it would allow me to practice anytime without having to schedule in advance and would allow me to spend as much time per week as I wanted practicing*. One parent, of participant 1111, indicated they would rather have speech lessons from a person, expressing preference for human connection in all speech therapy interactions.

Parents: How do you think the use of artificial intelligence in speech therapy, generally, would impact daily life for children and young adults with speech sound disorders?

Three of the four parents, plus the adult participant, indicated they foresaw a neutral-to-positive impact on daily life for children and young adults with speech sound disorders. Responses ranged from 50 (*neutral*) to 99 (*make daily life better*) on the visual analogue scale,

with an average response of 83.5. When asked to elaborate, themes that arose included accessing speech therapy without stigma and/or in a lower-pressure environment, increasing access to services, the benefit of repeated/home practice, and only using computerized treatment with those it is a good fit for. One parent omitted an answer to this question on the computerized survey.

Parents: Is there anything else we should know about your thoughts on computerized speech therapy?

One parent provided a response, indicating that it was "hugely helpful" to have the clinician review PERCEPT's predictions from the previous session at the start of the following session. The parent felt that this alleviated their child's frustration with the computer letting them know their production was not quite right.

Participants: Would you rather have speech lessons from a person, a computer or sometimes a person/sometimes a computer?

Three of the child participants indicated they would rather have a balance of person-led and computer-led sessions (multiple choice selection: *Sometimes a Person/Sometimes a Computer*). As one of our participants put it: *the person makes you good, and the computer tells you how good you are.*

One child participant indicated a full preference for person-led sessions (multiple choice selection: *Person*), explaining *being in-person is a nice experience for everyone, even if it's a bit of a drive*. As a note, this participant was homeschooled.

Participants: How often was the speech app awesome/terrible?

These questions were rated with a Likert scale: never (1), sometimes (2), often (3), always (4). The average participant response to "the app was awesome" fell between

"sometimes" and "often" (2.6). The average participant response to "the app was terrible" fell between "never" and "sometimes" (1.4).

Participants: What were the three best/three worse things about the website?

A variety of themes were cited as one of the three best things about the website, including: *it was easy to use, nothing was wrong with it, helped me get better at /ɹ/, the computer said the sound and also the prompt, I liked my emoji and the drawings of the clinician, it was accurate, it adjusts to the learner,* and *it offers multiple difficulty levels.* Some of the same themes were repeated for the three worst things, including: *it was sometimes slow; I didn't always think it was accurate; it would always tell me I was wrong; the [prosody prompts] were confusing; I wasn't able to use it on my own outside of the study, view my progress, or choose which sounds I wanted to work on; it told me I was correct too frequently*, and *it only said "correct" or "not quite", never "in-between".*

Discussion

This study aimed to compare perceptual improvements in rhoticity of /ɹ/ following an AI-assisted speech therapy package in children who could produce the /ɹ/ sound some of the time. Separate research questions examined the immediate effect of ChainingAI on listeners' perception of /ɹ/ production improvements in treated words, the ongoing effect of the entire treatment package on /ɹ/ perceptual improvement in untreated words, the agreement of PERCEPT-R ratings with clinician ratings, and stakeholder perspectives regarding computerized treatment with speech analysis for speech sound disorders. The goal of this line of research is to promote the development of an evidence-based AI tool that provides clinical-level practice and

feedback to learners at home, between visits with a clinician, in order to narrow the existing intervention intensity gap.

Study data support our first hypothesis: masked listeners rated /ɹ/ in practiced chains to have significantly more rhoticity after ChainingAI than directly after human-clinician led prepractice, before ChainingAI. The overall measured size of this effect was small, around 6 percentage points on a 100-point scale. Complicating interpretation of ChainingAI's impact, however, is the clinical observation that speech improvement isn't necessarily linear, especially as participants learn to bring speech motor plans to the level of conscious control that allows the practice of different vocal tract configurations (that are sometimes less ambiguously derhotic). Furthermore, the principles of motor learning that form the basis of ChainingAI de-emphasize in-session accuracy in favor of long-term, generalized learning. As this was primarily a feasibility study, future between-group studies can supplement these within-subject data to more fully illustrate the effect of ChainingAI.

Study data also provide support for our second hypothesis, showing increases in perceptual ratings of /ɹ/ on untreated words after the introduction of the AI-assisted treatment package compared to the no-treatment baseline phase. An immediate and abrupt increase in level and trend in perceptual rating of /ɹ/ in untreated words, accompanied by significant nonoverlap, is particularly compelling for one participant: 1111. Holistically, all participants demonstrated a raw mean-level increase from baseline to treatment, with significant nonoverlap for three of the five participants indicating a treatment response to the combined package. Furthermore, comparisons between the post-treatment phase and the no-treatment baseline indicate a statistically significant increase in rhoticity for five participants (including the two participants whose treatment phase performance largely overlapped with the no-treatment baseline). This

change reached our threshold for clinical significance for two participants. Note, however, that rising and unstable baselines warrant cautious interpretation of study results for participants other than 1111. Even though this research question evidences the entire AI-assisted treatment package and not the isolated effect of ChainingAI, the fact that these improvements were seen following clinician feedback on ~25 syllable level practice attempts thrice per week further bolsters the findings of Research Question 1 and suggests that ChainingAI did have some influence on overall treatment progress. The average raw improvement of 30.0% is almost identical to that seen in a previous study of clinician-led Speech Motor Chaining (Preston, Leece, & Maas, 2017). Individual effect sizes from pre- to post-treatment, however, were lower in the present study. This is likely because the present study excluded participants with low accuracy and low baseline variance, resulting in the same raw percent change calculating to a smaller effect size in the present study due to the impact of higher variance at baseline on the effect size statistic.

Study data also support our third hypothesis, showing that PERCEPT-Clinician agreement (i.e., F1-score) was largely within the range of agreement seen between human clinicians for four of five participants. However, the range and amount of agreement between well-calibrated expert raters surprised us. Low interrater reliability was previously seen in marginal /ɹ/ tokens by Li et al. (2023), but note that meaningful reliability comparisons between their study and the current study cannot be made because the ratings by Li et al. (2023) also include tokens from typical speakers. Reliability levels throughout the study were lower than in our previous studies using the same general methodology for /ɹ/ ratings (e.g., Benway et al., 2021, and in our unpublished pilot work with ChainingAI). This supports our intuition that the study inclusion criteria selected participants with more ambiguous /ɹ/ productions than in previous clinical trials (which typically exclude participants above a certain level of accuracy at

baseline). The out-of-box performance for the PERCEPT-R Classifier in Table 1 (specifically regarding the prevalence of false positives) might even suggest that these participants had more ambiguous feature spaces than the average participant in the PERCEPT-R Classifier validation and test sets, where average out-of-box F1-score performance was .76 (Table 5; Benway, Preston, Salekin, & McAllister, under review).

Even in the context of the wide variation of F1-score for the present study—including PERCEPT predictions falling near chance for two participants when compared to the mode of human clinicians—all participants demonstrated statistically significant improvements in rhoticity in untreated words after ten 40-minute sessions assisted with ChainingAI. It is likely that the variables impacting treatment outcomes operate on multiple timescales. This finding raises questions about what aspects of the prompt-production-feedback dosing structure—and latent variables that interact with it (such as motivation, attention, and self-monitoring)—are most important for the long-term arc of treatment progress. For example, the participant for whom PERCEPT performed around chance compared to the mode of clinician ratings is one of the two participants who had the highest accuracy at post-treatment (with the other highest-accuracy participant having the best PERCEPT performance). Further work with additional participants can investigate if interesting paradoxes may be at hand: that those with occasionally/marginally correct /ɹ/ who are most suitable for independent practice may also be most prone to low rating reliability, but stimulability (and, perhaps self-monitoring capacity) might bestow the potential for improvement even in the context of feedback that does not always match the judgment of any one clinician.

Lastly, exploration of survey data indicates that parent and participants largely feel that computerized intervention can positively impact service delivery for children with speech sound

disorders, most frequently mentioning hybrid clinician-AI models in which computerized systems facilitate at-home practice. Future survey research on this topic, however, might expand stakeholder polling beyond a self-selected group of people who would seek to enroll in a research study with computerized speech lessons. Even so, comparison of responses herein indicates that participants have differing views on ChainingAI, which supports clinical intuition that automated treatment may not meet everyone's personal preferences or speaker profile. Future studies can elucidate the social, emotional, and motivational preferences that make a learner a candidate for computerized treatment, and our ongoing work will adapt the ChainingAI interface into an interactive game for participants. The specific feedback the participants provided about the ChainingAI interface will guide the ongoing development of the tool.

Clinical Implications

This study provides evidence that, for some children, improvement in /ɹ/ production can occur in response to ~25 human-led practice trials with ~30 minutes of supplementation with an AI clinician, thrice per week. This supports the feasibility of our long-term objective to use AI-driven speech therapy to help remedy the intensity gap while also remaining within our own ethical guidelines that treatment must always be overseen by a clinician. Larger scale treatment studies will be necessary to in pursuit of this goal.

Because this study employed single case experimental design, we provided detailed clinical interpretation in the results section prior. Although few in number, the participants who were eligible for this study represented a wide range of error patterns that permits some general speculation about which clinical profiles may be most appropriate for AI-assisted intervention. For example, the participant who approached the ceiling of measurement on untreated words, with the highest PERCEPT-clinician agreement, is the participant who, at baseline, could

produce a fully rhotic /ɹ/ in all stressed syllable contexts and was working to minimize underarticulation in unstressed syllables. It may be that generalization to unstressed syllables might be an appropriate target for AI-assisted practice. This participant also showed some perceptual improvement just from participation in the baseline word lists, indicating sufficient self-monitoring abilities at baseline might be an important skill to maximize gains in (AI-assisted) independent practice. Participants who plateaued in treatment may benefit more from a mispronunciation detection algorithm that is retrained during the course of treatment to reflect changing speech performance.

Conversely, the participant experiencing the least effect (i.e., nonoverlap) between the combined treatment package and no-treatment baseline had an atypical pattern of derhoticity with no previous history of speech therapy. He was also older than the most frequently occurring ages in the training dataset for the PERCEPT-R Classifier. Similarly, the participant whose family eventually disclosed a diagnosis of ADHD had the most variable response and expressed frustration. It is likely that ChainingAI in its current form is least useful for novice learners or those who would require motivational/attentional support during practice. Lastly, it may also be that mispronunciation detection does better with underarticulation or derhoticity versus onglides and offglides, due to the way temporal salience interacts with the underlying speech technology. Future studies might seek to better identify the profile of participant, the timepoint in the overall arc of treatment, and the acoustic manifestations for which AI-assisted treatment is most appropriate.

Limitations and Future Directions

Although there are many strengths to this novel study, there are also limitations. The largest two limitations are the lack of geographical, racial, and ethnic diversity in the clinical

sample and the observed inter-rater reliability among our well-trained expert listeners. First, our ongoing work sets a long-term goal of validating the ChainingAI with a diverse set of speakers for whom a fully rhotic American English /ɹ/ is dialect appropriate, beginning with nationwide data collection to directly addresses racial and ethnic underrepresentation in the training data for the PERCEPT-R Classifier. These concurrent projects specifically aim to increase the representation of speakers of fully rhotic dialects of American English who identify as Black, Indigenous, Hispanic, Asian/Pacific Islander, or multi-racial. Second, low rater reliability that may be extant in the case of marginal productions from stimulable speakers that fall between fully derhotic and fully rhotic (Li et al., 2023). These speakers and productions, however, are clinically valid, and may represent a unique challenge to the interpretation of clinical speech technology efficacy studies, in general, when viewed from the position that AI-assisted treatment is most appropriate for speakers with emerging productions of a target sound. This challenge is also reminiscent of previous investigations of perceptual severity rating in speech sound disorders, broadly (Flipsen et al., 2005). To address this limitation, our ongoing work continues to refine scale and training characteristics that maximize reliable ratings for stimulable speakers and marginally-rhotic productions. In fact, it may be (clinically) more viable to evaluate PERCEPT based on one "gold-standard" clinician's ratings, than the mode of "tin-standard" panel ratings, and this methodological difference may partly explain PERCEPT performance disparity between Table 1 (F1-score when predicting one clinician's ratings, immediately after baseline) and Table 5 (F1-score predicting the mode of listener ratings, throughout the course of treatment). Lastly, a third limitation impacting interpretation in the context of this single case experiment were the unstable and rising baselines for participants 1107 and 1121. These baseline characteristics may represent threats to internal validity in the present study, which our ongoing

research is addressing through randomized, between group study designs to facilitate interpretation of the isolated effect of AI-assisted treatment.

Regarding future directions for the development and empirical testing of ChainingAI and the PERCEPT-R Classifier: first, the PERCEPT-R Classifier did not adapt its perceptual strictness as participants progressed, which may explain the negative session-by-time interaction in Research Question 1 and provides an alternate explanation for the disparity between classifier performance in Table 1 and Table 5. Our ongoing work will examine the ways in which to adapt PERCEPT performance to account for participants' incremental improvement in /ɹ/ during the course of intervention. Secondly, our probe strategy for the direct effect of ChainingAI in Research Question 1 was insensitive to potential mechanisms regarding overnight consolidation after learning (e.g., Breton & Robertson, 2017), which may be considered in future study designs. Thirdly, we observed that most participants nearly never heeded the ChainingAI prompts for prosodic variation, even after text-to-speech prompts demonstrating this variation were added to the website based on observations from two pilot participants (whose case studies are not reported here). The ineffectiveness of the current prosodic prompts would, theoretically, result in generalization that was lower than if the prosodic variation was included through the presumed mechanism of practice variability (Preston, Leece, McNamara, et al., 2017). Future development of ChainingAI will focus on how strengthening the salience of prosodic cues for learners. Next, as discussed for one participant, PERCEPT-R performance may have been confounded by instrumentation error in the case of word-initial targets because in-session productions were often lengthened and significant portions of the rhotic interval were missed during forced alignment (Benway & Preston, under review). Our ongoing work is developing different methods for improving identification of the rhotic associated interval within the word.

Finally, the results presented here do not provide any indication on the efficacy or efficiency of AI-assisted treatment in more intense or differently planned practice schedules, such as using AI in between sessions with a clinician. Our ongoing work will explore this in more detail.

Conclusion

This study provides the first evidence of participant improvement for /ɹ/ in untreated words in response to an AI-assisted treatment package. This treatment package included a human clinician delivering a 40-minute "Orientation to /ɹ/" and, at most, 10 minutes of syllable prepractice, three times per week, for three weeks. The 30 minutes following clinician-led prepractice was facilitated by Speech Motor Chaining, with the PERCEPT Engine simulating clinician perceptual judgment of the /ɹ/ production. Perceptual ratings of /ɹ/ in treated Chains were perceived to have significantly more rhoticity after ChainingAI than directly after human-clinician led prepractice. Perceptual ratings of /ɹ/ on untreated words showed significant nonoverlap with ratings from the no-treatment baseline phrase for three of the five participants, indicating a treatment-associated response to the AI-assisted package. All five participants demonstrated statistically significant improvements in /ɹ/ from pretreatment to post-treatment, with standardized effect sizes ranging from .36-1.6 and a mean of 30% improvement over baseline accuracy. PERCEPT-clinician agreement (i.e., F1-score) was largely within the range of agreement seen between human clinicians for four of five participants, but note that overall agreement between clinicians was lower than anticipated and warrants some caution with interpreting the present results. Exploration of survey data indicated that parents and participants largely felt that computerized intervention could positively impact service delivery for children with speech sound disorders, most frequently mentioning hybrid models in which computerized systems facilitate at-home practice. Future work will continue exploring the potential for AI-

assisted speech therapy and how these technologies can be personalized to maximize participant improvement.

Acknowledgements

The authors would like to thank participants and families for their enthusiastic participation in this study. Profound gratitude is due to Nathan Preston, for his full stack software expertise, and to members of the Speech Production Lab for masked data processing, including Megan Leece, Nicolle Caballero, Benny Herbst, Danielle Kealy, Stephanie Reeves, Kerry McNamara, Martine Schultheiss, Michela Eivers, Rachel Koury, Chyanne Lesher, and Abigail Matejko-Lima. This research was supported through an internal grant (CUSE II-14-2021; J. Preston, PI) and computational resources (NSF ACI-1341006; NSF ACI-1541396) provided by Syracuse University. This project was also supported by a Research Excellence Doctoral Funding Fellowship from Syracuse University.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining,

ASHA. (2018). *School practice mini- survey summary report: Number and type of responses.*

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Benway, N. R., Hitchcock, E., McAllister, T., Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual /ɹ/ errors in American English: A single case randomization design. *American Journal of Speech-Language Pathology*.

Benway, N. R., & Preston, J. L. (under review). Prospective Validation of Motor-Based Intervention with Automated Mispronunciation Detection of Rhotics in Residual Speech Sound Disorders.

Benway, N. R., Preston, J. L., Hitchcock, E. R., Rose, Y., Salekin, A., Liang, W., & McAllister, T. (in press). Reproducible Speech Research with the Artificial-Intelligence-Ready PERCEPT Corpora *Journal of Speech, Language, and Hearing Research*.

Benway, N. R., Preston, J. L., Salekin, A., & McAllister, T. (in preparation). Automated detection of rhoticity of American English /ɹ/ in children with residual speech sound disorders: The PERCEPT-R Classifier

Benway, N. R., Preston, J. L., Salekin, A., Xiao, Y., Sharma, H., & McAllister, T. (under review). Classifying Rhoticity of /ɹ/ in Speech Sound Disorder using Age-and-Sex Normalized Formants.

Berisha, V., Krantsevich, C., Stegmann, G., Hahn, S., & Liss, J. (2022). Are reported accuracies in the clinical speech machine learning literature overoptimistic? Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,

Boyce, S. E. (2015). The articulatory phonetics of /r/ for residual speech errors. *Seminars in Speech and Language*, *36*(4), 257-270. https://doi.org/10.1055/s-0035-1562909

Brandel, J., & Frome Loeb, D. (2011). Program intensity and service delivery models in the schools: SLP survey results. *Language Speech and Hearing Services in Schools*, *42*(4), 461-490. https://doi.org/10.1044/0161-1461(2011/10-0019)

Breton, J., & Robertson, E. M. (2017). Dual enhancement mechanisms for overnight motor memory consolidation. *Nat Hum Behav*, *1*(6). https://doi.org/10.1038/s41562-017-0111

Chen, Y.-P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A., Doube, W., & Morris, M. E. (2016). Systematic review of virtual speech therapists for speech disorders. *Computer Speech & Language, 37*, 98-128. https://doi.org/https://doi.org/10.1016/j.csl.2015.08.005

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, 6*(44), 29.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., & Alwan, A. (2000). Acoustic modeling of American English /r/. *Journal of the Acoustical Society of America, 108*(1), 343-356. https://doi.org/10.1121/1.429469

Flipsen, P., Jr. (2015). Emergence and prevalence of persistent and residual speech errors. *Seminars in Speech and Language, 36*(4), 217-223. https://doi.org/10.1055/s-0035-1562905

Furlong, L., Erickson, S., & Morris, M. E. (2017). Computer-based speech therapy for childhood speech sound disorders. *Journal of Communication Disorders, 68*, 50-69. https://doi.org/https://doi.org/10.1016/j.jcomdis.2017.06.007

Furlong, L., Morris, M., Serry, T., & Erickson, S. (2018). Mobile apps for treatment of speech disorders in children: An evidence-based analysis of quality and efficacy. *PLoS One, 13*(8), e0201513. https://doi.org/10.1371/journal.pone.0201513

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189-211. https://doi.org/10.1080/19345747.2011.618213

Goldman, R., & Fristoe, M. (2015). *Goldman Fristoe Test of Articulation - Third Edition*. Pearson.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10*(1), 14-23.

Guadagnoli, M., & Lee, T. (2004). Challenge point, a framework for conceptualizing the effects of varius practive conditions in motor learning. *Journal of Motor Behavior*.

Gupta, S., & DiPadova, A. (2019, June). Deep Learning and Sociophonetics: Automatic Coding of Rhoticity Using Neural Networks. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, Minneapolis, Minnesota.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Gwet, K. L. (2019). *irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC).* (Version 1.0)

Hair, A., Ballard, K. J., Markoulli, C., Monroe, P., Mckechnie, J., Ahmed, B., & Gutierrez-Osuna, R. (2021). A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World. *ACM Trans. Access. Comput., 14*(1), Article 3. https://doi.org/10.1145/3433607

Harel, D., & McAllister, T. (2019). Multilevel Models for Communication Sciences and Disorders. *Journal of Speech, Language, and Hearing Research, 62*(4), 783-801. https://doi.org/10.1044/2018_JSLHR-S-18-0075

Health Workforce Australia. (2014). *Speech Pathologists in Focus* (Australia's Health Workforce Series, Issue.

Hedlund, G., & Rose, Y. (2019). *Phon [Computer Software].* (Version 3.0.6-beta.4) Retrieved from https://phon.ca.

Hitchcock, E. R., Harel, D., & McAllister Byun, T. (2015). Social, Emotional, and Academic Impact of Residual Speech Errors in School-Aged Children: A Survey Study. *Semin Speech Lang, 36*(4), 283-294. https://doi.org/10.1055/s-0035-1562911

Hitchcock, E. R., Swartz, M. T., & Lopez, M. (2019). Speech sound disorder and visual biofeedback intervention: A preliminary investigation of treatment intensity. *Seminars in Speech and Language, 40*(02), 124-137.

Ingram, D. (2002). The measurement of whole-word productions. *Journal of Child Language, 29*(4), 713-733.

Jacko, J. A. (2012). Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications.

Kaipa, R., & Peterson, A. M. (2016). A systematic review of treatment intensity in speech disorders. *International Journal of Speech Language Pathology, 18*(6), 507-520. https://doi.org/10.3109/17549507.2015.1126640

Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048*.

Katz, L. A., Maag, A., Fallon, K. A., Blenkarn, K., & Smith, M. K. (2010). What Makes a Caseload (Un)Manageable? School-Based Speech-Language Pathologists Speak.

Language, Speech, and Hearing Services in Schools, 41(2), 139-151. https://doi.org/10.1044/0161-1461(2009/08-0090)

Klein, H. B., McAllister Byun, T., Davidson, L., & Grigos, M. I. (2013). A Multidimensional Investigation of Children's /r/ Productions: Perceptual, Ultrasound, and Acoustic Measures. *American Journal of Speech-Language Pathology, 22*(3), 540-553. https://doi.org/10.1044/1058-0360(2013/12-0137)

Koegel, L. K., Koegel, R., L., & Ingham, J. C. (1986). Programming Rapid Generalization of Correct Articulation through Self-Monitoring Procedures. *Journal of Speech and Hearing Disorders, 51*(1), 24-32. https://doi.org/10.1044/jshd.5101.24

Koegel, R., L., Koegel, L. K., Ingham, J. C., & Van Voy, K. (1988). Within-Clinic versus Outside-of-Clinic Self-Monitoring of Articulation to Promote Generalization. *Journal of Speech and Hearing Disorders, 53*(4), 392-399. https://doi.org/10.1044/jshd.5304.392

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*(3-4), 445-463. https://doi.org/10.1080/09602011.2013.815636

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications, 7*(1), 13619. https://doi.org/10.1038/ncomms13619

Li, S. R., Dugan, S., Masterson, J., Hudepohl, H., Annand, C., Spencer, C., Seward, R., Riley, M. A., Boyce, S., & Mast, T. D. (2023). Classification of accurate and misarticulated /ɑr/ for ultrasound biofeedback using tongue part displacement trajectories. *Clinical Linguistics & Phonetics*, *37*(2), 196-222. https://doi.org/10.1080/02699206.2022.2039777

Maas, E., & Farinella, K. A. (2012). Random versus blocked practice in treatment for childhood apraxia of speech. *Journal of Speech Language and Hearing Ressearch, 55*(2), 561-578. https://doi.org/10.1044/1092-4388(2011/11-0120)

Maas, E., Robin, D. A., Austermann Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology*, *17*(3), 277-298. https://doi.org/10.1044/1058-0360(2008/025)

MacDowell, M., Glasser, M., Fitts, M., Nielsen, K., & Hunsaker, M. (2010). A national view of rural health workforce issues in the USA. *Rural and remote health, 10*(3), 1531.

Matthews, T., Barbeau-Morrison, A., & Rvachew, S. (2021). Application of the Challenge Point Framework During Treatment of Speech Sound Disorders. *Journal of Speech Language and Hearing Research, 64*(10), 3769-3785. https://doi.org/10.1044/2021_jslhr-20-00437

McAllister, T., Hitchcock, E. R., & Ortiz, J. A. (2020). Computer-Assisted Challenge Point Intervention for Residual Speech Errors. *Perspectives of the ASHA Special Interest Groups*. https://doi.org/10.1044/2020_PERSP-20-00191

McAllister, T., Preston, J. L., Hitchcock, E. R., & Hill, J. (2020). Protocol for Correcting Residual Errors with Spectral, ULtrasound, Traditional Speech therapy Randomized Controlled Trial (C-RESULTS RCT). *BMC pediatrics, 20*(1), 66.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). *Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi* INTERSPEECH 2017: Proceedings of the 18th Annual Conference of the International Speech Communication Association (ISCA), Stockholm, Sweeden.

McCormack, J., McLeod, S., McAllister, L., & Harrison, L. J. (2009). A systematic review of the association between childhood speech impairment and participation across the lifespan. *International Journal of Speech-Language Pathology, 11*(2), 155-170. https://doi.org/10.1080/17549500802676859

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Murray, E., McCabe, P., & Ballard, K. J. (2020). The influence of type of feedback during tablet-based delivery of intensive treatment for childhood apraxia of speech. *Journal of Communication Disorders*, 106026. https://doi.org/https://doi.org/10.1016/j.jcomdis.2020.106026

McLeod, S., Ballard, K. J., Ahmed, B., McGill, N., & Brown, M. I. (2020). Supporting Children With Speech Sound Disorders During COVID-19 Restrictions: Technological Solutions. *Perspectives of the ASHA Special Interest Groups*. https://doi.org/doi:10.1044/2020_PERSP-20-00128

Miccio, A., Elber, M., & Forrest, K. (1999). The relationship between stimulability and phonological acquistion in children with normally developing and disordered phonologies. *American Journal of Speech-Language Pathology, 8*, 347-363.

Miller, P. (2016). Itinerancy between attractor states in neural systems. *Current Opinion in Neurobiology, 40*, 14-22. https://doi.org/https://doi.org/10.1016/j.conb.2016.05.005

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. FAT* '19: Conference on Fairness, Accountability, and Transparency, Atlanta, Georgia, USA.

Nagy, N., & Irwin, P. (2010). Boston (r): Neighbo(r)s nea(r) and fa(r). *Language Variation and Change, 22*(2), 241-278. https://doi.org/10.1017/S0954394510000062

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: nonoverlap of all pairs. *Behav Ther, 40*(4), 357-367. https://doi.org/10.1016/j.beth.2008.10.006

Preston, J. L., Benway, N. R., Leece, M. C., Hitchcock, E. R., & McAllister, T. (2020). Tutorial: Motor-based Treatment Strategies for /r/ Distortions. *Language, Speech, and Hearing Services in Schools, 54*, 966-980.

Preston, J. L., Caballero, N. F., Leece, M. C., Wang, D., Herbst, B. M., & Benway, N. R. (under review). A Randomized Controlled Trial of Treatment Distribution and Biofeedback Effects on Speech Production in School-Aged Children with Apraxia of Speech.

Preston, J. L., Leece, M. C., & Maas, E. (2017). Motor-based treatment with and without ultrasound feedback for residual speech-sound errors. *International Journal of Language & Communication Disorders, 52*(1), 80-94. https://doi.org/10.1111/1460-6984.12259

Preston, J. L., Leece, M. C., McNamara, K., & Maas, E. (2017). Variable practice to enhance speech learning in ultrasound biofeedback treatment for childhood apraxia of speech: A single case experimental study. *American Journal of Speech-Language Pathology, 26*(3), 840-852. https://doi.org/10.1044/2017_AJSLP-16-0155

Preston, J. L., Leece, M. C., & Storto, J. (2019). Tutorial: Speech motor chaining treatment for school-age children with speech sound disorders. *Language, Speech, and Hearing Services in Schools, 50*(3), 343-355. https://doi.org/10.1044/2018_LSHSS-18-0081

Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., & Maas, E. (2014). Ultrasound visual feedback treatment and practice variability for residual speech sound

errors. *Journal of Speech, Language, and Hearing Research, 57*(6), 2102-2115. https://doi.org/10.1044/2014_JSLHR-S-14-0031

Preston, J. L., Preston, N. J., & Benway, N. R. (2022). *Speech Motor Chaining Web-App.*

Pring, T., Flood, E., Dodd, B., & Joffe, V. (2012). The working practices and clinical experiences of paediatric speech and language therapists: a national UK survey [https://doi.org/10.1111/j.1460-6984.2012.00177.x]. *International Journal of Language & Communication Disorders, 47*(6), 696-708. https://doi.org/https://doi.org/10.1111/j.1460-6984.2012.00177.x

R Core Team. (2013). R: A language and environment for statistical computing.

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research.* (Version 1.9.12) Northwestern University, Evanston, Illinois.

Robey, R. R. (2004). A five-phase model for clinical-outcome research. *J Commun Disord, 37*(5), 401-411. https://doi.org/10.1016/j.jcomdis.2004.04.003

Ruscello, D. M. (1995). Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders, 28*(4), 279-302.

Rvachew, S., & Brosseau-Lapré, F. (2016). *Developmental Phonological Disorders: Foundations of Clinical Practice*. Plural Publishing.

Shields, R., & Hopf, S. C. (2023). Intervention for residual speech errors in adolescents and adults: A systematised review. *Clinical Linguistics & Phonetics*, 1-24. https://doi.org/10.1080/02699206.2023.2186765

Silverman, F. H., & Paulus, P. G. (1989). Peer reactions to teenagers who substitute /w/ for /r/. *Language, Speech, and Hearing Services in Schools, 20*(2), 219-221.

Sugden, E., Baker, E., Munro, N., Williams, A. L., & Trivette, C. M. (2018). Service delivery and intervention intensity for phonology-based speech sound disorders. *Int J Lang Commun Disord, 53*(4), 718-734. https://doi.org/10.1111/1460-6984.12399

Thoonen, G., Maassen, B., Wit, J., Gabreëls, F., & Schreuder, R. (1996). The integrated use of maximum performance tasks in differential diagnostic evaluations among children with motor speech disorders. *Clinical Linguistics & Phonetics, 10*(4), 311-336. https://doi.org/10.3109/02699209608985178

Tiede, M. K., Boyce, S. E., Holland, C. K., & Chou, K. A. (2004). A new taxonomy of American English /r/ using MRI and ultrasound. *J. Acoust. Soc. Am., 115*(5), 2533.

Verdon, S., Wilson, L., Smith-Tamaray, M., & McAllister, L. (2011). An investigation of equity of rural speech-language pathology services for children: A geographic perspective. *International Journal of Speech-Language Pathology, 13*(3), 239-250.

Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: a missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews, 13*(1), 70-77. https://doi.org/10.1002/mrdd.20139

Wiig, E., Semel, E., & Secord, W. (2013). Clinical evaluation of language fundamentals. *Bloomington, MN: Pearson*.

Wilbert, J., & Lüke, T. (2023). *Scan: Single-case data analyses for single and multiple baseline designs.*

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology, 13*(1), 61. https://doi.org/10.1186/1471-2288-13-61

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron, 102*(6), 1096-1110. https://doi.org/https://doi.org/10.1016/j.neuron.2019.04.023

CHAPTER 2 - AUTOMATED DETECTION OF RHOTICITY OF AMERICAN ENGLISH /ɹ/ IN CHILDREN WITH RESIDUAL SPEECH SOUND DISORDERS: THE PERCEPT-R CLASSIFIER

Nina R. Benway, Department of Communication Sciences & Disorders, Syracuse University, Syracuse, NY, USA

Jonathan L. Preston, Department of Communication Sciences & Disorders, Syracuse University, Syracuse, NY, USA; Haskins Laboratories, New Haven, CT, USA

Asif Salekin, Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, USA

and

Tara McAllister, Department of Communicative Sciences & Disorders, New York University, New York, NY, USA

Corresponding author:

Nina R Benway, nrbenway@syr.edu, Ph: 1-315-443-3143, Dept of Communication Sciences & Disorders, 621 Skytop Road, Suite 1200, Syracuse, NY 13244

Conflict of Interest:

The design of the PERCEPT Engine is patent pending: US Patent Application No. 63,450,762 (Benway, Preston, and Salekin).

Funding:

This research was supported through an internal grant (CUSE II-14-2021; J. Preston, PI) and computational resources (NSF ACI-1341006; NSF ACI-1541396) provided by Syracuse University, and by the National Institute on Deafness and Other Communication Disorders (NIH R01DC017476-S2; T. McAllister, PI).

Prologue to Chapter 2

Chapter 1 has provided the first evidence of participant improvement for /ɹ/ in untreated words in response to an AI-assisted treatment package. It is likely that the tool driving the automated portion of the treatment package, ChainingAI, has demonstrated therapeutic efficacy where other tools have not because of its strong foundation in multidisciplinary expertise. From a clinical perspective, ChainingAI automates a validated evidence-based practice. From a technical perspective, the PERCEPT-R Classifier that underlies ChainingAI meets several of the reproducibility guidelines advocated by Berisha et al. (2022), and Kapoor and Narayanan (2022). The points of Berisha et al. (2022) relate specifically to clinical replicability of speech technologies: building systems around low-dimension features that are validated in the acoustic phonetics literature and actively collected in clinically-relevant speech tasks. Kapoor and Narayanan (2022) emphasize the ways that poor experimental control may lead to between-dataset leakage that may bias experimental results in machine learning. The acoustic, clinical, and experimental factors that have motivated the development of the PERCEPT-R Classifier are detailed in this chapter and the associated appendix.

Abstract

Purpose: Supervised machine learning experiments were conducted during the development of the PERCEPT-R Classifier, an automated speech analysis system that predicts perceptual judgment of /ɹ/ in American English. Study outcomes reflect only stimulable participants with speech sound disorders, whose /ɹ/ was theorized to be clinically valid yet perceptually ambiguous compared to non-stimulable participants or typically developing speakers.

Method: 351 participants from the PERCEPT-R Corpus were split into training, validation, and test datasets. Formant features and Mel-frequency cepstral coefficient features were extracted from /ɹ/ within each recorded utterance. Shallow and deep neural networks were trained to associate input feature patterns with PERCEPT-R Corpus labels indicating perceptual judgment of /ɹ/ (i.e., correct/fully rhotic, incorrect/derhotic). Classifiers were evaluated by F1-score. Results were replicated. SHAP analysis estimated relative feature importance, and performance bias was explored.

Results: Age-and-sex normalized formant features significantly increased F1-score versus other feature sets. The best-performing classifier, a gated recurrent neural network, achieved a mean participant-specific F1-score of .81 after replication (σ_x= .10, med = .83, n = 48). The third formant most influenced classifier predictions, aligning with acoustic phonetic descriptions of /ɹ/. Post-hoc exploration indicated classifier performance was not systematically biased by age or sex of the speaker in the present dataset.

Conclusions: This article presents an age-and-sex normalized formant extraction methodology for the classification of fully rhotic versus derhotic /ɹ/ in the context of stimulable speakers with speech sound disorder that outperformed Mel-frequency cepstral coefficient-based

classifiers. The best-performing classifier exceeded our performance threshold for clinical utility. Clinical validation of a computerized intervention automated by this classifier is ongoing.

