10-16-2018

Selecting An Acoustic Correlate for Automated Measurement of American English Rhotic Production in Children

Heather Campbell
New York University

Daphna Harel
New York University

Elaine Hitchcock
Montclair State University, hitchcocke@montclair.edu

Tara McAllister Byun
New York University

International Journal of Speech-Language Pathology

ISSN: 1754-9507 (Print) 1754-9515 (Online) Journal homepage: https://www.tandfonline.com/loi/iasl20

Selecting an acoustic correlate for automated measurement of American English rhotic production in children

Heather Campbell, Daphna Harel, Elaine Hitchcock & Tara McAllister Byun

To cite this article: Heather Campbell, Daphna Harel, Elaine Hitchcock & Tara McAllister Byun (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, DOI: 10.1080/17549507.2017.1359334

To link to this article: https://doi.org/10.1080/17549507.2017.1359334

Published online: 10 Aug 2017.

International Journal of Speech-Language Pathology, 2018; 20: 635–643

Selecting an acoustic correlate for automated measurement of American English rhotic production in children

HEATHER CAMPBELL[1], DAPHNA HAREL[2], ELAINE HITCHCOCK[3] & TARA MCALLISTER BYUN[1]

[1]*Department of Communicative Sciences and Disorders, NYU Steinhardt School of Culture, Education, & Human Development, New York, NY, USA,* [2]*Center for the Promotion of Research Involving Innovative Statistical Methodology, New York, NY, USA, and* [3]*Department of Communication Sciences and Disorders, Montclair State University, Montclair, NJ, USA*

Abstract

Purpose: A current need in the field of speech–language pathology is the development of reliable and efficient techniques to evaluate accuracy of speech targets over the course of treatment. As acoustic measurement techniques improve, it should become possible to use automated scoring in lieu of ratings from a trained clinician in some contexts. This study asks which acoustic measures correspond most closely with expert ratings of children's productions of American English /ɹ/ in an effort to develop an automated scoring algorithm for use in treatment targeting rhotics.
Method: A series of ordinal mixed-effects regression models were fit over a large sample of children's productions of words containing /ɹ/ that had previously been rated by three trained clinicians. Akaike/Bayesian Information Criteria were used to select the best-fitting model.
Result: Controlling for age, sex, and allophonic contextual differences, the measure that accounted for the most variance in speech rating was F3–F2 distance normalised relative to a sample of age- and sex-matched speakers.
Conclusion: We recommend this acoustic measure for use in future automated scoring of children's production of American English rhotics. We also suggest that computer-based treatment with automated scoring should facilitate increases in treatment dosage by improving options for home practice.

Keywords: *Human speech*; *biofeedback therapy*; *linear-mixed effects models*; *ordinal regression analysis*; *speech sound disorders*; *speech pathology*

Introduction

Speech sound disorders (SSD) affect up to 10% of pre-school and school-aged children (American Speech-Language-Hearing Association, 2011) and can impede participation in social and academic activities (Gibbon & Paterson, 2006; Hitchcock, Harel, & McAllister Byun, 2015). While some children with SSD resolve their errors spontaneously, others require long-term clinical intervention before their speech deficits can be considered fully remediated (Flipsen, 2015). Speech sound errors that persist beyond 8–9 years of age, when the developmental sound inventory is expected to be complete, are referred to as residual speech sound errors (RSE) (Preston et al., 2014). Roughly 30% of children with a history of SSD continue to exhibit speech errors at 9 years of age, while 9% continue to show errors at 12–18 years of age (Lewis & Shriberg, 1994). In light of the challenge that these persistent cases present, 40% of school-based speech–language pathologists (SLPs) report having discharged children with RSEs from their caseloads even though full remediation had not yet been achieved (Ruscello, 1995).

In American English, misarticulation of the rhotic sound /ɹ/ (henceforth /r/) is one of the most prevalent RSEs and is considered the most challenging to remediate (Shuster, Ruscello, & Toth, 1995). The English /r/ is among the latest-acquired speech sounds, with an age of mastery as late as eight years (Smit, Hand, Freilinger, Bernthal, & Bird, 1990). Part of the reason it is acquired later than other sounds is its articulatory complexity, given that accurate production requires nearly simultaneous anterior and posterior lingual constrictions (Espy-Wilson, 1992) that can be achieved with a variety of lingual contours (Delattre & Freeman, 1968).

Correspondence: Heather Campbell, Department of Communicative Sciences and Disorders, NYU Steinhardt School of Culture, Education, & Human Development, 665 Broadway, Floor 9, New York, NY 10012, USA. Email: heather.campbell@nyu.edu

ISSN 1754-9507 print/ISSN 1754-9515 online
Published by Informa UK Limited, trading as Taylor & Francis Group
DOI: 10.1080/17549507.2017.1359334

