

Supplemental Material S1. Supplemental methods.

Participants and Operationalizing Stimulability/Clinical Indication for AI-Assisted Treatment

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 for a hybrid in-person/telepractice study who could produce /ɹ/ in some syllables/words but not others.

To pass study pre-screening, interested families reported that candidate participants had difficulty producing the American English rhotic /ɹ/ and were within the study's age range as of the date of consent. The study welcomed children and young adults aged 9; 0–20; 11 to reflect the range of ages represented during PERCEPT-R Classifier development. Pre-screening exclusionary criteria emphasized characteristics that might confound therapeutic response, such as a neurodevelopmental disorder (e.g., Autism spectrum disorder, Oppositional Defiant Disorder), permanent hearing loss, and American English not being among the child's dominant language(s) learned before the age of 3 (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 also 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. A total of 21 families responded to the pre-screening survey. Two families became ineligible after self-reporting a history of neurodevelopmental diagnosis and one family became ineligible after self-reporting orthodontia blocking the roof of the mouth. Of the remaining 18 families, 15 elected to schedule consent/assent video conferences.

Baseline stimulability for /ɹ/ and clinical indication for AI-assisted treatment were the most important eligibility requirements. These factors were evaluated at the consent/assent video conference before a dedicated eligibility session was scheduled. Stimulability was operationalized as > 20% baseline accuracy on syllable repetition lists (either prevocalic /ɹ/ subsets, postvocalic /ɹ/ subsets, or both). Nine participants did not qualify to schedule the remaining eligibility session because they were below the floor criterion for syllable stimulability. Clinical indication of Speech Motor Chaining was operationalized as < 40% accuracy for /ɹ/ on monosyllable/multisyllabic word reading lists. All participants were initially assessed using the same 100-item evaluation word list, which was balanced with regard to syllable count, position of /ɹ/ in word, and frontness/backness of the adjacent vowel. Three participants were eligible for the study based on this wordlist. Note, however, that if a participant's average 100-word accuracy was higher than the 40% inclusionary ceiling, 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 under the 40% accuracy ceiling. In these cases, a second, custom, 100-item wordlist was made (e.g., focusing on 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 complex words and one participant remaining above the threshold for study eligibility. The treatment targets and outcome measures selected for all eligible participants were customized to reflect each participant's baseline stimulability, and are described in other sections that follow.

All word/syllable list accuracy for eligibility tasks was rated by the first author and rerated by a research clinician, Megan Leece. There were no disagreements regarding inclusion/exclusion of participants based on these criteria.

The five stimuable participants meeting the definition for clinical indication of AI-assisted treatment completed a full eligibility evaluation to examine characteristics relative to inclusionary/exclusionary speech-language criteria in finer 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 in-person study data were collected with a shock-mounted Sennheiser MKE 600 super-cardioid microphone and Focusrite Scarlet audio interface. Each participant 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 tip 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 stimuable 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 stimuable 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 Linguistics Articulation Test-Normative Update Apraxia Screening (Bowers & Huisinigh, 2018) along with fewer than four transcoding errors on the Syllable Repetition Task (Shriberg et al., 2009). Descriptive information was also collected from participant families regarding speech sound disorder and previous intervention history, but this information was not exclusionary. All five participants remained eligible for the treatment phase of the study. These participants were randomized to intervention start points, from among the predetermined number of possible baseline visits (5–10).

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 (4 male and 1 female) were 10-19 years (\bar{x} = 12.7, σ_x = 3.6), and 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

Probe stimuli were selected from a custom list of 2,361 single /ɹ/ words with rhotic phonemes derived from the LIBRISPEECH speech recognition dictionary and Phon (Hedlund & Rose, 2019). From this custom list, subsets were 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 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.

Personalization of PERCEPT 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 ($\bar{x} = 497.2$, $\sigma_{\bar{x}} = 242$, $\min = 229$, $\max = 888$). 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 et al., 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, as done in Benway et al. (2021). 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 et al. (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 S-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-score¹, a common summary of the 2x2 contingency table (i.e., confusion matrix) that reflects the harmonic mean of precision and recall.

¹ The abbreviation "F1" is used differently in machine learning literature and speech science literature. This paper uses "F1-score" at every relevant instance to distinguish the common machine learning performance metric, the harmonic mean of classifier precision and recall, from "F1," the first formant.

Table S-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.

Figure S-1: Adaptive Speech Motor Chaining Algorithm for Structured Chaining

Figure: Adaptive Speech Motor Chaining algorithm for Structured Chaining. Productions from lower levels of linguistic complexity receive relatively more feedback than productions from higher levels. The relative frequency of approximated knowledge of performance feedback reduces in higher levels of linguistic complexity as well. Practice is blocked so each level of linguistic complexity is practiced multiple times before moving to the next block (e.g., [row, row, row, row], [rodeo, rodeo, rodeo, rodeo], [rodeo clown, ...], etc).

