Interspeech 2022
18-22 September 2022, Incheon, Korea

PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of /ɹ/

Nina R Benway[1], *Jonathan L. Preston*[1,2], *Elaine Hitchcock*[3], *Asif Salekin*[4], *Harshit Sharma*[4], *Tara McAllister*[5]

[1]Communication Sciences and Disorders, Syracuse University, New York, USA
[2]Haskins Laboratories, New Haven, Connecticut, USA
[3]Communication Sciences and Disorders, Montclair State University, New Jersey, USA
[4]Electrical Engineering and Computer Science, Syracuse University, New York, USA
[5]Communicative Sciences and Disorders, New York University, New York, USA

nrbenway@syr.edu

Abstract

We present the PERCEPT-R corpus, a labeled corpus of child speakers of American English with typical speech and residual speech sound disorders affecting rhotics. We demonstrate the utility of age-and-gender normalized formants extracted from PERCEPT-R in training support vector classifiers to predict ground-truth perceptual judgments of "rhotic" (i.e., dialect-typical) and clinical "derhotic" /ɹ/ for novel speakers (mean of participant-specific f-metrics = .83; SD = .18, N = 281).

Index Terms: clinical speech, child speech, open access dataset, mispronunciation detection, /ɹ/

1. Introduction

Children with speech sound disorder show diminished accuracy and intelligibility in spoken communication and may thus experience negative impacts on academic, socioemotional and socioeconomic outcomes [1-3]. While most speech errors resolve by the late school-age years, between 2-5% of speakers exhibit residual speech sound disorder (RSSD) that persists until adulthood [4]. We focus on /ɹ/, the most common and challenging residual speech deviation in American English [5].

Effective intervention can reduce the lifetime burden of RSSD for impacted individuals. Motor-based intervention is one evidence-based practice for RSSD involving adaptive delivery of auditory-visual models and verbal cues for articulator placement [6]. However, not all children have access to motor-based intervention due to clinician shortages [7]. Even when intervention for RSSD is secured, treatment intensity might be lower than required for successful intervention because of clinician caseload size [8, 9]. Such access barriers could potentially be mitigated by computerized therapy with automated mispronunciation detection [10], but no existing system is sufficient for clinical use [11]. The three fundamental issues impacting available systems are the lack of examples of RSSD speech for system training, low accuracy when analyzing sounds produced incorrectly, and no empirical assessment of therapeutic benefit [11, 12].

1.1. Motivation and Contributions

The development of child speech technologies is hindered by the lack of large-scale, labeled child speech corpora [13], doubly so for child clinical speech technologies [11]. Corpora are emerging to address the need for publicly available child clinical data, such as SEED [14] and the three datasets of Ultrasuite [15]. These corpora contain spoken utterances elicited during the course of evaluation and treatment (Table 1). However, to our knowledge, no publicly available clinical corpus provides labeled training data for the audio classification of rhotic/derhotic /ɹ/ in the context of RSSD.

Table 1: *Notable public clinical speech corpora.*

Corpus	Tokens	Speakers	Notes
SEED [14]	~16,000 words and sentences	58 children, 34 adults	Data from 16 speech tasks
Ultrasuite [15]	~14,500 phones, words, sentences, & nonspeech	86 children	Ultrasound treatment data
PERCEPT-R 2.2.1p	105,232 words	281 children, 1 young adult	Introduced herein

In the absence of high-quality data, mispronunciation detection algorithms are not sufficient for clinical use. McKechnie and colleagues [11] systematically reviewed automatic child speech analysis tools for clinical use or second language phonetic acquisition. The studies reviewed report percent agreement with human judgment ranging from 45.7% to 95.67%; however, the authors caution that these numbers may be misleading when the test datasets contain few exemplars of mispronounced speech. As such, no available tool reviewed, including audio classifiers, met the authors' accuracy threshold for identifying *incorrectly* produced words (>.8). Also, none of the studies specifically investigated rhotic classification in children. Gupta and DiPadova [16] later provided evidence that support vector machines were suitable for a related task, the classification of tokens showing typical sociophonetic variation with regards to rhoticity. This task, however, is not entirely analogous to the classification of clinical speech because of potential articulatory (i.e., acoustic) differences between dialect-typical /ɹ/ in non-rhotic dialects and the types of clinical /ɹ/ deviations observed in RSSD.