Introduction

There are many barriers restricting access to sufficiently intense speech therapy worldwide, including for those whose speech sound production difficulties continue past the age of 8 (residual speech sound disorders; RSSD). Computerized versions of clinically validated evidence-based practices could help narrow the gap between the higher treatment intensities shown to enhance speech learning and the lower treatment intensities available in everyday practice (Brandel & Frome Loeb, 2011; Hair et al., 2021; Kaipa & Peterson, 2016). Computerized practice may be particularly potent when combined with validated speech analysis algorithms that deliver clinical-grade feedback and adapt practice difficulty based on a learner's performance; however, no available published mispronunciation detection system focuses on American English /ɹ/. The present study, therefore, reviews the development of a speech analysis algorithm for classifying /ɹ/ production accuracy in American English RSSD, which we call the PERCEPT-R (*Perceptual Error Rating for the Clinical Evaluation of Phonetic Targets-Rhotics)* Classifier. As a mispronunciation detection[2] technology, the primary goal for the PERCEPT-R Classifier is to predict clinician perceptual judgment of /ɹ/ rhoticity in speech therapy practice trials. We focus here on fully rhotic dialects of American English because speakers whose /ɹ/

[2] Speech sound disorders involve much more than *mispronunciation*, but the term *mispronunciation detection* is used throughout this article to connect this work with a larger literature base that includes speech analysis for second-language learning and clinical purposes. Together, we refer to clinically validated, evidence-based practices with speech analysis algorithms such as mispronunciation detection as *clinical speech technologies*.

pronunciation is attributable to dialect differences would not meet the definition of RSSD nor be appropriate for clinical ratings of "correct" or "incorrect".

Multiple systematic reviews have highlighted several related factors that have previously inhibited the development of efficacious clinical speech technologies, particularly for child speakers (Chen et al., 2016; Furlong et al., 2017; Furlong et al., 2018; McKechnie et al., 2018). This study, and a companion clinical trial, accounts for these hindrances. First, there has been inadequate availability of child speech corpora for system training, with data scarcity compounded for clinical speech (Shahin et al., 2020). The development of the PERCEPT-R Classifier, however, is made possible by the existing open-access PERCEPT-R Corpus (Benway, Preston, Hitchcock, & McAllister, 2022; Benway et al., in press). The present study focuses on /ɹ/, often characterized as the most frequently impacted sound among American English speakers with RSSD (Lewis et al., 2015; Ruscello, 1995).

Second, no available automatic speech analysis tool has demonstrated acceptable accuracy in identifying incorrectly produced words from children with speech sound disorders (McKechnie et al., 2018). Given the breadth and depth of perceptually labeled /ɹ/ in the PERCEPT-R Corpus and the constraint of our research question to a single sound, we can fill this gap by engineering a mispronunciation classifier, specifically a rhoticity classifier. Classification has been shown to outperform probabilistic mispronunciation detection algorithms (i.e., goodness of pronunciation) when assumptions regarding training data quantity and task constraints are met (Strik et al., 2009; Yang et al., 2014), as they are in this study.

In this study, we accomplish classification by first quantifying theoretically motivated acoustic features expected to associate with clinician perception of fully rhotic /ɹ/. Then, we train algorithms to predict associations between these acoustic features and perceptual judgment using

experimental designs that emphasize replicability and external validity to the clinical setting. Lastly, we examine the relative importance of the features in classifier predictions to evaluate if the training aligns with the features theorized to be integral to the task (or, otherwise indicating that a confounding factor may be present). The present study addresses these essential feature selection and classification principles in the development of the PERCEPT-R Classifier, while recently completed work investigates the therapeutic efficacy of artificial intelligence assisted RSSD treatment using the tool (Benway & Preston, in preparation).

Automated Clinical Mispronunciation Detection

McKechnie and colleagues (2018) systematically reviewed the technical factors of speech analysis tools that had been designed to evaluate or modify child speech. The tools from the 32 reviewed articles most frequently attempted speech classification from word-level stimuli. Systems most often employed Gaussian Mixture Models – Hidden Markov Models (GMM-HMM) identification using Mel-frequency cepstral coefficient (MFCC) features for the task, and tested performance on a median of 37 samples. Accuracy of speech analysis tools, broadly, is determined by comparing the tool's predictions (e.g., correct/incorrect speech sound, correct/incorrect lexical stress, substitution/omission error) to a human listener's ground-truth classification of the speech. The systems analyzed by McKechnie and colleagues (2018) reported a wide range of accuracy, but the review authors cautioned that summarizing percent accuracy may be misleading when the tools are tested on few exemplars of errored speech. To repeat their example, if a test corpus contains 95% correct speech and 5% errored speech, a tool with zero sensitivity for errored speech could still have 95% overall accuracy.

No population studied, including speakers with SSD, met McKechnie and colleagues' (2018) benchmark for clinical utility: 80% agreement threshold for incorrectly produced words

(i.e., specificity). The author's rationale for this metric was that accurate feedback on incorrectly produced words is important in the context of modifying speech, with the 80% threshold reflecting previous percent agreement on speech judgments and reliability standards for agreement when re-rating the same behavior. In general, the reviewed studies reported higher accuracy when classifying correctly produced phonemes and lower accuracy for incorrectly produced phonemes. The best-performing tools were trained on in-domain speech samples from speakers from the population to be tested, forming the basis for McKechnie and colleagues' (2018) suggestion that future clinical tools must be validated on datasets that include many exemplars of incorrect speech. No attempts at rhoticity classification were reported for any of the 13 languages represented in the review's included studies.

More recently, however, Ribeiro et al. (2021) performed lab testing for the development of an ultrasound image processing system that predicts clinician perceptual judgment during speech therapy, including for Scottish English /ɹ/. Because the authors lacked enough annotated speech from children with SSD for system training, they built a convolutional neural network to estimate goodness of pronunciation scores relative to typical child and adult speakers. Test set participants with SSD were from the Ultrasuite dataset (Eshky et al., 2018). Each test-set token received an expert-derived ground-truth accuracy rating, but the authors only retained the tokens rated as clearly correct or clearly glided for the analysis (i.e., excluding tokens with ambiguous ratings). Reanalysis of the 8 test speakers with speech sound disorder in Table 2 of Ribeiro et al. (2021) suggests a participant-specific F1-score $\bar{x}$= .64 (σ_x = .25) for classification of /ɹ/. Note that the authors urge caution in interpreting the results because ground-truth agreement between expert raters was low for /ɹ/ (Krippendorf's α = .05 for binary ratings for rhotic phones). The authors attribute a portion of rater disagreement to the broad perceptual space for appropriate

rhotic realization in dialects of United Kingdom English, as well as the rarity with which rhotics are selected as clinical targets in these dialects.

Separately, Li et al. (2023) used automatically generated quantifications of ultrasound images to train support vector classifiers for the eventual purpose of generating a simplified visual display of the dorsal tongue surface during ultrasound biofeedback therapy for /ɹ/. The authors assessed classification accuracy relative to expert-listener ground truth for correct and misarticulated rhotics in the /ɑɹ/ syllable context for individuals with RSSD in a private dataset consisting of 23 children with RSSD and 17 children with typical speech. Li et al. (2023) reported percent accuracies in lab testing exceeding 89% for the classification of /ɑɹ/ from probe-elicited words, with the most misclassifications happening on tokens that had average ground-truth ratings falling in the middle of the 10-point perceptual rating scale (i.e., ratings between 2-8). The authors report that rater agreement for these more-ambiguous tokens from the middle of the scale was much lower than the agreement for the full dataset, which also included fully-correct /ɑɹ/ from typical speakers ($ICC_{ambiguous\ tokens}$ = .39, 95% CI [.22-.48]) versus $ICC_{full\ dataset}$ = .90, 95% CI [.89-.91). Precise classification performance, including the class balance of correct versus misarticulated /ɹ/ and participant-specific F1-score, was not reported by Li et al. (2023).

These examples highlight techniques common to a breadth and depth of clinical mispronunciation detection systems. These systems may commonly use MFCC features, including for /ɹ/ by Ribeiro et al. (2021). The rationale supporting MFCC-based classifiers has been that the Mel scale transforms frequency to align with human perception of pitch. Log-Mel filter banks have also been employed with great success in a number of well-performing proprietary commercial speech recognition frameworks (e.g., Liao et al., 2015). However, these successful speech recognition systems also contain language models that process semantic and

syntactic information, which can offset insufficiencies in acoustic modeling. Because mispronunciation detection algorithms do not contain language models, the quality of the selected acoustic feature is far more important in mispronunciation detection than in speech recognition (Leung et al., 2019). This raises the question if a feature more sensitive to the unique characteristics of /ɹ/ might improve rhoticity classification performance beyond the common MFCC-based systems or recent image-based systems.

Optimizing Mispronunciation Detection in /ɹ/

Much of the existing acoustic phonetics literature quantifies /ɹ/ with formants instead of MFCCs. Indeed, human speech perception has been shown to involve neural populations in the human superior temporal gyrus that are sensitive to formant structures (Mesgarani et al., 2014). Formants, frequency-wise bands of energy, are theorized to reflect the glottal source harmonics amplified by a given vocal tract configuration (Chilba & Kajiyama, 1941). Specifically for American English /ɹ/, we expect the third formant to be a salient feature for rhoticity classification that encodes perceptually relevant information about rhotic vocal tract configuration. Our preliminary work on this topic has shown that (age-and-sex normalized) formant features can successfully predict clinician judgment of rhotics in children with typical speech and speech sound disorders (Benway, Preston, Hitchcock, Salekin, et al., 2022), with an average participant-specific F1-score of .89 (σ_x= .18, n = 281). The present investigation expands upon this prior work by directly comparing performance of formant features to MFCC features and exploring if relative feature importance in the trained model aligns with the features theorized to be important for human perception. This investigation also examines overall F1-score performance in a more clinically representative (i.e., stimulable) subset of participants

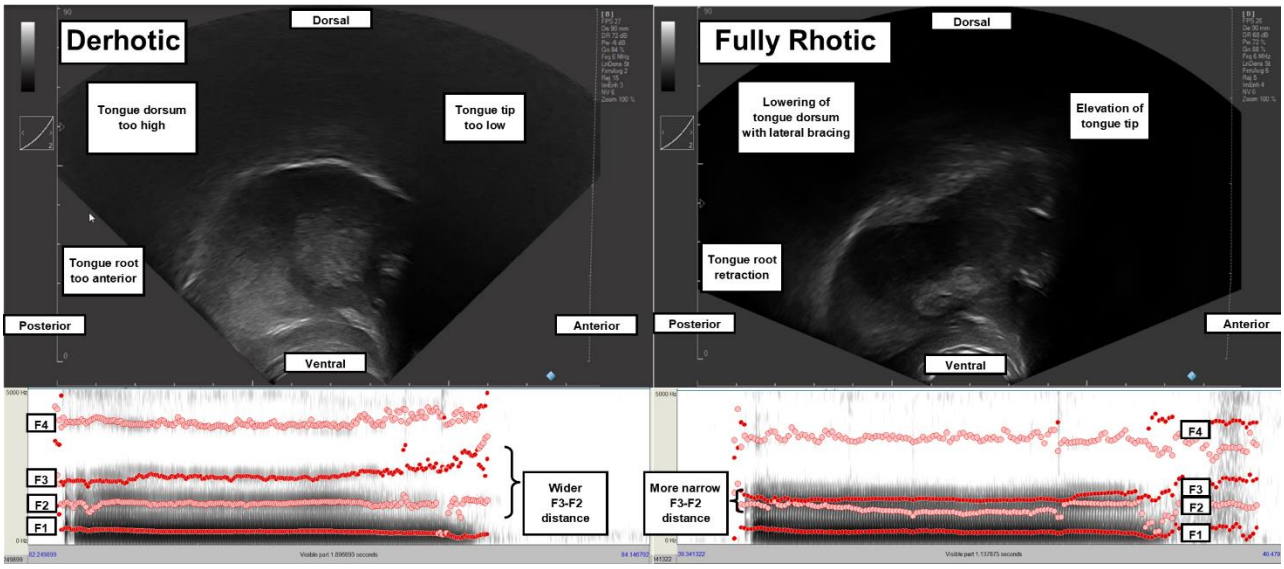
whose productions may be more ambiguous and accompanied by lower reliability of ground-truth ratings, as seen by Li et al. (2023).

Articulatory and Formant Characteristics of /ɹ/

Vocal tract configuration for American English /ɹ/ is complex, involving up to five coordinated gestures in the oral and pharyngeal cavities. These quasi-dependent articulations have been described as yielding either a predominantly "bunched" or "retroflexed" tongue shape (Delattre & Freeman, 1968; Preston et al., 2020). Common to these tongue shapes is a pharyngeal constriction of the tongue root, a low tongue dorsum, lateral tongue bracing, and a constriction of the oral cavity near the palate. In a bunched shape the oral cavity constriction is formed by a posterior blade/anterior dorsum of the tongue that is generally higher than the tip, while in a retroflexed shape the tip is generally higher than the blade. Either bunched or retroflexed vocal tract configurations yield “clinically correct” speech sounds in fully rhotic dialects of American English, and these two canonical configurations can be considered endpoints of an /ɹ/ tongue shape continuum (Boyce, 2015). Although there is between and within-speaker variation in the vocal tract configurations that generate a perceptually correct /ɹ/ in fully rhotic dialects of American English, such configurations yield a spectral envelope marked acoustically by a relatively high second formant (F2; Delattre & Freeman, 1968) and a relatively low third formant (F3; Espy-Wilson et al., 2000). The salience of F3 in the context of fully rhotic /ɹ/ is reinforced by acoustic investigation with real-time imaging that has shown that (narrower) degree of palatal constriction imparts a significant, medium-to-large effect on lowering of F3 (Harper et al., 2020). A low F3 and a high F2 results in the average F3-F2 distance of fully rhotic /ɹ/ being much narrower than the average F3-F2 distance produced by a neutral vocal tract or a derhotic vocal tract configuration.

Derhotic productions of /ɹ/ in fully rhotic dialects do not meet the perceptual standard to be considered clinically correct. Perceptually, derhotic /ɹ/ may often have characteristics of other approximant sounds (i.e., /w/, /l/, or /ɰ/), fricatives (i.e., /ɣ/ or /ʁ/), or have a vowel-like quality (i.e., /ʊ/, /ə/); furthermore, derhoticity could be minimal, moderate, or significant (e.g., Ball, 2017). Motor plans generating derhotic /ɹ/ commonly include one or more articulatory characteristics of a more neutral vocal tract: lower tongue tip, higher tongue dorsum, and/or insufficient tongue root retraction (Preston et al., 2020). Excessive lip rounding may or may not also be present. The F3-F2 distance of rhotics perceived as clinically incorrect is much greater than in rhotics perceived as clinically correct, all else being equal (Shriberg et al., 2001). The composite image in Figure 1 illustrates this principle for one participant (6103) reanalyzed from the work of Benway et al. (2021).

Figure 2-1. Vocal Tract Configuration and Formant Patterns for /ɹ/



Note: Inset text annotates ultrasound images for derhotic and fully rhotic (retroflexed) /ɹ/ as well as F3-F2 distances for the same individual.

Normalization of Feature Input

Formant values interact not only with vocal tract configuration but also vocal tract size. For instance, data from Lee et al. (1999) show F3-F2 distance of a neutral vocal tract varying from 1197 Hz (males, 19 years of age) to more than 2000 Hz (females and males, 5 or 6 years of age). For a fully rhotic /ɹ/, this average F3-F2 distance is markedly lower: varying in Lee et al. (1999)'s participants between 343 Hz (males, 19 years of age) to 797 Hz (males, 5 years of age). It may be that the raw F3-F2 feature space associated with fully rhotic productions in a younger child, with a smaller vocal tract, overlap with the F3-F2 feature space that would be associated with derhotic productions in an older adolescent with a larger vocal tract. Indeed, age-and-sex normalized F3-F2 difference has been found to better model expert listener ratings of clinically correct /ɹ/ than unnormed F3, F2, F3-F2, and F3/F2 (as well as normed F3, F2, and F3/F2; Campbell et al., 2018), likely because of the ability of normalization to minimize formant interactions arising from size and shape of the vocal tract cavity resonator. Therefore, the current investigation will include measures of age-and-sex normalized formants. Because feature normalization is an important consideration in the development of machine learning experiments, the proposed age-and-sex normalized features will be compared to a baseline condition in which feature values are z-standardized according to values in the self-same utterance.

Considerations for Automated Measurement

Although hand–corrected, age-and-sex normalized F3-F2 distance has been shown to index clinical perceptual judgment (Benway et al., 2021), it is unknown whether the same is true for formants generated automatically without hand-correction. This caveat is particularly salient in the context of rhotic F3-F2 distance; lowering of F3 in rhotics is important insofar as it creates

a single prominent band of energy in the region of F2 (F3-F2 approaching 0 Hz; e.g., Heselwood & Plug, 2011). An automated linear predictive coding (LPC) formant estimator that tracks only one formant in the context of a merged F3-F2 would either inappropriately fit F2 too low (likely tracking the second harmonic) or fit F3 too high (likely tracking F4). Each of these scenarios would erroneously predict a large F3-F2 distance for a fully rhotic production. Such errors are typically flagged and hand-corrected during human-supervised measurement, but there is no opportunity to do so in a fully automated workflow. Given this, it is possible that MFCCs may have an advantage over formants after all by reducing this measurement error, perhaps because MFCC features do not require a linear predictive coding algorithm to identify two peaks within a potentially merged F3-F2. Indeed, MFCCs have been reported to outperform formants in the sociophonetic classification of /ɹ/ in rhotic versus non-rhotic dialects, although the authors do not specifically attribute the superiority of MFCCs to any of the mechanisms posited here (Gupta & DiPadova, 2019).

Classification in Ambiguous Feature Spaces for RSSD

The performance of clinical speech technology must be analyzed in a way that emphasizes external validity and generalizability (i.e., machine learning replicability) of results for future clinical use. This study's emphasis on the evaluation of low-dimension formant features that are validated in the acoustic phonetics literature and actively collected in relevant speech tasks aligns well with reproducibility guidelines summarized by Berisha et al. (2022). Furthermore, replicable machine learning experiments for clinical speech technology must measure system performance with the speech representative of clinical end use. We adopt the stance that the use of an automated mispronunciation detection algorithm in treatment contexts is only clinically ethical for individuals who can occasionally produce fully rhotic exemplars (i.e.,

those who are clinically *stimulable*). In other words, if a participant cannot demonstrate the target skill at all, it would be clinically inappropriate to have that participant engage in practice without a clinician. It could potentially be misleading to report performance based on participants who consistently produce fully derhotic (or fully rhotic) productions, as these participants likely produce maximally different feature spaces for the target /ɹ/ that do not necessarily represent the rhotic/derhotic feature space of stimulable participants. Indeed, reanalysis of Benway et al. (2021) Figure 3 shows that motor plan improvements for stimulable participants result in incremental narrowing of F3-F2, and reanalysis of Benway et al., (2021) Figure 4 shows that narrower F3-F2 space is not necessarily directly indicative of complete perceptual resolution of derhoticity.

Taken together, these points suggest that significantly reduced F3-F2 distance may correspond to an incremental transition from an unambiguously-derhotic perceptual space to an ambiguous perceptual space, rather than categorical resolution of derhoticity. This may be problematic for defining salient feature spaces upon which to train a high-performing classifier. Therefore, a second goal of this study is to demonstrate classifier performance that exceeds the .8 threshold for clinical utility (e.g., McKechnie et al., 2010) in participants more reflective of clinical end, who likely have more ambiguous relationships between feature spaces and class labels than the participants previously evaluated by Benway, Preston, Hitchcock, Salekin, et al. (2022).

Purpose and Research Questions

The present investigation seeks to determine the acoustic features that optimize binary prediction of perceptual judgment of rhoticity (i.e., fully rhotic, derhotic) in word-level productions from children with RSSD, in context of the following research questions. **Research**

Question 1 is a preliminary investigation with shallow neural networks that examines the feature extraction and normalization techniques that optimize classification: **do age-and-sex normalized formants improve F1-score relative to utterance-normalized formants and utterance-normalized MFCCs?** Because of the salient F3-F2 signature for fully rhotic /ɹ/, we hypothesize that formant features will outperform MFCCs for rhoticity classification. Furthermore, because F3-F2 distance for fully rhotic /ɹ/ is age-and-sex dependent, we hypothesize that age-and-sex normalization of formant features relative to a published reference dataset will outperform formant features that are centered and scaled relative to the values present in the self-same utterance. **Research Question 2 takes the best-performing feature set from Research Question 1 and employs deep neural networks to maximize overall classifier performance: can the mean participant-specific F1-score exceed our .8 threshold for clinical acceptability in participants with more perceptually ambiguous feature spaces?** We exceeded this threshold in our preliminary work with shallow classifiers when predicting rhoticity in children with typical speech and fully incorrect tokens from children with /ɹ/ errors (Benway, Preston, Hitchcock, Salekin, et al., 2022); however, it is unknown if this accuracy can be replicated in participants who have emerging motor plans for /ɹ/ – and, likely, more ambiguous derhotic feature spaces and lower reliability in ground-truth class labels. We will also explore if output predictions from the best-performing classifier are systematically biased with regard to participant age-and-sex. **Research Question 3 is a post-hoc examination of the interpretability of the best-performing classifier relative to the salient acoustic features of /ɹ/: what is the relative importance of the individual acoustic features within the PERCEPT-R Classifier?** Valid machine learning models must not only perform well relative to performance metrics but also must be interpretable relative to theoretically important features. In

other words, if the features that were most influencing model predictions had little theoretical relevance for /ɹ/, it might indicate that PERCEPT-R was confounded in some way (e.g., by background noise that appears only in derhotic exemplars or vice versa). Following from previous acoustic modeling of perceptual judgment (Campbell et al., 2018), we hypothesize that F3-F2 will contribute the most to individual predictions made by the model, providing evidence that the model is tracking previously validated salient perceptual features for /ɹ/.

Methods

Experimental Design

We used a supervised learning framework (Singh et al., 2016) in which a predictive model learns to associate patterns in feature-based, independent variables for each file (Berisha et al., 2022) with a ground-truth rating provided by humans (i.e., listeners' perceptual judgments). We also took specific steps to emphasize replicability (Kapoor & Narayanan, 2022); for example, separate data splits were used for model training, validation, and testing. Appendix A includes a model card detailing model building and replicability details in the style of Mitchell et al. (2019).

Corpus Data

Speech data come from the private PERCEPT-R v2.2.2 Corpus, which is an extension of the open-access PERCEPT-R Corpus v2.2.2p that also includes participants excluded from the open-access subset following review of participant permissions. General characteristics of the corpus are detailed extensively in other publications and in PhonBank (Benway, Preston, Hitchcock, & McAllister, 2022; Benway et al., in press); see Benway et al. (in press) for a Dataset Datasheet (Gebru et al., 2018). Briefly, PERCEPT-R v2.2.2 consists overwhelmingly of single-word citation speech audio data recorded during 27 longitudinal clinical trials of children

with RSSD impacting /ɹ/ and cross-sectional studies of age-matched peers with typical speech. In contrast with the smaller open-access subset, the PERCEPT-R v2.2.2 Corpus contains 179,076 labeled utterances in 662 single-rhotic words and phrases from 413 child, adolescent, and young adult speakers of fully rhotic dialects of American English. Forty-six (11%) speakers in the PERCEPT-R v2.2.2 Corpus self-identified as Black/African American (n = 10), Asian (n = 7), American Indian or Alaska Native (n = 2), More than one race (n = 22), or Other (n = 5) according to the NIH race reporting framework. The present investigation excludes participants from the childhood apraxia of speech subset of the corpus, leaving 351 participants. We report additional demographic details of these participants in a subsequent section.

As in the open-access version 2.2.2p, the full PERCEPT-R Corpus contains audio files that are matched to a ground-truth label reflecting a listener judgment of rhoticity. Although the ground-truth labeling is detailed by Benway et al. (in press), it is briefly described here for current readers. Each utterance was rated by either a panel of expert listeners or a panel of lay listeners from a crowdsourcing platform (most frequently, three expert raters or nine crowdsourced raters; McAllister Byun et al., 2016). Raters were instructed to rate the perceptual accuracy of the /ɹ/ in the word with a binary rating relative to a fully rhotic dialect standard: 0 (incorrect/derhotic) or 1 (correct/fully rhotic). Each utterance is associated with a listener-average rating as well as a binary class label. The binary class label was derived from the listener-average rating, with ≥ .66 serving as the floor for class 1 (the fully rhotic class) to reflect that there is often not full agreement between expert raters in the context of RSSD (Klein et al., 2013). All utterances with a listener-average rating < .66 were assigned to class 0 (the derhotic class). There is an imbalance in the source data favoring derhotic tokens and male speakers, as

expected given that the data sources for the PERCEPT Corpus largely represent clinical trials for a disorder more prevalent in males (Wren et al., 2016).

Dataset Design for Clinical Replicability

The development of mispronunciation detection classifiers involves partitioning data into different experimental sets to prevent overfitting. Overfitting of a model to a specific dataset is a common pitfall of machine learning in which the model generated lacks replicability/external validity. Overfitting can happen because of sampling bias, redundant input features, or a lack of architecture parsimony. To monitor and minimize overfitting threats to external validity, corpora of exemplars are commonly split into *training, validation,* and *test* sets. Training sets contain the data from which the algorithm learns to associate patterns in the input with the ground-truth outcome variable. Validation sets are used to evaluate performance during training to provide feedback on learning and guide tuning of neural network hyperparameters. The final predictive performance reports the classifier's ability to accurately predict the ground-truth outcomes for data from the test set, which the model has not yet seen. For lab-demonstrated performance to have a chance to generalize to yet-unseen participants in a real-world clinical setting, it is important that neither individual tokens nor individual participants are included in more than one experimental set (most importantly, the training set).

There is also evidence that lab testing often overstates the predictive performance of clinical speech technology (Berisha et al., 2022). Because the decimal listener-average ratings of the PERCEPT-R Corpus reflect, in part, rater agreement, the likely result of the following dataset curation was the prioritization of speakers with the most ambiguous tokens, which, as discussed above, have previously yielded lower reliability in ground truths. However, from a replicability standpoint, it would be important that the current study optimizes hyperparameter tuning in the

context of the feature space the classifier would encounter clinically. Therefore, we took specific measures to strengthen the replicability of our results. For instance, we hand-crafted the validation and test datasets in the present investigation to reflect the subset of participants in the PERCEPT-R Corpus for whom automated independent practice would be clinically appropriate. Whereas PERCEPT-R v2.2.2 contains data from participants with RSSD and those with typical speech, audio classification of tokens in a clinical context is only relevant for individuals with RSSD. So, first, all speakers *without* speech sound disorder were excluded from validation and test subsets (but were included in training). We also excluded RSSD participants from the validation and test subsets with perceptual ratings averaging > 80%, as these participants may not reflect the average individual presenting for computerized speech therapy with mispronunciation detection. Second, we ensured that the validation and test sets only included stimulable participants who could occasionally produce a fully rhotic /ɹ/. The stimulability threshold was set as fully rhotic/derhotic proportion > .33 according to the heuristic that speakers who produce more than two derhotic productions per every fully rhotic production may not be candidates for independent practice with automated speech analysis. Because the feature space for a fully rhotic /ɹ/ is perhaps more salient than ambiguous/derhotic /ɹ/, we believed our dataset design measures would result in lower performance compared to mispronunciation detection investigations that include typical speakers in the test set (as done by Benway, Preston, Hitchcock, Salekin, et al., 2022, who performed leave-one-out cross validation including typical speakers). However, we also expect these decisions to increase external validity compared to testing on SSD participants with less ambiguous feature spaces.

In addition to selecting stimulable participants, we downsampled utterance data (reducing derhotic tokens) in the training, validation, and test sets for participants who had more than 200

tokens in the analysis. This was done to prevent training class imbalance and to increase replicability by simulating the ratio of data expected in clinical use. The downsample ratio in test and validation was based on the 2:1 ratio of derhotic: rhotic tokens seen in reanalysis of 229,934 previous practice trials with stimulable participants using computerized intervention (Preston et al., under review; Preston, Leece, & Maas, 2017). The training subset was downsampled (reducing derhotic tokens) to achieve an approximately 1:1 class ratio for balance of exemplars during training and tuning. The details regarding each experimental dataset are shown in Table 1 and Figure 2.

Age and Sex Fairness Exploration

We used this group of stimulable participants with SSD as a candidate pool from which the test and validation datasets were drawn by age-and-sex stratified random allocation without replacement. In other words, we tested the classifier on the broadest possible range of ages to evaluate performance regarding a range of demographic characteristics for potential participants. Filling the test set first, however, exhausted some of the age-and-sex strata. As a result, the test set had a broader representation of ages than the validation set. The proportion of participants selected for each set was tuned such that participants assigned to the training set accounted for 70% of corpus utterances, and participants in the validation/test sets accounted for 15% of utterances each. A 70:15:15 utterance-level data split is common, but, because each participant contributed different amounts of utterances in the PERCEPT-R Corpus, the number of participants in each set does not follow the 70:15:15 split. More information about this is included in the Replication section that follows. These three training, validation, and subsets were verified to have no participant overlap, which is important for replicability of performance to yet-unseen participants who would present to the clinic. Because males were represented at

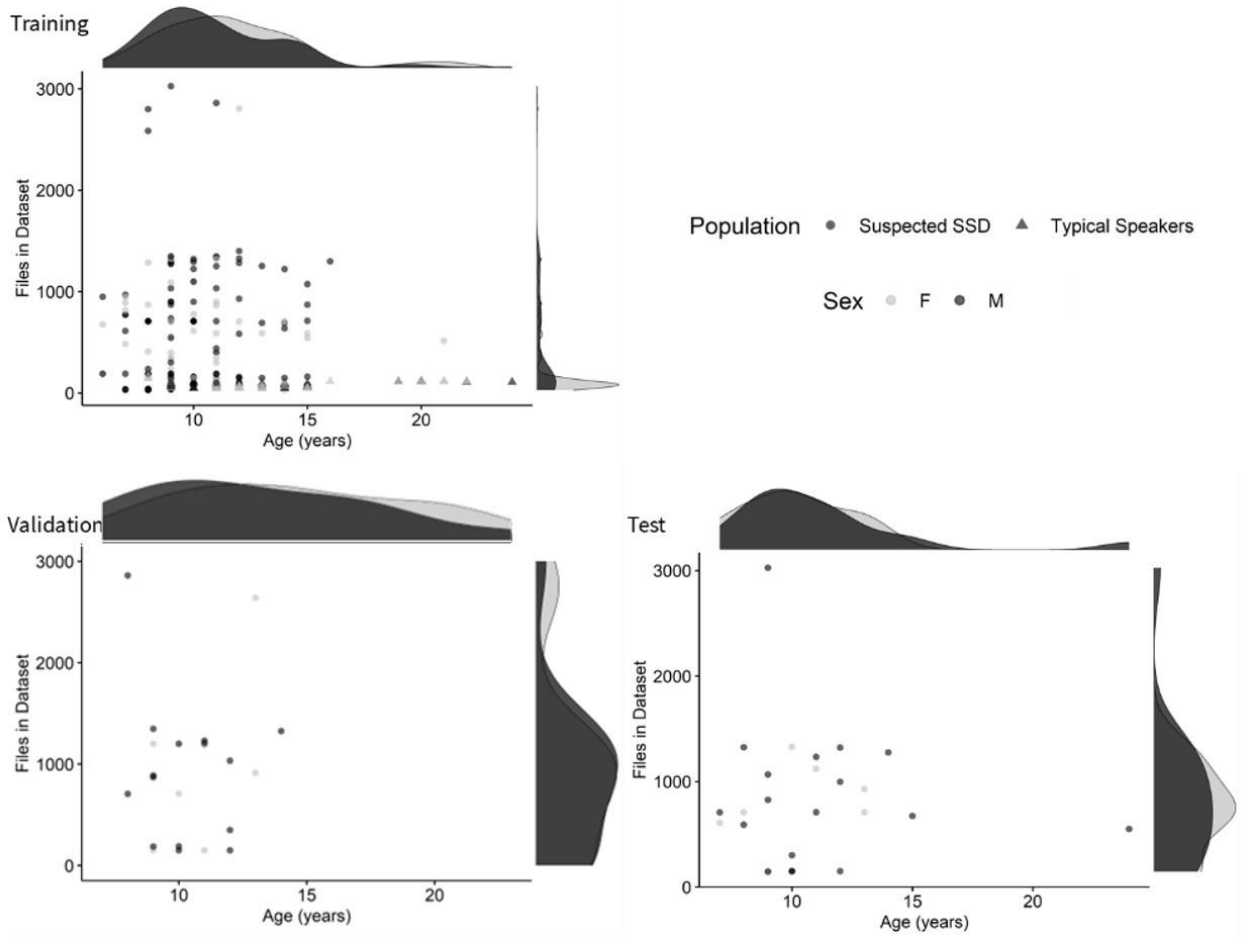
the desired ~2:1 frequency among the candidate participants, nothing further was needed to approach the clinically expected 2:1 ratio in the test and validation sets (note: the training set was more closely balanced with regard to sex because the training set sampled from the set of typical female speakers). The PERCEPT-R Corpus participant IDs assigned to each dataset are available at the Open Science Framework page for the PERCEPT project (https://osf.io/nqzd9/).

Table 2-1. Experimental Datasets

Subset	**Total Participants**	**Participants with SSD**	**Female Participants**	**Participant Age**	**Number of Derhotic Exemplars**	**Number of Fully Rhotic Exemplars**
Training	312	193	142	11.4 (3.5) [6-36]	36,979	32,705
Validation	22	22	7	10.5 (1.7) [8-14]	5,179	9,626
Test	26	26	8	10.8 (3.4) [7-24]	4,849	9,183

Note. Age is reported as $\bar{x}$ (σ_x) [range]. Because our stimulability criteria were associated with participants from longitudinal studies, the participant-level, training: validation: test proportions were different than the utterance-level, 70:15:15 training: validation: test proportions

Figure 2-2. Distribution of Age and Sex within Datasets



Note. Not shown is one participant in the training set (male, 36 years).

Feature Extraction

As a part of standard corpus processing, PERCEPT utterances were converted to one channel and standardized to an average intensity of 70 dB with a sampling rate of 44.1 kHz.

Estimation of Rhotic Interval Timepoints

We used GMM-HMM forced alignment to estimate the timestamps of the rhotic-associated target interval in each utterance. These estimates were generated using the Montreal Forced Aligner v2.0.0rc3 wrapper for the Kaldi Speech Recognition Toolkit (McAuliffe et al., 2017; Povey et al., 2011). Forced alignment preprocessing involved generating a Praat TextGrid (Boersma & Weenink, 2019) with a single labeled "Orthography" tier for every corpus file using

the PraatIO package in Python v3.7.6 (Mahrt, 2016), and extending the LIBRISPEECH dictionary with study-specific stimuli and ARPABET transcriptions (e.g., “erp”). Alignments were generated with default LIBRISPEECH pre-trained adult American English acoustic models that were then adapted to reflect GMM means observed in PERCEPT-R Corpus v 2.2.2 (e.g., Povey, 2012). Alignments for the predicted start and end of the rhotic target within the word were successfully generated for a total of 168,614 utterances. Each rhotic-associated interval was then extracted from the audio file, with a 10ms buffer to counteract edge effects during formant estimation. To assess automated performance, no alignments were hand-corrected.

Formant Estimation and Normalization

Formants were extracted from these rhotic-associated intervals using the Praat “To Formants: Robust” algorithm. Five formants were estimated from 5ms Gaussian-like windows, with a 5ms step between analysis frame centers and pre-emphasis above 50 Hz. Robust refinement of formant estimates used default settings: selected weighting of samples started ± 1.5 standard deviations from the mean and stopped after five iterations or if the relative change in variance was less than 1e-5. Praat function calls were facilitated by the Parselmouth API (Jadoul et al., 2018).