Despite its articulatory variability, perceptually accurate /r/ has consistent acoustic properties, characterised by a low third formant frequency (F3) relative to other vocalic sounds, in addition to a relatively high second formant frequency (F2) (Delattre & Freeman, 1968; Hagiwara, 1995). Taking advantage of this acoustic consistency, researchers have explored the efficacy of treatment for rhotic misarticulation that incorporates visual-acoustic biofeedback in a variety of forms, including electropalatography (EPG), ultrasound, and visual-acoustic biofeedback using an acoustic spectrogram or spectrum. The form of visual biofeedback that is the focus in the current study is spectral acoustic biofeedback, which uses real-time linear predictive coding (LPC) to form an acoustic spectrum representing the resonant frequencies of the vocal tract (McAllister Byun & Campbell, 2016; McAllister Byun & Hitchcock, 2012). In this kind of visual-acoustic biofeedback, the speech–language pathologist (SLP) familiarises the client with a typical formant configuration for /r/ and then cues the client to adjust his/her output during /r/ production to achieve a closer match with a visual target superimposed on the real-time LPC spectrum. Figure 1 presents two LPC spectra that might be used as examples for biofeedback intervention, drawing the client's attention to the lower height of F3 in a perceptually accurate /r/.

While visual-acoustic biofeedback can be effective in eliciting correct /r/ from children who have not responded to previous forms of treatment (McAllister Byun & Campbell, 2016; McAllister Byun & Hitchcock, 2012), gains made through biofeedback treatment do not automatically generalise to a context in which enhanced feedback is not available. A long duration of treatment may be needed before gains generalise outside of the therapy setting. Thus, the incorporation of biofeedback into therapy settings alone may not be sufficient to reduce the strain that clients with residual speech errors place on SLP resources. Home practice of speech targets is one way to increase the dosage of speech intervention. However, home practice comes with the risk that, without feedback from a trained observer, the child will counterproductively reinforce incorrect speech patterns. This is an issue of particular concern for children who have difficulty perceptually discriminating correct versus incorrect rhotics (Shuster, 1998). Given recent advances in speech recognition technology, it should be possible to use automated algorithms to monitor accuracy and provide appropriate feedback during home practice of rhotic targets.

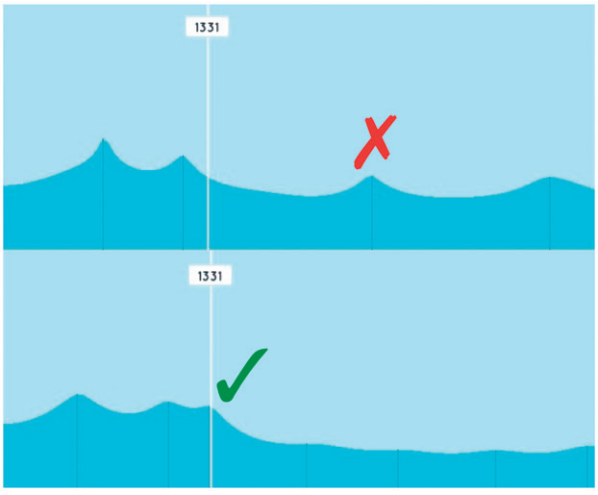

Figure 1. Formant frequencies represented as peaks of an LPC spectral display, with line representing an accurate rhotic target, currently set at 1646 Hz. Incorrect /r/ (top panel) is characterised by a relatively high F3, while correct /r/ (bottom panel) is characterised by a relatively low F3. Images from the "staRt" app (McAllister Byun et al., 2017).

App-based acoustic biofeedback therapy is a novel way to integrate the recently demonstrated efficacy of visual-acoustic biofeedback with the increased use of apps in clinical practice. Speech Therapist's App for /r/ Treatment (staRt) is an iOS app currently in development at New York University that uses mobile technology to generate a real-time LPC spectrum that clients can use to match an acoustic target representing correct /r/ (McAllister Byun et al., 2017). In its current iteration, the app is intended for use under the supervision of an SLP, who scores each production as correct or incorrect based primarily on auditory-perceptual judgment. The visual display provides an additional source of information, but it does not automatically classify productions as correct or incorrect rhotics. However, one goal in future versions of staRt is to incorporate features including a home practice routine and automated scoring. The app already generates continuous real-time estimates of formant frequencies in the user's speech; therefore, if a particular range of frequencies could be defined to represent a "correct /r/", using these LPC values as the basis for automated scoring would be a straightforward extension. In a maximally simple hypothetical scenario, since /r/ has a characteristically low F3, a threshold frequency could be defined such that all speech outputs with an F3 below this threshold would be classified as accurate /r/ productions. However, previous research suggests that other important considerations need to be addressed prior to implementation of an automated scoring algorithm.

Several acoustic measures may be considered for rhotics. F3 is recognised as the primary acoustic cue to rhoticity (Espy-Wilson, Boyce, Jackson, Narayanan, & Alwan, 2000; Idemaru & Holt, 2013), since the low height of F3 differentiates /r/ from acoustically similar sounds such as /l/ and /w/ (Polka & Strange, 1985). Rhotic sounds are further differentiated from these sounds by their relatively high F2, which is considered a secondary acoustic cue to rhoticity in English (Boyce & Espy-Wilson,

1997; Polka & Strange, 1985). Derived measures, including the distance between F3 and F2 (F3–F2) and the ratio of F3 and F2 (F3/F2), reflect the influence of both cues simultaneously (Lee, Potamianos, & Narayanan, 1999). Flipsen, Shriberg, Weismer, Karlsson, and McSweeny (2001) explored the relationships of raw versus derived acoustic measures in typically developing children of different ages and sexes. While significant correlations were found between raw formant frequencies (F2, F3) and both age and sex for most /r/ contexts, no relationships were found for the derived measures (F3–F2 or F3/F2). This suggests that the derived acoustic measures provide some correction for differences in raw formant height related to a child's age and sex. However, whether this correction is sufficient is unknown.