This paper offers two contributions. First, we present the PERCEPT-R corpus (*P*erceptual *E*rror *R*ating for the *C*linical *E*valuation of *P*honetic *T*argets-R). PERCEPT-R is an order of magnitude larger than other publicly available child clinical

10.21437/Interspeech.2022-10785

speech corpora, and provides labeled training data for rhotic/derhotic /ɹ/ in the context of RSSD. To this end, we also demonstrate the utility of using formant-based measures extracted from the corpus to train classifiers that predict human perceptual judgment of clinical speech.

2. Corpus Description

The PERCEPT-R corpus focuses on the speech of children from fully-rhotic American English dialects with RSSD that primarily impacts /ɹ/ (with a "rhotic" /ɹ/ being perceptually unmarked and "derhotic" being marked). The present description reflects the first public release of PERCEPT-R, v 2.2.1p, with all included participants having provided parental permission and assent (and, for adult participants, consent) for audio data sharing outside of the original study of enrollment. PERCEPT-R v 2.2.1p currently contains 32.47 hours of citation speech recordings reflecting 105,232 word-level utterances. Data were collected between 2006 and 2020 at Syracuse University, Montclair State University, and New York University during 22 separate studies. Procedures for data collection were approved by the Institutional Review Boards of the relevant university or through the Biomedical Research Alliance of New York (BRANY). The release of this corpus was considered by BRANY as not human subject research activity (i.e., exempt secondary data analysis, 21-038-524).

2.1. Participant Description

The 281 participants included in PERCEPT-R 2.2.1p range in age from 72 months to 288 months ($\bar{x} = 135.82$, $\sigma_{\bar{x}} = 30.44$). Of the 281 participants, 121 are females. The imbalance between males and females in the corpus reflects the increased prevalence of RSSD observed among males [17]. 78 of the 281 participants were recorded through studies of typically developing speakers. Figure 1 shows the distribution of ages within the corpus, grouped by gender and speaker group.

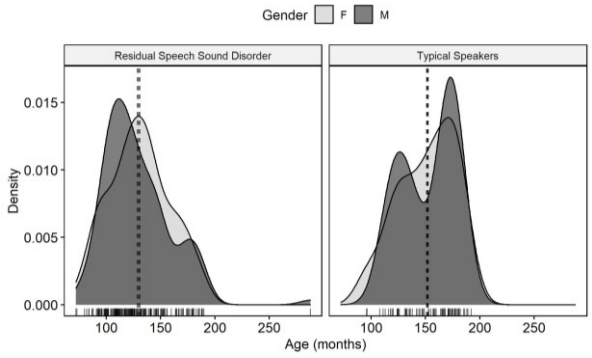

Figure 1: *Distribution of age and gender in corpus*

2.2. Data Acquisition

Data were collected directly by members of the originating study teams, most often a research-trained speech-language pathologist. The original purpose of treatment experiments, broadly, was to test the effect of modulations of motor learning intervention parameters on acoustic or perceptual speech accuracy. The original purpose of typical speech studies was to provide comparison speech data versus RSSD.

2.2.1. Recording Environment

Most of the originating studies were lab-based studies in which audio was recorded using a participant-worn headset (e.g., AKG C520) or lavalier mic (e.g., Sennheiser MKE 2). For telepractice sessions (i.e., spring 2020), audio recording was completed with a study-provided headset without network audio transfer (i.e., local to the participant). All files were collected as 16-bit PCM lossless audio in WAV containers.