LPC coefficients were calculated using the Burg algorithm (Childers, 1978). In Praat, the LPC filter order is controlled by setting *Number of Formants* and *Formant Ceiling*. Because formant estimation algorithms are sensitive to the parameters entered, we customized the formant ceiling for each speaker (Derdemezis et al., 2016). Different customization methods were used for training set participants and validation/test set participants because of the large number of participants in the training set. For training set participants, each utterance in PERCEPT-R v2.2.2 was processed by a custom parallelization wrapper for FastTrack (Barreda,

2021) to find, through grid search, the formant ceiling value that reduced regression residuals for the estimated formants. The original grid search space was largely unconstrained: five formants within a formant ceiling from 4500 to 7500 Hz, in 500 Hz steps. A maximally broad formant range was used (rather than rule-of-thumb heuristics for male/female/child; Barreda, 2021) because we had little direct data on which children/adolescents have a more adult-like vocal tract configuration versus a child-like configuration. The formant ceilings estimated during the first grid search were used constrain a second search space that was ±1.5 standard deviations from a participant's mean from the first search. Each participant was processed through Fast Track for a second grid search using this participant-specific search constraint (with 10 equally spaced steps in Hertz per speaker). A total of 69,340 training set utterances had formant ceilings determined in this way. For the 344 utterances for whom FastTrack failed, participant-specific average ceilings were determined from successfully measured tokens and passed to Praat for formant estimation. Note that although FastTrack generates formant estimates during utterance processing, these estimates were not used further in this study to maximize parity between the lab testing and use of the tool in production, which would not involve FastTrack.

In the validation and test subsets, formant ceilings were determined manually for each participant by observing formant tracks generated with different ceiling values as done by Benway et al. (2021), to mimic the manually-selected formant ceiling methodology planned for future clinical validation of the tool. Preliminary explorations showed the manual method enhanced test set performance.

We retained Praat time series estimates of F1, F2, and F3 from the rhotic-associated intervals for the analysis. Two formant transformations were generated: Euclidean F3-F2 distance and F3-F2 deltas calculated as the difference between F3-F2 at time *i* and time *i + 1*.

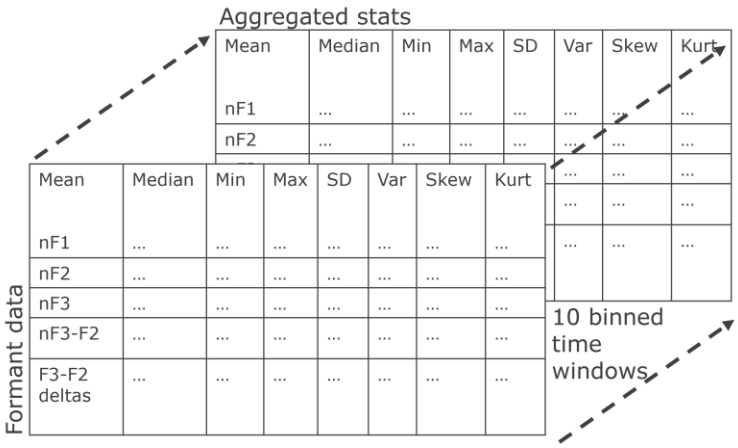
Undefined values in the Praat estimates were interpolated using the average of the preceding and following estimate in the relevant formant time series, except for missing starting and ending values which were edge-padded. No formant estimates were hand-corrected to enhance replicability to performance in automated clinical speech technology.

We created two feature sets from these data: one feature set normalized relative to age-and-sex-specific values for the same features for correct /ɹ/ from published reference data (Lee et al., 1999; age-and-sex normalization condition), and one normalized relative to the magnitude of an utterance's own formant values (utterance-normalized condition). For age-and-sex normalization, F1, F2, F3, and F3-F2 were centered and scaled (i.e., z-standardized) according to an age-and-sex-matched mean F1, F2, F3, or F3-F2 for typical /ɹ/. Note that the Lee dataset has coverage for individuals aged to 19, and the 19-year-old norms were used for eight individuals older than 19 in PERCEPT v2.2.2. For the utterance-normalized condition, F1, F2, F3, and F3-F2 were z-standardized according to the mean and standard deviation of the extracted rhotic-interval values for the self-same utterance. In both conditions, the F3-F2 delta transforms underwent linear conversion, to be scaled between -10 and +10, to meet neural network assumptions about consistency of data magnitude between features.

The following methods were used for both formant feature sets. Because formant estimates were taken from 5ms windows, the number of samples in the formant estimates varied with the length of the rhotic-associated interval. The time series for each formant and transformed formant was summarized into 10 bins, with each bin containing the median, mean, standard deviation, minimum, maximum, variance, skewness, and kurtosis of the 10% of formant estimates being grouped into a given bin. These bins were arranged such that each rhotic-associated interval was represented by a 3D NumPy array of the shape [5, 10, 8] representing [5

formants and transforms * 10 time windows * 8 summary statistics] for use with deep neural networks, as illustrated in Figure 3 below. Each array was checked to ensure it was the correct shape and to screen for zeros/NANs generated during processing. For shallow neural networks in Research Question 1, the means for each formant/transform-time window for a given utterance were flattened to an array of shape [1, 50] for row-wise stacking of utterances to meet data frame-style formatting conventions.

Figure 2-3. Representation of Input Feature Shape for Formants Features



MFCC Estimation

MFCCs were included as a baseline comparison. These features were generated in analogous fashion to that detailed prior, except with a function call to the Praat "To MFCCs" algorithm in place of the formant estimation function call. Window length and timestep were identical to the settings used during formant extraction. Thirteen MFCCs were computed with default filter bank parameters. The z-standardization of each utterance relative to its self-same values was identical to the utterance normalization formants. Timeseries standardization and quality control were also identical to the methods described for formants, yielding a 3D array [13

MFCCs * 10 time windows * 8 summary statistics] for use with deep neural networks. This 3D array was flattened to an utterance-wise data frame for use with shallow neural networks in Research Question 1 as previously described.

Statistical Comparison of Features

We quantified the impact of feature sets with linear mixed-effects models fit with the lmer function in the lme4 R package (Bates et al., 2014). Fixed effect terms modeled *classifier type* (random forest versus stochastic gradient descent), *retraining timepoint* (out-of-box versus after retraining, discussed in more detail below), *feature input* (MFCCs versus utterance-normalized formants versus age-and-sex normalized formants), and all interactions. Random intercepts were included for *participants*; no random slopes were examined. *Classifier type* was included in the model to better isolate the effect of *feature input* due to patterns observed in raw values depending on *classifier type*, *feature input*, and *retraining timepoint*. Parameters were estimated by restricted maximum likelihood. Categorical variables were effects coded. MFCCs served as the reference level for *feature input*, out-of-box testing was the reference level for *retraining timepoint*, and the random forest performance was the reference level for *classifier type*.

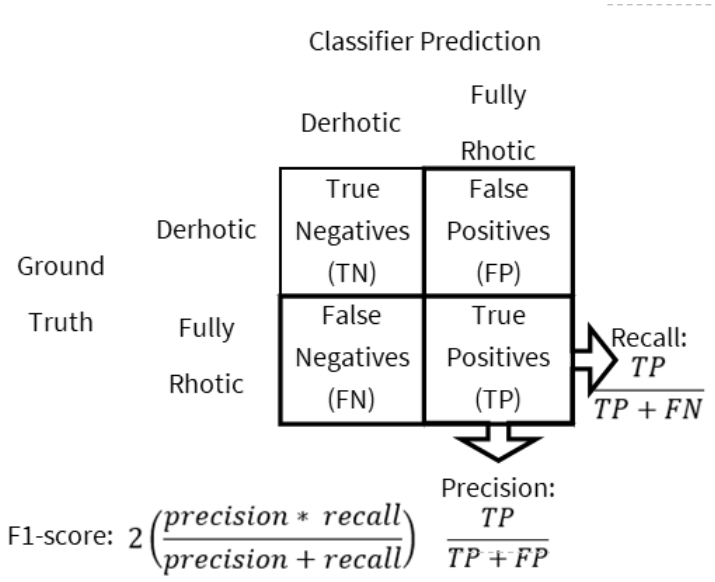
Classifier Architectures, Training, and Tuning

Machine learning classifiers model the relationships between a set of input features and a dependent variable by minimizing a loss function (i.e., the residuals) between the model-predicted outcome and ground-truth-observed outcomes. The choice of a machine learning classification algorithm is an empirical one. As is common, we frame the classifiers discussed herein within two categories: shallow neural networks and deep neural networks (e.g., Robles Herrera et al., 2022). Shallow neural networks are two-layer networks that use linear

combinations of functions comprised of feature variables and their weights. Deep neural networks extend shallow networks beyond two linear layers, with outputs of one (nonlinear) layer acting as inputs to the next (nonlinear) layer. The weight for each node is fit with an optimizer algorithm (e.g., gradient descent) to estimate the local minimum of a loss function. Then, the weight and bias is scaled using a nonlinear activation function that decides whether the information from that neuron will be passed on to neurons in subsequent layers. These nonlinear activation functions constrain the output to ensure the models run to completion (i.e., avoid gradient vanishing, gradient explosion, or failure to converge).

Whether shallow or deep, neural networks can perform classification or regression (i.e., prediction of a binary or continuous outcome). As the goal of the present study is to predict a binary clinical judgment of /ɹ/ (fully rhotic/correct versus derhotic/incorrect), we frame the task as classification and will evaluate predictive performance relative to ground truth using a confusion matrix (i.e., contingency table; Figure 4). This study uses F1-score, whereas McKechnie and colleagues (2018) evaluated performance with specificity $\left(\frac{true\ negatives}{true\ negatives+false\ positives}\right)$. F1-score is the harmonic mean of precision (i.e., positive predicative value) and recall (i.e., sensitivity). In addition to being commonly used, the inclusion of recall in the F1-score captures information on false negatives not represented by the specificity metric. System tuning that considers precision and recall may also be more desired in a clinical context when correct discoveries of correct productions advance stimulable participants to higher levels of linguistic complexity during practice. The threshold value used by McKechnie and colleagues, 80%, is retained in this study as reflects previous percent agreement on speech judgments, reliability standards for agreement when re-rating the same behavior, and a psychometric benchmark for a valid tool (Plante & Vance, 1994).

Figure 2-4. Analyzing Classifier Performance with F1-Score



We approached the feature set comparison in Research Question 1 with shallow neural networks to reserve computational time and energy resources. For shallow classification, we employed a random forest classifier and a stochastic gradient descent classifier with different loss functions using Scikit-Learn (Pedregosa et al., 2011). These classifiers were selected to provide coverage for both node-based and function-based architectures that allowed for warm-start retraining with new data. Training occurred with 8 CPU computer cores in a dedicated academic virtual environment at Syracuse University. All seeds were set to 24601, except during retraining when the seed was set to the cross-validation iteration (1-5).

For Research Question 2, we designed and trained deep neural networks using the best-performing input feature set from Research Question 1. This was done using Pytorch (Paszke et al., 2019). We compared a convolutional neural network (CNN), a gated recurrent neural network (GRNN), and a joint CNN-GRNN. CNNs perform feature mapping on input data that is

able to capture local correlations in frequency and time that would be salient for speech processing (Huang et al., 2015). In contrast, (G)RNNs can capture temporal dependencies on longer timescales to extract patterns from unstructured time-series data; this is also salient for speech, as this long-range context would include information about /ɹ/ over the course of the phone (Graves et al., 2013). We expected the CNN-GRNN would best capture the normalized frequency differences between the fully rhotic and derhotic time series because the architecture has convolutional feature extraction filters followed by gated recurrent units, which account for both frequency and local/long-range time domains.

Deep and shallow algorithms alike include hyperparameters, such as the number of nodes in a decision tree, the learning rate, the tolerance for early stopping, the number of convolutional filters, and the number of neurons in a hidden layer. The values chosen for hyperparameters and the manner in which they are chosen can greatly influence model performance and replicability. These hyperparameter values, however, are not necessarily theoretically motivated by the research question or feature set. Therefore, to find the set of hyperparameters that optimize each model architecture at hand we employed the Optuna hyperparameter optimization framework (Akiba et al., 2019). Models were each tuned with 50 Optuna trials, with the goal of finding the hyperparameters that maximized participant-specific mean F1-score in the validation set. Tuning constraints for shallow and deep neural networks appear in Tables 2 and 3. The architecture for the best performing algorithm is shown in Figure 5.

Table 2-2. Hyperparameter Tuning for Shallow Neural Networks

Classifier	Parameter	Possible Values	Tuned Value
Random Forest	The number of trees	$50 \leq x \leq 1000$	n_estimators = 80
	The loss function evaluating splits	GINI, entropy	criterion = GINI
	Fraction of samples required to split an internal node	$.1 \leq x \leq .9$	max_samples_split = .10
	Fraction of samples required to be at a leaf	$0 \leq x \leq .5$	max_samples_leaf = .005
	The number of features considered when splitting nodes	Square root, log2, None	max_features = None
Stochastic Gradient Descent	The loss function against which predictions are evaluated	Hinge, log, Huber, modified Huber, squared hinge, perceptron, squared error, epsilon insensitive, squared epsilon insensitive	loss = Huber
	The regularization term to be used	L2, L1, elasticnet	penalty = L2
	Constant multiplier of regularization term	1e-5, $\leq x \leq .5$	alpha = .22
	The learning rate scheduling algorithm	Constant, optimal, invscaling, adaptive	learning_rate = adaptive
	The initial learning rate	$0 \leq x \leq 1$	eta0 = .781
	Early stopping tolerance	$0 \leq x \leq 1$	tol = .55

Note. Huber loss considers mean square error and mean absolute error and reduces sensitivity to dataset outliers.

Training for deep networks was constrained to 25 epochs with early stopping after 5 epochs without decreasing validation loss. Training and validation batch sizes were set to 64 and were processed through Pytorch Dataloaders. The cut scores used to delineate the continuous class predictions in the last layer of the model were determined separately for each trained/tuned model based on a grid search maximizing F1-score in validation. Training occurred on 32 CPU

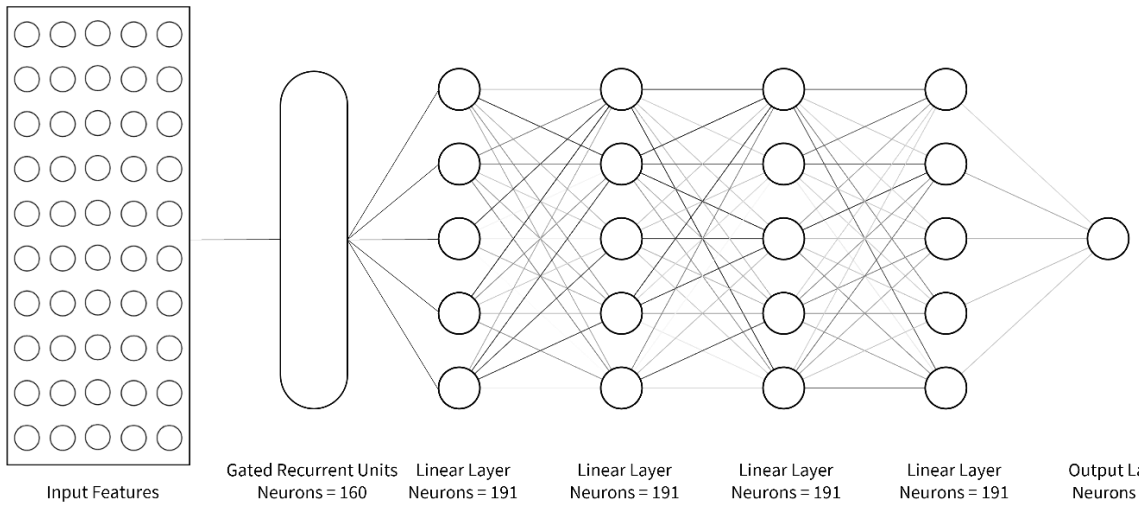
cores and 4 GPU cores split between the OrangeGrid and SUrge distributed computing systems at Syracuse University, with job scheduling managed by HTCondor (Thain et al., 2005).

Table 2-3. Hyperparameter Tuning for Deep Neural Networks

Parameter	**Possible Values**	**Tuned Value in GRNN**
Dropout	$.2 \leq x \leq .5$	dropout = .283
Learning Rate	$1e\text{-}5 \leq x \leq 1e\text{-}1$	lr = 3.7e-4
Weight Decay	$1e\text{-}5 \leq x \leq 1e\text{-}1$	weight_decay = 1.68e-5
Neuron Type	ReLU, Gelu, Sigmoid, Tanh, Hardswish, ELU, Hardsigmoid, Rrelu, LogSoftmax	torch.nn.Hardswish()
Optimizer	Adam, RMSprop, SGD, ASGD	torch.optim.RMSprop()
CNN/GRNN Layers	$1 \leq x \leq 4$	num_layers = 1
Neurons CNN/GRNN	$16 \leq x \leq 1024$	hidden_size = 160
Linear Layers	$1 \leq x \leq 4$	4 * torch.nn.Linear()
Neurons Linear Layers	$8 \leq x \leq 1024$	191

Note. Values selected as a result of hyperparameter tuning of the best-performing gated recurrent neural network (GRNN) are in bold. Hardswish is a piecewise activation (e.g., link) function that transforms neuron output (x) to zero when $x \leq -3$, does not transform neuron output when $x \geq +3$, and otherwise transforms output by $x(x+3)/6$. RMSprop is an adaptive algorithm that divides gradients by the root of the moving average of the square of neuron gradients.

Figure 2-5. Architecture for the Best Performing Neural Network.



Note. This gated recurrent neural network includes one recurrent layer, four fully connected linear layers, and an output classification layer.

Out of the Box Testing and Participant Fine-Tuned Personalization

After the models were trained and tuned, we tested model performance on the test-set subjects whose speech the models had not yet encountered. The trained model from each neural network architecture was used to predict listener perceptual judgment for each utterance in the test set, one participant at a time. We first did this using the *out-of-the-box* hyperparameters and model decision threshold that maximized validation set performance during model training. Next, to reflect the customization possible in a clinical setting, the candidate algorithms were *personalized*, one test set participant at a time. To do this, data for each test participant was re-partitioned into separate re-training, re-validation, and test sets using 5-fold cross validation. The participant-average number of recorded utterances used to personalize the system for test set participants ($\overline{x}$ = 803.3, $\sigma_{\overline{x}}$ = 138.2) could be collected in just over eight instances of 5-minute, 100-item word list recordings (possibly spread out over 1-2 weeks, as in an ongoing clinical trial with this tool). Model hyperparameters during personalization were not retuned, meaning the hyperparameters for participant-specific re-training were identical to those maximizing the

validation set for each architecture except that, in deep neural networks, the learning rate was fixed at 1e-3, weight decay was fixed at zero, class-decision thresholds were updated for each participant, and batch size was lowered. Gradients were fixed for all but the last two layers of each neural network. For shallow neural networks in Research Question 1, 100 participant-specific trees were added to the random forest and support gradient descent classifiers were trained for an additional 10 epochs.

Explainable AI

We performed two post-hoc analyses with the best performing model from Research Question 2 to explain 1) the presence of age and sex effects on classifier performance and 2) the relative importance of formant features to model predictions.

We fit a (second) linear mixed-effects model to explore the presence of age and sex effects on classifier performance for the best-performing classifier. This statistical model was also fit with restricted maximum likelihood estimation, again using the lmer function in lme4. Fixed effects for *age, sex,* and an *age-by-sex interaction* were included, with random intercepts for *participants*. *Sex* was effects coded with *male* serving as the reference level. The outcome variable was *participant-specific F1-score*.

Feature importance was analyzed as Research Question 3, for which we conducted a SHAP analysis (SHapley Additive exPlanations; Lundberg & Lee, 2017) to explain the relative contribution of each acoustic feature to the model prediction. SHAP draws from game theory, specifically Shapley Values, to determine the marginal contribution of an individual feature to final model performance (Yang, 2019). Shapley Value contributions can be directly computed through a series of leave-one-out permutations in which model output with the target feature is compared to the output of models lacking that feature, considering both the main effect of the

left-out feature and the interactions of the target feature with other features. To estimate Shapley Values in this analysis, we employed the GradientShap algorithm in the Captum interpretability library for Pytorch. SHAP analyses are performed with the (best-performing) trained model, but operate on background and input datasets instead of training/validation/test datasets. Background sets define the priors against which the model predictions from input-set tokens with left-out features are compared (Yuan et al., 2022). Because we wanted to explain the impact of different features (i.e., formants, /ɹ/ time window) on the predictions the model makes, the background set was comprised of features from 1,973 utterances with ambiguous ratings from expert listeners, representing 84 participants (average of expert listener ratings = $.3 < x < .7$). In contrast, the input set included a randomly sampled set of 2,405 utterances that received unanimous 0 or 1 ground-truth ratings from expert or crowdsourced listeners, representing all participants. With this experimental design, we could estimate which features the PERCEPT-R Classifier had learned to differentiate ground-truth derhotic or fully rhotic productions from ambiguous input. Because our feature input for the 2,405 utterances in the explainability set was of the shape [5 formants, 10 time windows, 8 statistical representations, the GradientShap output was a 4-dimensional array of shape [2,405 utterances * 5 formants * 10 time windows * 8 statistical representations]. We were interested in *global importance* of the features, so we averaged the output to a 3-dimensional array of shape [5, 10, 8] representing the utterance-wise mean Shapley estimate for the features. We then examined the global importance for formants/formant transforms by averaging across time windows and statistical representations, and the global importance for time windows by averaging across formants/formant transforms and statistical representations.

Results

Research Question 1: Do age-and-sex normalized formants improve F1-score relative to utterance-normalized formants and utterance-normalized MFCCs?

The performance of shallow neural networks trained on age-and-sex normalized formants was compared to the performance of classifiers trained on utterance-normalized formant feature sets and MFCC feature sets (Table 4). Each stochastic gradient descent classifier and random forest of decision trees was tuned individually. Age-and-sex normalized features outperformed utterance-normalized formants and utterance-normalized MFCCs at all three stages of model development. Table 4 shows participant mean-performance averaged across both shallow classifier architectures.

A linear mixed model evaluating the result sin Table 4, fit by restricted maximum likelihood estimation, converged with no warnings. The model contained significant main effects for: *feature input,* indicating that participant-specific F1-score improved by .10 with age-and-sex normalized formants over utterance-normalized MFCCs ($\hat{\beta}$ = .105, *SE* = .026, *df* = 275, *t* = 4.07 $p < .001$); *classifier type*, indicating participant-specific F1-score improved by .10 with the stochastic gradient descent classifier over the random forest classifier ($\hat{\beta}$ = .100, *SE* = .026, *df* = 275, *t* = 3.9 $p < .001$); and *retraining timepoint*, indicating participant-specific F1-score improved by .06 after retraining versus out-of-the-box predictions ($\hat{\beta}$ = .060, *SE* = .026, *df* = 275, *t* = 4.07 *p* = .02). The pairwise comparison between utterance-normalized formants and utterance-normalized MFCCs was not significant ($\hat{\beta}$ = .03, *SE* = .026, *df* = 275, *t* = 1.26, *p* = .21). Significant interactions involving *classifier type* and *feature input* indicate that the stochastic gradient descent classifier performed worse than the random forest classifier in the context of utterance-normalized formants ($\hat{\beta}$ = -.11, *SE* = .036, *df* = 275, *t* = -2.93, $p < .01$) and age-and-sex normalized formants ($\hat{\beta}$ = -.09, *SE* = .036, *df* = 275, *t* = -2.58, *p* = .01). Lastly,

significant interactions involving *classifier type* and *timepoint* indicate that the advantage of the stochastic gradient descent classifier is tempered during retraining ($\hat{\beta}$ = -.09, *SE* = .036, *df* = 275, *t* = -2.7 *p* < .01). These results support the study's first hypothesis.

Table 2-4. Shallow Neural Network Feature Comparison.

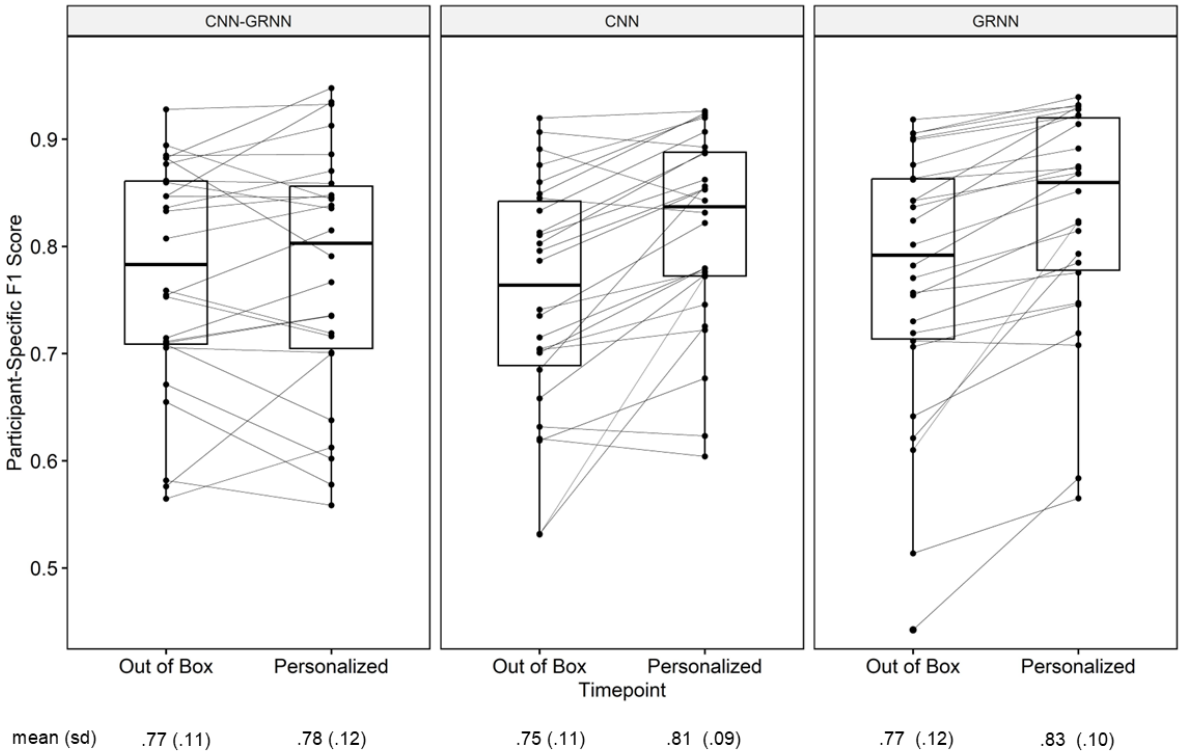
Feature	**Validation Performance**	**Test – Out-of-Box**	**Test – Personalization**
MFCCs	.66 (.13)	.68 (.16)	.69 (.17)
Formants (utterance-normalized)	.63 (.10)	.66 (.09)	.70 (.10)
Formants (age-and-sex normalized)	.73 (.11)	.74 (.13)	.77 (.12)

Note. Performance reported as Participant-Specific F1-Scores ($\overline{x}$, σ_x)

Research Question 2: Can the mean participant-specific F1-score exceed our .8 threshold for clinical acceptability in participants with more perceptually ambiguous feature spaces?

Research Question 2 maximized overall classifier performance by training deep neural networks using the best-performing feature set from Research Question 1. The primary classification outcome is the mean participant-specific F1-score (each itself the mean of 5-fold cross validation within participant) for the prediction of perceptual judgement of "fully rhotic" or "derhotic" /ɹ/, after retraining-based personalization to yet-unseen test participants. Models were only trained on age-and-sex normalized formant feature sets, the best performing features in Research Question 1, to reserve computational resources. The GRNN was the best performing architecture both out of the box and after personalization (Figure 6).

Figure 2-6. Participant-Specific F1-Scores in the Test Set.



Note. CNN = convolutional neural network, GRNN = gated recurrent neural network, sd = standard deviation.

Replication: Swapping Validation and Test Sets

The validation and test sets represent ~30% of the total utterances but only ~15% of the total available participants because the stimulability criteria for the test and validation sets happened to prioritize participants with many tokens. We also noticed that the test set consistently outperformed the validation set with both shallow and deep classifiers. Therefore, to be able to measure performance on the broadest set of participants possible and increase external validity of this study to future clinical scenarios, we retrained/retuned a GRNN on the age-and-sex normalized formant feature set after swapping validation and test sets (Table 5; Table 6). Furthermore, we noticed our personalization procedures lowered performance versus out-of-box testing for five of the participants. Because we would be able to flag these cases in clinical use, we present a final condition that represents the best performance per speaker from the out-of-box

and personalized testing timepoints. This combined, final performance is our reported result for the research question at hand. Of note, in all experiments, the median value was greater than the mean value presented in the tables ($median_{originalfinal}$ = .86, $median_{replicationfinal}$ = .80, $median_{combinedfinal}$ = .83). Overall, these results provide support that classification accuracy can exceed our .8 threshold for clinical acceptability for the average participant with a more perceptually ambiguous feature spaces than in Benway, Preston, Hitchcock, Salekin, et al. (2022).

Table 2-5. Mean and Standard Deviations of Participant-Specific F1-Scores: Age-and-Sex Normalized Formants with GRNN

Experiment	**Validation Performance**	**Test Out of Box**	**Test (CV) Personalized**	**Test Final**
Original	.75 (.11)	.77 (.13)	.83 (.08)	.83 (.11)
Replication	.80 (.09)	.75 (.11)	.79 (.09)	.80 (.09)
Combined	.77 (.10)	.76 (.12)	.81 (.10)	.81 (.10)

Note. CV = cross-validated

Table 2-6. Participant-Weighted Confusion Matrix for Final, Combined Experiment

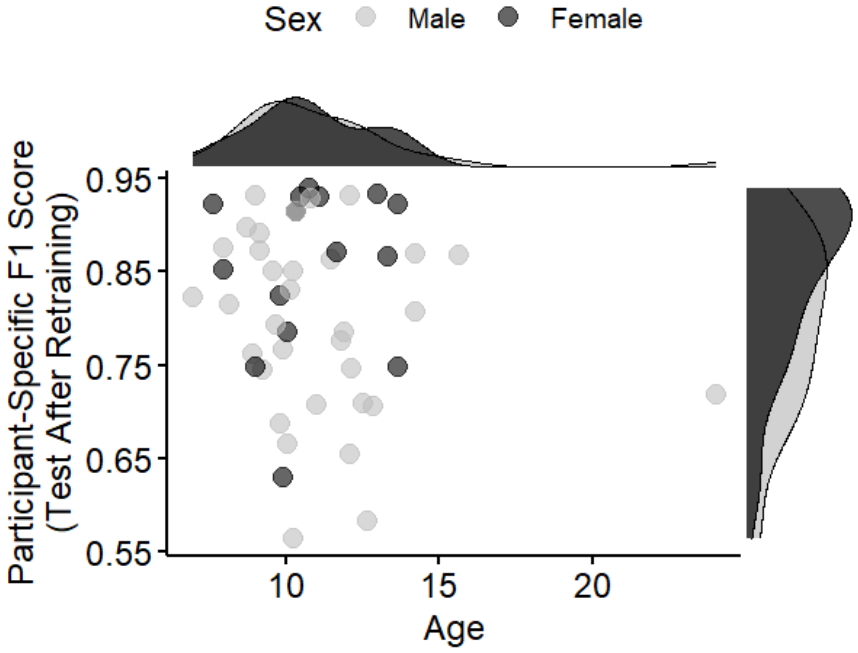
GRNN Prediction	**Ground-Truth Derhotic**	**Ground-Truth Fully Rhotic**
Derhotic	.70	.30
Fully Rhotic	.12	.88

Note. Average participant-specific F1-score = .81; $\sigma_{\bar{x}}$= .10; med = .83, n = 48, standardized by ground-truth rating.

Exploration of Model Fairness Regarding Age and Sex

The validation and test sets were hand crafted to compare model performance for male and female participants meeting our stimulability criteria at a range of ages. This exploration helps inform whether we should further develop normalization methods or demographic-specific classifiers for end use of the PERCEPT-R Classifier in the clinic. A linear mixed model fit on the combined test dataset (n = 48 left-out participants; Figure 7) indicates that neither the fixed effects of age ($\hat{\beta}$ = -0.020, *SE* = .018, *t* = -1.13, *p* = .265) nor sex ($\hat{\beta}$ = -0.083, *SE* = .16, *t* = -.52, *p* = .606) nor the age-sex interaction ($\hat{\beta}$ = 0.013, *SE* = .0115, *t* = .904, *p* = .371) were significant. These results do not provide evidence that classifier performance was systematically biased for or against individuals of a particular age or sex in the present dataset, and do not raise model fairness/ethical concerns regarding clinical use of the current version of the model relative to these demographic characteristics.

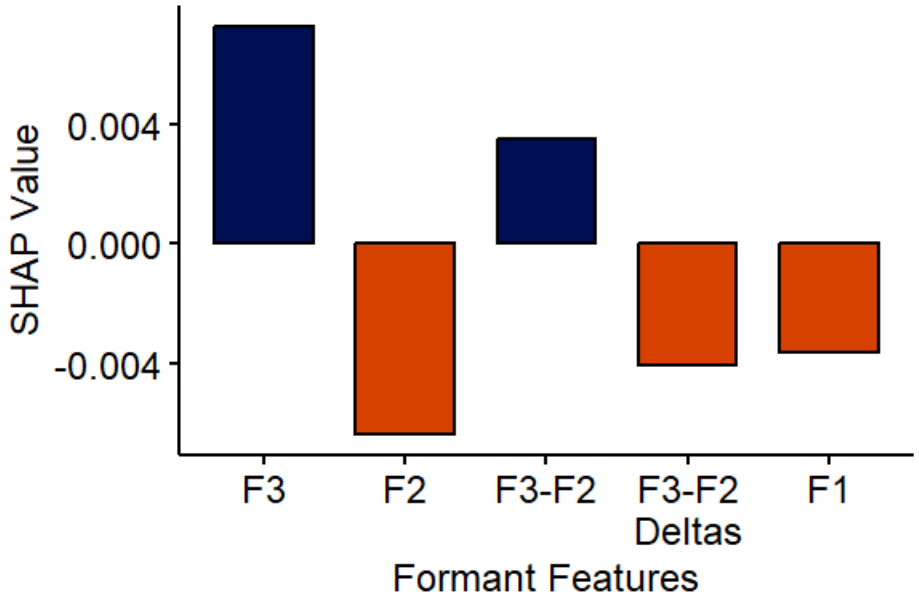
Figure 2-7. GRNN F1-Scores, by Participant Age and Sex



Research Question 3: What is the relative importance of the individual acoustic features within the PERCEPT-R classifier?

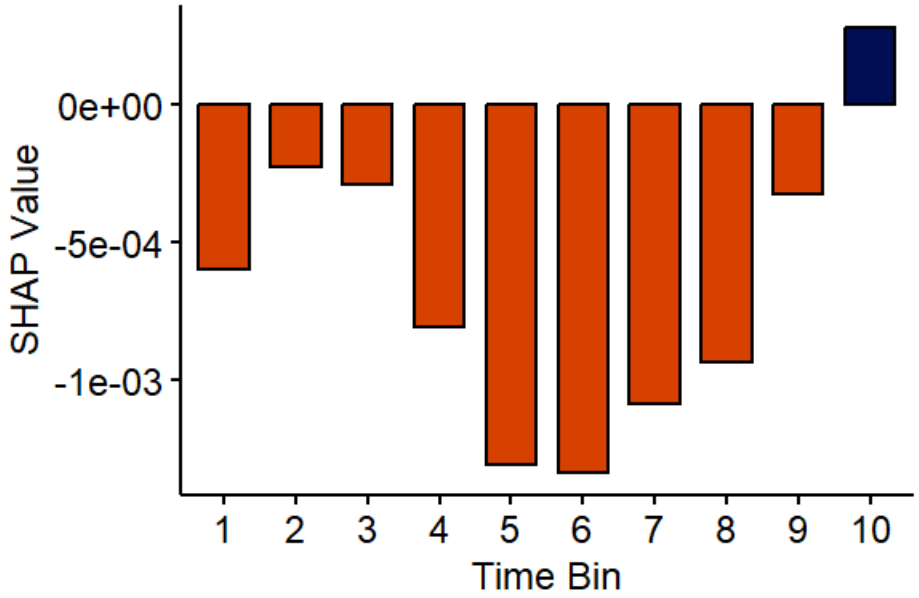
We estimated the marginal contributions for induvial formant features/transforms and time intervals. SHAP estimates for F3 and F3-F3 were positively signed (Figure 8). Because positive SHAP estimates indicate relative influence for pushing predictions toward class 1 (i.e., fully rhotic/clinically correct), we interpret this as evidence that age-and-sex normalized F3 and F3-F2 distance are salient for differentiating unanimous fully rhotic productions from ambiguous input, independent of all other formant features. Conversely, SHAP estimates were negatively signed for age-and-sex normalized F2, F3-F2 deltas, and F1. Because negative SHAP estimates indicate relative influence for pushing predictions toward class 0 (i.e., derhotic/clinically incorrect), we interpret this as evidence that, independent of age-and-sex normalized F3 and F3-F2, values for age-and-sex normalized F2, F3-F2 deltas, and F1 were salient for differentiating unanimous derhotic productions from ambiguous input. The high absolute SHAP estimate for age-and-sex normalized F3 indicated that this feature contributed the most overall importance to PERCEPT-R Classifier predictions, partly supporting our hypothesis that age-and-sex normalized F3-F2 would be the most salient factor in our model (e.g., Campbell et al., 2018). Although the F3-F2 hypothesis was not fully supported, the importance of age-and-sex normalized F3 in the PERCEPT-R Classifier has a strong theoretical grounding from the acoustic phonetics of /ɹ/ and these results do not raise concerns that the model was confounded in some way.