Another metric considered by Flipsen et al. (2001) involved external standardisation of the acoustic measures to normative data. Raw F2 and F3 normative distributions were reported in Lee et al. (1999) while derived F3–F2 and F3/F2 normative distributions were calculated from the same raw data and reported in the appendix of Flipsen et al. (2001). Their results suggested that age and sex differences were better accounted for by the normalised versions of F2, F3, F3–F2 and F3/F2 than the raw formant values or derived measures. In a companion paper, Shriberg, Flipsen, Karlsson, and McSweeny (2001) found that normalised values of F3–F2 distinguished productions from children with typically developing /r/ versus children with /r/ errors, and also distinguished children with /r/ errors and a history of SSD versus those with isolated /r/ errors (Shriberg et al., 2001), suggesting their utility for evaluating children's rhotics.

However, the currently available normative data have several important limitations. First, the data provided in Lee et al. (1999) and used in Flipsen et al. (2001) are based on a relatively small sample size. Specifically, norms were calculated from a sample of 436 children between the ages of five and 17 years old, with 9–25 children representing each age and sex combination. Therefore, all normative data and resulting means and standard deviations calculated may be compromised by sampling error. Second, the normative data were drawn from a limited geographic region, the upper Midwestern USA, and therefore may not be generalisable to other regions. To this point, Campbell and McAllister Byun (2017) found significant formant differences, including differences in F3, between the normative group from Lee et al. (1999) and a sample of children from the northeast. Finally, the data in Lee et al. (1999) are based exclusively on measures of stressed syllabic /r/ in the word *bird*, whereas the current speech sample represents a range of /r/ contexts, including onset /r/ as in *red*, /r/ in a consonant cluster as in *tree*, and rhotic diphthongs as in *door, care,* and *fear*. Both Flipsen et al. (2001) and McAllister Byun and Tiede (2017) found significant differences in F3 and F3–F2 values in connection with different /r/ contexts. In consideration of these limitations of the normative data available, it remains unclear whether acoustic measures normalised using these data can be expected to outperform raw measures in predicting the perceptually rated accuracy of children's rhotics.

Finally, there is evidence that acoustics may interact with age and/or sex in determining listeners' expectations of the acoustic properties associated with accurate /r/ production. Munson, Edwards, Schellinger, Beckman, and Meyer (2010) found differences in how speech sound productions were categorised based on the perceived age of children between the ages of two and five. Their study focussed on acquisition of /s/, another late-developing sound. When the same acoustically intermediate fricative was embedded in a carrier phrase in a young child's voice, listeners were less likely to classify it as a correct /s/ sound than when it was presented in a carrier phrase with an older child's voice. Other research has reported differences in perception of sibilants when produced by males versus females (Dart, 1991; Zimmerman, Steiner., & Pond, 2002). These findings suggest that listeners may bring age- and sex-based expectations to a speech rating task that have the potential to interact with the properties of the raw acoustic signal.

The existing literature thus presented a number of questions open for investigation. The current study examined the relationship between trained listener judgments and several acoustic measures of /r/ sounds produced by children receiving treatment for rhotic misarticulation, with the goal of finding the best acoustic measure for use in automated scoring algorithms in this specific therapeutic context. To investigate these questions, the present study compared models that include normalised and non-normalised versions of both raw (F2, F3) and derived (F3–F2, F3/F2) acoustic values, with and without interactions of acoustic measures with age and sex. The goal of this investigation was to select an optimal model for use in automated scoring of children's /r/ sounds.

Method

Sample and procedure

This study utilised a data set compiled from /r/ productions by children over the course of participation in three different 8–10 week intervention studies targeting /r/ misarticulation in children acquiring English as a first language. These studies were an acoustic biofeedback treatment study (McAllister Byun & Hitchcock, 2012), an ultrasound biofeedback treatment study (McAllister Byun, Hitchcock, & Swartz, 2014), and an EPG biofeedback treatment study (Hitchcock, McAllister

Table I. This table shows the number of participants, ages and sexes of participants, and the number of productions pooled across these participants within each study.

Study	Number of children	Sex breakdown	Age range (mean)	Number of productions
Acoustic (2012)	11	1 female 10 male	6;0–11;9 (9;0)	2109
EPG (2015)	5	3 female 2 male	6;10–9;10 (7;8)	2926
Ultrasound (2014)	6	4 female 2 male	6;1–10;9 (8;0)	1040
Total	22			6075

Byun, Swartz, & Lazarus, in press). Although the original criteria for inclusion differed slightly among the three studies (see individual articles for specific inclusionary criteria from each study), all participants were classified as exhibiting /r/ misarticulation and were judged to fall within normal limits in a hearing screening and an examination of the oral mechanism.