2.2.2. Data Collection Timepoints

The corpus includes longitudinal recordings of participants before, during, and after speech treatment, as well as cross-sectional samples from age-matched children with typical speech. Audio samples from pre-treatment and post-treatment evaluation sessions were collected during probe tasks that included direct imitation, reading, and picture naming. Audio samples from within treatment sessions were solicited using the general trial-by-trial structure of motor-based practice (i.e., presentation of prompt, utterance, clinical feedback). All corpus items represent citation speech (isolated phrases, words, and syllables).

2.2.3. Audio Processing

All corpus audio was recorded at the level of the session or the task and study-specific target utterances were manually segmented from session- or task-length audio. Segment boundaries and orthographic labels were annotated using Praat TextGrids. In the case of more recent studies these segmentation labels were placed automatically using intensity detection or by our treatment software, CPP [18], in which case the automatically generated segmentations and transcripts were manually verified by trained research assistants. The utterances were extracted using Praat [19] or the Parselmouth API [20] for Praat, depending on the originating study. All corpus files have been standardized to 1 channel, 44.1 kHz audio scaled to an average intensity of 70 dB using Parselmouth.

2.3. Corpus Labels and Metadata

The PERCEPT-R corpus is formatted for Phon software [21], which contains scripts for detailed lexical and phonological analysis of PERCEPT-R. Important points are detailed below. Each record in PERCEPT-R is linked with a participant, orthographic transcription of the utterance, IPA and ARPABET target transcriptions of the utterance, the number of unique listeners' perceptual ratings, the number of "rhotic" ratings, the calculated average rating, the originating study, and the timepoint of data collection.

2.3.1. Transcripts and Phonological Coverage

Each audio file in PERCEPT-R 2.2.1p represents one utterance consisting of one syllable or word or, in a minority of cases, compound words or short phrases (e.g., "candy bar", "burn up"; 0.8% of total). The lexical referent for all utterances was known and transcribed at the time of recording. In the 2.2.1p release, all utterances have one target rhotic.

There are 471 unique target utterances represented in PERCEPT-R 2.2.1p; most frequently monosyllable words (n = 78,749) or disyllable words (n = 6250). Of these 471 unique target word forms, 41 are phonotactically-probable target nonwords (e.g., /ɑɹd/, /kɝ/). The five most frequent target words in the corpus are beard (n=2098), turn (n=2089), nurse (n=1971), ladder (n=1957), and chair (n=1955). The five most frequent target nonwords in the corpus are /ɝ/ (n=1743), /ɹɑ/, (n=1725), /ɹi/ (n=1570), dɝ (n=165), and /ɝp/ (n=144). There are 93 different C-V word shapes represented in the corpus,

most frequently ˈCVC (n=18,589), ˈCVCC (n=13,023) and ˈCCVC (n=9576). There are 32,969 singleton onset /ɹ/ and 23,375 singleton coda /ɹ/. There are 23 unique onset clusters represented in the corpus (e.g., "st", 3641), and 30 unique coda clusters in the corpus (e.g., "rd", 6340).

2.3.2. Rhotic Perceptual Labels

Perceptual rating of /ɹ/ tokens is not straightforward; expert listeners rating /ɹ/ in RSSD have previously shown 85% agreement [22]. The imperfect agreement seen for /ɹ/ ratings is likely due to factors known to influence speech perception: phonetic, lexical, and prosodic context, and, perhaps most relevant, *expectation* of a sound [23].

Perceptual ratings for utterances in PERCEPT-R 2.2.1p were derived from either crowdsourced listening tasks or expert listening tasks, depending on the study of origin for a given utterance. We estimate that 72,126 utterances have crowdsourced listener ratings and 33,106 have expert ratings. Crowdsourced ratings were obtained using the Amazon MTurk platform. In originating studies that used expert listener rating, ratings were obtained from licensed speech-language pathologists or speech-language pathologists in training who had completed coursework in speech sound disorders. Both rating platforms used the same general approach: utterances with orthographic labels were randomized to batches with different speakers and different study timepoints (i.e., pre-treatment, during treatment, and post-treatment), with participant and timepoint information unknown to the rater. The perceptual accuracy of an individual utterance was calculating by summing of listener responses (0 = derhotic and 1 = rhotic) and dividing by the number of raters (e.g., (0 + 1 + 0)/3 = .33).