Figure 2-8. SHAP Estimates, Ordered Left to Right by Overall Feature Importance



We also explored the relative feature importance of the temporal bins to inform future development of the PERCEPT-R Classifier. Time windows 1-9 (Figure 9) had negative SHAP estimates, with those in the middle of the time series having the most influence; we interpret this as evidence that, when PERCEPT-R makes a prediction of class 0, it is most often because of what is happening in the middle of the rhotic-associated interval. Conversely, SHAP values were positive for the last time window, indicating relative influence for this interval in pushing predictions toward class 1. We interpret this as evidence that the end of the rhotic-associated interval is salient for fully rhotic predictions and justifies the use of sequential models like the best-performing GRNN. Note, however, the low absolute magnitude of the SHAP value for time window 10 compared to time windows 1-9. Because SHAP values are additive, the importance of this time window is likely canceled out by the other time windows.

Figure 2-9. SHAP Values, Ordered by Time



Discussion

This study documents the technical development of the PERCEPT-R Classifier and evaluates the acoustic features that optimize the binary prediction of a listener's perceptual judgment (i.e., fully rhotic/derhotic) of /ɹ/ in words produced by children with SSD. In the long term, we hope this research leads to the development of validated clinical speech technology tools that provide feedback within a motor learning framework and adapt the difficulty of the session to the client's performance. Indeed, our ongoing clinical trials are evaluating the use of the PERCEPT-R Classifier for this purpose. We believe such tools can narrow the intensity gap between research-based recommendations for a high number of trials in frequently spaced sessions and the traditional intensity of services available with current delivery models. We foresee the main use of these tools to be through at-home practice with clinical-grade feedback that adapts difficulty based on a learner's performance.

Machine learning experiments supported the study's first hypothesis: we found that age-and-sex normalization of formants greatly improved shallow neural network classifier accuracy versus utterance-normalized formants and utterance-normalized MFCCs. Because none of our formant features were hand-corrected, we interpret the higher performance of age-and-sex normalized formants versus MFCCs as evidence that meaningful formant features can be extracted in an automated fashion and can better capture perceptual judgement than MFCCs, despite the known pitfalls of LPC formant estimation. The advantage of formants, however, is only seen in the context of age-and-sex normalization (versus utterance normalization), which we interpret as evidence that this normalization technique can better reconcile perception of rhoticity in a way that is less confounded by vocal tract size. Notably, the non-superior performance of utterance-normalized formants in out-of-box testing versus MFCCs was previously reported in sociophonetic classification of /ɹ/ in rhotic versus non-rhotic dialects (Gupta & DiPadova, 2019).

Our second hypothesis was also supported: we found that that the best-performing GRNN classifier, trained on the best-performing feature set from Research Question 1, surpassed our participant-specific F1-score threshold for clinical utility (.8) in participants theorized to have more perceptually ambiguous feature spaces than previously tested by Benway, Preston, Hitchcock, Salekin, et al. (2022). We also strengthened the external generalizability of these findings through replication, by switching the original validation and test subsets, rerunning the experiment, and presenting the average of the two experiments as the final outcome for this research question: average test-participant-specific F1-score = .81 (σ_x= .10; med = .83, n = 48). As a whole, the work surpasses the average test-participant-specific F1-score for state-of-the-art rhotic classification in speech sound disorder from that previously shown in the literature ($\bar{x}$ =

.64, σ_x= .25; Ribeiro et al., 2021). Of note is that Ribeiro et al. (2021) excluded tokens with non-unanimous ratings from their analysis, a methodological decision that was also undertaken in previous investigations of mispronunciation detection (Strik et al., 2009), sociophonetic rhoticity classification (Gupta & DiPadova, 2019). A particular strength of the present study, however, is that we did *not* exclude such tokens. While omitting these tokens would increase the reported performance of the PERCEPT-R Classifier — indeed, F1-score in test, using only the subset of unanimously rated PERCEPT-R Corpus tokens, out-of-the-box, shows $\bar{x}_{\text{F1-score}}$ = .88 — ambiguous tokens with rater disagreement are clinically encountered and the exclusion of these tokens during model development would have lowered the external validity and clinical utility of the work presented here.

Although we had expected the CNN-GRNN to outperform the other model architectures, this expectation was not realized. We interpret this, as well as the lack of large differences between tested neural network architectures, as evidence for the saliency of age-and-sex normalized formants and formant transforms for the detection of rhoticity. Theoretically, the convolutional filters of a CNN can learn to extract relevant, frequency-based information from audio. Further work can determine if we can train a CNN to effectively extract formant-analogous features from the audio itself, requiring less hands-on customization than is required by LPC formant estimation.

The lower F1-score seen in replication might possibly be due to participant or audio differences within the original validation and test subsets. Acoustician confidence in manual formant ceiling estimation was informally noted to be lower in the validation set, possibly due to participant-specific audio quality issues at the time of data collection. Recently completed work evaluates classifier performance in a third, prospectively collected group of individuals with

RSSD (Benway & Preston, under review). This ongoing work also examines clinical efficacy of clinical speech technology that uses the PERCEPT-R Classifier.

In the present study, we also explored the effect of age and sex on classifier performance, which provided statistical evidence that the GRNN predictions were not significantly influenced by sex or age across the range covered by the original-test and replication-test subsets in the context of age-and-sex normalized features. We interpret this as evidence that the classification performance demonstrated herein can extend to individuals who present to the clinic during prospective validation and are demographically and/or acoustically similar to those in the validation or test sets. It remains to be seen if age and/or sex effects exists in explorations with larger sample sizes and higher power to detect such differences, particularly when increasing the sample size at every age range also allows for testing with more young adults than our current corpus affords. Note that it is possible that the method we used to create the validation and test sets – ensuring that if there was only a handful of participants at a particular age-sex stratum, they were assigned to the validation and tests sets – created a selection bias in our training set. Future data collection can recruit additional young adult participants to allow for a more robust test of age and/or sex effects on classifier performance. Evaluation of performance with regard to speaker ethnicity was not possible in the current dataset, and is discussed as a limitation below.

Our third hypothesis was partially supported: a SHAP experiment with the best-performing classifier indicated that (age-and-sex normalized) F3 was the most influential feature on individual predictions made by the model when differentiating ambiguous tokens from unanimously rated fully rhotic and unanimously rated derhotic tokens. The importance of F3, a formant well-established in the acoustic phonetics literature as a reflection of rhoticity, in this post-hoc analysis provides a “sanity check” that the classifier did indeed learn something

perceptually relevant about /ɹ/. We consider our hypothesis to only be partly supported, however, because we had expected the transform of F3, F3-F2 distance, to be the most influential feature for the reasons established prior. It is possible that the neural network was able to model the F3-F2 relationship effectively using non-transformed F3 and F2. It may also be possible that F3 is more important than F3-F2 specifically during formant transitions. We still take the importance of F3 as evidence that the PERCEPT-R Classifier is learning previously validated, perceptually salient features for /ɹ/. This evidence reduces the likelihood that the model performance was confounded in some way that would not generalize to use with audio collected from future participants.

Clinical Interpretation, Limitations, and Future Directions

The major limitation from this paper is that lab testing, although important, likely overestimates predictive performance of clinical speech technology systems (Berisha et al., 2022). Because of this, we reserve clinical interpretation of these results and subsequent clinical validation after a recently completed clinical trial (Benway & Preston, in preparation), and a fourth, companion article (Benway & Preston, under review).

The limitation of lab testing is particularly salient when the lab data do not permit exploration of model performance with regard to demographic characteristics that may be underrepresented in the training data. Specifically, the current composition of the PERCEPT-R Corpus did not allow for the meaningful evaluation of model performance and/or model fairness relative to race or ethnicity. Our concurrent projects begin to address this paramount concern by increasing PERCEPT Corpus representation of typical speakers and speakers with RSSD from fully rhotic dialects of American English who identify as Black, Indigenous, Hispanic, Asian/Pacific Islander, multi-racial, and/or other underrepresented communities. We are

particularly motivated by this goal, as these communities are often most underserved with regards to intervention intensity and well designed, fair, and ethical clinical speech technologies may be a particularly useful tool for clinicians seeking to narrow the intervention intensity gap in these communities.

An additional limitation related to clinical utility of the present results is the amount of hand crafting that would be involved for feature generation and labeling for retraining in a prospective clinical context, and, eventually, at scale in a production version of the software. Our ongoing research seeks to meet or exceed the present performance with more automated measures and facilitate this process for clinicians with our existing suite of clinical software, including with the Speech Motor Chaining web application (Preston et al., 2022).

Another limitation of the present study is that, although measures of central tendency (i.e., mean, median) for participant-specific F1 in our original experiment and replication surpassed our threshold for clinical utility, the entire range of participant-specific F1-score performance did not fall above this threshold. Future work can identify participant- or context-specific factors that may maximize the number of participants for whom PERCEPT-R can be clinically useful, such as the subtype of /ɹ/ distortion or the number of files available for PERCEPT-R personalization.

Finally, it may be possible to increase performance of MFCC feature sets through age-and-sex normalization analogous to that performed by (Lee et al., 1999), which future studies can explore with PERCEPT Corpus participants.

Conclusions

This study details the development and validation of a mispronunciation detection algorithm for speech sound disorders impacting /ɹ/ in American English, the PERCEPT-R

Classifier. This article presents an age-and-sex normalized formant extraction methodology that outperforms MFCC features for the classification of fully rhotic vs derhotic /ɹ/ in the context of speech sound disorder mispronunciation detection. The lab-tested GRNN trained on these formant features outperformed participant-specific average F1-score from the literature by 17 points ($\bar{x}$ = .81, σ_x= .10, med = .83, n = 48). An explainability analysis indicated that F3 is most influential feature in classifier predictions, in line with acoustic phonetic descriptions of /ɹ/. Exploration of model performance regarding age and sex of participants did not highlight fairness issues in the current set of participants. Our ongoing work examines the clinical impact of the PERCEPT-R Classifier for /ɹ/-based speech sound disorder.

Acknowledgements

We are grateful to Harshit Sharma, Yi Xiao, and Dr. Michael McAuliffe, who provided valuable technical feedback throughout these experiments. We are also grateful to PERCEPT-R Corpus participants and their families. This research was supported through an internal grant (CUSE II-14-2021; J. Preston, PI) and computational resources (NSF ACI-1341006; NSF ACI-1541396) provided by Syracuse University, and by the National Institute on Deafness and Other Communication Disorders (NIH R01DC017476-S2; T. McAllister, PI).

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining,

Ball, M. J. (2017). Transcribing rhotics in normal and disordered speech. *Clinical Linguistics & Phonetics, 31*(10), 806-809. https://doi.org/10.1080/02699206.2017.1326169

Barreda, S. (2021). Fast Track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard, 7*(1).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Benway, N. R., Hitchcock, E., McAllister, T., Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual /ɹ/ errors in American English: A single case randomization design. *American Journal of Speech-Language Pathology*.

Benway, N. R., & Preston, J. L. (in preparation). Artificial Intelligence Assisted Speech Therapy for /ɹ/ using Speech Motor Chaining and the PERCEPT Engine: a Single Case Experimental Clinical Trial with ChainingAI.

Benway, N. R., & Preston, J. L. (under review). Prospective Validation of Motor-Based Intervention with Automated Mispronunciation Detection of Rhotics in Residual Speech Sound Disorders.

Benway, N. R., Preston, J. L., Hitchcock, E. R., & McAllister, T. (2022). *PERCEPT-R Corpus*. https://doi.org/10.21415/0JPJ-X403

Benway, N. R., Preston, J. L., Hitchcock, E. R., Rose, Y., Salekin, A., Liang, W., & McAllister, T. (in press). Reproducible Speech Research with the Artificial-Intelligence-Ready PERCEPT Corpora *Journal of Speech, Language, and Hearing Research*.

Benway, N. R., Preston, J. L., Hitchcock, E. R., Salekin, A., Sharma, H., & McAllister, T. (2022). *PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of /ɹ/* INTERSPEECH 2022: Proceedings of the 23rd Annual Conference of the International Speech Communication Association (ISCA), Incheon, Republic of Korea.

Berisha, V., Krantsevich, C., Stegmann, G., Hahn, S., & Liss, J. (2022). Are reported accuracies in the clinical speech machine learning literature overoptimistic? Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,

Boersma, P., & Weenink, D. (2019). *Praat [Computer Software]*. (Version 6.1.38) https://www.fon.hum.uva.nl/praat/

Boyce, S. E. (2015). The articulatory phonetics of /r/ for residual speech errors. *Seminars in Speech and Language*, *36*(4), 257-270. https://doi.org/10.1055/s-0035-1562909

Brandel, J., & Frome Loeb, D. (2011). Program intensity and service delivery models in the schools: SLP survey results. *Language Speech and Hearing Services in Schools, 42*(4), 461-490. https://doi.org/10.1044/0161-1461(2011/10-0019)

Campbell, H., Harel, D., Hitchcock, E., & McAllister Byun, T. (2018). Selecting an acoustic correlate for automated measurement of American English rhotic production in children. *International Journal of Speech Language Pathology*, *20*(6), 635-643. https://doi.org/10.1080/17549507.2017.1359334

Chen, Y.-P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A., Doube, W., & Morris, M. E. (2016). Systematic review of virtual speech therapists for speech disorders. *Computer Speech & Language*, *37*, 98-128. https://doi.org/https://doi.org/10.1016/j.csl.2015.08.005

Chilba, T., & Kajiyama, M. (1941). *The Vowel, its Nature and Structure*. Tokyo-Kaiseikan Publishing Company Ltd.

Childers, D. G. (1978). *Modern spectrum analysis*. IEEE Computer Society Press.

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, 6*(44), 29.

Derdemezis, E., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., & Bolt, D. M. (2016). Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *American Journal of Speech-Language Pathology, 25*(3), 335-354. https://doi.org/doi:10.1044/2015_AJSLP-15-0020

Eshky, A., Ribeiro, M. S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J. M., & Wrench, A. (2018). *UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions* INTERSPEECH 2018: Proceedings of the 19th Annual Conference of the International Speech Communication Association (ISCA),

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., & Alwan, A. (2000). Acoustic modeling of American English /r/. *Journal of the Acoustical Society of America, 108*(1), 343-356. https://doi.org/10.1121/1.429469

Furlong, L., Erickson, S., & Morris, M. E. (2017). Computer-based speech therapy for childhood speech sound disorders. *Journal of Communication Disorders, 68*, 50-69. https://doi.org/https://doi.org/10.1016/j.jcomdis.2017.06.007

Furlong, L., Morris, M., Serry, T., & Erickson, S. (2018). Mobile apps for treatment of speech disorders in children: An evidence-based analysis of quality and efficacy. *PLoS One, 13*(8), e0201513. https://doi.org/10.1371/journal.pone.0201513

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Graves, A., Mohamed, A., & Hinton, G. (2013, 26-31 May 2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing,

Gupta, S., & DiPadova, A. (2019, June). Deep Learning and Sociophonetics: Automatic Coding of Rhoticity Using Neural Networks. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, Minneapolis, Minnesota.

Hair, A., Ballard, K. J., Markoulli, C., Monroe, P., Mckechnie, J., Ahmed, B., & Gutierrez-Osuna, R. (2021). A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with Apraxia World. *ACM Trans. Access. Comput., 14*(1), Article 3. https://doi.org/10.1145/3433607

Harper, S., Goldstein, L., & Narayanan, S. (2020). Variability in individual constriction contributions to third formant values in American English /ɹ/. *The Journal of the Acoustical Society of America, 147*(6), 3905-3916. https://doi.org/10.1121/10.0001413

Heselwood, B., & Plug, L. (2011). The Role of F2 and F3 in the Perception of Rhoticity: Evidence from Listening Experiments. ICPhS,

Huang, J.-T., Li, J., & Gong, Y. (2015). An analysis of convolutional neural networks for speech recognition. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics, 71*, 1-15. https://doi.org/https://doi.org/10.1016/j.wocn.2018.07.001

Kaipa, R., & Peterson, A. M. (2016). A systematic review of treatment intensity in speech disorders. *International Journal of Speech Language Pathology, 18*(6), 507-520. https://doi.org/10.3109/17549507.2015.1126640

Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048*.

Klein, H. B., McAllister Byun, T., Davidson, L., & Grigos, M. I. (2013). A Multidimensional Investigation of Children's /r/ Productions: Perceptual, Ultrasound, and Acoustic Measures. *American Journal of Speech-Language Pathology, 22*(3), 540-553. https://doi.org/10.1044/1058-0360(2013/12-0137)

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105*(3), 1455-1468. https://asa-scitation-org.libezproxy2.syr.edu/doi/pdf/10.1121/1.426686

Leung, W., Liu, X., & Meng, H. (2019, 12-17 May 2019). CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Lewis, B. A., Freebairn, L., Tag, J., Ciesla, A. A., Iyengar, S. K., Stein, C. M., & Taylor, H. G. (2015). Adolescent outcomes of children with early speech sound disorders with and without language impairment. *American Journal of Speech-Lanugage Pathology, 24*(2), 150-163. https://doi.org/10.1044/2014_AJSLP-14-0075

Li, S. R., Dugan, S., Masterson, J., Hudepohl, H., Annand, C., Spencer, C., Seward, R., Riley, M. A., Boyce, S., & Mast, T. D. (2023). Classification of accurate and misarticulated /ɑr/ for ultrasound biofeedback using tongue part displacement trajectories. *Clinical Linguistics & Phonetics*, *37*(2), 196-222. https://doi.org/10.1080/02699206.2022.2039777

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q. M., Sainath, T. N., Senior, A., Beaufays, F., & Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. INTERSPEECH 2015: Proceedings of the 16th Annual Conference of the International Speech Communication Association (ISCA), Dresden, Germany.

Mahrt, T. (2016). *PraatIO*. https://github.com/timmahrt/praatIO

McAllister Byun, T., Harel, D., Halpin, P. F., & Szeredi, D. (2016). Deriving gradient measures of child speech from crowdsourced ratings. *Journal of Communication Disorders*, *64*, 91-102. https://doi.org/10.1016/j.jcomdis.2016.07.001

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). *Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi* INTERSPEECH 2017: Proceedings of the 18th Annual Conference of the International Speech Communication Association (ISCA), Stockholm, Sweeden.

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech Language Pathology*, *20*(6), 583-598. https://doi.org/10.1080/17549507.2018.1477991

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*(6174), 1006-1010. https://doi.org/10.1126/science.1245994

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. FAT* '19: Conference on Fairness, Accountability, and Transparency, Atlanta, Georgia, USA.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Plante, E., & Vance, R. (1994). Selection of Preschool Language Tests. *Language, Speech, and Hearing Services in Schools*, *25*(1), 15-24. https://doi.org/10.1044/0161-1461.2501.15

Povey, D. (2012). *train_map.sh.*

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., & Schwarz, P. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding,

Preston, J. L., Benway, N. R., Leece, M. C., Hitchcock, E. R., & McAllister, T. (2020). Tutorial: Motor-based Treatment Strategies for /r/ Distortions. *Language, Speech, and Hearing Services in Schools, 54*, 966-980.

Preston, J. L., Caballero, N. F., Leece, M. C., Wang, D., Herbst, B. M., & Benway, N. R. (under review). A Randomized Controlled Trial of Treatment Distribution and Biofeedback Effects on Speech Production in School-Aged Children with Apraxia of Speech.

Preston, J. L., Leece, M. C., & Maas, E. (2017). Motor-based treatment with and without ultrasound feedback for residual speech-sound errors. *International Journal of Language & Communication Disorders, 52*(1), 80-94. https://doi.org/10.1111/1460-6984.12259

Ribeiro, M. S., Cleland, J., Eshky, A., Richmond, K., & Renals, S. (2021). Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors. *Speech Communication, 128*, 24-34. https://doi.org/https://doi.org/10.1016/j.specom.2021.02.001

Robles Herrera, S., Ceberio, M., & Kreinovich, V. (2022). When is deep learning better and when is shallow learning better: qualitative analysis. *International Journal of Parallel, Emergent and Distributed Systems, 37*(5), 589-595.

Ruscello, D. M. (1995). Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders, 28*(4), 279-302.

Shahin, M., Zafar, U., & Ahmed, B. (2020). The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 400-412. https://doi.org/10.1109/JSTSP.2019.2959393

Shriberg, L. D., Flipsen Jr, P., Karlsson, H. B., & McSweeney, J. L. (2001). Acoustic phenotypes for speech-genetics studies: An acoustic marker for residual/ɜ/distortions. *Clinical Linguistics & Phonetics, 15*(8), 631-650.

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom),

Strik, H., Truong, K., de Wet, F., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication, 51*(10), 845-852. https://doi.org/https://doi.org/10.1016/j.specom.2009.05.007

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the Condor experience. *Concurrency and computation: practice and experience, 17*(2-4), 323-356.

Wren, Y., Miller, L. L., Peters, T. J., Emond, A., & Roulstone, S. (2016). Prevalence and Predictors of Persistent Speech Sound Disorder at Eight Years Old: Findings From a Population Cohort Study. *Journal of Speech, Language, and Hearing Research, 59*(4), 647-673. https://doi.org/10.1044/2015_JSLHR-S-14-0282

Yang, X., Loukina, A., & Evanini, K. (2014, 7-10 Dec. 2014). Machine learning approaches to improving pronunciation error detection on an imbalanced corpus. 2014 IEEE Spoken Language Technology Workshop (SLT),

Yuan, H., Liu, M., Krauthammer, M., Kang, L., Miao, C., & Wu, Y. (2022). An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. *arXiv preprint arXiv:2204.11351*.

CHAPTER 3 – REPRODUCIBLE SPEECH RESEARCH WITH THE ARTIFICIAL-INTELLIGENCE-READY PERCEPT CORPORA

Nina R. Benway, Department of Communication Sciences & Disorders, Syracuse University, Syracuse, NY, USA

Jonathan L. Preston, Department of Communication Sciences & Disorders, Syracuse University, Syracuse, NY, USA; Haskins Laboratories, New Haven, CT, USA

Elaine Hitchcock, Department of Communication Sciences & Disorders, Montclair State Uni., Montclair, NJ

Yvan Rose, Department of Linguistics, Memorial University, Newfoundland and Labrador, Canada

Asif Salekin, Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, USA

Wendy Liang, Department of Communicative Sciences & Disorders, New York University, New York, NY, USA

and

Tara McAllister, Department of Communicative Sciences & Disorders, New York University, New York, NY, USA

Corresponding author: Nina R Benway nrbenway@syr.edu, Ph: 1-315-443-4485 Dept of Communication Sciences & Disorders, 621 Skytop Road, Suite 1200, Syracuse, NY 13244

Conflict of Interest: None

Funding: Funding for corpus compilation has been provided by National Institute on Deafness and Other Communication Disorders (NIH R01DC017476-S2, T. McAllister, PI). This research was supported in part through computational resources provided by Syracuse University (NSF ACI-1341006; NSF ACI-1541396).

Current citation: Benway, N. R., Preston, J. L., Hitchcock, E. R., Rose, Y., Salekin, A., Liang, W., & McAllister, T. (in press). Reproducible Speech Research with the Artificial-Intelligence-Ready PERCEPT Corpora. Journal of Speech, Language, and Hearing Research.

Prologue to Chapter 3

Chapter 2 documents the development of the PERCEPT-R Classifier, which predicts human perceptual judgment of fully rhotic versus derhotic /ɹ/ in the context of stimulable speakers. In line with reproducibility suggestions advocated by Berisha et al. (2022), these experiments demonstrated that age-and-sex normalized formant features, which have acoustic and perceptual motivation for relevance to /ɹ/ (particularly for the third formant; Campbell et al., 2017; Espy-Wilson et al., 2000), significantly increased predictive performance over nonspecific Mel-frequency cepstral coefficient features. The final results of this study are strengthened by representing the mean of the original experiment and a replication. Post-hoc exploration did not raise concerns regarding systematic bias of PERCEPT-R Classifier performance based on age or sex of the speaker.

It may be likely that the success of the PERCEPT-R Classifier is due to the depth and breadth of the PERCEPT-R Corpus, one of the corpora discussed in the present chapter. The PERCEPT-R Corpus represents approximately one and a half decades of clinical trial data arising from the labs of Dr. Elaine Hitchcock of Montclair State University, Dr. Tara McAllister of New York University, and Dr. Jonathan Preston of Syracuse University. These corpora have been published as open-access datasets to offset the general lack of large-scale, labeled child speech corpora (Fainberg et al., 2016; Kennedy et al., 2017; Liao et al., 2015; Shahin et al., 2020; Yeung & Alwan, 2018). In fact, Yeung & Alwan (2018) consider child speech recognition, generally, to be "riddled with errors" (p. 1661) due to the lack of appropriate datasets for model building. The chapter that follows describes how the PERCEPT-Corpora might address these needs, as well as separate needs for reproducible research and clinical perceptual training.

Abstract

Background: Publicly-available speech corpora facilitate reproducible research by providing open-access data for participants who have consented/assented to data sharing among different research teams. Such corpora can also support clinical education, including perceptual training and training in the use of speech analysis tools.

Purpose: In this Research Note, we introduce the PERCEPT-R and PERCEPT-GFTA corpora, which together contain over 36 hours of speech audio (> 125,000 syllable, word, and phrase utterances) from children, adolescents, and young adults aged 6-24 with speech sound disorder (primarily residual speech sound disorders impacting /ɹ/) and age-matched peers. We highlight PhonBank as the repository for the corpora and demonstrate use of the associated speech analysis software, Phon, to query the corpus. A worked example of research with PERCEPT-R, suitable for clinical education and research training, is included as an appendix. Support for end users and information/descriptive statistics for future releases of the PERCEPT Corpus can be found in a dedicated Slack channel. Finally, we discuss the potential for PERCEPT corpora to support the training of artificial intelligence clinical speech technology appropriate for use with children with speech sound disorders, the development of which has historically been constrained by the limited representation of either children or individuals with speech impairments in publicly available training corpora.

Conclusion: We demonstrate the use of PERCEPT corpora, PhonBank, and Phon for clinical training and research questions appropriate to child citation speech. Increased use these tools has the potential to enhance reproducibility in the study of speech development and disorders.

Introduction

Reproducible research is supported by the existence of open-access datasets composed exclusively of data from participants who have consented to sharing of their data outside of the original study context. Such datasets mitigate many practical issues related to the sharing of participant data across multiple research teams, and can also be used to support aspects of clinician training. This Research Note covers multiple aims related to promoting reproducibility in research on speech development and disorders through open-access datasets. First, we discuss existing open-access corpora of clinical child speech. Second, we introduce two novel corpora under the PERCEPT project (Perceptual Error Rating for the Clinical Evaluation of Phonetic Targets), which were developed to support reproducibility in research, clinical training, and the training of speech technology tools for the recognition and classification of clinical child speech. The PERCEPT-R and PERCEPT-GFTA audio corpora are centered around the speech of children and young adults with residual speech sound disorder (RSSD) impacting /ɹ/, as well as age-matched peers. Together, the corpora (currently in version 2.2.2p, with *p* denoting the public subset of participants) are unique in their large size, representing 125,632 recorded syllable, word, and phrase speech utterances from 453 participants (to date) across 34 published and unpublished clinical speech studies. Third, we culminate with a discussion of PhonBank, the open-access speech corpora repository through which PERCEPT is distributed, as well as the PhonBank-associated analysis software, Phon. Appendix A includes a worked example of a reproducible research analysis that can be completed with PERCEPT-R and Phon.

Reproducibility and Open Access Speech Corpora

Open-access datasets support reproducible research by mitigating resource, institutional, and researcher-imposed obstacles to the sharing of data. The prevalence of data sharing in the

biomedical sciences has improved over the last decade; however, over 81% (85/104) of PubMed publications sampled between 2015 and 2017 by Wallach et al. (2018) omit a data sharing statement. Furthermore, a data sharing statement does not guarantee that data will actually be made available to third parties. In their study, Gabelica et al. (2022) were unable to access original study data for 93.2% of open-access BioMed Central articles containing a "data available upon request" statement (i.e., 1,416/1,792 study authors did not respond to the data request; 132/1,972 responded "no"; 122/1,792 shared their data). Reasons for not sharing the data are detailed by Gabelica et al. (2022) and most frequently included: original study authors ceasing communication with Gabelica et al. before data were shared (including after non-disclosure agreements were signed), original study authors citing a lack of informed patient consent for data sharing, original study authors indicating loss of access to the data, or original study authors not wanting to share without specific understanding of how the data would be used by Gabelica et al.

Publicly-hosted open-access datasets represent a more reliable alternative to post-hoc data sharing. One repository for such data is PhonBank (https://phon.talkbank.org/). Since its inception in 2006, the PhonBank database project has developed methods and technologies for speech corpus building and phonological/phonetic analysis. The PhonBank repository of speech corpora, and the accompanying open-source software Phon (https://www.phon.ca)[3], were

---

3 The early development of Phon was funded by grants from the Social Sciences and Humanities Research Council of Canada, the Canada Fund for Innovation, as well as by a Petro-Canada Award for Young Innovators. Since 2006, the development of Phon and PhonBank

developed for the study of language acquisition and have become increasingly relevant to the study of speech disorders (e.g., McAllister Byun & Rose, 2016; Rose & Stoel-Gammon, 2015). The PhonBank database now includes a series of clinical corpora documenting speech patterns across populations of typical and atypical language learners. PhonBank is a component of the larger TalkBank database system (https://talkbank.org), alongside the long-standing CHILDES database (Child Language Data Exchange System; MacWhinney & Snow, 1985) and other databases such as AphasiaBank and FluencyBank. The TalkBank project was founded to support and actively contribute to the Open Science movement, with important goals such as increased scientific transparency, re-use of data, cooperation, accountability, and reproducibility for research (https://openscience.org/).

Open-access datasets, such as those on PhonBank, allow analyses to be replicated by third parties, supporting the reliability of research results arising from use of the data. They also permit researchers to address novel research questions relevant to the data that has already been collected. The value of data sharing is magnified in a field such as communication sciences and disorders, where speaker recruitment may be challenging due to low disorder prevalence and slow/costly study procedures, particularly in the context of longitudinal studies and/or clinical trials. In addition to benefitting researchers, open-access datasets also provide instructional opportunities for undergraduate and graduate researchers-in-training, who can practice a breadth of data analysis and statistical analysis techniques without undertaking novel data collection.

---

has been funded primarily through grants from the National Institutes of Health (USA; R01 HD051698, R01 HD051698-06A1, R01 HD051698-11, and R01 HD051698-16).

The benefits of open-access datasets for reproducibility extend past the laboratory context, particularly when a dataset is well-annotated and formatted for use with standardized analysis tools. Open speech corpora can be used to support clinician training, including perceptual training and instruction in the use of speech analysis tools that may be of utility in clinical practice. Additionally, speech corpora serve as the training data underlying the development of clinical speech technologies, such as those involved in automated mispronunciation detection or the forced alignment of speech samples. In the paragraphs below we review two existing speech corpora that meet the following criteria: publicly available, contain the speech of individuals with communication disorders, and overlap with the customary age range after which a speech sound disorder is considered an RSSD, the population of focus in our ongoing research. Note that readers interested in the speech of younger children with speech sound disorder are referred the PhonBank datasets used by Shahin et al., 2020.

Clinician Training and the Speech Exemplar and Evaluation Database. The practice of speech-language pathology relies heavily on perceptual judgment of fine-grained differences in the speech stream. Open-access corpora can support the process of fine-tuning the perceptual mechanism of future and current speech-language pathologists (SLPs) by making a wide variety of speech exemplars available for clinician training. Such corpora may be particularly valuable when accompanied by expert ratings. One such open-access labeled corpus specifically constructed for the perceptual training of SLPs is the Speech Exemplar and Evaluation Database (SEED; Speights Atkins et al., 2020). The SEED corpus offers high-quality audio representing the speech of individuals across the lifespan with speech sound disorders, as well as matched participants without communication disorders. As of 2020, the SEED corpus contained ~16,000 words and sentences from 58 children and 34 adults engaged in 16 different standardized speech

tasks relevant to the diagnosis of speech disorders, such as the Goldman-Fristoe Test of Articulation (Goldman & Fristoe, 2000, 2015), the Rainbow Passage, and the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster et al., 2009). The participants represented in SEED are speakers of a dialect of American English, and were recruited from the communities around Cincinnati, Ohio, USA and Auburn, Alabama, USA. The corpus is available on OSF (https://osf.io/ygc8n/).

Speech Technology and Ultrasuite. Speech corpora can be used to train speech technology systems, such as commercially-available voice-activated assistants (e.g., Apple Siri, Amazon Alexa). For some time, researchers have explored the idea that accurate speech technology embedded in computerized treatment could be used to augment SLP services; such technology may help clients with speech disorders to receive SLP-overseen treatment with sufficient frequency despite barriers such as overburdened caseloads (Campbell et al., 2017; McKechnie et al., 2020; McLeod et al., 2020). Systematic review completed by McKechnie et al. (2018) – describing a decade of technological developments and clinical validity in automatic mispronunciation[4] analysis tools (i.e., review of populations tested, stimuli elicited, system accuracy, and clinical change) – found that existing speech analysis systems focusing on mispronunciation detection in children, the topic of central importance to our research, fall short of clinically acceptable levels of accuracy when evaluating sounds produced *incorrectly*. This contrasts with the robust performance of commercially-available speech technologies on the

[4] Note that mispronunciation detection speech technology can encompass both clinical applications and second language learning applications, both of which were reviewed by McKechnie et al. (2018).

speech of typical adult speakers of mainstream dialects of American English. The same general principles of speech technology development are relevant across contexts (adult versus child speech, typical versus clinical speech); the fundamental difference between these cases is the lack of large-scale, labeled speech corpora on which to train speech technology for child and/or clinical populations (Fainberg et al., 2016; Kennedy et al., 2017; Liao et al., 2015; Shahin et al., 2020; Yeung & Alwan, 2018). Because accurate speech analysis systems must be trained on a critical mass of data that is highly similar to the speech meant to be analyzed, child clinical speech analysis systems can only be expected to meet a clinically-useful threshold if trained on enough speech data that is representative of its future use. However, sufficiently large datasets for child speech – especially for children with speech sound disorders – are rare.

Ultrasuite (Eshky et al., 2018) is one such open-access dataset specifically developed for the purpose of training machine learning systems for automated analysis of ultrasound tongue shape during assessment and treatment of speech sound disorders. Ultrasuite is centered upon acoustic and ultrasound image data, rather than phonetic transcriptions, for two datasets from children with speech sound disorders and one from age-matched peers. Across the three datasets, Ultrasuite contains an estimated 18.67 hours of speech representing approximately 14,500 phones, words, sentences, and nonspeech recordings from 86 children aged 5;8 to 13;4. Ultrasuite represents a valuable resource for individuals seeking articulatory or ultrasound image datasets in the context of RSSD.