Overall, the pooled dataset featured 22 children ranging in age from 6;0 to 11;9, with eight females and 14 males. All productions were isolated words elicited in probes administered over the course of treatment. Word probes varied in length from 20 to 64 words; the number of probes elicited from a child over the course of a study ranged from a minimum of three to a maximum of ten. The 6075 productions were subdivided into phonetic categories, including post-vocalic ($n = 1532$), syllabic ($n = 808$), singleton onset ($n = 774$) and cluster onset ($n = 2961$). See Table I for a detailed breakdown of the data from each study and the Appendix for a compiled list of words represented in this analysis.

Measures

In order to identify an automated measure of rhotic accuracy that can be incorporated into app-based treatment, it is important that the automated ratings match clinicians' perceptual accuracy ratings as closely as possible. In each of the three studies that contributed to the present data set, binary perceptual ratings for each token were acquired in a blinded randomised fashion from three certified SLPs using E-Prime 2.0 (Psychology Software Tools) and Praat (Boersma & Weenink, 2014). Mean ratings for each production were calculated such that each of the 6075 productions had an ordinal rating of 0, 0.33, 0.67 or 1. Within a given study, raters were required to achieve at least 80% pairwise agreement with one another; raters who did not meet this criterion were retrained or replaced. Ratings obtained from three trained listeners who achieve a minimum of 80% pairwise agreement have been described as the minimum "industry standard" for blinded listener ratings reported in intervention studies (McAllister Byun, Halpin, & Szeredi, 2015).

Formant frequencies from each child's productions were measured by three trained graduate students using Praat acoustic software (Boersma & Weenink, 2014). Speech samples from all studies had been recorded using the Computerized Speech Lab (CSL, KayPentax, Model 4150B) at 16-bit encoding with the microphone placed approximately five inches from each participant's mouth. All recordings were collected in a sound-shielded room. Recordings from the acoustic biofeedback study from 2012 were collected at a sampling frequency of 48 000 Hz, while recordings from the other two studies were collected at a sampling frequency of 44 100 Hz. After an optimal LPC filter order was determined for each participant (Vallabha & Tuller, 2004), the formant frequencies of each token of the /r/ sound were extracted from a point that was visually judged to represent the minimum F3 in the rhotic interval. Derived measures (F3–F2, F3/F2) were calculated from the raw acoustic measures for each production. To identify outlier data points that might reflect measurement error, the means and standard deviations of F3 and F3–F2 distance for all repetitions of each category were plotted for all participants. Plotted values that stood out on visual inspection were re-measured in order to correct any potential formant tracking errors. See Hitchcock et al. (in press), McAllister Byun and Hitchcock (2012) and McAllister Byun et al. (2014) for more detail about the procedure for re-measuring outliers and establishing reliability.

Measures of F3 and F2 were normalised by calculating z-scores relative to the mean and standard deviation from the appropriate subgroup of the normative sample collected by Lee et al. (1999). The same normalisation was performed for F3–F2 and F3/F2 using the additional normative data reported in Flipsen et al. (2001). Because normative values were only available for measurements in Hertz, throughout this analysis we report all measures in Hertz as opposed to a psychoacoustic scale such as Bark or mel.

Statistical analyses

A series of regression models, each examining one acoustic measure, was used to model the expected mean perceptual rating score aggregated over the clinician ratings on each production as a function of word-level and child-level predictors. Because the perceptual outcome measure was ordered but not continuous (possible values were 0, 0.33, 0.67 and 1), ordinal regression models were used.

A total of 32 total models were considered. All models included one of the eight previously mentioned acoustic measures as the predictor, including the four raw-data measures (F2, F3, F3–F2 and F3/F2) and the four normed measures (z-scores of F2, F3, F3–F2 and F3/F2, respectively). Each model adjusted for a word-level fixed effect of /r/ context (post-vocalic, syllabic, singleton onset, and cluster onset), and also included fixed effects for age (in months) and sex. All models also included child-level and token-level random effects to adjust for additional variability introduced by any inherent differences not captured by the aforementioned variables.

For each of the eight acoustic measures considered, a set of four models was considered that differed based on which first-order interactions were included. Since the relationship between perceptual ratings and acoustic measures may depend on the age and sex of the speaker (Munson et al., 2010), models including an interaction between the acoustic measure of interest and participant-level factors were examined. Specifically, each model included: either no interaction terms, an interaction between the acoustic measure and age only, an interaction between the acoustic measure and sex only, or the interaction between the acoustic measure and both age and sex. See Figure 2 for a visual summary of the 32 models evaluated.

The Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978) were used to select the regression model among the 32 candidate models that best explains the variation in the mean perceptual ratings. The AIC and BIC are indices used to compare non-nested models of a dataset that take into account the number of predictors in each model (Cohen, Cohen, West, & Aiken, 2013). Both the AIC and BIC penalise the log-likelihood of the data by accounting for the cost of estimating the parameters that are included in the model. The model with the lowest AIC and BIC values is chosen.

All analyses were conducted in RStudio version 0.99.879 (RStudio Team, 2017). The dataset was compiled and visualised using the "tidyverse" set of packages (Wickham, 2016), and regression models were fit using the "clmm" function in the 'ordinal' package in R (Christensen, 2015).