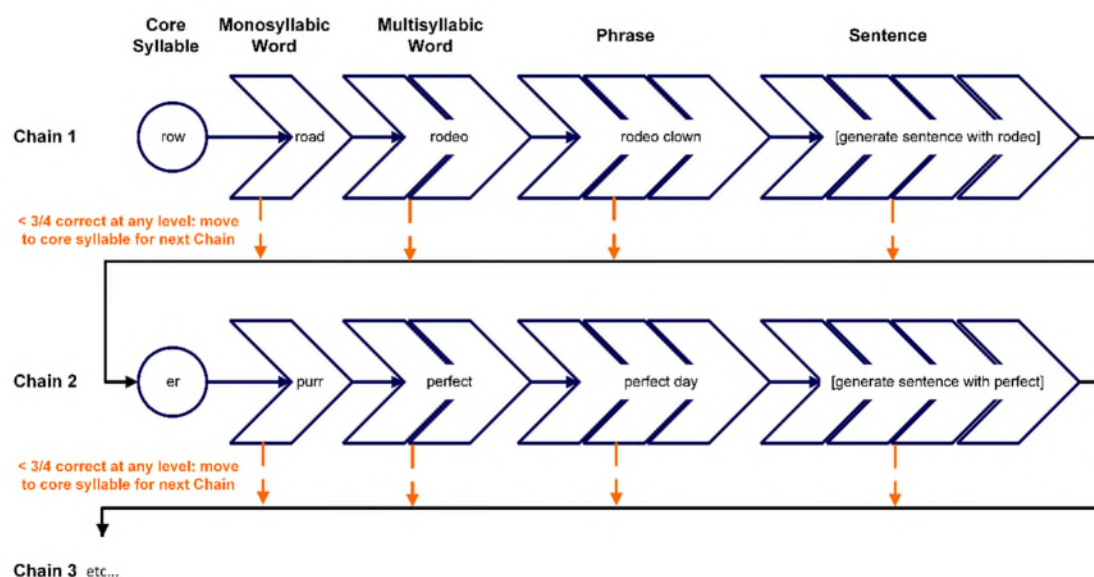
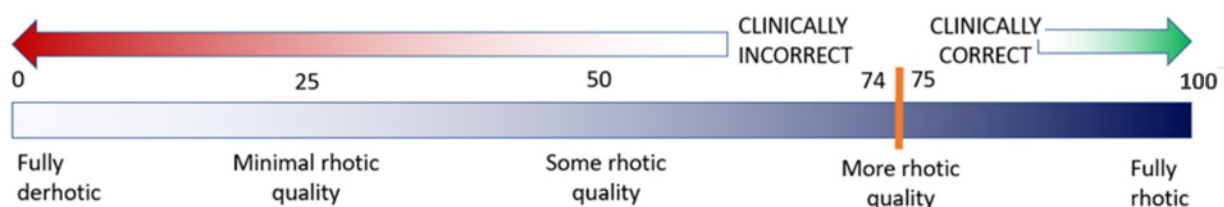


Figure S-2: Outcome Rating Scale Schematic



Methodological Reporting – 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	<ul style="list-style-type: none"> • 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	<p>Participant computer with internet connection</p> <p>Researcher computer</p> <p>Speech Motor Chaining Web App</p> <p>Participant Smartphone</p> <p>Shure MV5 cardioid digital condenser mic (20 Hz to 20 kHz)</p> <p>Sennheiser MKE600 super-cardioid digital condenser mic (40 Hz to 20 kHz)</p>
Intervention	<p>Artificial intelligence driven Speech Motor Chaining web app (Chaining-AI)</p> <ul style="list-style-type: none"> • Prepractice: < 10 minutes or 16 correct productions • Block size: 4

	<ul style="list-style-type: none"> • Number of chains: 4 • Targets per chain: 2 • Block accuracy criterion: 3/4 • Random practice: 5 minutes
Procedural Fidelity	<p>ChainingAI is inherently high-fidelity with regard to the therapeutic parameters specified above.</p> <p>Fidelity also evaluated for participant interaction with ChainingAI</p> <ul style="list-style-type: none"> • Frequency of redirection • Frequency of technical support • Total prepractice productions • Total ChainingAI productions • Average minutes: seconds spent in practice • ChainingAI productions per minute
Analyses	<ul style="list-style-type: none"> • 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.

Feedback delay due to PERCEPT processing time

We explored the amount of time required for PERCEPT to process predictions submitted to the PERCEPT server through ChainingAI. The standard Python package “time” was used to timestamp the moment that ChainingAI made a request to PERCEPT and the moment that PERCEPT handed the prediction back to ChainingAI, and these timestamps were printed to the PERCEPT logs for each file processed during the study. This time, however, does not account for the time it would take to upload the audio file from the participant’s browser to the server that hosts PERCEPT and ChainingAI. To calculate the round trip time from the participant computer to the PERCEPT server and back, we estimated the transfer time for the average-sized file at a range of internet speeds using an established, freely available tool (<https://www.meridianoutpost.com/resources/etools/calculators/calculator-file-download-time.php>). To provide an estimate of the entire time that automation was engaged, we also calculated the duration of the KR and KP text to speech feedback prompts.

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. We asked a general question: 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?