There is an imbalance favoring derhotic tokens in PERCEPT-R 2.2.1p, as the data (by design) come overwhelmingly from recordings of children with speech distortions. 31,106 utterances were unanimously rated as "derhotic", representing 167 participants. An additional 30,840 utterances were rated derhotic by most, but not all, raters (average rating = $0 < x < .5$; n=204 participants). 16,550 tokens received unanimous ratings of "rhotic" (n=232 participants). Finally, 26,736 tokens were rated "rhotic" by half or more raters, but fell short of unanimous ratings (average rating = $.5 \leq x < 1$; n=269 participants). 155 unique participants contributed tokens rated fully rhotic as well as tokens rated fully derhotic.

2.3.3. Corpus Distribution and Future Corpus Development

The PERCEPT corpus is publicly available through partnership with PhonBank [24], a NIH-funded data-sharing platform for speech-language research. PhonBank, as well as the larger TalkBank project it belongs to, is committed to making speech and language data Findable, Accessible, Interoperable, and Reusable (FAIR). PhonBank maintains the open-source software Phon, which enables indexing and analysis based on phonological characteristics and participant characteristics.

For each participant and session, the word-level utterances have been concatenated into a single audio file with pauses between words. These are time-aligned with an orthographic transcript saved in two formats: Phon-readable XML and Praat TextGrids. Record metadata (see: section 2.3) is attached to each utterance in Phon and can be exported using standard Phon functionality. Corpus-level metadata and datasheets [25] are published on PhonBank as well.

Future versions of the public PERCEPT-R corpus will include pretrained acoustic models that can be used with the Montreal Forced Aligner [26], as well as time-aligned phonetic segmentation. Furthermore, to increase the utility of the corpus for general-purpose child and clinical speech recognition beyond the phoneme /ɹ/, we are in the process of adding labeled corpus data from the Goldman-Fristoe Test of Articulation-Third Edition [27], a standardized speech instrument that elicits single words representing a wide range of phonetic targets in English (e.g., *bath*, *jumping*). In addition to children with RSSDs and typically developing children, these new recordings include a sizable sample of children with apraxia of speech (CAS) or other SSD.

3. Corpus Demonstration

Classification experiments with leave-one-participant-out validation demonstrate the ability of the PERCEPT-R 2.2.1p dataset to train classifiers that predict perceptual judgment of rhotic accuracy in novel individuals.

3.1. Feature Extraction

This demonstration used formants to quantify the speech signal because of the known association of formants with perceptual judgments of /ɹ/ accuracy for clinical populations [28, 29]. In rhotic dialects of American English, /ɹ/ is marked acoustically by a relatively high second formant (F2; [30]) and a relatively low third formant (F3; [31]). This results in a much narrower average F3-F2 distance in rhotics than in derhotics, all else being equal [32]. For this reason, we retained all formants F3 and lower, as well as the calculated F3-F2 distance, for this corpus demonstration.

Speaker-specific LPC settings (e.g., Praat "maximum formant" values) were used to adapt the analysis to each speaker (see: [33]), with personalized settings estimated using a implementation of the Praat FastTrack plugin [34] customized for the HTCondor framework [35] on the OrangeGrid computing environment at Syracuse University. Formants were estimated using the Praat Formant (robust) function with function calls automated using the Parselmouth API. Robust formants provide estimates using the autocorrelation method with robust linear prediction that is meant reduce variance and bias versus the Burg method of estimation [36]. Five formants were estimated from 25 ms windows with a 25% overlap. Selective weighting of samples associated with the robust method began at 1.5 standard deviations with 5 refinement iterations. Formant value time series were estimated for the entire word to minimize estimation error due to edge effects within the relatively short /ɹ/ intervals. Missing formant samples were imputed by mean-interpolation given the previous and following samples in the timeseries for that formant. Extracted formants were z-standardized (e.g., [37]) with regard to /ɹ/ age-and-gender specific formant in the sample collected by Lee and colleagues [38].