Result

Table II displays the AIC and BIC values for all 32 models fit. In all models, normalised F3–F2 was the acoustic value that was associated with the lowest AIC and BIC across all of the interaction possibilities (see bolded values in table). For both AIC and BIC, the lowest value across all 32 models was the model that included the normalised F3–F2 and the interactions of this acoustic measure with both age and sex. In this model, as expected, a higher normalised F3–F2 distance was associated with significantly lower accuracy ratings ($\beta = -1.21$, SE = 0.08, $p < 0.0001$), and phonetic context was statistically significant ($\chi^2 = 26.3$, $p < 0.0001$), indicating that the mean perceptual accuracy differed across each context. Although neither the main effects for age ($\beta = -0.008$, SE = 0.01, $p = 0.58$) nor sex ($\beta = -0.43$, SE = 0.60, $p = 0.48$) were statistically significant after controlling for all other variables in the model, both variables were significant in their

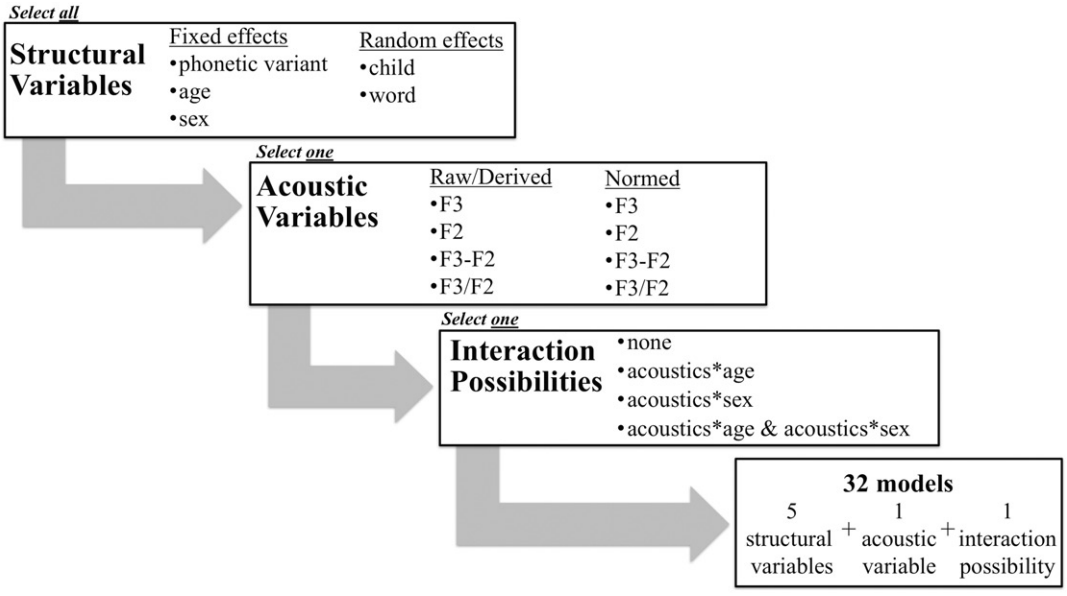

Figure 2. All models predicting mean perceptual rating of accuracy. All models included five structural variables. In addition to this base, one acoustic variable was added to each model. Each of the eight acoustic variables was run with either one of four interaction possibilities, for a total of 32 models.

Table II. This table shows the AIC and BIC for all 32 models: Lowest AIC and BIC for each interaction combination is marked in bold (comparison considering all models) and lowest AIC and BIC overall are marked with an asterisk.

	Main effects		Main effects + acoustics*age		Main effects + acoustics*sex		Main effects + acoustics*age + acoustics*sex	
Acoustic measure included in model	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
F2	9213.6	9287.5	9212.2	9292.8	9188.4	9268.9	9183.3	9270.5
F3	7920.9	7994.7	7871.1	7951.7	7912.8	7993.3	7851.6	7938.9
F3–F2	7752.0	7825.8	7739.7	7820.2	7753.9	7834.5	7741.7	7828.9
F3/F2	8056.9	8130.7	8027.2	8107.7	8058.8	8139.3	8029.2	8116.4
Normalised F2	9203.6	9277.5	9203.2	9283.8	9177.4	9257.9	9170.7	9258.0
Normalised F3	7960.9	8034.8	7900.2	7980.7	7914.3	7994.8	7800.2	7887.4
Normalised F3–F2	**7704.9**	**7778.7**	**7680.4**	**7760.9**	**7672.0**	**7752.5**	**7617.3***	**7704.5***
Normalised F3/F2	7971.6	8045.5	8051.4	8132.0	7949.4	8029.9	8156.7	8243.9

AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion.

Table III. This table shows the full output of the best overall model, including coefficients, standard errors (SE), and *p* values (significant effects marked with asterisks) for all fixed effects and variance for all random effects.