Current Needs for Speech Corpora. SEED and Ultrasuite are two high-quality corpora addressing different facets of the need for open-access clinical datasets of child speech. However, they are still fairly modest in size relative to corpora typically used to train models for recognition of typical adult speech. SEED, specifically, represents a broad range of disordered

speech patterns, but with a relatively limited number of participants representing each pattern. The PERCEPT-R corpus presented here aims to fill a complementary need by providing deep coverage for a single sound that is a common target of speech intervention in rhotic dialects of American English (i.e., /ɹ/). Specifically, in connection with our team's interest in the development of efficacious interventions for children with RSSDs, we sought to build a corpus sufficient to train an automated mispronunciation detection system that would be able to classify children's productions of /ɹ/ as perceptually typical or perceptually atypical with a level of accuracy suitable for use in clinical practice. The related PERCEPT-GFTA corpus addresses the need for breadth in speech corpora by sampling a range of phonemes besides /ɹ/. Furthermore, these two corpora can be combined to train forced alignment systems such as the Montreal Forced Aligner wrapper to the Kaldi Speech Recognition Toolkit (McAuliffe et al., 2017; Povey et al., 2011), which is a commonly-used research tool that automatically predicts the boundary locations between phonemes in words. Because these tools are only pretrained in American English for adult speakers, researchers wishing to use models specific to child/clinical speech would need a critical mass of speech and the knowledge to update the training of the tools.

The PERCEPT Corpora have been curated specifically to increase utility for machine learning/AI applications: audio pre-processing has been standardized, high quality machine-readable ground-truth metadata and class labels are available for training, and a critical mass of content (thought colloquially to be ~ 40 hours) is available. Furthermore, a derived version of the PERCEPT corpora will make the corpus data available in file formats easily readable by common AI development languages (e.g., Python, Torch). In the sections that follow, we present the properties of the PERCEPT corpora, describe how they can be accessed through the Phon

database system using the open-source PhonBank repository, and discuss how these data can be used in clinical, research, and pedagogical applications.

Description of the PERCEPT Corpora

Data for PERCEPT-R and PERCEPT-GFTA were collected during 34 separate cross-sectional and longitudinal studies at Syracuse University, Montclair State University, and New York University between 2006 and 2021. Appendices B and C contain summaries of the studies included in these corpora as well as associated citations. Appendix D contains a summary of the corpus using the "Datasheet for Datasets" framework, which emphasizes transparency, accountability, and ethical AI (Gebru et al., 2018). All data collection and data management procedures were approved by the Institutional Review Boards associated with the relevant university, with recent multisite studies receiving approval through the Biomedical Research Alliance of New York (BRANY). The curation and release of this corpus was classified by BRANY as not human subject research activity (i.e., exempt secondary data analysis; protocol number 21-038-524). PERCEPT-R v2.2.2p and PERCEPT-GFTA v2.2.2p include only participants whose documentation of parental consent and child assent (or, for adult participants, only their own consent) permits audio data sharing outside of the original study of enrollment; this is specifically indicated by the "p" designator in the version number of this release and future releases of PERCEPT corpora. The PERCEPT project is distributed for non-commercial use through PhonBank and additional information and/or support for end users can be found at the PERCEPT page at the Open Science Framework (https://osf.io/nqzd9/) as well as in the PERCEPT channel of the Slack workspace for corpus phonetics: Phon Corps (tinyurl.com/2tnm2vkw).

PERCEPT-R

PERCEPT-R v2.2.2p contains 32.0 hours of citation-speech recordings reflecting 107,281 word-level utterances that contain the phoneme /ɹ/. Each utterance is encoded as left-channel, 44.1 kHz audio and labeled with perceptual ratings of the /ɹ/ and the orthographic transcript of the utterance, in addition to other metadata that we describe in detail in a subsequent section. The PERCEPT-R corpus focuses on the speech of children from rhotic American English dialects who exhibit RSSDs that primarily impact /ɹ/, as well as age-matched peers with typical speech. In these dialects, a "fully rhotic" /ɹ/ is considered perceptually typical and a "derhotic" /ɹ/ is considered atypical. For research participants seeking intervention for RSSD, who constitute the majority of speakers represented in the PERCEPT-R corpus, the goal of intervention is the acquisition and generalization of a motor plan for a fully rhotic /ɹ/ that is perceptually typical relative to the customary /ɹ/ of the individual's local community.

PERCEPT-GFTA

PERCEPT-GFTA v 2.2.2p is a related corpus containing 4.3 hours of citation-speech recordings representing 20,041 word-level utterances. While PERCEPT-R is focused on words containing rhotics, the PERCEPT-GFTA corpus is comprised of recordings elicited in the administration of the Sounds-in-Words subtest of the Goldman-Fristoe Test of Articulation (GFTA, editions 2 and 3; Goldman & Fristoe, 2000; Goldman & Fristoe, 2015) and thus encompasses a diverse set of target phonemes. The two corpora share a large number of overlapping participants, because most of the studies represented in PERCEPT-R administered the GFTA-2 or GFTA-3 as part of the initial assessment of eligibility. However, PERCEPT-GFTA also contains records from studies of children without RSSD affecting /ɹ/, including studies of children with suspected Childhood Apraxia of Speech (CAS; Preston et al., 2013;

Preston et al., 2016; Swartz & Hitchcock, 2021). Unlike the PERCEPT-R corpus, records in the PERCEPT-GFTA corpus do not contain accuracy labels, only an orthographic transcript of each utterance. Note that although participants overlap between PERCEPT-R and PERCEPT-GFTA, utterances do not overlap.

Participants

The 105,591 utterances in PERCEPT-R v2.2.2p were collected from 280 child, adolescent, and young adult speakers of American English from the Northeastern United States, aged 6;0 – 24;0 ($\bar{x}$ =11;4, σ_x =2;6), engaging in word-level citation speech at several time points during assessment and treatment. Of the 280 participants, 128 are females (ages for females: $\bar{x}$ =11;8, σ_x =2;5, min = 6;1, max = 17;3; ages for 152 males: $\bar{x}$ =11;1, σ_x =2;7, min = 6;0, max = 24;0). No participants were known to be transgendered. Thirty-three corpus speakers (12%) self-identified as Black/African American, Asian, More than one race, or Other according to the NIH race reporting framework (see also: Appendix B). The imbalance between males and females reflects the increased prevalence of RSSD observed among males (Wren et al., 2016). Of the 280 participants, 95 were recruited to studies of typically-developing speakers[5], 22 were recruited to studies based on a history of preschool SSD, and the remaining 163 participants were recruited to studies of individuals with RSSD. Figure 1 shows the distribution of ages (in years) within the PERCEPT-R corpus, grouped by sex and speaker group. The mean number of utterances contributed by the 280 speakers in PERCEPT 2.2.2p is 417.5 (σ_x =555.7, min = 29,

[5]N.B.: *being recruited to a study* meant that informal screening indicated the participant may meet the study eligibility criteria; however, the corpora metadata does not indicate whether the participant met all study-level inclusionary criteria for RSSD or typical speech.

max = 3025). Some participants (i.e., those in studies of typical development) were recorded at only one time point; others were tracked longitudinally over the course of treatment and were recorded at as many as 55 unique time points (see details below). Figure 2 shows the number of wav files by age, sex, and speaker group in PERCEPT-R.

Figure 3-1. Distribution of participants in the PERCEPT-R corpus.

Data grouped by sex and speaker group. F = Female, M = Male.

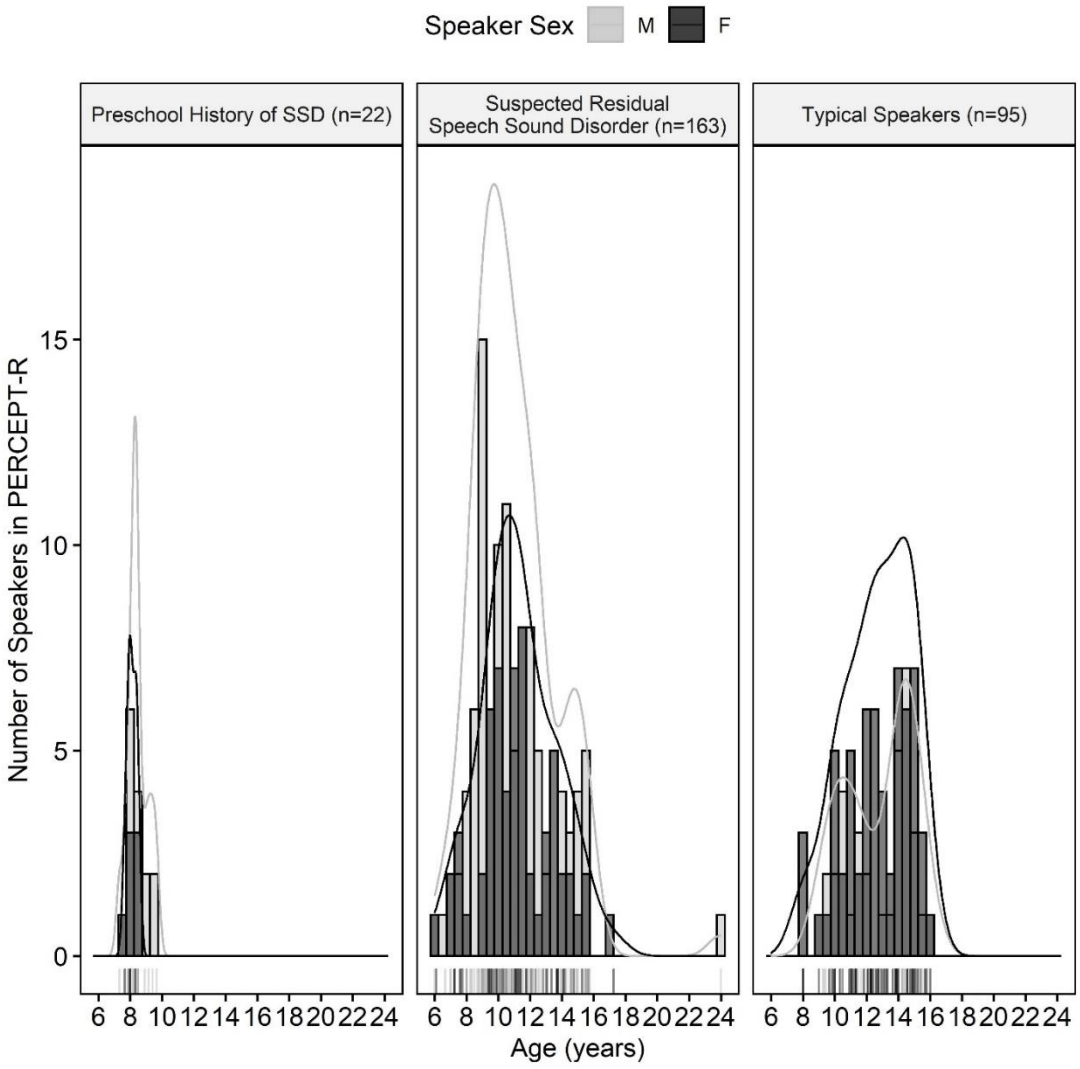
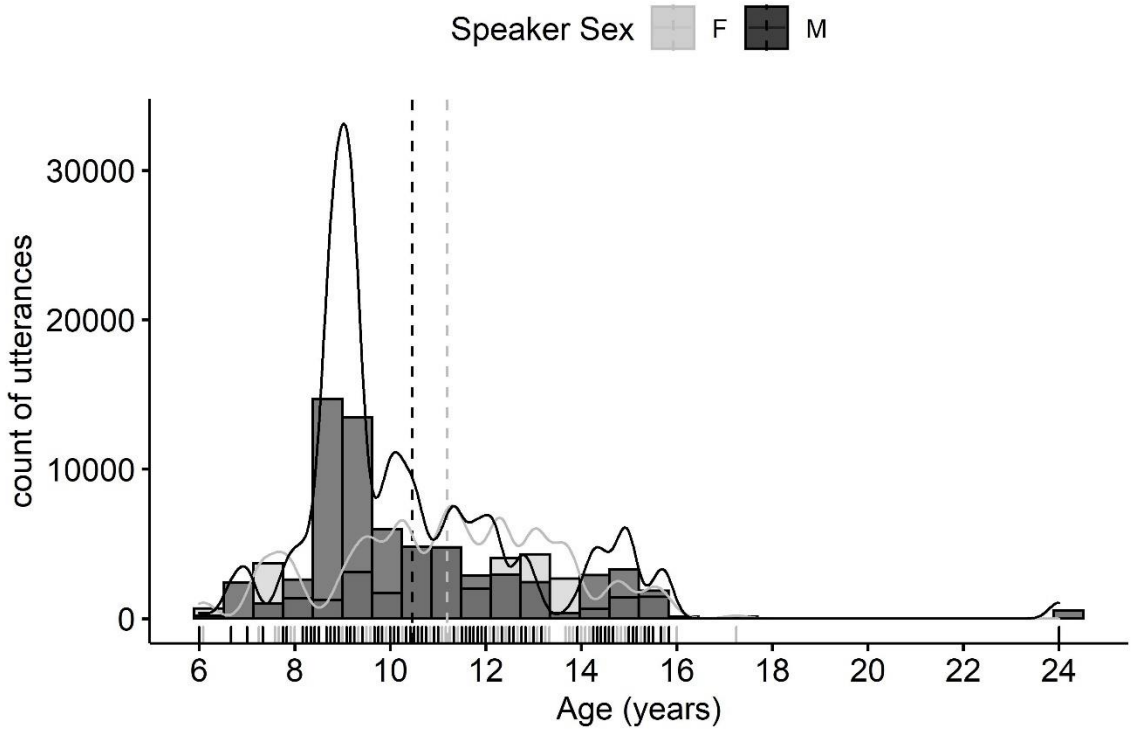


Figure 3-2. Distribution of audio files in the PERCEPT-R corpus

Data grouped by age, sex, and speaker group. F = Female, M = Male.

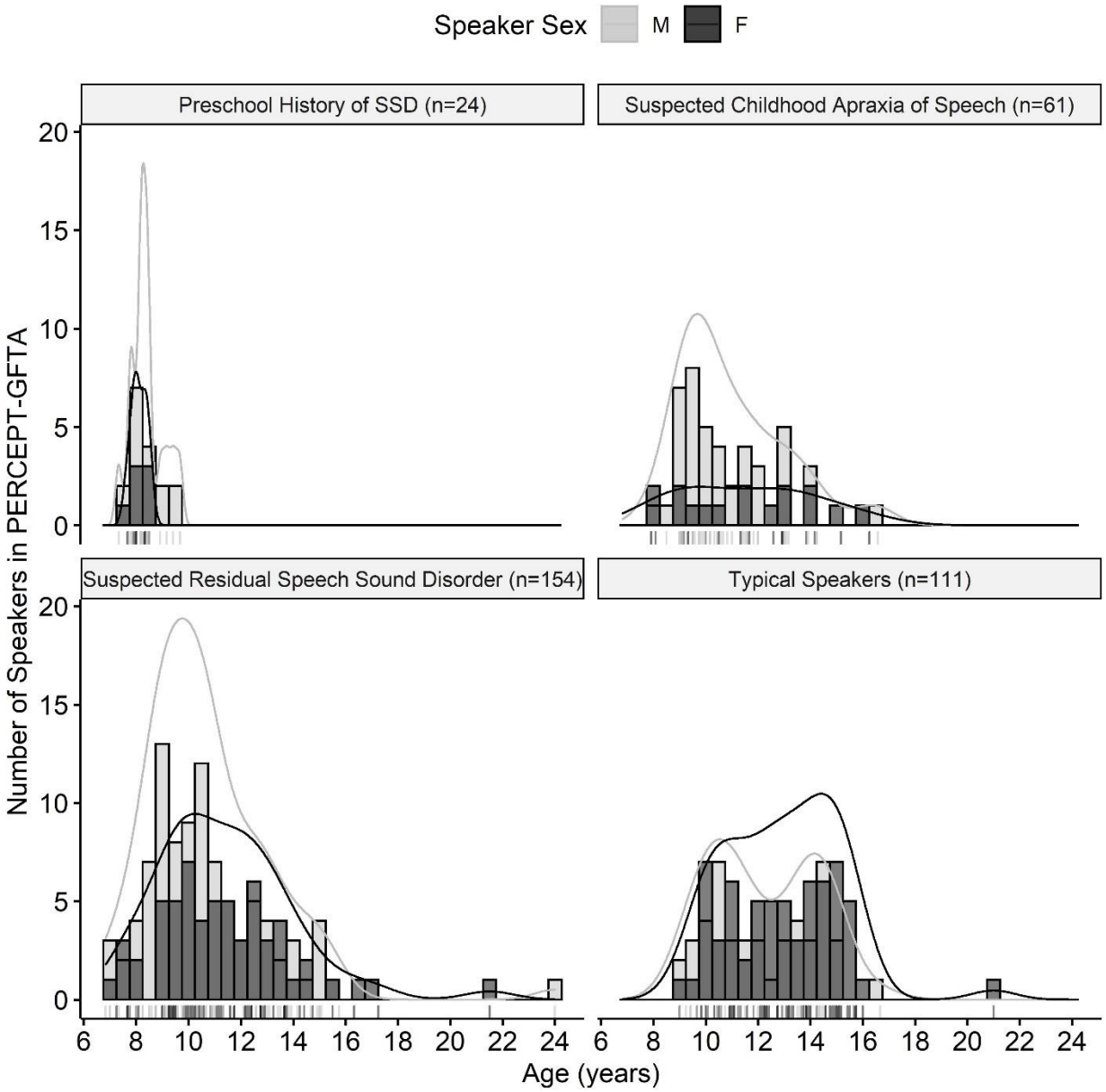


The 350 participants included in PERCEPT-GFTA v 2.2.p are between the ages of 6;10 – 24;0 ($\bar{x}$ =11;4, σ_x =2;6). Note that many participants in this corpus overlap with the speakers of PERCEPT-R, and the speakers that do overlap have the same identification number across both studies. Of the 350 PERCEPT-GFTA participants, 147 are females (distribution for females: $\bar{x}$ =11;11, σ_x =2;7, min = 7;3, max = 21;6; distribution for 203 males: $\bar{x}$ =10;11, σ_x =2;4, min = 6;10, max = 24;0). No participants were known to be transgendered. Fifty-two corpus speakers (15%) self-identified as Black/African American, Asian, More than one race, or Other according to the NIH race reporting framework (see also: Appendix C). Twenty-four of the participants in PERCEPT-GFTA were recruited based on a history of SSD when aged 4-5 years, 61 were recruited for CAS, 154 were recruited for RSSD, and 111 were recruited for studies of typical speech

development. All but two participants are represented in the PERCEPT-GFTA corpus with only a single timepoint of data collection. 106 speakers completed the GFTA-2 and 244 speakers completed the GFTA-3. Figure 3 shows the distribution of ages within the PERCEPT-GFTA corpus, grouped by participant sex and speaker group.

Figure 3-3. Distribution of participants in the PERCEPT-GFTA corpus.

Data grouped by sex and speaker group. F = Female, M = Male.



Data Acquisition

Original Data Collection Purposes. Data were collected during standardized assessment and intervention tasks for lab-based studies of speech sound production. The original purpose of treatment studies, broadly, was to evaluate changes in perceptual or speech sound accuracy in response to the controlled delivery of speech sound intervention built upon the principles of motor learning (Maas et al., 2008). The original purpose of studies recruiting typical speakers was to provide speech data to serve as a point of comparison for speakers with RSSD. Data were collected directly by the study team associated with the originating data collection site according to standardized procedures that, in the case of multi-site studies, were shared across sites.

Data Collection Timepoints and Speech Tasks. Speech samples from children with RSSD in PERCEPT-R 2.2.2p were collected longitudinally before, during, and after speech treatment. Speech samples from children with typical speech were collected at a single study timepoint. Most speech samples were collected through probe tasks that elicited isolated phrases, words, and syllables as citation speech, the speech register used when a speaker is asked to produce an isolated utterance in its canonical form. Productions for probe tasks were obtained via direct imitation of the research SLP, reading, and/or picture naming. Feedback on production accuracy was generally not provided in the probe context. A subset of audio samples represented in the corpus was collected within the course of treatment delivery (i.e., in a context where clinical cueing and feedback were provided). Elicitation methods in this context could include direct imitation or reading, as indicated by the treatment protocol for a given study and level of treatment complexity.

Audio samples for PERCEPT-GFTA were collected during evaluation sessions. For children enrolled in a treatment study, the GFTA was administered prior to the initiation of

treatment. Following the standardized procedures for the instrument, productions were elicited through picture naming and response to clinical prompts about the pictures. Both the second and third editions of the GFTA are represented in the corpus, depending on the originating study.

Recording Environment. For both PERCEPT-R and PERCEPT-GFTA, most of the originating studies were conducted in the lab setting, with audio recorded from a participant-worn headset (e.g., AKG C520) or lavalier mic (e.g., Sennheiser MKE 2). Generally, audio was digitized external to the recording device using an audio interface (e.g., Steinberg UR22c) and simultaneously recorded to the clinician's computer and an external solid-state recorder (e.g., Marantz PMD620-MKII). The files curated for the corpus represent the "primary" audio analyzed during the original study (e.g., the computer-recorded audio, except in the cases of poor audio at the computer; then, the external digital recorder would have been used). Despite the site differences in recording, all recording environments and audio equipment were of a similar specification. (While the amount of background noise may differ from utterance to utterance, this variability can help support the training of generalizable speech analysis algorithms.) A small minority of data were collected via telepractice, either when in-person studies were conducted remotely during COVID-19 shutdowns, or as part of planned studies of telepractice treatment. In this case, participants were provided with a study-provided headset with built-in microphone (e.g., Plantronics Blackwire C225) and instructed in recording to a voice capture software (e.g., iOS Voice Memos) on their local device, then securely uploading audio to the study team. This process for making local recordings allowed the originating study teams to analyze participant audio without potential confounds introduced by network transfer of the audio (e.g., dropped samples, latency, compression; see Sanker et al., 2021). In other words, all corpus audio was

captured using a recording device that was in the same room as the participant, whether the participant was seen for an in-person study or for telepractice.

Post-Processing of Collected Audio. All corpus audio was recorded at the level of the task or, in the case of intervention trials, at the level of the treatment session. The session- or task-length audio was manually or semi-manually segmented into target utterances by trained research assistants using TextGrids in Praat (Boersma & Weenink, 2022) to insert segment boundaries and orthographic labels. In the case of more recent studies (e.g., McAllister et al., 2020), TextGrid segmentation boundaries were initially placed automatically using intensity detection in Praat or a custom audio-event detection algorithm designed by this paper's first author that parsed audio markers generated by our treatment software, Challenge Point Program (McAllister et al., 2021). The identified target utterances were then extracted from the task-level or session-level audio using Praat (Boersma & Weenink, 2019) or the Parselmouth API (Jadoul et al., 2018) for Praat, depending on the originating study.

All instances of segmentations and orthographic transcripts were manually verified by trained research assistants during the original study and a subset were again verified using a hybrid Google Speech-manual review method during PERCEPT-R corpora curation. (PERCEPT-GFTA did not receive additional review because those utterances were segmented and labeled by trained research assistants specifically for the purpose of PERCEPT corpus creation, whereas PERCEPT-R files were segmented and labeled for the purpose of the originating studies between which segmentation and utterance conventions might have differed slightly). During this hybrid transcript verification process each utterance in PERCEPT-R was transcribed by Google Speech speech-to-text service to obtain an orthographic transcript of the audio within the file that was then compared to the orthographic transcript that was assigned to

the file at the time of original data collection. The following conditions resulted in automated verification: (1) when the Google Speech orthographic transcripts matched the study-generated label verbatim, and when the Phon-generated phonetic transcription of the Google Speech transcripts matched the Phon-generated phonetic transcription of the study-generated label (2) verbatim and (3) when /ɹ/ phones were mistranscribed by Google Speech as /w/ or /l/ (i.e., phones that may have similar features to derhotic /ɹ/). A total of 73,781 out of 179,076 utterances in the total (private and public) PERCEPT-R corpus met this standard. The transcripts that did not meet the threshold for automated verification were then triaged to select an additional 31,089 files for manual review. Files were triaged for manual review using the Levenshtein distance to identify files having the largest string difference between the Google Speech transcript and the transcript assigned at the time of data collection. Manual review was completed by trained research assistants who listened to each file and the transcript assigned to the file the time of data collection. Listeners indicated whether the audio in the file was unsuitable for corpus processing, whether there were extraneous words, cross-talk, or non-speech noises (e.g., laughing, coughing) adjacent to an otherwise-usable piece of the target audio, or whether the audio matched the study-generated orthographic transcript verbatim. This process was facilitated by custom Praat and Python scripts written by the first author. In total 581/31,089 manually reviewed files lost candidacy for the public and private versions of the PERCEPT-R corpus due to this review. The most common reason for excluding audio after manual review was due to background noise or clinician cross talk that would have been permissible in audio analysis procedures for the original studies but not for the more-automated steps required to preprocess utterances for PERCEPT. Because the 1.87% of files that lost candidacy after manual review were sampled from the

corpus stratum representing lowest confidence in transcript accuracy, we expect the overall rate of low-quality files in the PERCEPT-R to be less than this value.

All verified utterances have been standardized to left channel, 44.1 kHz audio scaled to an average utterance intensity of 70 dB using Parselmouth, and concatenated into silence-buffered audio using a custom Groovy script (Hedlund, 2022). These concatenated utterances represent an alphabetical indexing of every utterance produced by that participant in a given study and session, with each utterance buffered by intervening silence. Such concatenation allows all data from the same speaker-timepoint to be simultaneously viewable as different records within the same Phon window, while alphabetical ordering facilitates user navigation of utterances within the sessions.

Corpus Labels and Metadata

The audio records in the PERCEPT corpora have been linked with several pieces of file metadata. Each record in PERCEPT is linked with a participant, originating study, and timepoint of data collection, as well as orthographic, IPA, and ARPABET transcriptions of the *target* utterance. (ARPABET is an ASCII-readable phonetic alphabet that is commonly encountered in speech signal processing, as ARPABET transcriptions contain no special Unicode characters that might hinder processing versus, e.g., IPA.) Note that individual utterances, most often word-level productions, are orthographically transcribed but not phonetically transcribed; however, users of the corpus can add and save their own phonetic transcriptions to their local copy using the Phon software (the process of which we describe in a following section). Records in PERCEPT-R also include information about perceptually rated accuracy of the target rhotic sound, including the number of unique listener ratings obtained for the token, the number of ratings indicating that it

was a typical (i.e., fully rhotic) production, and the calculated average rating for the /ɹ/ within the utterance.

Transcriptions and Phonological Coverage

PERCEPT-R v2.2.2p provides coverage for 499 unique target utterances, including real words and phototactically licit target nonwords (e.g., / ɑ͡ɚd/, /kɝ/). The five most frequently-appearing target words in the PERCEPT-R corpus are *beard* (n=2096), *turn* (n=2192), *nurse* (n=1969), *ladder* (n=1957), and *chair* (n=1954). The five most frequent target nonwords in the PERCEPT-R corpus are /ɹɑ/, (n=1725), /ɝ/ (n=1689), /ɹi/ (n=1624), /dɝ/ (n=165), and /ɝp/ (n=144). The v 2.2.2p release of PERCEPT-R contains only utterances with a single rhotic per utterance, which avoids complexities that arise when a single word-level record needs to be linked to multiple records of perceptually rated accuracy. Rhotic sounds are represented according to the phonological distribution in Table 1, which also demonstrates the Phonex queries used to extract this information from the corpus. The target utterances of PERCEPT-GFTA are congruent with the word lists for the GFTA-2 (n = 53 utterances) and GFTA-3 (n = 60 utterances). Stimulus characteristics of these standardized tests have been extensively detailed by Macrae (2017), and we direct interested readers to that work.

Table 3-1. Phonological distribution of PERCEPT-R target utterances, across all syllable stress types.

Description	Phonex Query Syntax	Plain English Query	Count in Corpus
Word-initial singleton rhotic onsets	(?<\b\s?)(r:O|ɹ:O)	Return all records with /r/ or /ɹ/ occurring in onsets, that are preceded by a word boundary and an optional stress marker.	30,833
Word-initial complex rhotic onsets	(?<\b\s?)\c+(r:O|ɹ:O)	Return all records with /r/ or /ɹ/ occurring in onsets, that are preceded by a word boundary, an optional stress marker, and one or more consonant sounds.	13,662
Syllabic rhotics	(ɝ:N|ɚ:N)	Return all records with nucleus entirely comprised of /ɝ/ or /ɚ/	24,159
Word-final post-vocalic rhotics	\v͡ɚ(?>\b)	Return all records with a rhotic schwa offglide after a monophthong, followed by a word boundary.	20,829
Post-vocalic rhotic followed by word-final consonant	\v͡ɚ\c+ (?>\b)	Return all records with a rhotic schwa offglide after a monophthong or diphthong, followed by one or more consonants and then a word boundary.	15,693

Note. Queries can be entered into a fillable form in the Phon software to grep relevant records. An expanded example is available as Appendix A. The grapheme "r" (versus "ɹ") is not used in the corpus but is included in these examples to demonstrate the capabilities of Phonex. Note that rhotics following diphthongs in words such as /a͡ʊɚ/ will be returned in the syllabic rhotics query rather than the post-vocalic query because Phon sets a syllable boundary between the /a͡ʊ/ and /ɚ/.

Perceptual Labels in PERCEPT-R. Perceptual ratings for tokens were derived by one of two general methods, depending on the study of origin. Perceptual ratings for select studies,

representing an estimated 75,746 tokens, were rated by untrained listeners recruited online (n < through the Amazon Mechanical Turk crowdsourcing platform, as described by McAllister Byun et al. (2015). The perceptual accuracy of an estimated 29,536 tokens from the remaining studies were evaluated by a panel of trained listeners, typically licensed SLPs or students in SLP programs who have completed coursework in speech sound disorders, as described e.g., in Preston and Leece (2017) and Benway et al. (2021). The mode of crowdsourced listeners per file was 9 and the mode of expert listeners per file was 3, following from McAllister Byun et al. (2015). All listeners were adults who passed catch trials meant to highlight listeners whose responses disagreed with expert consensus for a handful of training tokens. The crowdsourced and expert rating methods followed the same general principles: tokens were randomized to listening modules that contained utterances from different speakers and different study timepoints (i.e., pre-treatment, within-treatment, and post-treatment). Raters were instructed to rate the accuracy of the /ɹ/ sound in each word, using a strict standard in which only fully rhotic, adult-like productions are scored as correct. Raters were provided with the orthographic transcript of the target but were masked with respect to the participant identity and the timepoint of elicitation of each token. Crowdsourced ratings also contained catch trials designed to flag and discard raters whose ratings were suggestive of inattention. The perceptual accuracy for individual rhotics in the corpus is the average of these expert or crowdsourced ratings: the sum of listener responses (0 = derhotic and 1 = fully rhotic), divided by the number of raters (e.g., (0 + 1 + 0)/3 = .3̄).

A summary of the derived ratings appears in Table 2. 31,156 tokens received unanimous ratings in which all raters agreed a production was atypical (i.e., derhotic). An additional 34,343 tokens were rated atypical by most, but not all, raters. A total of 23,232 tokens were rated as

typical (i.e., fully rhotic) by more than half of raters, but fell short of a unanimous rating. Finally, 16,549 tokens received unanimous ratings indicating typical production. There is a imbalance favoring derhotic tokens in the PERCEPT-R corpus, which reflects the intention that the data come overwhelmingly from recordings of children with RSSD.

Table 3-2. Distribution of rhotic perceptual ratings within PERCEPT-R.

Perceptual category	**Average perceptual rating of rhotic**	**Records in category**	**Participants in category**
Unanimously derhotic	0	31,156	168
Consensus derhotic	$0 < x \leq .5$	34,343	211
Consensus fully rhotic	$.5 < x < 1$	23,232	267
Unanimously fully rhotic	1	16,549	231

Accessing the PERCEPT Corpora: PhonBank and Phon

The PERCEPT corpora are publicly available through partnership with PhonBank (Rose & MacWhinney, 2014), an NIH-funded data-sharing platform for speech-language research. Within the TalkBank family, PhonBank is unique in its organization around the Phon software program; all of the other TalkBank databases center around the CLAN program (Computerized Language Analysis; dali.talkbank.org/clan). While Phon and CLAN offer similar basic functionality (such as time-aligned records, a standardized annotation system, and query functions), Phon differs from CLAN in that Phon offers specialized functions for the study of phonetics and phonology (Rose & MacWhinney, 2014). Below we describe the features of Phon and demonstrate their application with a sample participant from PERCEPT-R.

Structure of Database

Within Phon, data transcripts, which contain both the linguistic data and the associated metadata (e.g., information about participants) are organized around a two-level nesting structure. The top level is the ‘project’ folder, which can contain one or more data corpora. Each ‘corpus’ folder contains one or more data transcript(s), which are organized as a list of data

records that contain five default tiers in addition to as many user-defined tiers as are needed for the research at hand. The naming conventions associated with each nested element of the Phon database system in the PERCEPT corpora are summarized in Table 3 The default tiers in each Phon record are as follows:

- Orthography: Primary textual data, typically representing the content of the current utterance
- IPA Target: Representation in the International Phonetic Alphabet (IPA) of a model pronunciation of the current utterance
- IPA Actual: Representation in IPA of the speaker's pronunciation of the current utterance[6]
- Segment: Time stamp corresponding to the time interval that contains the current utterance on the recorded media file (if any)
- Notes: Generic tier to record notes as needed

[6] Because the PERCEPT corpora are not phonetically transcribed, the default IPA Actual tier does not appear in corpus records. However, in corpora where the tier does appear, Phon algorithmically encodes phone-by-phone alignments between Target and Actual transcriptions.

Table 3-3. Mapping of Phon elements to PERCEPT-R and PERCEPT-GFTA naming conventions.

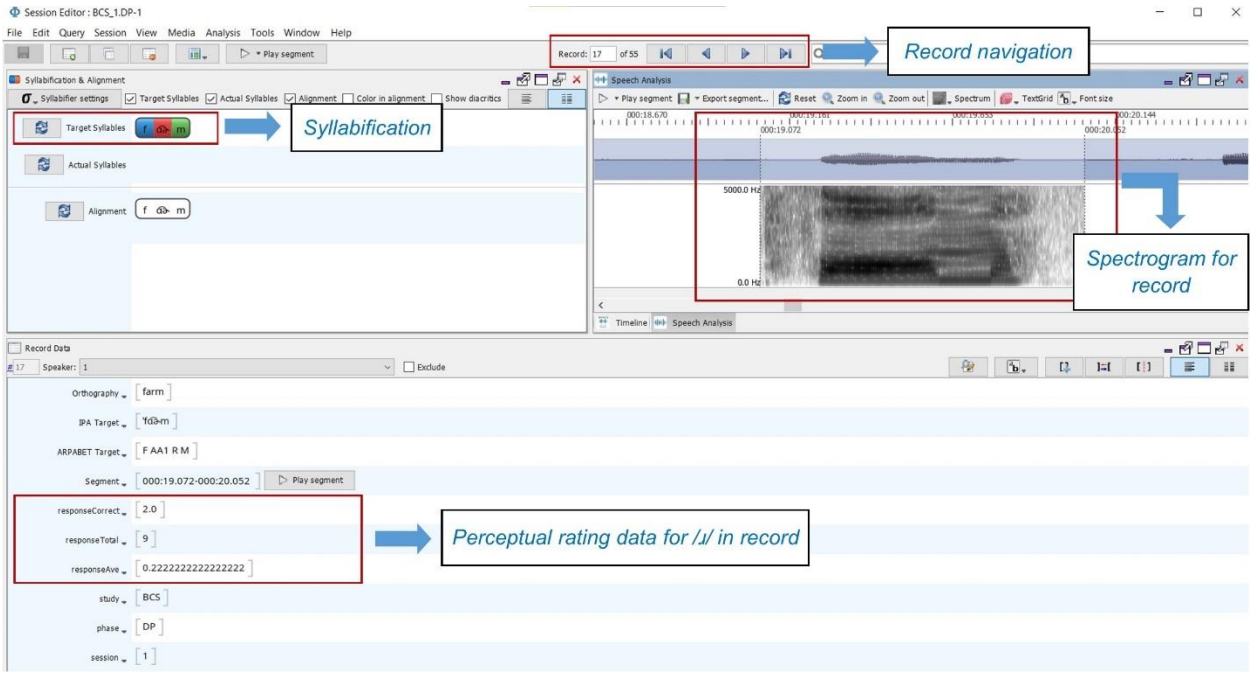
Phon Element	**PERCEPT-R Naming Convention**	**PERCEPT-R Example**	**PERCEPT-GFTA Naming Convention**	**PERCEPT-GFTA Example**
Corpora	[Study Name]_[ID]	CRESULTSSCED_33	[Study Population]	SuspectedCAS
Session	[Study Phase]_[Study Session Number]	PRE_1	[Study Name]_[ID]	PrestonCASR15_179
Record	Utterance audio and metadata	indexed within the session file from 1-n	Utterance audio and metadata	indexed within the session file from 1-n

Implicit in this structure is the fact that Phon relies on different levels of alignment within each record to allow transcribed units (e.g., orthographic words, IPA Target phonemes) to be linked with phonological data based on positions within the utterance, word, and syllable. At the utterance level, words are implicitly marked as utterance-initial, utterance-medial, or utterance-final, and language-specific algorithms mark each phone present in the phonetic transcription with its syllable position. In both cases, the encodings derived through algorithms remain fully modifiable by the user.