The section of the formant timeseries that pertained to the /ɹ/ was extracted using boundaries determined by the Montreal Forced Aligner wrapper [26] for the Kaldi Speech Recognition Toolkit [39]. For the purposes of this demonstration, default pre-trained American English acoustic models were used (i.e., Librispeech [40]). Rhotic boundaries were expanded by 25 ms on either side to offset aligner errors. These rhotic-associated formant time series were then binned into three temporal windows reflective of a phone and transition model [41]: early (average of the first third of windows), middle (average of the middle third of windows), and late (average of the last third of windows). These windows were stacked vertically into an array where each utterance was one row, 12 features wide (i.e., nF1,

nF2, nF3, nF3-F2, all at each averaged at three timepoints). The univariate ability of nF3-F2 to separate the binary classes in the dataset at the three timepoints is shown in Figure 2.

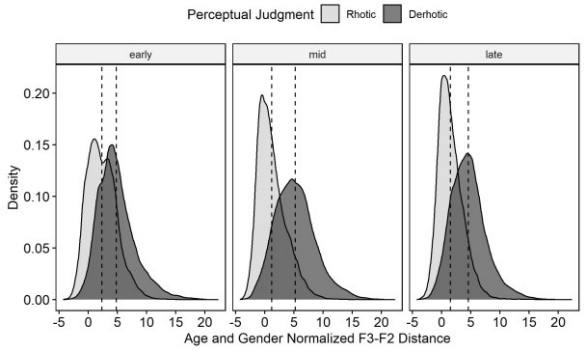

Figure 2. *Univariate class separation with nF3-F2.*

3.2. Leave-One-Participant-Out Validation

We ran two demonstrations: the first for the classification of rhotics in words with unanimous ratings of "rhotic"/"derhotic", for comparison with previous non-clinical classification attempts for /ɹ/ [16], and the second for the classification of all rhotics in the corpus, as ambiguous data are often encountered during clinical intervention. For ambiguous data, average listener ratings < .7 were considered derhotic. Each demonstration fit 281 separate models using leave-one-participant-out cross validation [42], mimicking clinical use.

Support vector classifiers (SVC) modeled the relationship between the formant-based features and perceptual rating. SVC models were fit using a radial basis function kernel in Scikit-Learn v. 1.0 in Python v. 3.8.12. Model performance was judged using the weighted f-metric, to reflect an imbalanced dataset and a use case where the classifier is expected to encounter more derhotic /ɹ/ than rhotic /ɹ/. We present the average f-metric for all cross validations as a preliminary index of suitability for predicting /ɹ/ accuracy in novel participants.

Table 2. *Average f-metric per subset reflecting 281 cross-validations, one for each participant.*

Subset	Average F-metric	Standard Deviation
Unanimous labels	.91	.15
All data	.83	.18

4. Results

The PERCEPT-R 2.2.1p corpus demonstrated utility for training a classifier with good accuracy for identifying the rhoticity of nonambiguous /ɹ/ ($\mu_{fmetric\ unanimous\ data} = .91$; $\sigma_\mu = .15$). Performance was lower, but still above the threshold of clinical utility, for the classification of the entire dataset ($\mu_{fmetric\ all\ data} = .83$; $\sigma_\mu = .18$). Distributions of all participant-specific f-metrics are shown in Figure 3.

5. Discussion

This study addresses the first barrier to the development of child clinical speech technologies: the lack of useful public data. The size of PERCEPT-R 2.2.1p compares favorably to existing publicly available corpora of children's clinical speech. The quality of corpus labels and integration with Phon allows for detailed phonological analyses to be completed on corpus items. We anticipate that these features will make the corpus a valuable resource for researchers in linguistics and communication disorders. We also anticipate that the data will be of interest to speech technology engineers interested in FAIR engineering, in addition to those interested in /ɹ/ classification.