Fixed effect	Coefficient	SE	*p*-value
Normalised F3–F2	−1.28	0.08	<0.0001**
Age (months)	−0.008	−0.01	0.58
Sex	−0.43	0.60	0.48
Vocalic /r/ (relative to cluster /r/)	−1.07	0.21	<0.0001**
Singleton onset /r/ (relative to cluster /r/)	−0.48	0.27	0.07
Syllabic /r/ (relative to cluster /r/)	−1.1	0.28	0.0001**
Interaction of *z*-score F3–F2 with age (months)	0.005	0.0007	<0.0001**
Interaction of *z*-score F3–F2 with sex	0.28	0.04	<0.0001**
Variance of random effects:			
word 0.43			
Child 1.68			

interaction with the acoustic variables ($\beta = 0.0053$, SE = 0.00070, $p < 0.0001$ for age and $\beta = 0.28$, SE = 0.036, $p < 0.0001$ for sex). This indicates that the relationship between the acoustic measure and perceived accuracy differed depending on the child's age and sex, and that the normalisation of F3–F2 did not fully account for these differences. Table III presents all coefficients from the regression model with the overall lowest AIC and BIC.

Discussion

This study presents an investigation of the relationship between trained listener judgments and several acoustic measures of /r/ sounds produced by children receiving treatment for rhotic misarticulation. The current analysis represents a novel contribution in that it is the first to use a large sample of child speech data to systematically select an acoustic measure that best explains clinicians' perceptual ratings of accuracy. These findings have immediate applications in automated scoring algorithms within speech treatment apps, including the staRt biofeedback app described above.

The comparison of the 32 models laid out above indicated that normalised F3–F2 distance was the best acoustic predictor of trained clinician response based on both AIC and BIC. This is consistent with previous research suggesting that a score representing the difference between F3 and F2 should outperform either F3 or F2 taken individually (Flipsen et al., 2001). The model comparison also indicated that a normalised difference score was a better predictor than the derived F3–F2 value. This result is slightly surprising in light of the limitations of the normative data available, but it does accord with previous research (Flipsen et al., 2001). Finally, interactions of normalised F3–F2 with both age and sex were found to improve overall model fit. It is noteworthy that this was true not only for AIC but also for BIC, which more strongly penalises the inclusion of additional parameters. This finding supports previous research by Munson et al. (2010) in finding that listeners may assign different accuracy ratings to the same acoustic signal depending on their understanding of the speaker's age and sex. Therefore, in contexts where the normative data from Lee et al. (1999) are considered appropriate, we advocate for the use of normalised F3–F2 distance in automated detection of /r/ accuracy.

In some cases, researchers may be working with a small sample size, raising concerns about whether a model that includes interactions with age and sex would be sufficiently powered. In such cases, it is reasonable to omit interactions and use normalised F3–F2 alone. In other cases, researchers may have reason to believe that the norms from Lee et al.

(1999) are not a good representation of the sample of interest. For example, participants may come from a region that differs significantly in dialect from those participants represented in Lee et al. (1999). In such a scenario, researchers may want to use the best-performing non-normalised model, which was found to feature F3–F2 distance in interaction with age only. A direction for future research is to collect more representative normative values, including a more geographically diverse sample.

In addition, a fully representative normative sample might include /r/ in phonetic contexts other than the syllabic rhotics measured by Lee et al. (1999). The present study did find significant differences in perceptual rating among the various /r/ contexts even while controlling for direct acoustic measures. Klein, Grigos, McAllister Byun, & Davidson (2012) noted that the shorter duration of consonantal relative to vocalic /r/ makes the task of assigning perceptual ratings more challenging in the former case. Klein et al. hypothesised that adult listeners tend to apply a less stringent standard in rating children's consonantal /r/ tokens due to the limited duration of acoustic information combined with a top-down expectation to hear /r/. This hypothesis could most directly be tested by including duration of the /r/ interval as a predictor in the model. We are currently obtaining durational measures of this data set in order to address this question.

A limitation of the present study is its use of only three expert clinician ratings as the basis for our accuracy measure. Averaging across three raters allowed for ordinal regression, but averaging across a larger number of clinicians would make it possible to treat mean rating as a continuous variable and model it with normal linear regression (McAllister Byun, Harel, Halpin, & Szeredi, 2016). We were limited in this regard by the available materials: our analysis required a large data sample, but due to the high cost of obtaining ratings from trained clinicians, it is rare for researchers who use trained listeners' ratings to exceed the "industry standard" of three raters. An alternative would be to commission a small number of experts to rate child speech productions using a visual analogue scale (McAllister Byun et al., 2016; Schellinger, Munson, & Edwards, 2016), since these measures can reflect more variation in perceived accuracy than aggregated binary ratings allow. We do not currently have gradient ratings on a large-scale corpus level, which points to a valuable direction for future research.