The PERCEPT corpora are formatted specifically for Phon software, allowing users to query the corpus audio based on phonological and phonemic information. Each record displays the audio data as well as the PERCEPT metadata for that token (i.e., participant, orthographic transcription of the utterance, IPA and ARPABET transcriptions of the target utterance, the originating study, the timepoint of data collection, and, for PERCEPT-R, the rhotic rating data). Each of these datapoints is viewable using the Phon Session Editor graphical user interface. It is important to note that records in the PERCEPT corpora do not contain information in the IPA

Actual tier, since these productions have not been phonetically transcribed. To access records in the PERCEPT corpora, a user should save the desired dataset(s) to the computer's Phon Workspace, where Phon will list each dataset as a Phon project. Opening a project in Phon will display the Phon corpora associated with that project, and selecting a corpus will display the Phon sessions associated with that corpus. Selecting a session will allow the user to navigate through all audio records and linked metadata participant, study, and session in question. An example of a session record, linked audio file, and linked metadata from the PERCEPT-R corpus appear in Figure 4 below.

Figure 3-4. Annotation of the Phon Session Editor window, showing one audio record and associated metadata.



Although formatted for Phon, all PERCEPT data are also accessible through the file system of the user's operating system (e.g., File Explorer in Windows 10). For PERCEPT-R, data for each study-participant are contained in a labeled directory. These study-participant directories contain session-level XML records encoding the Phon data elements described above,

as well as a *media* subdirectory containing the concatenated session WAV audio as well as Praat TextGrids with utterance-level segmentations and orthographic transcripts in an interval tier. This file structure allows for the PERCEPT corpora to be parsed with tools in addition to Phon (e.g., Python). We also plan to release a derived version of the corpora to increase ease of use with artificial intelligence tools such as Pytorch.

Querying the PERCEPT Corpora in Phon

In line with the tenets of Open Science, the Phon project aims not only to support data sharing through PhonBank, but also to facilitate reproducible analysis of phonological data. Historically, phonological research has relied heavily on manual encoding and querying of data, which is slow, effortful, and difficult to reproduce; see Rose & MacWhinney (2014) for a summary of previous solutions in this area. Phon was designed to provide a user-friendly interface that allows researchers to systematically query phonological data and reproduce the same analyses with the same or different data in future research. Instead of scripting, Phon guides the user to program each analysis using a fillable form. Different forms are designed to meet different needs, which range from generic query forms targeting user-determined data tiers to special forms executing standard clinical analyses such as calculation of percent consonants correct in the case of phonetically transcribed corpora. Each phone present can also be referenced through a set of descriptive phonological features, which enables the user to query a corpus with reference to natural classes of speech sounds (e.g., all labial consonants; all voiceless obstruents). Beyond textual, phonological, and clinical analyses, Phon also supports acoustic data analysis through the integration of Praat libraries, including full support for TextGrid data annotations.

Once filled, the forms can also be saved as separate files for later use, either within the current project or as external files that can be shared with other scholars. Custom analysis reports can also be saved and shared in a similar way. Finally, while all analyses in Phon are based on scripts hidden from the user through fillable forms, users with advanced programming skills can also directly access the contents of the scripts for further customization. This also holds true of acoustic analyses, as Praat custom scripts can be easily adapted for use into Phon.

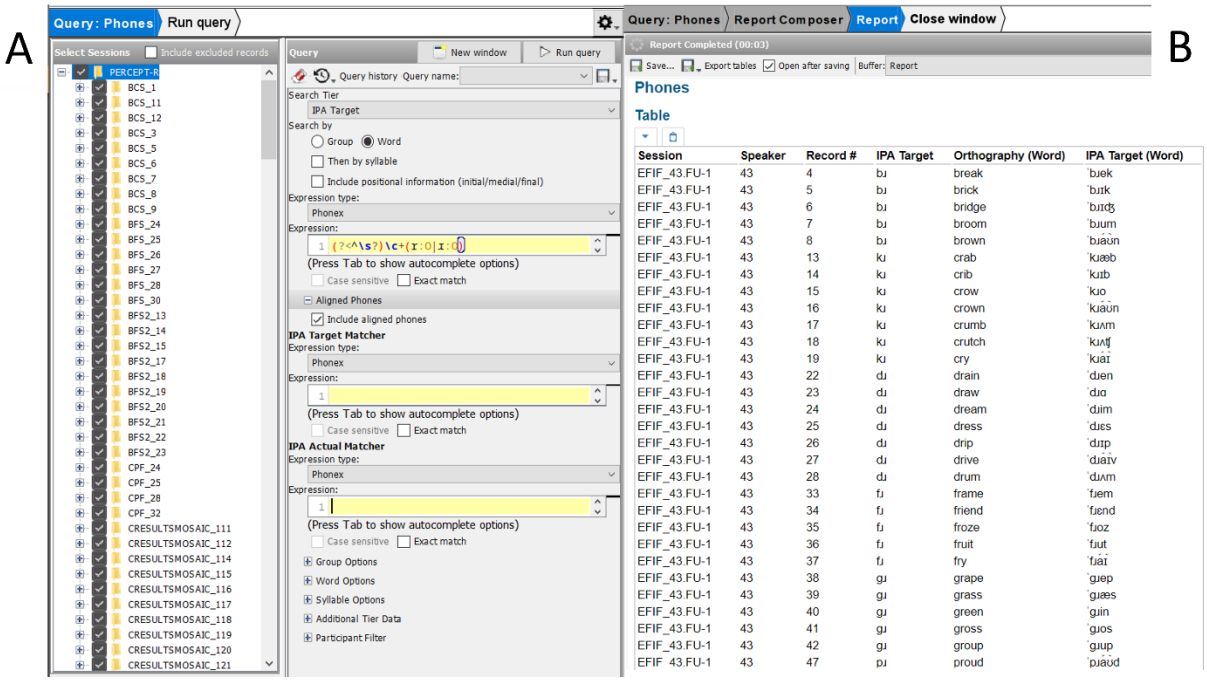
A sample query using fillable forms and PERCEPT-R is shown in the two panels of Figure 5. In this example, a user has selected the fillable form Query: Phones…. Records in the database can be queried using plaintext (e.g., return all records with the grapheme "r" on a given tier), regular expression (e.g., return all records with the grapheme "r" followed by a space on a given tier), or Phonex, a regular expression language that can also represent phonological information in queries (e.g., return all records with syllable-onset rhotics[7] on a given tier). For example, the Phonex expression (?<^\s?)\c+(r:O|ɹ:O) returns all audio records containing /ɹ/ in a complex onset, as shown previously in row 2 of Table 1. Panel A shows record selection, query parameters, and the Run Query button. A sample formatted query output report appears in Panel B. The other query strings referenced in Table 1 would be conducted in a similar fashion to the example illustrated in Figure 5. Note, however, that PERCEPT 2.2.2p is not poised to use all Phon capabilities, as Phon can also perform relational analyses (i.e., percent correct, phonological processes) comparing the canonical IPA Target and realized IPA Actual tiers in

[7] A note to the user: The theory underling Phon and Phonex is language-neutral, and the Phonex term "rhotic" will return alveolar taps/flaps in such queries, given that the symbol is present in the queried tier.

corpora containing IPA Actual transcriptions in addition to the IPA Target transcriptions available in PERCEPT. For interested readers, a worked example of the methodology associated with a PERCEPT-relevant research question is included as Appendix A and available on the PERCEPT Open Science Framework page (https://osf.io/nqzd9/). This example explores the extent to which syllable position (i.e., pre-vocalic/post-vocalic) and neighboring vowel articulation (i.e., vowel height-frontness/backness) are associated with perception of rhoticity in the PERCEPT-R corpus. Educators who adapt Appendix A as a classroom or research training exercise may be interested in the derived PERCEPT-R Corpus Sample, also distributed on the Open Science Framework page, which would allow for a portion of the corpus to be used with lower storage needs and processing times. The PERCEPT-R Corpus Sample reflects the 6 participants with audio permissions from Benway et al., (2021).

Figure 3-5. Example Phon query and output for grepping records.

This search returns records containing word-initial complex rhotic onsets, as seen in row 2 of Table 1. Panel A shows the selection of participant-sessions to search while Panel B shows the output.



Discussion

We have presented two corpora in the PERCEPT project, which in the publicly available version 2.2.2p contain a combined 36.3 hours of audio encompassing 125,632 tokens from a currently unprecedented 453 speakers, mostly with RSSD. These data are formatted for Phon and distributed through the TalkBank platform. Updates to the corpus and/or new corpora through the PERCEPT project will be made available on these same platforms, with general discussion and support available through the PERCEPT channel in the Phon Corps workspace on Slack. It is our intention that the PERCEPT corpora will have utility for research questions concerning child citation speech. For instance, Phon queries of the PERCEPT-R corpus could enable a large-scale study elucidating how different target vowels may serve as more or less facilitative contexts for rhotic production across individuals. The PERCEPT corpora are also expected to have value for clinician training activities involving speech perception and the use of speech analysis tools. For example, PERCEPT-R can be filtered to show only productions unanimously-rated fully rhotic or derhotic, and examples representing different phonetic contexts could be selected for use in

perceptual practice. It is also possible to sample a range of accuracy levels within a single participant; as one example, PERCEPT-R BCS_1 produced /r/ with nearly 0% accuracy in session 1 and nearly 100% accuracy in session 55, while sessions numbered between 35 and 40 reflect a mix of correct and incorrect productions. Furthermore, examination of productions with intermediate ratings (neither unanimously correct nor unanimously incorrect) across multiple participants could allow comparison of different phonetic realizations of rhotic distortions. Finally, phonetic transcriptions (i.e., IPA Actual) can be added to sessions in the PERCEPT Corpora by end users, providing opportunities for students in phonetics or speech sound disorder classes to transcribe speech from speakers recruited to studies of RSSD, childhood apraxia or speech, or typical speech. An interesting use of Phon would be that its analyses can also be used to facilitate efficient grading of such transcription exercises, making a high number of transcription practice trials feasible in the course of a semester. For example, students can be directed to enter their transcription responses in Phon sessions that include a hidden tier containing the instructor's ground-truth transcription, from which the instructor can automatically calculate student/instructor percent correct statistics as well as identify subsets of phones often transcribed in error.

Limitations of the PERCEPT Corpora

While the corpora are strongly positioned to address research questions and clinical training needs relating to child clinical speech, there are limitations related to the use of these corpora. From a sociolinguistic perspective, the participants represented to the corpus are overwhelmingly white individuals from the Northeastern United States and the age distribution of these participants is skewed toward the younger limit of the age range. The sociolinguistic composition of the corpora reflects the demographics of the participants who presented for the

clinical trials during which these data were collected. Ongoing work in our labs specifically seeks to increase the representation of speakers who identify as Black, Indigenous, Hispanic, Asian/Pacific Islander, and/or multi-racial. Additionally, the corpus labels are limited to perceived rhoticity of /ɹ/, and transcriptions of the actual productions (i.e., IPA Actual Phon Tier) are not available in the corpora. The nature of Phon, however, allows end users desiring transcriptions to be able to save their own transcriptions with their version of the corpus downloaded to their computer.

Future Speech Technology Directions for PERCEPT

We also believe the PERCEPT corpora will play a role in the training, adapting, and fine-tuning of research-relevant and clinically-relevant speech technologies such as algorithms for forced alignment and mispronunciation detection in child speech. Because child speech changes during the course of development, we see the PERCEPT corpora being of most use for technological applications in which the end use case is most similar to the age range and speech task represented in the corpora. Such speech technologies represent future directions for our own research and are expanded upon below.

Forced alignment is a common technique in which an utterance with a known transcript is submitted to a Gaussian Mixture Model and Hidden Markov Model (GMM-HMM) aligner that estimates segmental boundaries within the utterance. That is, when provided an utterance, orthographic transcript, and a dictionary mapping of orthographic word-level entries to phonemes, the aligner will estimate the timestamps for word boundaries and phoneme boundaries). One widely-used tool for forced alignment is the Montreal Forced Aligner (McAuliffe et al., 2017), which is a wrapper to the Kaldi Speech Recognition Toolkit (Povey et al., 2011). The Montreal Forced Aligner, and similar forced alignment tools, predict phoneme

boundaries based on referential acoustic models that summarize the feature space for a given phoneme in the target language. The similarity of these models to the speech being aligned influences the accuracy of segmental boundary information (Knowles et al., 2018; Mahr et al., 2021). Several pretrained acoustic models are available for the Montreal Forced Aligner, but none specifically for child speech alignment. The PERCEPT corpus provides a robust dataset for creating pretrained child speech acoustic models that can be used by the broader child speech research community.

As reviewed above, mispronunciation detection algorithms, in the context of clinically efficacious interventions overseen by SLPs, may someday play a role in overcoming access barriers to RSSD interventions. Existing mispronunciation detection algorithms for child speech have not demonstrated sufficient accuracy for clinical use (McKechnie et al., 2018). Automated speech analysis systems are more accurate when trained on large datasets that are highly similar to the speech meant to be analyzed (Kennedy et al., 2017; Liao et al., 2015; Yeung & Alwan, 2018); however, sufficiently large datasets for child speech – especially for children with speech sound disorders – are rare (e.g., Shahin, 2020). The PERCEPT corpora are well-positioned to overcome two factors that have limited the success of clinical speech technology thus far: the scarcity of atypical exemplars for algorithm training and inadequate technical description of the tools. To this end, we have developed audio classification algorithms for the detection of rhotic/derhotic /ɹ/ (Benway, Preston, Hitchcock, Salekin, et al., 2022) and are piloting the use of these algorithms clinically.

Conclusion

The PERCEPT-R and PERCEPT-GFTA corpora contain > 36 hours of speech audio (125,632 utterances) from children, adolescents, and young adults aged 6-24 with speech sound disorder (primarily RSSDs impacting /ɹ/) and their age-matched peers. These corpora are distributed through the PhonBank repository and are formatted for the database software Phon. As open-access data, the PERCEPT corpora directly encourage reproducible research by removing barriers associated with sharing datasets across research groups. The tenets behind reproducible research are present in other related uses of the PERCEPT corpora: clinical perceptual training, instruction in the use of tools for speech analysis, and development of speech technologies to facilitate acoustic analysis and automated classification of clinical child speech.

Acknowledgements

We wish to thank our participants and their families, as well as our many research speech language pathologists (most notably, Megan Leece), whose dedication, time, and ingenuity have generated the data for this corpus. We are also appreciative of the research assistants involved in corpus curation, including Kelly Garcia, Allison Corsetti, and Michela Eivers (annotation, verification), as well as Felicia Pace (Google Speech Data). Much gratitude is also given to Elizabeth Roepke, who provided thought-provoking insight related to the use of PERCEPT in an academic context. Funding for corpus compilation has been provided by National Institute on Deafness and Other Communication Disorders (NIH R01DC017476-S2, T. McAllister, PI). This research was supported in part through computational resources provided by Syracuse University (NSF ACI-1341006; NSF ACI-1541396).

Data Availability Statement

The open-access versions of the PERCEPT Corpora are published at https://phonbank.talkbank.org/. Scripts and data associated with the Phon demonstration of the

corpus are available through the PERCEPT page at the Open Science Framework: https://osf.io/nqzd9/. Assistance with these resources is available through correspondence with the first author or in the PERCEPT channel of the Slack workspace for corpus phonetics: Phon Corps (tinyurl.com/2tnm2vkw)[8].

[8] For transparency, the full URL is: https://join.slack.com/t/phoncorps/shared_invite/zt-pr9g3f5d-5i8mf7ts73e8TZDxSl_q3Q

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining,

Altman, D. G. (1990). *Practical Statistics for Medical Research*. CRC press.

ASHA. (2018). *School practice mini- survey summary report: Number and type of responses.*

Ayala, S. A., Eads, A., Kabakoff, H., Swartz, M. T., Shiller, D. M., Hill, J., Hitchcock, E. R., Preston, J. L., & McAllister, T. (2023). Auditory and Somatosensory Development for Speech in Later Childhood. *Journal of Speech, Language, and Hearing Research*, 1-22.

Ball, M. J. (2017). Transcribing rhotics in normal and disordered speech. *Clinical Linguistics & Phonetics, 31*(10), 806-809. https://doi.org/10.1080/02699206.2017.1326169

Barreda, S. (2021). Fast Track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard, 7*(1).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology review*, *16*(4), 161-169. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2366174/pdf/nihms-22298.pdf

Benway, N. R., Hitchcock, E., McAllister, T., Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual /ɹ/ errors in American English: A single case randomization design. *American Journal of Speech-Language Pathology*.

Benway, N. R., & Preston, J. L. (in preparation). Artificial Intelligence Assisted Speech Therapy for /ɹ/ using Speech Motor Chaining and the PERCEPT Engine: a Single Case Experimental Clinical Trial with ChainingAI.

Benway, N. R., & Preston, J. L. (under review). Prospective Validation of Motor-Based Intervention with Automated Mispronunciation Detection of Rhotics in Residual Speech Sound Disorders.

Benway, N. R., Preston, J. L., Hitchcock, E. R., & McAllister, T. (2022). *PERCEPT-R Corpus*. https://doi.org/10.21415/0JPJ-X403

Benway, N. R., Preston, J. L., Hitchcock, E. R., Rose, Y., Salekin, A., Liang, W., & McAllister, T. (in press). Reproducible Speech Research with the Artificial-Intelligence-Ready PERCEPT Corpora *Journal of Speech, Language, and Hearing Research*.

Benway, N. R., Preston, J. L., Hitchcock, E. R., Salekin, A., Sharma, H., & McAllister, T. (2022). *PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of /ɹ/* INTERSPEECH 2022: Proceedings of the 23rd Annual Conference of the International Speech Communication Association (ISCA), Incheon, Republic of Korea.

Benway, N. R., Preston, J. L., Salekin, A., & McAllister, T. (under review). Automated detection of rhoticity of American English /ɹ/ in children with residual speech sound disorders: The PERCEPT-R Classifier

Benway, N. R., Preston, J. L., Salekin, A., Xiao, Y., Sharma, H., & McAllister, T. (under review). Classifying Rhoticity of /ɹ/ in Speech Sound Disorder using Age-and-Sex Normalized Formants.

Berisha, V., Krantsevich, C., Stegmann, G., Hahn, S., & Liss, J. (2022). Are reported accuracies in the clinical speech machine learning literature overoptimistic? Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,

Boersma, P., & Weenink, D. (2019). *Praat [Computer Software]*. (Version 6.1.38) https://www.fon.hum.uva.nl/praat/

Bowers, L., & Huisingh, R. (2018). *LAT-NU: LinguiSystems Articulation Test–Normative Update* Pro-Ed.

Boyce, S. E. (2015). The articulatory phonetics of /r/ for residual speech errors. *Seminars in Speech and Language, 36*(4), 257-270. https://doi.org/10.1055/s-0035-1562909

Brandel, J., & Frome Loeb, D. (2011). Program intensity and service delivery models in the schools: SLP survey results. *Language Speech and Hearing Services in Schools, 42*(4), 461-490. https://doi.org/10.1044/0161-1461(2011/10-0019)

Breton, J., & Robertson, E. M. (2017). Dual enhancement mechanisms for overnight motor memory consolidation. *Nat Hum Behav, 1*(6). https://doi.org/10.1038/s41562-017-0111

Campbell, H., Harel, D., Hitchcock, E., & McAllister Byun, T. (2018). Selecting an acoustic correlate for automated measurement of American English rhotic production in children. *International Journal of Speech Language Pathology, 20*(6), 635-643. https://doi.org/10.1080/17549507.2017.1359334

Campbell, H. M., Harel, D., & Byun, T. M. (2017). Selecting an acoustic correlate for automated measurement of /r/ production in children. *The Journal of the Acoustical Society of America, 141*(5), 3572-3572. https://doi.org/10.1121/1.4987592

Chen, Y.-P. P., Johnson, C., Lalbakhsh, P., Caelli, T., Deng, G., Tay, D., Erickson, S., Broadbridge, P., El Refaie, A., Doube, W., & Morris, M. E. (2016). Systematic review of virtual speech therapists for speech disorders. *Computer Speech & Language, 37*, 98-128. https://doi.org/https://doi.org/10.1016/j.csl.2015.08.005

Chilba, T., & Kajiyama, M. (1941). *The Vowel, its Nature and Structure*. Tokyo-Kaiseikan Publishing Company Ltd.

Childers, D. G. (1978). *Modern spectrum analysis*. IEEE Computer Society Press.

Delattre, P., & Freeman, D. C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics, 6*(44), 29.

Derdemezis, E., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., & Bolt, D. M. (2016). Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *American Journal of Speech-Language Pathology, 25*(3), 335-354. https://doi.org/doi:10.1044/2015_AJSLP-15-0020

Eshky, A., Ribeiro, M. S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J. M., & Wrench, A. (2018). *UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions* INTERSPEECH 2018: Proceedings of the 19th Annual Conference of the International Speech Communication Association (ISCA),

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., & Alwan, A. (2000). Acoustic modeling of American English /r/. *Journal of the Acoustical Society of America, 108*(1), 343-356. https://doi.org/10.1121/1.429469

Fainberg, J., Bell, P., Lincoln, M., & Renals, S. (2016). Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation. INTERSPEECH 2016: Proceedings of the 17th Annual Conference of the International Speech Communication Association (ISCA), San Fransisco, USA.

Flipsen, P., Jr. (2015). Emergence and prevalence of persistent and residual speech errors. *Seminars in Speech and Language, 36*(4), 217-223. https://doi.org/10.1055/s-0035-1562905

Flipsen, P., Jr., Hammer, J. B., & Yost, K. M. (2005). Measuring severity of involvement in speech delay: segmental and whole-word measures. *American Journal of Speech-Lanugage Pathology, 14*(4), 298-312. https://doi.org/10.1044/1058-0360(2005/029)

Furlong, L., Erickson, S., & Morris, M. E. (2017). Computer-based speech therapy for childhood speech sound disorders. *Journal of Communication Disorders, 68*, 50-69. https://doi.org/https://doi.org/10.1016/j.jcomdis.2017.06.007

Furlong, L., Morris, M., Serry, T., & Erickson, S. (2018). Mobile apps for treatment of speech disorders in children: An evidence-based analysis of quality and efficacy. *PLoS One, 13*(8), e0201513. https://doi.org/10.1371/journal.pone.0201513

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: mixed-methods study. *Journal of clinical epidemiology*. https://doi.org/https://doi.org/10.1016/j.jclinepi.2022.05.019

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189-211. https://doi.org/10.1080/19345747.2011.618213

Goldman, R., & Fristoe, M. (2000). *GFTA-2: Goldman Fristoe Test of Articulation, Second Edition*. American Guidance Service.

Goldman, R., & Fristoe, M. (2015). *Goldman Fristoe Test of Articulation - Third Edition*. Pearson.

Graves, A., Mohamed, A., & Hinton, G. (2013, 26-31 May 2013). Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing,

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10*(1), 14-23.

Guadagnoli, M., & Lee, T. (2004). Challenge point, a framework for conceptualizing the effects of varius practive conditions in motor learning. *Journal of Motor Behavior*.

Gupta, S., & DiPadova, A. (2019, June). Deep Learning and Sociophonetics: Automatic Coding of Rhoticity Using Neural Networks. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, Minneapolis, Minnesota.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Gwet, K. L. (2019). *irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC).* (Version 1.0)

Hair, A., Ballard, K. J., Markoulli, C., Monroe, P., Mckechnie, J., Ahmed, B., & Gutierrez-Osuna, R. (2021). A Longitudinal Evaluation of Tablet-Based Child Speech Therapy with

Apraxia World. *ACM Trans. Access. Comput., 14*(1), Article 3. https://doi.org/10.1145/3433607

Harel, D., & McAllister, T. (2019). Multilevel Models for Communication Sciences and Disorders. *Journal of Speech, Language, and Hearing Research, 62*(4), 783-801. https://doi.org/10.1044/2018_JSLHR-S-18-0075

Harper, S., Goldstein, L., & Narayanan, S. (2020). Variability in individual constriction contributions to third formant values in American English /ɹ/. *The Journal of the Acoustical Society of America, 147*(6), 3905-3916. https://doi.org/10.1121/10.0001413

Health Workforce Australia. (2014). *Speech Pathologists in Focus* (Australia's Health Workforce Series, Issue.

Hedlund, G. (2022). *PERCEPT_import.groovy*.

Hedlund, G., & Rose, Y. (2019). *Phon [Computer Software]*. (Version 3.0.6-beta.4) Retrieved from https://phon.ca.

Heselwood, B., & Plug, L. (2011). The Role of F2 and F3 in the Perception of Rhoticity: Evidence from Listening Experiments. ICPhS,

Hitchcock, E. R., Harel, D., & McAllister Byun, T. (2015). Social, Emotional, and Academic Impact of Residual Speech Errors in School-Aged Children: A Survey Study. *Semin Speech Lang, 36*(4), 283-294. https://doi.org/10.1055/s-0035-1562911

Hitchcock, E. R., Swartz, M. T., & Lopez, M. (2019). Speech sound disorder and visual biofeedback intervention: A preliminary investigation of treatment intensity. *Seminars in Speech and Language, 40*(02), 124-137.

Huang, J.-T., Li, J., & Gong, Y. (2015). An analysis of convolutional neural networks for speech recognition. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Ingram, D. (2002). The measurement of whole-word productions. *Journal of Child Language, 29*(4), 713-733.

Jacko, J. A. (2012). Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics, 71*, 1-15. https://doi.org/https://doi.org/10.1016/j.wocn.2018.07.001

Kaipa, R., & Peterson, A. M. (2016). A systematic review of treatment intensity in speech disorders. *International Journal of Speech Language Pathology, 18*(6), 507-520. https://doi.org/10.3109/17549507.2015.1126640

Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048*.

Katz, L. A., Maag, A., Fallon, K. A., Blenkarn, K., & Smith, M. K. (2010). What Makes a Caseload (Un)Manageable? School-Based Speech-Language Pathologists Speak. *Language, Speech, and Hearing Services in Schools, 41*(2), 139-151. https://doi.org/10.1044/0161-1461(2009/08-0090)

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18*(2), 124-132.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: evaluations and recommendations. Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction,

Klein, H. B., McAllister Byun, T., Davidson, L., & Grigos, M. I. (2013). A Multidimensional Investigation of Children's /r/ Productions: Perceptual, Ultrasound, and Acoustic

Measures. *American Journal of Speech-Language Pathology, 22*(3), 540-553. https://doi.org/10.1044/1058-0360(2013/12-0137)

Knowles, T., Clayards, M., & Sonderegger, M. (2018). Examining factors influencing the viability of automatic acoustic analysis of child speech. *Journal of Speech, Language, and Hearing Research, 61*(10), 2487-2501. https://doi.org/10.1044/2018_JSLHR-S-17-0275

Koegel, L. K., Koegel, R., L., & Ingham, J. C. (1986). Programming Rapid Generalization of Correct Articulation through Self-Monitoring Procedures. *Journal of Speech and Hearing Disorders, 51*(1), 24-32. https://doi.org/10.1044/jshd.5101.24

Koegel, R., L., Koegel, L. K., Ingham, J. C., & Van Voy, K. (1988). Within-Clinic versus Outside-of-Clinic Self-Monitoring of Articulation to Promote Generalization. *Journal of Speech and Hearing Disorders, 53*(4), 392-399. https://doi.org/10.1044/jshd.5304.392

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*(3-4), 445-463. https://doi.org/10.1080/09602011.2013.815636

Ledford, J. R., Lane, J. D., & Severini, K. E. (2018). Systematic Use of Visual Analysis for Assessing Outcomes in Single Case Design Studies. *Brain Impairment, 19*(1), 4-17. https://doi.org/10.1017/BrImp.2017.16

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105*(3), 1455-1468. https://asa-scitation-org.libezproxy2.syr.edu/doi/pdf/10.1121/1.426686

Leonard, M. K., Baud, M. O., Sjerps, M. J., & Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nature Communications, 7*(1), 13619. https://doi.org/10.1038/ncomms13619

Leung, W., Liu, X., & Meng, H. (2019, 12-17 May 2019). CNN-RNN-CTC Based End-to-end Mispronunciation Detection and Diagnosis. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Lewis, B. A., Freebairn, L., Tag, J., Ciesla, A. A., Iyengar, S. K., Stein, C. M., & Taylor, H. G. (2015). Adolescent outcomes of children with early speech sound disorders with and without language impairment. *American Journal of Speech-Lanugage Pathology, 24*(2), 150-163. https://doi.org/10.1044/2014_AJSLP-14-0075

Li, S. R., Dugan, S., Masterson, J., Hudepohl, H., Annand, C., Spencer, C., Seward, R., Riley, M. A., Boyce, S., & Mast, T. D. (2023). Classification of accurate and misarticulated /ɑr/ for ultrasound biofeedback using tongue part displacement trajectories. *Clinical Linguistics & Phonetics*, *37*(2), 196-222. https://doi.org/10.1080/02699206.2022.2039777

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q. M., Sainath, T. N., Senior, A., Beaufays, F., & Bacchiani, M. (2015). Large vocabulary automatic speech recognition for children. INTERSPEECH 2015: Proceedings of the 16th Annual Conference of the International Speech Communication Association (ISCA), Dresden, Germany.

Maas, E., & Farinella, K. A. (2012). Random versus blocked practice in treatment for childhood apraxia of speech. *Journal of Speech Language and Hearing Ressearch, 55*(2), 561-578. https://doi.org/10.1044/1092-4388(2011/11-0120)

Maas, E., Robin, D. A., Austermann Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology, 17*(3), 277-298. https://doi.org/10.1044/1058-0360(2008/025)

MacDowell, M., Glasser, M., Fitts, M., Nielsen, K., & Hunsaker, M. (2010). A national view of rural health workforce issues in the USA. *Rural and remote health, 10*(3), 1531.

Macrae, T. (2017). Stimulus Characteristics of Single-Word Tests of Children's Speech Sound Production. *Language, Speech, and Hearing Services in Schools, 48*(4), 219-233. https://doi.org/10.1044/2017_LSHSS-16-0050

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language, 12*(2), 271-295.

Mahr, T. J., Berisha, V., Kawabata, K., Liss, J., & Hustad, K. C. (2021). Performance of Forced-Alignment Algorithms on Children's Speech. *Journal of Speech, Language, and Hearing Research, 64*(6S), 2213-2222. https://doi.org/10.1044/2020_JSLHR-20-00268

Mahrt, T. (2016). *PraatIO*. https://github.com/timmahrt/praatIO

Matthews, T., Barbeau-Morrison, A., & Rvachew, S. (2021). Application of the Challenge Point Framework During Treatment of Speech Sound Disorders. *Journal of Speech Language and Hearing Research, 64*(10), 3769-3785. https://doi.org/10.1044/2021_jslhr-20-00437

McAllister Byun, T., Halpin, P. F., & Szeredi, D. (2015). Online crowdsourcing for efficient rating of speech: A validation study. *Journal of Communication Disorders, 53*, 70-83. https://doi.org/10.1016/j.jcomdis.2014.11.003

McAllister Byun, T., Harel, D., Halpin, P. F., & Szeredi, D. (2016). Deriving gradient measures of child speech from crowdsourced ratings. *Journal of Communication Disorders, 64*, 91-102. https://doi.org/10.1016/j.jcomdis.2016.07.001

McAllister Byun, T., & Rose, Y. (2016). Analyzing Clinical Phonological Data Using Phon. *Seminars in Speech and Language, 37*(2), 85-105. https://doi.org/10.1055/s-0036-1580741

McAllister, T., Hitchcock, E. R., & Ortiz, J. A. (2021). Computer-Assisted Challenge Point Intervention for Residual Speech Errors. *Perspectives of the ASHA Special Interest Groups*. https://doi.org/10.1044/2020_PERSP-20-00191

McAllister, T., Preston, J. L., Hitchcock, E. R., & Hill, J. (2020). Protocol for Correcting Residual Errors with Spectral, ULtrasound, Traditional Speech therapy Randomized Controlled Trial (C-RESULTS RCT). *BMC pediatrics, 20*(1), 66.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). *Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi* INTERSPEECH 2017: Proceedings of the 18th Annual Conference of the International Speech Communication Association (ISCA), Stockholm, Sweeden.

McCormack, J., McLeod, S., McAllister, L., & Harrison, L. J. (2009). A systematic review of the association between childhood speech impairment and participation across the lifespan. *International Journal of Speech-Language Pathology, 11*(2), 155-170. https://doi.org/10.1080/17549500802676859

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech Language Pathology, 20*(6), 583-598. https://doi.org/10.1080/17549507.2018.1477991

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Murray, E., McCabe, P., & Ballard, K. J. (2020). The influence of type of feedback during tablet-based delivery of intensive treatment for childhood apraxia of speech. *Journal of Communication Disorders*, 106026. https://doi.org/https://doi.org/10.1016/j.jcomdis.2020.106026

McLeod, S., Ballard, K. J., Ahmed, B., McGill, N., & Brown, M. I. (2020). Supporting Children With Speech Sound Disorders During COVID-19 Restrictions: Technological Solutions. *Perspectives of the ASHA Special Interest Groups*. https://doi.org/doi:10.1044/2020_PERSP-20-00128

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*(6174), 1006-1010. https://doi.org/10.1126/science.1245994

Miccio, A., Elber, M., & Forrest, K. (1999). The relationship between stimulability and phonological acquistion in children with normally developing and disordered phonologies. *American Journal of Speech-Language Pathology, 8*, 347-363.

Miller, P. (2016). Itinerancy between attractor states in neural systems. *Current Opinion in Neurobiology, 40*, 14-22. https://doi.org/https://doi.org/10.1016/j.conb.2016.05.005

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. FAT* '19: Conference on Fairness, Accountability, and Transparency, Atlanta, Georgia, USA.