Our demonstration indicates that meaningful features can be extracted from PERCEPT-R for the purpose of binary classification of /ɹ/ in children with RSSD and is in line with previous classification attempts for /ɹ/. The distribution of participant-specific f-metrics indicates that some voices may be a better fit for automated speech analysis than others, likely because of difficulties with formant extraction. Future development on classifiers trained on the PERCEPT-R corpus can identify speech biomarkers associated with high classification accuracy and methodologies/features/classifier architectures to increase accuracy for ill-fitting subjects.

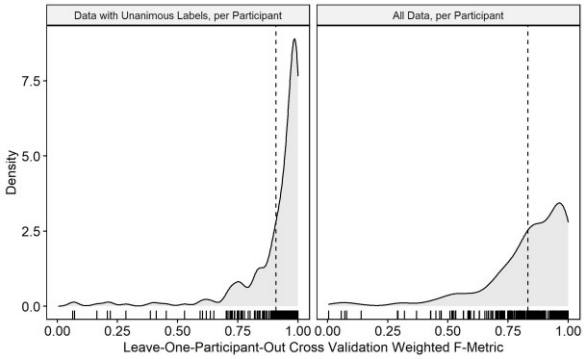

Figure 3. *Distribution of participant-specific f-metrics.*

Importantly, our results surpass the threshold set by McKechnie and colleagues [11] for clinically useful speech technology. Further work can embed optimized versions of these classifiers into existing motor-based intervention software. The planned experimental validation of such tools will address the second and third largest barriers to the success of child clinical speech technology: insufficient technical description and lack of clinical efficacy data for these systems.

6. Conclusions

We have presented the first public release of the PERCEPT-R corpus, 2.2.1p. The corpus is the first dataset to demonstrate utility for training clinically-acceptable classifiers for the prediction of /ɹ/ perceptual judgment in children with RSSDs.

7. Acknowledgements

The authors wish to thank the participants and their families. 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).

8. References

[1] Hitchcock, E.R., Harel, D., and McAllister Byun, T. (2015). Social, emotional, and academic impact of residual speech errors in school-aged children: A survey study. *Seminars in Speech and Language, 36*(4), 283-94.

[2] McCormack, J., McLeod, S., McAllister, L., and Harrison, L.J. (2009). A systematic review of the association between childhood speech impairment and participation across the lifespan. *International J. of Speech-Language Pathology, 11*(2), 155-170.

[3] Hall, B.J.C. (1991). Attitudes of fourth and sixth graders toward peers with mild articulation disorders. *Language, Speech, and Hearing Services in Schools, 22*(1), 334-340.

[4] Flipsen, P., Jr. (2015). Emergence and prevalence of persistent and residual speech errors. *Seminars in Speech and Language, 36*(4), 217-23.

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

[6] Preston, J.L., Benway, N.R., Leece, M.C., Hitchcock, E.R., and McAllister, T. (2020). Tutorial: Motor-based treatment strategies for /r/ distortions. *Language, Speech, and Hearing Services in Schools, 54*, 966-980.

[7] Sugden, E., Baker, E., Munro, N., and Williams, A.L. (2016). Involvement of parents in intervention for childhood speech sound disorders: A review of the evidence. *International Journal of Language & Communication Disorders, 51*(6), 597-625.

[8] Sugden, E., Baker, E., Munro, N., Williams, A.L., and Trivette, C.M. (2018). Service delivery and intervention intensity for phonology-based speech sound disorders. *International Journal of Language and Communication Disorders, 53*(4), 718-734.

[9] Katz, L.A., Maag, A., Fallon, K.A., Blenkarn, K., and 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.

[10] McLeod, S., Ballard, K.J., Ahmed, B., McGill, N., and Brown, M.I. (2020). Supporting children with speech sound disorders during COVID-19 restrictions: Technological solutions. *Perspectives of the ASHA Special Interest Groups*.

[11] McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., and 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.

[12] Shahin, M., Zafar, U., and 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.