Another caveat is that expert ratings may not reflect how everyday listeners would respond to these speech samples. The goal of this study, for reasons articulated above, was to predict trained clinicians' ratings on the basis of acoustic measures. In other contexts, however, naïve listeners' ratings may be of equal or greater importance. Previous research suggests that clinicians and naïve listeners may attend to different acoustic cues when rating speech accuracy. Klein, Grigos, McAllister Byun, and Davidson (2012) reported that inexperienced listeners' ratings of children's /r/ sounds are more strongly correlated with F3, while ratings from experienced listeners are more strongly correlated with F3–F2. They interpreted this in conjunction with the suggestion from McGowan, Nittrouer, and Manning (2004) that some children lack the articulatory skill necessary to lower F3 to adult-like levels, and, therefore, rely on F2 to signal a rhotic production. In one possible interpretation of the finding in Klein et al. (2012), clinicians' extensive exposure to young speakers might lead to a shift in the relative weighting of F2 and F3 to reflect the greater importance of F3–F2 distance in children's rhotics. Inexperienced listeners, who are exposed primarily to the acoustic cues in adult speech, might be expected to continue to rely more heavily on F3. Munson et al. (2010) also found that listeners' amount of exposure to child speech was a significant predictor of response patterns on a speech rating task. Other research comparing trained versus untrained listeners has found that naïve listeners tend to be more lenient than experts; that is, they are more likely to rate an ambiguous speech production as correct than trained listeners. To generate automated predictions of how children's /r/ sounds would be perceived by naïve listeners, future research could repeat the process from the present study using mean ratings aggregated across non-expert listeners as the dependent variable. As described in McAllister Byun et al. (2015), crowd sourcing could represent an efficient option to obtain naïve listener ratings.

Conclusion

The primary focus of the current study was to select the acoustic measure that best predicted trained clinicians' perceptual ratings of accuracy. We found that a normalised F3–F2 value in interaction with the age and sex of children was the best predictor, a finding that can be implemented into automated scoring algorithms within speech treatment applications. Despite potential limitations with respect to the limited normative sample and differences between trained clinicians and naïve listeners, the current study addresses a need in the field of speech–language pathology to establish efficient and accurate ways of selecting appropriate acoustic goals for individuals enrolled in biofeedback therapy.

Funding

This work was supported by the National Institutes of Health (NIH) under [grant #R01DC013668]. The authors are involved in the development of staRt, an app to provide visual-acoustic biofeedback treatment. They do not receive financial compensation for their role in developing the app.

Declaration of interest

No potential conflict of interest was reported by the authors.

ORCID

Heather Campbell http://orcid.org/0000-0002-6072-5850

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. doi:10.1109/TAC.1974.1100705

American Speech-Language-Hearing Association. (2011). Speech-Language Pathology Medical Review Guidelines. [Online] Retrieved from http://www.asha.org/uploadedFiles/SLP-Medical-Review-Guidelines.pdf.

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer (Version 5.3.84) [Computer program]. Retrieved from http://www.fon.hum.uva.nl/praat/.

Boyce, S., & Espy-Wilson, C.Y. (1997). Coarticulatory stability in American English /r/. *Journal of the Acoustical Society of America*, *101*, 3741–3753. doi:10.1121/1.418333

Campbell, H., & McAllister Byun, T. (2017). Deriving individualised /r/ targets from the acoustics of children's non-rhotic vowels. *Clinical Linguistics & Phonetics*. Advance online publication. doi:10.1080/02699206.2017.1330898

Christensen, R.H.B. (2015). 'ordinal' package in R: Regression Models for Ordinal Data via Cumulative Link (Mixed) Models (Version 2015.1-21). Retrieved from https://cran.r-project.org/web/packages/ordinal/index.html.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York, NY: Routledge.

Dart, S.N. (1991). *Articulatory and acoustic properties of apical and laminal articulations* (Ph.D. dissertation), UCLA Working Papers in Phonetics, Los Angeles, CA.

Delattre, P., & Freeman, D.C. (1968). A dialect study of American r's by x-ray motion picture. *Linguistics*, *6*, 29–68. doi:10.1515/ling.1968.6.44.29

Espy-Wilson, C.Y. (1992). Acoustic measures for linguistic features distinguishing the semivowels/wjrl/in American English. *Journal of the Acoustical Society of America*, *92*, 736–757. doi:10.1121/1.403998

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*, 343–356. doi:10.1121/1.429469

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

Flipsen, P., Shriberg, L.D., Weismer, G., Karlsson, H.B., & McSweeny, J.L. (2001). Acoustic phenotypes for speech-genetics studies: Reference data for residual/ɜ·/distortions. *Clinical Linguistics & Phonetics*, *15*, 603–630. doi:10.1080/02699200110069410

Gibbon, F.E., & Paterson, L. (2006). A survey of speech and language therapists' views on electropalatography therapy outcomes in Scotland. *Child Language Teaching and Therapy*, *22*, 275–292.Retrieved from http://journals.sagepub.com/doi/abs/10.1191/0265659006ct308xx

Hagiwara, R. (1995). *Acoustic realizations of American /r /as produced by women and men*. (Ph.D. dissertation), UCLA, Los Angeles, CA.