Nagy, N., & Irwin, P. (2010). Boston (r): Neighbo(r)s nea(r) and fa(r). *Language Variation and Change*, *22*(2), 241-278. https://doi.org/10.1017/S0954394510000062

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: nonoverlap of all pairs. *Behav Ther*, *40*(4), 357-367. https://doi.org/10.1016/j.beth.2008.10.006

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Plante, E., & Vance, R. (1994). Selection of Preschool Language Tests. *Language, Speech, and Hearing Services in Schools*, *25*(1), 15-24. https://doi.org/10.1044/0161-1461.2501.15

Povey, D. (2012). *train_map.sh.*

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., & Schwarz, P. (2011). The Kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding,

Preston, J. L., Benway, N. R., Leece, M. C., & Caballero, N. F. (2021). Concurrent Validity Between Two Sound Sequencing Tasks Used to Identify Childhood Apraxia of Speech in School-Age Children. *American Journal of Speech-Language Pathology, 30*(3S), 1580-1588. https://doi.org/10.1044/2020_AJSLP-20-00108

Preston, J. L., Benway, N. R., Leece, M. C., Hitchcock, E. R., & McAllister, T. (2020). Tutorial: Motor-based Treatment Strategies for /r/ Distortions. *Language, Speech, and Hearing Services in Schools, 54*, 966-980.

Preston, J. L., Caballero, N. F., Leece, M. C., Wang, D., Herbst, B. M., & Benway, N. R. (under review). A Randomized Controlled Trial of Treatment Distribution and Biofeedback Effects on Speech Production in School-Aged Children with Apraxia of Speech.

Preston, J. L., & Leece, M. C. (2017). Intensive treatment for persisting rhotic distortions: A case series. *American Journal of Speech-Language Pathology, 26*(4), 1066-1079. https://doi.org/10.1044/2017_AJSLP-16-0232

Preston, J. L., Leece, M. C., & Maas, E. (2016). Intensive treatment with ultrasound visual feedback for speech sound errors in childhood apraxia. *Frontiers in human neuroscience, 10*, 1-9. https://doi.org/10.3389/fnhum.2016.00440

Preston, J. L., Leece, M. C., & Maas, E. (2017). Motor-based treatment with and without ultrasound feedback for residual speech-sound errors. *International Journal of Language & Communication Disorders, 52*(1), 80-94. https://doi.org/10.1111/1460-6984.12259

Preston, J. L., Leece, M. C., McNamara, K., & Maas, E. (2017). Variable practice to enhance speech learning in ultrasound biofeedback treatment for childhood apraxia of speech: A single case experimental study. *American Journal of Speech-Language Pathology, 26*(3), 840-852. https://doi.org/10.1044/2017_AJSLP-16-0155

Preston, J. L., Leece, M. C., & Storto, J. (2019). Tutorial: Speech motor chaining treatment for school-age children with speech sound disorders. *Language, Speech, and Hearing Services in Schools, 50*(3), 343-355. https://doi.org/10.1044/2018_LSHSS-18-0081

Preston, J. L., McCabe, P., Rivera-Campos, A., Whittle, J. L., Landry, E., & Maas, E. (2014). Ultrasound visual feedback treatment and practice variability for residual speech sound errors. *Journal of Speech, Language, and Hearing Research, 57*(6), 2102-2115. https://doi.org/10.1044/2014_JSLHR-S-14-0031

Preston, J. L., Preston, N. J., & Benway, N. R. (2022). *Speech Motor Chaining Web-App.*

Pring, T., Flood, E., Dodd, B., & Joffe, V. (2012). The working practices and clinical experiences of paediatric speech and language therapists: a national UK survey [https://doi.org/10.1111/j.1460-6984.2012.00177.x]. *International Journal of Language & Communication Disorders, 47*(6), 696-708. https://doi.org/https://doi.org/10.1111/j.1460-6984.2012.00177.x

R Core Team. (2013). R: A language and environment for statistical computing.

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research.* (Version 1.9.12) Northwestern University, Evanston, Illinois.

Ribeiro, M. S., Cleland, J., Eshky, A., Richmond, K., & Renals, S. (2021). Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors. *Speech Communication, 128*, 24-34. https://doi.org/https://doi.org/10.1016/j.specom.2021.02.001

Robey, R. R. (2004). A five-phase model for clinical-outcome research. *J Commun Disord, 37*(5), 401-411. https://doi.org/10.1016/j.jcomdis.2004.04.003

Robles Herrera, S., Ceberio, M., & Kreinovich, V. (2022). When is deep learning better and when is shallow learning better: qualitative analysis. *International Journal of Parallel, Emergent and Distributed Systems, 37*(5), 589-595.

Rose, Y., & MacWhinney, B. (2014). The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford Handbook of Corpus Phonology*.

Rose, Y., & Stoel-Gammon, C. (2015). Using PhonBank and Phon in studies of phonological development and disorders. *Clinical Linguistics & Phonetics, 29*(8-10), 686-700. https://doi.org/10.3109/02699206.2015.1041609

Ruscello, D. M. (1995). Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders, 28*(4), 279-302.

Rvachew, S., & Brosseau-Lapré, F. (2016). *Developmental Phonological Disorders: Foundations of Clinical Practice*. Plural Publishing.

Shahin, M., Zafar, U., & Ahmed, B. (2020). The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 400-412. https://doi.org/10.1109/JSTSP.2019.2959393

Shields, R., & Hopf, S. C. (2023). Intervention for residual speech errors in adolescents and adults: A systematised review. *Clinical Linguistics & Phonetics*, 1-24. https://doi.org/10.1080/02699206.2023.2186765

Shriberg, L. D., Flipsen Jr, P., Karlsson, H. B., & McSweeney, J. L. (2001). Acoustic phenotypes for speech-genetics studies: An acoustic marker for residual/ɜ/distortions. *Clinical Linguistics & Phonetics, 15*(8), 631-650.

Shriberg, L. D., Lohmeier, H. L., Campbell, T. F., Dollaghan, C. A., Green, J. R., & Moore, C. A. (2009). A nonword repetition task for speakers with misarticulations: the Syllable Repetition Task (SRT). *Journal of Speech, Language, and Hearing Research, 52*(5), 1189-1212. https://doi.org/10.1044/1092-4388(2009/08-0047)

Silverman, F. H., & Paulus, P. G. (1989). Peer reactions to teenagers who substitute /w/ for /r/. *Language, Speech, and Hearing Services in Schools, 20*(2), 219-221.

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom),

Speights Atkins, M., Bailey, D. J., & Boyce, S. E. (2020). Speech exemplar and evaluation database (SEED) for clinical training in articulatory phonetics and speech science. *Clinical Linguistics & Phonetics, 34*(9), 878-886. https://doi.org/10.1080/02699206.2020.1743761

Strik, H., Truong, K., de Wet, F., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication, 51*(10), 845-852. https://doi.org/https://doi.org/10.1016/j.specom.2009.05.007

Sugden, E., Baker, E., Munro, N., Williams, A. L., & Trivette, C. M. (2018). Service delivery and intervention intensity for phonology-based speech sound disorders. *Int J Lang Commun Disord, 53*(4), 718-734. https://doi.org/10.1111/1460-6984.12399

Swartz, M. T., & Hitchcock, E. R. (2021). *Visual-acoustic Biofeedback and Auditory Masking Intervention for RSE in Children With CAS: A Case Series* American Speech-Language-Hearing Association National Convention, Washington, DC.

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the Condor experience. *Concurrency and computation: practice and experience*, *17*(2-4), 323-356.

Tiede, M. K., Boyce, S. E., Holland, C. K., & Chou, K. A. (2004). A new taxonomy of American English /r/ using MRI and ultrasound. *J. Acoust. Soc. Am., 115*(5), 2533.

Verdon, S., Wilson, L., Smith-Tamaray, M., & McAllister, L. (2011). An investigation of equity of rural speech-language pathology services for children: A geographic perspective. *International Journal of Speech-Language Pathology*, *13*(3), 239-250.

Wallach, J. D., Boyack, K. W., & Ioannidis, J. P. A. (2018). Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. *PLOS Biology, 16*(11), e2006930. https://doi.org/10.1371/journal.pbio.2006930

Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: a missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews, 13*(1), 70-77. https://doi.org/10.1002/mrdd.20139

Wiig, E., Semel, E., & Secord, W. (2013). Clinical evaluation of language fundamentals. *Bloomington, MN: Pearson*.

Wilbert, J., & Lüke, T. (2023). *Scan: Single-case data analyses for single and multiple baseline designs.*

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology, 13*(1), 61. https://doi.org/10.1186/1471-2288-13-61

Wren, Y., Miller, L. L., Peters, T. J., Emond, A., & Roulstone, S. (2016). Prevalence and Predictors of Persistent Speech Sound Disorder at Eight Years Old: Findings From a

Population Cohort Study. *Journal of Speech, Language, and Hearing Research, 59*(4), 647-673. https://doi.org/10.1044/2015_JSLHR-S-14-0282

Yang, X., Loukina, A., & Evanini, K. (2014, 7-10 Dec. 2014). Machine learning approaches to improving pronunciation error detection on an imbalanced corpus. 2014 IEEE Spoken Language Technology Workshop (SLT),

Yeung, G., & Alwan, A. (2018). On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children. INTERSPEECH 2018: Proceedings of the 19th Annual Conference of the International Speech Communication Association (ISCA), Hyderabad, India.

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron, 102*(6), 1096-1110. https://doi.org/https://doi.org/10.1016/j.neuron.2019.04.023

Yuan, H., Liu, M., Krauthammer, M., Kang, L., Miao, C., & Wu, Y. (2022). An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. *arXiv preprint arXiv:2204.11351*.

EPILOGUE

This dissertation has brought together a committee of experts in RSSD, speech analysis, and machine learning to directly consider three fundamental issues that have hindered the prior development of effective clinical speech technology systems. Chapter 1 includes an empirical demonstration of the therapeutic benefit of an AI-assisted motor-based intervention, ChainingAI, that combines Speech Motor Chaining and the PERCEPT-R Classifier. Chapter 2 explains several machine learning experiments that facilitated the technical development of the PERCEPT-R Classifier. This development was permitted by the corpus presented in Chapter 3, which has also been published as an open-access dataset to begin to offset the paucity of clinically-relevant child speech samples for system training.

Taken together, this dissertation accelerates the development of paradigm-shifting AI-driven precision treatment that is personalized to a child's speech patterns. The validation of ChainingAI as an effective driver of speech sound learning in RSSD foretells future service delivery models in which the intervention intensity gap may someday be narrowed through hybrid clinician-AI treatment programs, with ChainingAI enabling effective evidence-based home practice that supplements face-to-face services. In this framework, clinically ethical AI-mediated speech therapy still demands that the clinician controls eligibility determination, target selection, treatment programming, stimulability cueing, and treatment monitoring, while also considering a child's own desire to modify their speech. However, the repetition of session-based practice and feedback in between visits with a clinician can be automated.

These studies have set the foundation for future research that will further our goal of expanding access to effective, theoretically-motivated, high-fidelity treatment. Ongoing work is continuing to drive these projects forward. Benway, Siriwardena, et al. (under review) has

piloted rhotic /ɹ/ classification with different feature encodings that may further increase PERCEPT predictive performance. Work is also underway in the Speech Production Lab at Syracuse University to build classifiers for /s/ and /z/, which, together with PERCEPT-R, will allow ChainingAI to target over 90% of the sound errors observed in individuals with RSSD (as demonstrated by reanalysis of Table 1 by Lewis et al., 2015). This five-year project (NIH 1R01DC020959-01, J. Preston, PI) will specifically be tasked to also increase the racial and ethnic diversity reflected in the speakers of the PERCEPT Corpora and the capabilities of the PERCEPT-R Classifier. Success in these projects and studies that could continue to arise from them may set the stage for additional development of classifiers for other speech sounds, clinical tasks, different ages, and other subtypes of speech sound disorders.

References

Benway, N. R., Siriwardena, Y. M., Preston, J. L., Hitchcock, E., McAllister, T., & Espy-Wilson, C. (under review). Acoustic-to-Articulatory Speech Inversion Features for Mispronunciation Detection in Child Speech Sound Disorders.

Lewis, B. A., Freebairn, L., Tag, J., Ciesla, A. A., Iyengar, S. K., Stein, C. M., & Taylor, H. G. (2015). Adolescent outcomes of children with early speech sound disorders with and without language impairment. *American Journal of Speech-Language Pathology, 24*(2), 150-163. https://doi.org/10.1044/2014_AJSLP-14-0075

APPENDICES

APPENDIX CHAPTER 1-A

Methodological reporting in the SCRIBE framework.

SCRIBE Factor	Description
Design	No treatment-treatment-no treatment (A-B-A) multiple baseline single case experimental design with a priori determination of phase changes
Procedural Changes	Participants 1121 and 1130 were treated with a classifier that used an updated procedure for participant-specific fine-tuning, as described in the text
Replication	5 subjects
Randomization	Concealed randomization: number of baselines (5-10)
Selection criteria	• Stimulable for /ɹ/ • GFTA-3 < 8th percentile • Pass CELF-5 Screening • Pass childhood apraxia of speech screening • Protrude tongue from mouth • No known history of neurodevelopmental disorder, neurological disorder, brain injury, voice, or fluency disorder • No major orthodontia that blocks tongue contact with hard palate
Participant selection characteristics	Children who can produce an adult-like /ɹ/ "some of the time" referred from advertisement to clinicians
Setting	Hybrid (in-person/remote)
Ethics Approval	Syracuse University (#21-370) and The College of Saint Rose (#4374)
Measures	Expert listener perceptual rating of /ɹ/ in practiced Chains and unpracticed words
Masking	Listeners for the primary outcome measure were masked to participant identity and timepoint of utterance
Equipment	Participant computer with internet connection Researcher computer Speech Motor Chaining Web App Participant Smartphone Shure MV5 cardioid digital condenser mic (20 Hz to 20 kHz) Sennheiser MKE600 super-cardioid digital condenser mic (40 Hz to 20 kHz)
Intervention	Artificial intelligence driven Speech Motor Chaining web app (Chaining-AI)

	• Prepractice: < 10 minutes or 16 correct productions • Block size: 4 • Number of chains: 4 • Targets per chain: 2 • Block accuracy criterion: 3/4 • Random practice: 5 minutes
Procedural Fidelity	ChainingAI is inherently high-fidelity with regard to the therapeutic parameters specified above. Fidelity also evaluated for participant interaction with ChainingAI • Frequency of redirection • Frequency of technical support • Total prepractice productions • Total ChainingAI productions • Average minutes: seconds spent in practice • ChainingAI productions per minute
Analyses	• Linear mixed models to examine if ChainingAI resulted in near-immediate improvement in the perceived rhoticity of /ɹ/ on practiced Chains. • Visual analysis of level, trend, and nonoverlap to determine if the total AI-assisted treatment package resulted in perceptual improvement in /ɹ/ on untreated words in post-session probes, compared to a no-treatment baseline. • Pre–post change with effect sizes • F1-score, the harmonic mean of precision and recall (i.e., positive predictive value and sensitivity), of PERCEPT predictions compared to clinician judgments • Survey exploration of parent and participant end-user experience with AI-assisted intervention

Note. SCRIBE = Single Case Reporting Guideline in Behavioral Interventions (Tate et al., 2016). Please see text for full descriptions of each factor.

APPENDIX CHAPTER 2-A

Model Card for Model Reporting and Model Reproducibility

Adapted from Mitchell et al., (2019) and Kapoor and Narayanan (2022)

Information about report

Benway, Nina R
Preston, Jonathan L
Salekin, Asif
McAllister, Tara

Automated detection of rhoticity of American English /ɹ/ in children with residual speech sound disorders: The PERCEPT-R Classifier

The PERCPET-R 2.2.2p Corpus is available under a CC BY-NC-SA license, through PhonBank. Note that the present investigation also included participants who were not included in the open-access release of the PERCEPT-R Corpus after review of participant consent/asset forms.

Email address of the corresponding author:
nrbenway@syr.edu

Model details and scientific claim(s) of interest

Does your paper make a generalizable claim based on the ML model?

The PERCEPT-R classifier is intended to be used to predict human perceptual judgment of /ɹ/ (i.e., fully rhotic, derhotic) in the context of stimulable participants with residual speech sound disorders impacting /ɹ/ in fully rhotic dialects of American English.

Is the scientific claim made about a distribution or population from which you can sample?

Population: child, adolescent, and young adult speakers of American English with residual speech sound disorders impacting /ɹ/ and their age-matched peers. Sample: Participants in the PERCEPT-R corpus v. 2.2.2. For the test and validation subsets we employed age-and-sex stratified random sampling without replacement from a subset of participants with residual speech sound disorders who had an fully rhotic: derhotic ratio between 4:1 and 1:4. The training subset included the balance of participants from the corpus, representing speakers with residual speech sound disorders and typical speakers.

Does the scientific claim only apply to certain subsets of the distribution mentioned in Q6?

Our model was tested against single-channel, 44.1 kHz wav audio from speakers of fully-rhotic dialects of American English. The model's predictions of /ɹ/ perceptual judgment may not generalize to 1) non-clinical settings, 2) non-fully rhotic dialects of American English, 3) low-quality audio (i.e., insufficient sampling rate, speaker cross-talk), 4) utterances containing more than one /ɹ/.

Ethical considerations and non-intended use cases

- PERCPET-R is not intended to be used outside of the direction of a speech-language pathologist.
- PERCEPT-R is not intended to rule in/rule out speech sound disorder in specific individuals.

- PERCEPT-R is not intended to predict perceptual judgement in dialects of American English that are not fully rhotic.
- PERCEPT-R is not intended to predict perceptual judgment in typical speakers.
- PERCEPT-R is not intended to predict perceptual judgment of speech sounds other than /ɹ/.
- PERCEPT-R is not intended to predict perceptual judgment in individuals outside of the validated age range (currently, 8 years to 24 years).

Metrics

PERCEPT-R was validated using participant-specific F1-score, the harmonic mean of precision and recall.

Criterion: Train-test split is maintained across all steps in creating the model

How was the dataset split into train and test sets? (For example, cross-validation; separate train and test sets).

The data were split into training, validation, and test sets at the level of the participant. We verified that data from one speaker was only included in one experimental set.

Are there duplicates in the dataset? If yes, explain how duplicates are handled to ensure the train-test split.

Although a handful of participants were recruited to more than one component study in the source data, when these data were curated into the PERCEPT corpora these participants received the same PERCEPT ID number. Because the development of the PERCEPT classifier referenced PERCEPT corpus IDs and not original study IDs, there are no duplicate participants across the different experimental subsets.

In case the dataset has dependencies (e.g., multiple rows of data from the same patient), describe how the dependencies were addressed.

There are multiple utterances from the same participant, but because we block-randomized participants to experimental subsets by participant ID, and verified that participants were not represented in more than one dataset, we expect dataset leakage to be mitigated.

List all the pre-processing steps used in creating your model (imputing missing data, normalizing feature values, selecting a subset of rows from the dataset for building the model).

- Missing ground truths were not imputed.
- Entire feature sets were not imputed in the case of missing audio.
- Missing points in the formant or MFCC time series were imputed by averaging the values of the timepoint trajectory immediately before and immediately after the missing values, except when the edge-padding was employed due to the missing values being the first or last values of the time series.
- In the utterance-normalized conditions, features were normalized relative to the other values in the feature time series for a given rhotic-associated interval.
- In the age-and-sex normalized condition, features were normalized relative to published third party means and standard deviations for /ɹ/ in speakers of American English.
- All data assigned to the training set was used in model building.

How was the train-test split observed during each pre-processing step? If applicable, use a separate line for each step mentioned in Q12.

In no instance were features generated, imputed, or normalized in a manner that referenced any other utterance in the same experimental subset or in different experimental subsets.

List all the modeling steps used in creating your model.

- Feature selection occurred with a priori empirical comparisons using shallow classifiers.
- Hyperparameter tuning was facilitated through the Optuna framework as described in the accompanying paper.
- Model selection occurred as described in the accompanying paper.

How was the train-test split observed during each modeling step?

- The validation set was used to guide model training.
- Hyperparameters were not retuned for out-of-box testing in the test set.
- During participant-specific testing, the out-of-box model was reloaded from disk for each participant to preclude carryover from one participant to the next.

List all the evaluation steps used in evaluating model performance.

Please see the accompany paper for a description. Out of sample testing is reported in an accompanying paper (Benway et al, in preparation).

How was the train-test split observed during each evaluation step?

Testing happened in a separate program after all training runs were complete, with data reloaded at each time of testing.

Criterion: Test set is drawn from the distribution of scientific interest.

Why is your test set representative of the population or distribution about which you are making your scientific claims?

The test set reflects individuals who presented for studies of residual speech sound disorders of /ɹ/ in the context of fully rhotic dialects of American English, and the guidelines for end use of the model are reflective of the inclusionary criteria from these original studies. We also switched the validation and test set, and present the final model performance as the average performance across all participants in validation or test. We therefore expect the sample herein to be clinically reflective of the end use population.

Explain the process for selecting the test set and why this does not introduce selection bias in the learning process.

Test set selection is described in detail in the accompanying paper. It is possible that selecting the test/validation sets in an age-and-sex stratified manner did introduce bias during training, as this by definition excluded certain age-and-sex strata from the training set when there were not enough participants in a given strata to overflow into the training set once the test and validation sets were assigned.

Criterion: Each feature used in the model is legitimate for the task

List the features used in the model, alongside an argument for their legitimacy.

All model features can be obtained from real-world information at the time a participant presents to the clinical setting: the individual's age, the individual's sex, and audio recordings of participants producing words with /ɹ/.

APPENDIX CHAPTER 3-A

Worked Example of Data Exploration with PERCEPT-R and Phon

The following is a minimal replicable example summarizing one use of the PERCEPT-R corpus, formatted as a methodological checklist to highlight the utility of the corpus for educational purposes. Editable and freely distributable versions of this checklist, the referenced Phon XML queries, and the referenced R code are available at the Open Science Foundation (OSF) page for PERCEPT (https://osf.io/nqzd9/).

This example is formatted for Phon version 3.4.3 and was tested in a Windows computing environment. Please note that educators using this example as an exercise may wish to update this document for the less resource-intensive PERCEPT-R Corpus Sample available on the Open Science Framework page.

- Introduction
  - ☐ Prevocalic and postvocalic rhotics are often treated differently in a clinical context (Boyce, 2015). It could be clinically useful to understand the extent to which syllable position (i.e., pre-vocalic/post-vocalic) and neighboring vowel articulation (i.e., vowel height-frontness/backness) are associated with perception of rhoticity in the PERCEPT-R corpus.
  - ☐ This example pulls records from the PERCEPT-R corpus that are in singleton pre-vocalic or post-vocalic positions. Instructors using this example for educational purposes may wish to have students extend this exercise to include comparisons of cluster /ɹ/ and/or syllabic /ɹ/.
- Methodology
  - ☐ Download and install R/R Studio from `https://posit.co/download/rstudio-desktop/`
  - ☐ Download the `PERCEPT_workedexample` directory from OSF and save it to your computer.
  - ☐ Install Phon from https://www.phon.ca/phon-manual/getting_started.html#download_phon
  - ☐ Download the PERCEPT-R corpus and unzip it to your computer's Phon workspace.
    - You may wish to download the full PERCEPT-R Corpus from `https://phon.talkbank.org/access/Clinical/PERCEPT-R.html`
    - Alternatively, you may wish to use the PERCEPT-R Corpus Sample which is inside the `PERCEPT_workedexample` directory from OSF
    - The file should be unzipped such that the directories are organized in the following manner:

- ☐ Phon Workspace Directory
- ☐ … PERCEPT-R 2.2.2p Directory
- ☐ ……BCS_1 Directory
- ☐ ……BCS_2 Directory
- ☐ ……et cetera

- Note: the path to your computer's Phon workspace is visible from the Phon Welcome Screen.

☐ Open the PERCEPT-R project in Phon.

☐ From the `Query` menu, select `Phones…`

☐ Repeat the following for each query:

- Set the query parameters using either option 1 OR option 2:

☐ Option 1 – load the queries that are included in the `PERCEPT_workedexample` directory

a. Save the XML queries on OSF to the analysis folder
b. From the Query window, select `Query > Browse…`
c. Load the XML query. A new window will open.

☐ Option 2 – manually set query parameters using fillable forms by ensuring the following are selected

a. Search by: Word, Then by syllable
b. Expression type: Phonex
c. Expression:
   i. For pre-vocalic stressed singleton rhotics, enter the following in the yellow box: `(?<\b\s?)ɹ\v`
      1. In plain language, this query indicates the following: "Return all vowels…but only if, looking behind (i.e., before) the vowel, there is a word boundary followed by an optional stress marker, then an /ɹ/"
      2. Note: putting the word boundary and stress marker within the look behind parentheses means that those symbols won't be captured and printed in the report with the /ɹV/ of interest.
   ii. For post-vocalic stressed singleton rhotics, enter the following in the yellow box: `\v͡ɚ(?>\b)`
      1. In plain language, this query indicates the following: "Return all vowels…but only if, looking ahead (i.e., after) the vowel, there is an /ɹ/ followed by a word boundary"
      2. Note: as above, putting the word boundary and stress marker within the look behind parentheses means that those symbols won't

be captured and printed in the report with the /Vɹ/ of interest.

iii. Keep the default settings for `Aligned Phones`, `Group Options`, and `Word Options`.

iv. Ensure the following are selected for Syllable Options: Search by syllable, Singleton syllables (words with only one syllable), Multiple syllables (initial, medial, final), Syllable stress (primary stress), Syllable type (Any syllable)

d. Ensure the following are selected for `Add aligned words: Orthography, IPA Target,` and enter the following tier names separated by ",": `responseAve, study, phase, session`

☐ Option 3 – extended/modified queries:

a. Instructors wishing to include, e.g., clusters and/or vocalic /ɝ/ in the analysis would have students adapt the above Phonex expressions to the relevant contexts.

- Select sessions to be queried
  - ☐ To reproduce this example, select all PERCEPT-R sessions. Note that fewer sessions can be queried if computational resources/processing time are concerns.
- Click `Run query` (top right)
- After the query has run, `Report Composer` will become available; click it (top left)
- Add the desired tables to the report
  - ☐ To reproduce this example, ensure only `Table(all results in one table)` is in the Report List.
  - ☐ Users might find, however, that other table formats are more appropriate for their needs.
- Click `Next: Report` (top left)
- The report will populate in this screen after it is compiled. It may take a few minutes; there is a timer in the top left. Note that if the report does not populate and you are on a computer managed by an Information Technology division, you may need to have Phon whitelisted.
- To save the report, select Export tables… Export tables as Excel workbook.
- Select the data you wish to save and click `Export`.
- Save the `.xls` file in the `PERCEPT_workedexample` directory.
- Confirm the export worked by visually inspecting the `.xls` file in a program such as Microsoft Excel
- Save the `.xls` file as a `.csv` using the `Save As…` menu, keeping the original filename.

☐ To reproduce this example, run `descriptiveStatsandDataViz.Rmd` from within the `PERCEPT_workedexample` directory.

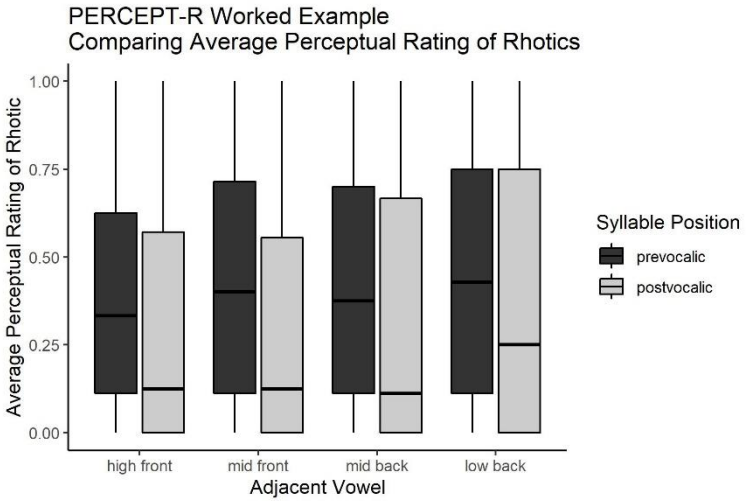
- R and R studio can be installed following these directions: `https://posit.co/download/rstudio-desktop/`
- Install the following required packages, following these directions `https://docs.posit.co/ide/user/ide/guide/ui/ui-panes.html#packages`:
- dplyr
- readr
- tidyr
- stringr
- ipa
- ggpubr

☐ The descriptiveStatsandDataViz.Rmd script will:

- Read in csv files in analysis directory and concatenate them to one dataframe
- Clean variable data types and names
- Convert IPA Unicode symbols to ASCII symbols to ensure compatibility with other R packages
- Code each observation with respect to syllable position (i.e., pre-vocalic/post-vocalic) and the articulatory context of the adjacent vowel (i.e., vowel height-frontness/backness)
- Subset the query results to participants from studies of residual speech sound disorder (i.e., filtering out participants from studies of typical speech)
- Subset the query results to vowels that appear in both syllable positions (e.g., filtering out /ɹæ/ as there is no post-vocalic /æɹ/ in encoded in the Target transcriptions)
- Print descriptive statistics:

Syllable Position	Vowel Context	Vowel (ARPABET)	Mean Rhotic Rating (Standard Deviation)	Number of Ratings
prevocalic	high front	IY, IH	0.373 (0.317)	4904
prevocalic	mid front	EY, EH	0.431 (0.336)	5328
prevocalic	mid back	OW, AH	0.425 (0.335)	5523
prevocalic	low back	AA	0.441 (0.346)	5037
postvocalic	high front	IH, IY	0.311 (0.358)	5448
postvocalic	mid front	EH	0.315 (0.356)	8073
postvocalic	mid back	AO	0.322 (0.372)	2247
postvocalic	low back	AA	0.370 (0.375)	3141

- Generate a box plot showing data distributions for vowel contexts, grouped by syllable position:



- Note: educators are encouraged to have students practice interpreting the generated descriptive statistics and plots as part of the exercise.

☐ Print the formatted data, boxplot, and descriptive statistics to the output subdirectory of the `PERCEPT_workedexample` directory.

- Discussion

☐ Users can extend the R script to include statistical analyses and effect size comparisons that can quantify the differences seen between perceptual /ɹ/ ratings and syllable position/vowel context.

☐ A priori experimentation can be designed to explore if the observed differences in perceptual /ɹ/ ratings are due to, among other mechanisms:

- A practice effect due to the frequency with which certain vowels and syllable contexts are selected as targets between pre-treatment and post-treatment timepoints.
- Rater perceptual tolerance for fully-rhotic /ɹ/ varying across syllable positions and vowel contexts.
- Articulatory facilitation of a fully-rhotic /ɹ/ by a given syllable positions and vowel contexts.

APPENDIX CHAPTER 3-B

Study-Level Summary of PERCEPT-R v2.2.2p

Table 1: Participant, rater, and stimulus characteristics for all studies included in corpus. Acronyms: SSD = speech sound disorder, TD = typically developing, unpb = unpublished, unk = unknown.

Corpus Name	Reference Number	Participant Type(s)	N of parti-cipants	N Self-Reported Repre-sentation of Black/African American, Asian, More than one race, and other People of Color	Age range	Rater Type	Mode of stimulus presentation
BCS	2	SuspectedSSD	9	1	8 - 13	Crowdsourced Listeners	Syllables: Imitation; Words: Reading
BFS	3	SuspectedSSD	6	1	9 - 15	Crowdsourced Listeners	Reading
BFS2	4	SuspectedSSD	10	3	9 - 15	Crowdsourced Listeners	Reading
CPF	5	SuspectedSSD	4	0	7 - 10	Crowdsourced Listeners	Reading
CRESULTSMOSAIC	6	TDChildren	74	10	9 - 15	Crowdsourced Listeners	Syllables: Imitation; Words: Reading
CRESULTSRCT	7	SuspectedSSD	16	3	9 - 15	Crowdsourced Listeners	Picture Naming; Reading
CRESULTSSCED	8	SuspectedSSD	6	0	9 - 15	Lab Listeners	Picture Naming; Reading

EFIF	9	SuspectedSSD	5	1	6 - 13	Crowdsourced Listeners	Syllables: Imitation; Words: Reading
EPG	10	SuspectedSSD	3	0	7 - 9	Both	Picture Naming; Reading
HFS	11	TDChildren	13	7	9 - 15	Lab Listeners	Syllables: Imitation; Elicitation within Treatment
Lillianne	12	SuspectedSSD	1	0	11 - 11	Lab Listeners	Picture Naming; Reading
PerceptionRCT	13	SuspectedSSD	29	4	7 - 15	Lab Listeners	Picture Naming; Reading
PrestonEdwards2007	14	SuspectedSSD	33	unk	10 - 15	Lab Listeners	Picture Naming
PrestonERP		TDAdultsandChildren	5	0	9 - 16	Lab Listeners	Reading
PrestonHullEdwards2013	15	PreKHistorySSD	22	unk	7 - 9	Lab Listeners	Picture Naming
PrestonIntensiveRSSD2017	16	SuspectedSSD	1	unk	17-17	Lab Listeners	Reading
PTR	17	SuspectedSSD	16	1	9 - 14	Crowdsourced Listeners	Reading
staRt	18	SuspectedSSD	4	unk	9 - 10	Crowdsourced Listeners	Reading
TD	unpb	TDAdultsandChildren	4	0	7-8	Lab Listeners	Syllables: Imitation; Words: Reading
TPT	unpb	SuspectedSSD	7	2	9 - 14	Crowdsourced Listeners	Reading
UnpublishedIntensives	unpb	SuspectedSSD	2	0	12 - 24	Lab Listeners	Imitation
US2014	19	SuspectedSSD	7	0	6 - 15	Lab Listeners	Picture Naming; Reading
VAB	20	SuspectedSSD	9	0	6 - 11	Lab Listeners	Picture Naming; Reading

Table 2: Characteristics of treatment for all studies that included a treatment component. Acronyms: US = Ultrasound, VAB = Visual-acoustic biofeedback, EPG = Electropalatography, unpb = unpublished,

Corpus Name	Reference Number	Treatment Type	Treatment Duration-Sessions	Treatment Duration - Weeks
BCS	2	Biofeedback-US-VAB-EPG	20	10
BFS	3	Biofeedback-VAB	20	10
BFS2	4	Biofeedback-VAB	20	10
CPF	5	Biofeedback-VAB, EPG	Flexible	Flexible
CRESULTSRCT	7	Biofeedback-US-VAB	19	10
CRESULTSSCED	8	Biofeedback-US-VAB	20	5
EFIF	9	Biofeedback-VAB	6 to 8	8
EPG	10	Biofeedback-EPG	16	8
Lillianne	12	Biofeedback-US	16	8.5
PerceptionRCT	13	Biofeedback-US	14	7
PrestonIntensiveRSSD2017	16	Biofeedback-US	14	1
PTR	17	Biofeedback-US	8	10
staRt	18	Biofeedback-VAB	16	6 to 8
TPT	unpb	Biofeedback-VAB	20	14
UnpublishedIntensives	unpb	Motor-No Biofeedback	8	1
US2014	19	Biofeedback-US	16	8
VAB	20	Biofeedback and Motor-Based Treatment	20	10

Appendix 3-B Citation information

1) Benway, N. R., Preston, J. L., Hitchcock, E. R., Salekin, A., Sharma, H., & McAllister, T. (2022). PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of /ɹ/. INTERSPEECH 2022: Proceedings of the 23rd Annual Conference of the International Speech Communication Association (ISCA), Incheon, Republic of Korea.
2) McAllister Byun, T., Hitchcock, E. R., & Ferron, J. (2017). Masked visual analysis: Minimizing type I error in visually guided single-case design for communication disorders. *Journal of Speech, Language, and Hearing Research*, *60*(6), 1455-1466.
3) McAllister Byun, T., & Campbell, H. (2016). Differential effects of visual-acoustic biofeedback intervention for residual speech errors. *Frontiers in Human Neuroscience*, *10*, 567.