[13] Yeung, G. and Alwan, A. (2018). *On the difficulties of automatic speech recognition for kindergarten-aged children*. in *INTERSPEECH 2018: Proceedings of the 19th Annual Conference of the ISCA*. Hyderabad, India.

[14] Speights Atkins, M., Bailey, D.J., and 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.

[15] Eshky, A., Ribeiro, M.S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J.M., and Wrench, A.. (2018). *ULTRASUITE: A repository of ultrasound and acoustic data from child speech therapy sessions*, in *INTERSPEECH 2018: Proceedings of the 19th Annual Conference of the ISCA*. p. 1888--1892.

[16] Gupta, S. and DiPadova, A. (2019). *Deep learning and sociophonetics: Automatic coding of rhoticity using neural networks*. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*.

[17] Wren, Y., Miller, L.L., Peters, T.J., Emond, A., and 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), 647-73.

[18] McAllister, T., Hitchcock, E.R., and Ortiz, J.A. (2020). Computer-assisted challenge point intervention for residual speech errors. *Perspectives of the ASHA Special Interest Groups*.

[19] Boersma, P. and Weenink, D. (2020). *Praat [computer software]*. Amsterdam: University of Amsterdam.

[20] Jadoul, Y., Thompson, B., and de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics, 71*, 1-15.

[21] Hedlund, G. and Rose, Y. (2019). *Phon [computer software]*.

[22] Klein, H.B., McAllister Byun, T., Davidson, L., and Grigos, M.I. (2013). A multidimensional investigation of children's /r/ productions: Perceptual, ultrasound, and acoustic measures. *American J. of Speech-Language Pathology, 22*(3), 540-553.

[23] Yi, H.G., Leonard, M.K., and Chang, E.F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron, 102*(6), 1096-1110.

[24] Rose, Y. and MacWhinney, B. (2014). The PhonBank project: Data and software-assisted methods for the study of phonology and phonological development. *Oxford Handbook of Corpus Phonology*.

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

[26] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M.. (2017). *Montreal forced aligner: Trainable text-speech alignment using kaldi*, in *INTERSPEECH 2017: Proceedings of the 18th Annual Conference of the ISCA*. Stockholm, Sweeden. p. 498-502.

[27] Goldman, R. and Fristoe, M. (2015). *Goldman Fristoe Test of Articulation - third edition*. Pearson.

[28] Campbell, H., Harel, D., Hitchcock, E., and McAllister Byun, T. (2018). Selecting an acoustic correlate for automated measurement of american english rhotic production in children. *International J. of Speech Language Pathology, 20*(6), 635-643.

[29] Dugan, S.H., Silbert, N., McAllister, T., Preston, J.L., Sotto, C., and Boyce, S.E. (2019). Modelling category goodness judgments in children with residual sound errors. *Clinical Linguistics and Phonetics, 33*(4), 295-315.

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

[31] Espy-Wilson, C.Y., Boyce, S.E., Jackson, M., Narayanan, S., and Alwan, A. (2000). Acoustic modeling of American English /r/. *Journal of the Acoustical Society of America, 108*(1), 343-56.

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

[33] Derdemezis, E., Vorperian, H.K., Kent, R.D., Fourakis, M., Reinicke, E.L., and 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.

[34] Barreda, S. (2021). Fast track: Fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard, 7*(1).

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

[36] Lee, C.-H. (1988). On robust linear prediction of speech. *IEEE Trans. Acoustics, Speech, & Signal Processing, 36*(5), 642-650.

[37] Benway, N.R., Hitchcock, E., McAllister, T., Feeny, G.T., Hill, J., and 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*.

[38] Lee, S., Potamianos, A., and 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.

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

[40] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). *Librispeech: An ASR corpus based on public domain audio books*. in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

[41] Hertz, S.R. (1991). Streams, phones and transitions: Toward a new phonological and phonetic model of formant timing. *Journal of Phonetics, 19*(1), 91-109.

[42] Cawley, G.C. and Talbot, N.L.C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res., 11*, 2079–2107.