Hitchcock, E., Harel, D., & 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*, 283–293. doi:10.1055/s-0035-1562911

Hitchcock, E., McAllister Byun, T., Swartz, M., & Lazarus, R. (in press). *Effectiveness of electropalatography for treating misarticulation of /r/. American Journal of Speech-Language Pathology.*

Idemaru, K., & Holt, L.L. (2013). The developmental trajectory of children's perception and production of English /r/-/l/. *Journal of the Acoustical Society of America*, *133*, 4232–4246. doi:10.1121/1.4802905

Klein, H.B., Grigos, M.I., McAllister Byun, T., & Davidson, L. (2012). The relationship between inexperienced listeners' perceptions and acoustic correlates of children's /r/ productions. *Clinical Linguistics & Phonetics*, *26*, 628–645. doi:10.3109/02699206.2012.682695

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, *105*, 1455–1468. doi:10.1121/1.426686

Lewis, B.A., & Shriberg, L.D. (November, 1994). Life span interrelationships among speech, prosody-voice, and nontraditional phonological measures. Talk presented at the Miniseminar at the American Speech–Language Hearing Association Convention, New Orleans, LA.

McAllister Byun, T., & Campbell, H. (2016). Differential effects of visual-acoustic biofeedback intervention for residual speech errors. *Frontiers in Human Neuroscience*, *10*, 1–17. doi:10.3389/fnhum.2016.00567

McAllister Byun, T., Campbell, H., Carey, H., Liang, W., Park, T.H., & Svirsky, M. (2017). Enhancing intervention for residual rhotic errors via app-delivered biofeedback: A case study. *Journal of Speech, Language, and Hearing Research*, *60*, 1810–1817. doi:10.1044/2017_JSLHR-S-16-0248

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. doi: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. doi:10.1016/j.jcomdis.2016.07.001

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*, 207–221. doi:10.1044/1058-0360(2012/11-0083)

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*, 2116–2130. doi:10.1044/2014_JSLHR-S-14-0034

McAllister Byun, T., & Tiede, M. (2017). Perception-production relations in later development of American English rhotics. *PLoS One*, *12*.10.1371/journal.pone.0172022

McGowan, R.S., Nittrouer, S., & Manning, C.J. (2004). Development of [j] in young, midwestern, American children. *Journal of the Acoustical Society of America*, *115*, 871. doi:10.1121/1.1642624

Munson, B., Edwards, J., Schellinger, S.K., Beckman, M.E., & Meyer, M.K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clinical Linguistics & Phonetics*, *24*, 245–260. doi:10.3109/02699200903532524

Polka, L., & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate /r/ and /l/. *Journal of the Acoustical Society of America*, *78*, 1187–1197. doi:10.1121/1.392887

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*, 2102–2115. doi:10.1044/2014_JSLHR-S-14-0031

RStudio Team. (2017). *RStudio: integrated development for R (Version 0.99.879) [Computer program]*. Boston, MA: RStudio, Inc. Retrieved from https://www.rstudio.com/products/rstudio/.

Ruscello, D.M. (1995). Visual feedback in treatment of residual phonological disorders. *Journal of Communication Disorders*, *28*, 279–302. doi:10.1016/0021-9924(95)00058-X

Schellinger, S.K., Munson, B., & Edwards, J. (2016). Gradient perception of children's productions of /s/ and /θ/: A comparative study of rating methods. *Clinical Linguistics & Phonetics*, *31*, 80–103. doi:10.1080/02699206.2016.1205665

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Shriberg, L.D., Flipsen, P., Karlsson, H.B., & McSweeny, J.L. (2001). Acoustic phenotypes for speech-genetics studies: An acoustic marker for residual /ɝ/ distortions. *Clinical Linguistics & Phonetics*, *15*, 631–650. doi:10.1080/02699200110069429

Shuster, L.I. (1998). The perception of correctly and incorrectly produced /r/. *Journal of Speech, Language, and Hearing Research*, *41*, 941–950. doi:10.1044/jslhr.4104.941

Shuster, L.I., Ruscello, D.M., & Toth, A.R. (1995). The use of visual feedback to elicit correct /r/. *American Journal of Speech-Language Pathology*, *4*, 37–44. doi:10.1044/1058-0360.0402.37

Smit, A.B., Hand, L., Freilinger, J.J., Bernthal, J.E., & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders*, *55*, 779–798. doi:10.1044/jshd.5504.779

Vallabha, G., & Tuller, B. (2004). Choice of filter order in LPC analysis of vowels. *From Sound to Sense*, *50*, B148–B163. Retrieved from http://www.rle.mit.edu/soundtosense/conference/pdfs/fulltext/Saturday%20Posters/SB-Vallabha-STS.pdf

Wickham, H. (2016). 'tidyverse' packages in R: Easily Install and Load 'Tidyverse' Packages (Version 1.0.0). Retrieved from https://github.com/hadley/tidyverse

Zimmerman, I.R., Steiner, V.G., & Pond, R.E. (2002). *Preschool Language Scale*, 4th ed. *(PLS-4) Spanish* Ed. San Antonio, TX: Psychological Corporation.

Appendix: Words included in the data set

bar	deer	frog	pear	scrap	string
bear	door	fruit	pray	scrape	stroll
bore	draw	fur	purr	scratch	strong
bread	dream	gear	rain	scream	tear
broom	drum	grape	ray	screw	tiger
brown	fair	green	read	scroll	train
car	far	grew	red	scrub	trash
core	fear	group	rip	shower	trip
crab	four	grow	rock	sir	troll
crack	fray	hair	root	sore	trout
crow	friend	hammer	row	star	truck
dear	fries	her	run	straw	water