4) McAllister Byun, T. (2017). Efficacy of visual–acoustic biofeedback intervention for residual rhotic errors: A single-subject randomization study. *Journal of Speech, Language, and Hearing Research*, *60*(5), 1175-1193.
5) McAllister, T., Hitchcock, E. R., & Ortiz, J. A. (2021). Computer-assisted challenge point intervention for residual speech errors. *Perspectives of the ASHA special interest groups*, *6*(1), 214-229.
6) Ayala, S.A., Eads, A., Kabakoff, H., Swartz, M., Shiller, D.M., Hill, J., Hitchcock, E.R., Preston, J.L., & McAllister, T. (Under review). Auditory and Somatosensory Development for Speech in Later Childhood. DOI 10.17605/OSF.IO/BKASM
7) McAllister, T., Preston, J. L., Hitchcock, E. R., & Hill, J. (2020). Protocol for correcting residual errors with spectral, ultrasound, traditional speech therapy randomized controlled trial (C-RESULTS RCT). *BMC Pediatrics*, *20*(1), 1-14.
8) Benway, N. R., Hitchcock, E. R., McAllister, T., Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual/ɹ/errors in American English: A single-case randomization design. *American Journal of Speech-Language Pathology*, *30*(4), 1819-1845.
9) McAllister Byun, T., Swartz, M. T., Halpin, P. F., Szeredi, D., & Maas, E. (2016). Direction of attentional focus in biofeedback treatment for /r/ misarticulation. *International Journal of Language & Communication Disorders*, *51*(4), 384-401.
10) Hitchcock, E. R., McAllister Byun, T., Swartz, M., & Lazarus, R. (2017). Efficacy of electropalatography for treating misarticulation of /r/. *American Journal of Speech-Language Pathology*, *26*(4), 1141-1158.
11) Campbell, H., Harel, D., Hitchcock, E., & McAllister Byun, T. (2017). Selecting an acoustic correlate for automated measurement of American English rhotic production in children. *International Journal of Speech-Language Pathology, 20*(6), 635-643.
12) Hitchcock, E. R., & McAllister Byun, T. (2015). Enhancing generalisation in biofeedback intervention using the challenge point framework: A case study. *Clinical Linguistics & Phonetics*, *29*(1), 59-75.
13) Preston, J. L., Hitchcock, E. R., & Leece, M. C. (2020). Auditory perception and ultrasound biofeedback treatment outcomes for children with residual/ɹ/distortions: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research*, *63*(2), 444-455.
14) Preston, J. L., & Edwards, M. L. (2007). Phonological processing skills of adolescents with residual speech sound errors. *Language, Speech, and Hearing Services in Schools*, *38*(4), 297-308.
15) Preston, J. L., Hull, M., & Edwards, M. L. (2013). Preschool speech error patterns predict articulation and phonological awareness outcomes in children with histories of speech sound disorders. *American Journal of Speech-Language Pathology, 22*(2), 173-184.
16) Preston, J. L., & Leece, M. C. (2017). Intensive treatment for persisting rhotic distortions: A case series. American Journal of Speech-Language Pathology, 26(4), 1066-1079.
17) McAllister, T., Eads, A., Kabakoff, H., Scott, M., Boyce, S., Whalen, D. H., & Preston, J. L. (2022). Baseline Stimulability Predicts Patterns of Response to Traditional and Ultrasound Biofeedback Treatment for Residual Speech Sound Disorder. *Journal of Speech, Language, and Hearing Research*, 1-21.
18) Peterson, L., Savarese, C., Campbell, T., Ma, Z., Simpson, K. O., & McAllister, T. (2022). Telepractice treatment of residual rhotic errors using app-based biofeedback: A pilot study. *Language, Speech, and Hearing Services in Schools*, *53*(2), 256-274.

19) McAllister Byun, T., Hitchcock, E. R., & Swartz, M. T. (2014). Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2116-2130.
20) McAllister Byun, T., & Hitchcock, E. R. (2012). Investigating the use of traditional and spectral biofeedback approaches to intervention for /r/ misarticulation. *American Journal of Speech-Language Pathology*, *21*(3), 207-221.

In accordance with TalkBank rules, any use of data from this corpus must be accompanied by at least one of the above references. Please cite all papers relevant to the portion of the corpus you used in your own research.

APPENDIX CHAPTER 3-C

Study-Level Summary of PERCEPT-GFTA v2.2.2p

Table 1: Participant and stimulus characteristics for all studies included in corpus. Acronyms: SSD = speech sound disorder, CAS = Childhood Apraxia of Speech, TD = typically developing, GFTA = Goldman Fristoe Test of Articulation, unpb = unpublished, unk = unknown.

Corpus Name	Reference Number	Corpus	N of participants	N Self-Reported Representation of Black/African American, Asian, More than one race, and other People of Color	Age range	GFTA Version
PrestonHullEdwards2013	2	PreKHistorySSD	24	unk	7 - 9	GFTA2
CAS		SuspectedCAS	5	1	7 - 13	GFTA3
CASAJSLPProsody	3	SuspectedCAS	7	2	8 - 16	GFTA2
PrestonBrickLandi2013	4	SuspectedCAS	5	unk	9 - 13	GFTA2
PrestonCASR15	unpb	SuspectedCAS	41	6	9 - 16	GFTA3
PrestonIntensiveCAS2017	5	SuspectedCAS	3	0	10 - 14	GFTA2
BCS	6	SuspectedSSD	9	1	7 - 12	GFTA2
BFS	7	SuspectedSSD	5	0	9 - 15	GFTA2
BFS2	8	SuspectedSSD	15	4	9 - 15	GFTA2
CPF	9	SuspectedSSD	1	0	10 - 10	GFTA2
CRESULTSRCT	10	SuspectedSSD	54	7	9 - 16	GFTA3
CRESULTSSCED	11	SuspectedSSD	1	0	11 - 11	GFTA3
EFIF	12	SuspectedSSD	6	1	6 - 13	GFTA2
PerceptionRCT	13	SuspectedSSD	37	7	7 - 15	GFTA3
PrestonIntensiveRSSD2017	14	SuspectedSSD	3	0	13 - 21	GFTA2
PTR	15	SuspectedSSD	19	1	9 - 14	GFTA2
Sjolie2017JCD	16	SuspectedSSD	1	0	7 - 7	GFTA2

UnpublishedIntensives		SuspectedSSD	2	0	12 - 24	GFTA3
US2014	17	SuspectedSSD	1	0	7 - 7	GFTA2
PrestonERP	unpb	TDChildrenandAdults	7	0	9 - 21	GFTA2
CRESULTSMOSAIC	18	TDChildrenandAdults	104	22	9 - 15	GFTA3

Appendix 3-C Citation information

1) Benway, N. R., Preston, J. L., Hitchcock, E. R., Salekin, A., Sharma, H., & McAllister, T. (2022). PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of /ɹ/. INTERSPEECH 2022: Proceedings of the 23rd Annual Conference of the International Speech Communication Association (ISCA), Incheon, Republic of Korea.
2) Preston, J. L., Hull, M., & Edwards, M. L. (2013). Preschool speech error patterns predict articulation and phonological awareness outcomes in children with histories of speech sound disorders. *American Journal of Speech-Language Pathology, 22*(2), 173-184.
3) Preston, J. L., Leece, M. C., McNamara, K., & Maas, E. (2017). Variable practice to enhance speech learning in ultrasound biofeedback treatment for childhood apraxia of speech: A single case experimental study. *American Journal of Speech-Language Pathology, 26*(3), 840-852.
4) Preston, J. L., Brick, N., & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood Apraxia of speech. *American Journal of Speech-Language Pathology*, *22*(4), 627-644.
5) Preston, J. L., Leece, M. C., & Maas, E. (2016). Intensive treatment with ultrasound visual feedback for speech sound errors in childhood apraxia. *Frontiers in Human Neuroscience, 440*. DOI: https://doi.org/10.3389/fnhum.2016.00440
6) McAllister Byun, T., Hitchcock, E. R., & Ferron, J. (2017). Masked visual analysis: Minimizing type I error in visually guided single-case design for communication disorders. *Journal of Speech, Language, and Hearing Research*, *60*(6), 1455-1466.
7) McAllister Byun, T., & Campbell, H. (2016). Differential effects of visual-acoustic biofeedback intervention for residual speech errors. *Frontiers in Human Neuroscience*, *10*, 567.
8) McAllister Byun, T. (2017). Efficacy of visual–acoustic biofeedback intervention for residual rhotic errors: A single-subject randomization study. *Journal of Speech, Language, and Hearing Research*, *60*(5), 1175-1193.
9) McAllister, T., Hitchcock, E. R., & Ortiz, J. A. (2021). Computer-assisted challenge point intervention for residual speech errors. *Perspectives of the ASHA special interest groups*, *6*(1), 214-229.
10) McAllister, T., Preston, J. L., Hitchcock, E. R., & Hill, J. (2020). Protocol for correcting residual errors with spectral, ultrasound, traditional speech therapy randomized controlled trial (C-RESULTS RCT). *BMC Pediatrics*, *20*(1), 1-14.
11) Benway, N. R., Hitchcock, E. R., McAllister, T., Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual/ɹ/errors in American English: A single-case randomization design. *American Journal of Speech-Language Pathology*, *30*(4), 1819-1845.

12) McAllister Byun, T., Swartz, M. T., Halpin, P. F., Szeredi, D., & Maas, E. (2016). Direction of attentional focus in biofeedback treatment for /r/ misarticulation. *International Journal of Language & Communication Disorders*, *51*(4), 384-401.
13) Preston, J. L., Hitchcock, E. R., & Leece, M. C. (2020). Auditory perception and ultrasound biofeedback treatment outcomes for children with residual/ɹ/distortions: A randomized controlled trial. *Journal of Speech, Language, and Hearing Research*, *63*(2), 444-455.
14) Preston, J. L., & Leece, M. C. (2017). Intensive Treatment for Persisting Rhotic Distortions: A Case Series. *American Journal of Speech-Language Pathology, 26*(4), 1066-1079.
15) McAllister, T., Eads, A., Kabakoff, H., Scott, M., Boyce, S., Whalen, D. H., & Preston, J. L. (2022). Baseline Stimulability Predicts Patterns of Response to Traditional and Ultrasound Biofeedback Treatment for Residual Speech Sound Disorder. *Journal of Speech, Language, and Hearing Research*, 1-21.
16) Sjolie, G. M., Leece, M. C., & Preston, J. L. (2016). Acquisition, retention, and generalization of rhotics with and without ultrasound visual feedback. *Journal of Communication Disorders*, *64*, 62-77.
17) McAllister Byun, T., Hitchcock, E. R., & Swartz, M. T. (2014). Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2116-2130.
18) Ayala, S.A., Eads, A., Kabakoff, H., Swartz, M., Shiller, D.M., Hill, J., Hitchcock, E.R., Preston, J.L., & McAllister, T. (Under review). Auditory and Somatosensory Development for Speech in Later Childhood. DOI 10.17605/OSF.IO/BKASM

In accordance with TalkBank rules, any use of data from this corpus must be accompanied by at least one of the above references. Please cite all papers relevant to the portion of the corpus you used in your own research.

APPENDIX CHAPTER 3-D

Datasheet for PERCEPT Corpora

Adapted from Gebru et al. (2018)

Motivation

For what purpose was the dataset created?

The PERCEPT corpora were created during the collection of research data regarding the speech of children with speech sound disorders and age-matched, typically-speaking peers. The corpora were standardized for the purpose of training audio classifiers for the automatic classification of fully rhotic versus derhotic instances of the speech sound /ɹ/ as spoken by children, adolescents, and young adults with residual speech sound disorder (RSSD).

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was assembled by the CRESULTS (ClinicalTrials.gov ID: NCT03737318) study team at Montclair State University, New York University, and Syracuse University.

Who funded the creation of the dataset?

Funding for the compilation of the PERCEPT corpora has been provided by the National Institute on Deafness and Other Communication Disorders (NIH R01DC017476-S1, T. McAllister, PI). This research was supported in part through computational resources provided by Syracuse University (NSF ACI-1341006; NSF ACI-1541396).

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Corpus instances are spoken audio utterances containing the speech sound /ɹ/, as well as ground-truth labels (perceptual ratings of the accuracy of the /ɹ/ in each utterance). Utterances range in linguistic complexity from syllables to multi-word phrases.

How many instances are there in total?

There are 125,632 utterances in PERCEPT 2.2.2p.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The 2.2.2p public release of the PERCEPT corpora include only those participants providing consent/assent for future use of study audio. Consent/assent was ascertained by manual review of consent and assent documents on file.

What data does each instance consist of?

Each instance is raw 16-bit PCM lossless audio stored in a WAV container along with a metadata XML file, designed to be read by the Phon Database (phon.ca)

Is there a label or target associated with each instance?

Each instance is accompanied by metadata and a ground-truth class label. Metadata includes a filename, orthographic transcript of the audio, study timepoint of data

collection, participant identifier, participant age, participant sex, and originating study. Ground-truth class labels concern the perceptual rating of /ɹ/ from the utterance and include number of listeners, number of listeners who rated the /ɹ/ as correct, and average listener rating.

Is any information missing from individual instances?

No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Yes. Each instance from the same participant is labeled with the same participant ID, across both corpora. Likewise, instances from the same time point are marked as such.

Are there recommended data splits (e.g., training, development/validation, testing)?

Recommended data splits will be included with future publications demonstrating empirical utility of the corpus for audio classification.

Are there any errors, sources of noise, or redundancies in the dataset?

None known.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset and any derived versions of the dataset are archived at PhonBank.
DOI for PERCEPT-R: 10.21415/0JPJ-X403
DOI for PERCEPT-GFTA: 10.21415/1H2C-8G56

The dataset can be best accessed through Phon: phon.ca

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

The transcripts of the audio do not contain confidential information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

The dataset contains no offensive content.

Does the dataset identify any subpopulations (e.g., by age, gender)?

PERCEPT-R v2.2.2p utterances were collected from 280 child, adolescent, and young adult speakers of American English, aged 6;0 – 24;0 (years; months, $\bar{x}$=11;4, σ_x=2;6); 128 participants are females (ages for females: $\bar{x}$=11;8, $\sigma_{\bar{x}}$=2;5, min = 6;1, max = 17;3; ages for 152 males: $\bar{x}$=11;1, σ_x=2;7, min = 6;0, max = 24;0). Of the 280 participants, 95 were recruited to studies of typically-developing speakers , 22 were recruited to studies based on a history of preschool speech sound disorder (SSD), and the remaining 163 participants were recruited to studies of individuals with RSSD. Thirty-three corpus speakers (12%) self-identified as Black/African American, Asian, More than one race, or Other according to the NIH race reporting framework (see also: Appendix B).

PERCEPT-GFTA v 2.2.p were collected from 350 participants between the ages of 6;10 – 24;0 ($\bar{x}$=11;4, σ_x=2;6). Note that

many participants in this corpus overlap with the speakers of PERCEPT-R, and the speakers that do overlap have the same identification number across both studies. Of the 350 PERCEPT-GFTA participants, 147 are females (ages for females: $\bar{x}$=11;11, σ_x=2;7, min = 7;3, max = 21;6; ages for 203 males: $\bar{x}$=10;11, σ_x=2;4, min = 6;10, max = 24;0). Twenty-four of the participants in PERCEPT-GFTA were recruited based on a history of preschool SSD, 61 were recruited for CAS, 155 were recruited for RSSD, and 111 were recruited for studies of typical speech development. Fifty-two corpus speakers (15%) self-identified as Black/African American, Asian, More than one race, or Other according to the NIH race reporting framework (see also: Appendix C).

The imbalance between males and females reflects the increased prevalence of RSSD observed among males (Wren et al., 2016).

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

It is unlikely, but theoretically possible, that participants can be recognized by familiar listeners through the sound of their voice and the metadata indicating speaker age and sex.

Does the dataset contain data that might be considered sensitive in any way?

The raw data contains audio of participant voices repeating words and short phrases with no other context.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)?

Audio instances were recorded directly from the participant by researchers, following the procedures of the originating study and manually reviewed for transcript accuracy at the time of audio processing in the original clinical trials.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

Specific details of audio capture differ by originating study. Generally, audio was captured by microphone (lapel or headset) within a quiet research location. Some studies with data collection circa 2010 operationalized audio capture at a lower sampling rate which was then upsampled for the purposes of corpus curation (e.g., McAllister Byun, T., & Hitchcock, E. R., 2012; Preston & Edwards, 2007).

Were any ethical review processes conducted (e.g., by an institutional review board)?

Each originating study was approved by either the Institutional Review Boards of New York University, Syracuse University, Montclair State University, the Biomedical Research Alliance of New York, or a combination of these boards.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Data was directly obtained from the participants by researchers associated with our laboratories.

Were the individuals in question notified about the data collection?

All participants were informed of data collection. Informed written consent was obtained from participants 18 and older, and from the parent/guardian of participants under 18. Informed written assent was obtained from participants under 18.

Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Audio was standardized to a common number of channels, sampling rate, and intensity level. No data were imputed.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Yes. Raw data were saved on a private, password protected virtual machine at Syracuse University.

Is the software used to preprocess/clean/label the instances available?

Yes: Python, Praat, and Phon.

Uses

Has the dataset been used for any tasks already?

The instances in the corpus have been used for primary data analysis as part of the originating studies. The curated data is currently being used for the development of the PERCEPT-R Audio Classification Engine and a clinical trial in which learners

Is there a repository that links to any or all papers or systems that use the dataset?

The Open Science Framework page for PERCEPT is: https://osf.io/nqzd9/.

Are there tasks for which the dataset should not be used?

The dataset should not be used to train automated clinical speech-language assessment or treatment tools unless the design and use of those tools is appropriately supervised by a certified speech-language pathologist.

Maintenance

Who is supporting/hosting/maintaining the dataset?

The corpora are hosted at PhonBank.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Nina R Benway MS CCC-SLP
(nrbenawy@syr.edu)
Slack support for PERCEPT:
tinyurl.com/2tnm2vkw

Is there an erratum?

Not at this time. Any errata will be posted on PhonBank and OSF.

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?

Yes.

If it relates to other ethically protected subjects, have appropriate obligations been met?

The corpus contains information from minors. All data collection materials were approved by an Institutional Review Board for study procedures involving minors.

If it relates to people, were there any ethical review applications/reviews/approvals?

Yes.

If it relates to people, could this dataset expose people to harm or legal action?

No.

If it relates to people, does it unfairly advantage or disadvantage a particular social group?

The individuals presenting for the originating studies overwhelmingly self-selected white, high socioeconomic status families. The speech patterns of individuals from other communities may not be reflected in the corpus. Our ongoing projects aim to expand representation to other linguistic communities.

VITA

Nina R Benway

Education

Doctor of Philosophy Candidate, Syracuse University — Anticipated 2023
Speech-Language Pathology
Interdisciplinary Graduate Neuroscience Concentration
Advisor: Dr. Jonathan L Preston
Dissertation: **Artificial Intelligence Assisted Motor Learning for Speech Sound Disorders Impacting /ɹ/: The PERCEPT Project**

Advanced Certificate, The State University of New York at Buffalo — 2013
Gifted and Talented Education

Master of Science in Education, The College of Saint Rose — 2011
Communication Sciences and Disorders

Bachelor of Arts, Cornell University — 2008
Linguistics

Employment

Pre-Doctoral Researcher, Syracuse University, Syracuse, New York	2018 – Present
Clinical Supervisor, The College of Saint Rose, Albany, New York	2018 – 2019
Visiting Faculty, The College of Saint Rose, Albany, New York	2017 – 2018
Adjunct Instructor, The College of Saint Rose, Albany, New York	2014 – 2017
Director of IGNITE (Student Support), Brown School, Schenectady, New York	2012 – 2017

Professional Licensure

American Speech-Language Hearing Association Cert. of Clinical Competence	2012-Present
New York State Licensure in Speech-Language Pathology	2012-Present

Research Interests

speech signal processing, speech motor learning, clinical trials, mispronunciation detection, machine learning, clinical speech technology

Research Support

Intensive Speech Motor Chaining Treatment and Artificial Intelligence Integration for Residual Speech Sound Disorders
Dates: 03/01/2023 – 02/29/2028
Grant: NIH 1R01DC020959-01; Percentile 9 Impact Score 23
Role: Co-Investigator; co-author of proposal
PI: J. Preston

Direct Costs: $1,547,791
The proposed research addresses the critical public health need for sufficiently intense speech treatments, without which ~6 million Americans adults live with unresolved speech sound disorders. The project's investigation of intensive motor-based treatment bootcamps and artificial intelligence-driven treatment potentiates an evidence-based paradigm shift in which combined speech-language pathologist/artificial intelligence service delivery overcomes existing access barriers to sufficiently intense treatment. The project is relevant to the NIDCD's and NIH's missions of conducting behavioral research to reduce disability, as those with unresolved speech sound disorders may experience reduced educational, social, emotional, and occupational outcomes throughout the lifespan.

Administrative Supplement to Support Collaborations to Improve the AI/ML-Readiness of NIH-Supported Data under Parent Award *Biofeedback-Enhanced Treatment for Speech Sound Disorder: Randomized Controlled Trial and Delineation of Sensorimotor Subtypes*.

Dates: 08/15/2021 – 08/14/2022
Grant: NIH 3R01DC017476-03S2
Role: Named Senior/Key Personnel on the Administrative Supplement; intellectual contributions to proposal
PI: T. McAllister
Direct Costs: $219,461, Administrative Supplement
Speech sound disorder in childhood poses a barrier to academic and social participation, with potentially lifelong consequences for educational and occupational outcomes. The parent R01 award aims to meet a public health need by conducting the first randomized controlled trial comparing the efficacy and efficiency of speech intervention with and without real-time visual biofeedback. The administrative supplement lays the groundwork for the development of automated speech recognition tools for speech sound disorder by modifying and augmenting an existing corpus of acoustic recordings of child speech in preparation for sharing with researchers in artificial intelligence and machine learning (AI/ML).

Innovative & Interdisciplinary Research Grant *Developing a Clinical Speech Recognition System for Childhood Speech Disorders*

Dates: 06/01/2021 – 05/31/2023
Grant: Syracuse University CUSE Grant II-14-2021
Role: Named Key Personnel; intellectual contributions to proposal
PI: J. L. Preston
Direct Costs: $20,000
This project optimizes a near real-time speech analysis algorithm (PERCEPT) to classify correct and incorrect /r/ sounds, then integrate this algorithm into existing, validated, computerized motor-based intervention software developed at Syracuse University (Speech Motor Chaining).

Patents

Benway, N.R., Preston., J. L., Salekin, A. "Clinical Speech Analysis System for Childhood Speech Disorders" (pending). US Patent Application No. 63,450,762. Filed March 8, 2023.

Publications

Coretta, S., Dokovova, M., Mridhula, M., and the **Open Science Framework Research Group**. (in press). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human speech analyses. Advances in Methods and Practices in Psychological Science.

Benway, N. R., Preston, J. L., Hitchcock, E. R., Rose, Y., Salekin, A., Liang, W., & McAllister, T. (in press). Reproducible Speech Research with the Artificial-Intelligence-Ready PERCEPT Corpora. Journal of Speech, Language, and Hearing Research.

Benway, N.R., Preston, J.L., Hitchcock, E, Salekin, A. Sharma, H., & McAllister, T. (2022). PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of /ɹ/. INTERSPEECH 2022: Proceedings of the 23rd Annual Conference of the International Speech Communication Association (ISCA), Incheon, Republic of Korea.

Benway, N.R., Hitchcock, E, McAllister, T, Feeny, G. T., Hill, J., & Preston, J. L. (2021). Comparing biofeedback types for children with residual /ɹ/ errors in American English: A single case randomization design. *American Journal of Speech-Language Pathology*. 30(4), 1819 – 1845.

Preston, J.L., **Benway, N.R.**, Leece, M.C., & Caballero, N.F. (2021). Concurrent validity between two sound sequencing tasks used to identify childhood apraxia of speech in school-age children. *American Journal of Speech-Language Pathology*. 18;30(3S), 1580-1588.

Benway, N.R., Garcia, K., Hitchcock, E., McAllister, T., Leece, M., Wang, Q., & Preston, J. L. (2021). Associations Between Speech Perception, Vocabulary, and Phonological Awareness Skill in School-Aged Children with Speech Sound Disorders. *Journal of Speech, Language, and Hearing Research*. 64(2), 452-463.

Benway, N. R., & Preston, J. L. (2020). Differences Between School-Aged Children with Apraxia of Speech and Other Speech Sound Disorders on Multisyllable Repetition. *Perspectives of the ASHA Special Interest Groups: Advances in Research and Clinical Management of CAS Forum*. 5(4), 794-808.

Preston, J. L., **Benway, N. R.**, Leece, M., Hitchcock, E., & McAllister, T. (2020). Tutorial: Motor-based Treatment Strategies for /r/ Distortions. *Language, Speech, and Hearing Services in Schools*. 54(966-980).

Papers Under Review

Benway, N. R., Preston, J. L., Salekin, A., & McAllister, T. (under review). Automated detection of rhoticity of American English /ɹ/ in children with residual speech sound disorders: The PERCEPT-R Classifier

Benway, N. R.*, Siriwardena, Y. M.*, Preston, J. L., Hitchcock, E., McAllister, T., & Espy-Wilson, C. (under review). Acoustic-to-Articulatory Speech Inversion Features for Mispronunciation Detection in Child Speech Sound Disorders. * denotes equal contributions

Benway, N. R., Salekin, A., Xiao, Y., Sharma, H., & Preston, J. L. (under review). Classifying Rhoticity of /ɹ/ in Speech Sound Disorder using Age-and-Sex Normalized Formants: PERCEPT-R Classifier.

Benway, N. R., & Preston, J. L. (under review). Clinical Efficacy and End User Acceptability of Motor-Based Intervention with Automated Mispronunciation Detection of Rhotics in Residual Speech Sound Disorders.

Preston, J.L., Leece, M.C., Cabellero, N.F., Wang, D.L., Herbst, B.H., and **Benway, N.R**. (under review). A Randomized Controlled Trial of Treatment Distribution and Biofeedback Effects on Speech Production in School-Aged Children with Apraxia of Speech

Ad Hoc Peer Reviews

American Journal of Speech-Language Pathology
Clinical Linguistics and Phonetics
International Journal of Language and Communication Disorders
International Journal of Speech-Language Pathology
Perspectives of the ASHA Special Interest Groups

Posters and Presentations

Benway, N. R., Preston, J. L., Hitchcock, E., Salekin, A., Sharma, H., Rose, Y., McAllister, T. (2022, November). 4560: PERCEPT-R and PERCEPT-GFTA: Two Artificial Intelligence-Ready Speech Corpora Focusing on Residual Speech Sound Disorders. Annual Convention of the American Speech-Language-Hearing Association, New Orleans, L.A.

Wruck, C., **Benway, N. R.**, Preston., J., & Munson, B. (2022, November). Social Biases in the Assessment of /s/ Accuracy. Annual Convention of the American Speech-Language-Hearing Association, New Orleans, L.A.

Benway, N. R., McAllister, T., Hitchcock, E., & Preston, J. L. (2021, November). 10325: Single Case Experimental Comparison of Biofeedback Types for Children with Residual Speech Sound Errors Impacting /ɹ/. Annual Convention of the American Speech-Language-Hearing Association, Washington D.C.

Benway, N. R. Owens, R., & Pavelko, S. (2021, November). 8521V: Introducing SPOON: Automated SUGAR Language Sample Analysis. Annual Convention of the American Speech-Language-Hearing Association, Washington D.C.

Caballero, N., **Benway, N. R.**, Leece, M., McNamara, K., Preston, J. L. (2021, November). Massed Versus Distributed Treatment via Teletherapy for School-age Children with Childhood Apraxia of Speech. Annual Convention of the American Speech-Language-Hearing Association, Washington D.C.

McAllister, T., Hitchcock, E.R., **Benway, N.R.**, Ochs, L., Leece, M., Preston, J. L. (2021, November). Telepractice Delivery of Traditional and Visual-Acoustic Biofeedback Treatment for Residual Speech

Errors Affecting /r/. Annual Convention of the American Speech-Language-Hearing Association, Washington D.C.

Benway, N. R., McAllister, T., Hitchcock, E., & Preston, J. (2020, December). Comparing Biofeedback Types for Children with Residual Speech Production Errors on /ɹ/. 12th International Seminar on Speech Production (ISSP 2020), Providence, RI (virtual conference).

Leece, M., **Benway, N. R.**, McAllister, T., & Preston, J. (2020, November). Try it for six sessions: Using machine learning to predict outcomes of SSD treatment. Proposal accepted at the Annual Convention of the American Speech-Language-Hearing Association, San Diego, CA (Convention cancelled).

Caballero, N., Leece, M., **Benway, N. R.**, & Preston, J. (2020, November). 10972: Measuring Features of CAS in School-Age Children: A Reliability Study. Proposal accepted at the Annual Convention of the American Speech-Language-Hearing Association, San Diego, CA (Convention cancelled).

Benway, N. R. & Preston, J. (2020, July). Clustering of CAS and Non-CAS SSD using Multiple Assessment Tools. Proposal accepted at the Research Symposium of Apraxia Kids, Plano, TX (Convention cancelled).

Benway, N. R., Campeas, R., & Preston, J. (2019, November). *Predicting CAS Diagnosis from Multisyllabic Word Repetitions in School-Age Children.* Poster session presented at the meeting of the Convention of the American Speech-Language-Hearing Association, Orlando, Florida.

Benway, N., Campeas, R., & Preston, J. (2019, May). *Differentiating Subtypes of Speech Sound Disorders Using Multisyllabic Word Repetitions*. Presentation at the International Child Phonology Conference, Montreal, Quebec, Canada.

Muldoon, D., **Benway, N.**, Shanock, A., & Alfonso, V. (2018, November). *A Review of the Psychometric Integrity of Preschool Language Tests: Findings & Implications for SLPs.* Presentation at the Convention of the American Speech-Language-Hearing Association, Boston, Massachusetts.

Alfonso, V., Shanock, A., Muldoon, D., **Benway, N.**, & Oades-Sese, G. (2018, May). *Psychometric Integrity of Preschool Speech/Language Tests: Implications for Diagnosis and Progress Monitoring of Treatment*. Poster session presented at the meeting of the Association for Psychological Science, San Francisco, California.

Benway, N. (2013, November). *Gifted with Special Needs: Identification, Differentiation, and Advocacy for Twice-Exceptional Children. Presentation* #1624 *at the Convention of the American Speech-Language-Hearing Association, Chicago, Illinois.*

Benway, N. (2012, April). *Turning Speech into Visual Feedback*. Presentation at the Convention of the NYS Speech-Language Hearing Association, Saratoga Springs, New York.

Benway, N. (2011, November) *Acoustic and Perceptual Changes in a Male to Female Transgender Individual. Poster session #8903 presented at the* Convention of the American Speech-Language-Hearing Association, San Diego, California.

Hertz, S., Gibson, M., **Glatthorn, N**., Hegde, P., Mills, H., and Spencer, I. (2008, April). The Role of Prosody in Speech Parsing. Poster Presentation at Experimental and Theoretical Advances in Prosody, Ithaca, New York.

Honors and Awards

All-University Doctoral Prize. Syracuse University. 2023

Research Excellence Doctoral Funding Fellowship. Syracuse University. 2022-2023

Graduate Dean's Award for Excellence in Research and Creative Work. Syracuse U. 2022

Doctoral Fellowship. Syracuse University. 2019-2020, 2021-2022

New Century Doctoral Scholarship. American Speech Language Hearing Foundation. 2021

Graduate Student Scholarship. American Speech Language Hearing Foundation. 2020

Meritous Poster: Measuring Features of CAS in School-Age Children: A Reliability Study. Annual Convention of the American Speech-Language-Hearing Association, San Diego, CA (Convention cancelled). 2020

Dr. Kathy Yorkston Student Travel Award. Motor Speech Conference. 2020

Dr. Rosemary S. Callard-Szulgit Gifted Education Student Award. University at Buffalo 2018

Alumna of Excellence. Communication Sci and Disorders, The College of Saint Rose 2017

Teaching Experience (Instructor of Record)

CSD 594: Clinical Speech Sound Disorders. The College of Saint Rose. 2018 – 2019

CSD 109: Phonetics. The College of Saint Rose. 2014 – 2019

CSD 316/616: Introduction to Applied Phonetics. Syracuse University 2018

CSD 219: Speech Sound Disorders. The College of Saint Rose. 2017 – 2018

CSD 472/548: Augmentative and Alt. Communication. The College of Saint Rose. 2014

Institutional Service

Student Representative, Meredith Professorship Review Committee. Syracuse Uni. 2021-2022

Thesis Reader, The Renée Crown University Honors Program. Syracuse University. 2021-2022

Member, CSD Curriculum Committee. The College of Saint Rose. 2017-2018

Clinical Supervision

CSD 589: Specialty Practicum Supervisor. Articulation Camp. The College of Saint Rose 2019

Clinical Fellowship Supervisor, Schenectady, New York. 2017– 2018

CSD 580: First Practicum. Evaluation Team Supervisor. The College of Saint Rose 2017

CSD 587: School Practicum. Field Supervisor. The College of Saint Rose 2014 – 2017

Professional Development Sessions

Invited Panelist, Speech Science Technical Committee, ASHA Convention 2023 Forthcoming
Ethical AI in Clinical Practice

Invited Speaker, Grand Rounds, The College of Saint Rose. Albany, NY Forthcoming
AI-Assisted Speech Therapy

Invited Speaker, Department of Linguistics, University of Potsdam, Germany Forthcoming
Half-day workshop: *Forced Phonetic Alignment with the Montreal Forced Aligner*

Invited Speaker, Supervisor's Workshop, The College of Saint Rose 2022
Two-hour workshop: *Motor Learning with Speech Motor Chaining*

Invited Co-Speaker, NSSLHA Professional Development at The College of Saint Rose 2017
Two-hour seminar: *Accents and Dialects: From Phonetics to Assessment.*

Invited Speaker, Sister Charlene Bloom Mini-Convention at The College of Saint Rose. 2015
One-hour seminar: *Supporting Students with Executive Functioning Deficits.*

Invited Speaker, Professional Development. The Brown School. Schenectady, NY 2015
One-hour seminar: *Hearing Across the Lifespan.*

Invited Speaker, Brain Burst Community Forum. Schenectady, NY 2014
Ten-minute talk: *Giftedness in School Age Children.*

Invited Speaker, Schenectady City School District. Schenectady, NY 2014
Half-day workshop: *Assessing Metacognition through Language and Behavior Ratings.*

Research Experience

Graduate Researcher, Syracuse Speech Production Lab, Syracuse University 2018 – Present

Graduate Researcher, Transgender Voice Modification, The College of Saint Rose 2010 – 2011

Graduate Researcher, Childhood Apraxia of Speech, The College of Saint Rose 2009

Research Assistant, NovaSpeech LLC, Ithaca, NY 2007 – 2009

Founding Organizer, Cornell Undergraduate Linguistics Colloquium. Ithaca, NY 2006 – 2008

Undergraduate Research Assistant, Department of Linguistics, Cornell University 2006 – 2007

Selective Leadership Development Coursework

Selected Participant, SU Women in Science and Engineering 2020 – 2022

Selected Participant, ASHA Leadership Development Program: Schools Cohort. 2015 – 2016

Selected Clinical Skills

Assistive Technology Optimization
Autism Spectrum Disorder
Childhood Speech Sound Disorders
Childhood Language Disorders
Comorbid Other Health Interaction (Anxiety, ADHD, Epilepsy, OCD)
Childhood Fluency Disorders
Childhood Sensorineural Hearing Loss
Childhood Swallowing Disorders
Giftedness and Twice-Exceptionality
Learning Disabilities: Reading, Writing, Math
Social Communication/Metacognition

Technical Skills

Clinical: Speech Motor Chaining, Ultrasound Biofeedback, Visual-Acoustic Biofeedback
Software: Praat, Phon, Microsoft Office Suite, Blackboard LMS, Canvas LMS, REDCap.
Scripting: Praat, Regular Expressions, Natural Language Toolkit, SPSS, R, ggplot, Python, Optuna, Pytorch, Scikit-learn, SAS, Jupyter Notebook, FFMPEG, Montreal Forced Aligner, Windows, Windows Subsystem for Linux, Debian/GNU Linux, HTCondor, Google Cloud, VertexAI, Google Cloud Speech to Text, Google Cloud Text to Speech, Speech Synthesis Markup Language