Listener Perceptions and Acoustic Characteristics of Children's Rhotic Vowels

Crystal Baines, Paula Frigon, Rebecca Ritz

Supervisors: Karen Pollock, Hyunju Chung

Short Header: Acoustics and Perception of Children's Rhotics

ABSTRACT

Many children have difficulty producing /r/ sounds. In this study we are interested in children's production of vocalic forms of /r/, otherwise known as rhotic vowels. Rhotic vowels can be monophthongs (/ɝ/and/ɚ/), as in words like *stir* and *tiger*, or diphthongs (e.g., /ɑ͡ɚ/ or /ɪ͡ɚ/*),* as in *car* and *ear.* The current study investigates the relationship between inexperienced listeners' perceptions of children's rhotic vowel productions and the acoustic correlates of these productions. Stimuli consisted of 198 vowels extracted from elicited single words produced by children aged three to five. The tokens included productions that were phonetically transcribed as fully rhotic (correct), partially or fully derhoticized (incorrect), or non-rhotic (control), by a trained listener. The tokens transcribed as partially or fully derhoticized were further categorized based on the type of substitution error made by the speaker. Twenty inexperienced listeners with no experience in phonetic transcription or acoustic analysis were asked to judge rhoticity of token sounds by clicking on one of four choices: definitely r, somewhat r-like, not very r-like, definitely not r. Acoustic properties of these rhotic vowels (F3-F2 and duration) were measured and compared to the phonetic transcriptions and inexperienced listener ratings to determine whether listeners perceive a difference between control vowels, correct, and incorrect productions of rhotic vowels and if this difference is related to acoustic correlates.

BACKGROUND

Previous findings

General Developmental Patterns. The development of the /r/ phoneme is a complex process for young children, including those with typically developing speech and those with speech sound disorders. This sound can be produced by different tongue configurations (i.e. bunched /r/ or retroflex /r/) and it varies depending on the vocalic context in which it occurs. McGowan, Nittrouer, and Manning (2004) found that toddlers acquire the ability to produce the /r/ phoneme at different ages depending on its position within a word. The complex and variable articulation of the /r/ phoneme coupled with the limitations in young children's articulatory control is the likely cause of this difference.

Vowels are an integral but variable and complex component of human speech, a component which takes years to master. Lee, Potamianos, and Naravanan (1999) looked at the vowel acoustic measurements of ten monophthong vowels (bead, bit, bet, bat, pot, ball, but, put, boot, bird) as a function of the age and gender of the speaker and found that within and between-subjects, spectral variability decreased with age. This decrease was particularly significant in the age range from five to nine and also from age ten to fourteen, suggesting that "children younger than ten years have not fully established their optimal or stable articulatory targets for vowels" (p.1465). Although not optimal, vowel productions by both typically developing children and children with speech sound disorders are perceived as correct (but childlike) approximations of the target vowel by 2 to 3 years of age. This suggests that vowels are acquired early, but are still being fine-tuned until the age of ten (Pollock & Berni, 2003).

Rhotic vowel development in children is a particular point of interest as it combines the developmental difficulties of vowel production with that of the /r/ phoneme. Rhotic vowels are typically among the last vowels to be acquired by children (Stoel-Gammon & Pollock, 2008). They change based on the context in which they occur within the word (i.e. the phonetic environment), resulting in many variations of vocal tract configurations. In a study that analyzed vowel errors in phonologically disordered children, Pollock and Keiser (1990) found rhotic vowels to be significantly less indicative of speech sound disorders than non-rhotic vowels. This further supports the idea that it is the physical limitations of the vocal tract rather than the general developmental curve that is the cause of the difficulty that children have with these sounds. In the aforementioned study by Lee et al. (1999), the researchers included one rhotic vowel, the common /ɝ/, in the word 'bird'. They found that, of all the monophthongs they looked at, the /ɝ/ sound was held for the longest length of time. This suggests that though young children might not be able to physically produce the /r/, they are aware of the presence of the /r/ component in rhotic vowels.

Acoustics and Perception. The variability in the articulation of the /r/ phoneme makes its measurement challenging. Many researchers have turned to acoustics as the best way to measure its properties. Regardless of manner of articulation (i.e. bunched vs. retroflex), the acoustic cue for /r/ is a steep downwards slope in the trajectory of the third formant frequency (F3). Furthermore, Guenther, Espy-Wilson, Boyce, Matthies, Zandipour, and Perkell (1999) found that within individual speakers, different articulatory configurations used to produce /r/ in different contexts had systematic tradeoffs that act to reduce acoustic variability, "…thus

allowing relatively large contextual variations in vocal tract shape for /r/ without seriously degrading the primary acoustic cue" (p. 2854).

In the area of vowel measurement, F1 and F2 have long been used to map out acoustic vowel space (Potter & Peterson, 1948). In the frequently cited study of the acoustics and perception of vowels by Peterson and Barney (1952), the researchers found that acoustic measurements of vowel sounds are not distributed in a random order but rather there is a strong relationship between the speaker's intended vowel and the formant frequency pattern. The listening study showed that the identification of vowels was higher for some than others; the vowels created in the "limit" positions of the articulatory mechanisms were more easily identified (i.e. the vowels where the tongue is in the highest or lowest and farthest front or back position). An article by Hillenbrand, Getty, Clark, and Wheeler (1995) states that the Peterson and Barney (1952) article is limited because no measurements of dynamic properties (e.g. vowel duration, pattern of spectral change over time) were made. In their study, Hillenbrand et al. (1995) found that vowels were more accurately classified when duration measures were considered along with spectral data.

The two aforementioned studies on vowel acoustics included the rhotic vowel /ɝ/. Peterson and Barney (1952) found that of all the vowels they looked at, /ɝ/ had a relatively high intelligibility among listeners. They also found that the /ɝ/ sound was easily distinguished from all the other vowels if the third formant frequency (F3) was used. The proximity of F3 to F2 for the /ɝ/ vowel as compared to non-rhotic monophthongs is also commented on in both studies. Findings in Flipsen, Shriberg, Weismer, Karlsson, and Mcsweeny (2001) indicate that in

adult speakers, /ɝ/ productions are best characterized acoustically by either F3 minus F2 or F3 divided by F2, with a smaller distance between F3 and F2 being indicative of rhoticity.

In a study by Klein, Grigos, Byun and Davidson (2012), the researchers recorded five male 6 and 7 year old children's imitations of a phrase ("say ___ again") that included "consonantal /r/" (prevocalic /r/ occurring in CVC syllables where the post vocalic consonant was a voiceless plosive) and monophthong "vocalic /r/" or rhotic vowels (monosyllables with stressed /ɝ/ preceding or following voiced plosives). They compared inexperienced listeners' perceptions of these sounds to those of two expert clinicians. In this study, inexperienced listeners were speech-language pathology students with previous coursework completed in the areas of acoustic and articulatory phonetics, but no clinical experience transcribing children's speech. Heights of F2 and F3 were also measured to calculate F3-F2 distance and acoustic properties of consonantal and rhotic vowels were compared. They found that inexperienced listeners had a lower sensitivity and specificity than expert clinicians when attempting to differentiate between "intermediate" and "fully correct" /r/ productions. This suggests that clinical experience may have an effect on the ability to hear whether an /r/ production is correct. It was noted, however, that the ratings for rhotic vowel monophthongs were more consistent than the ratings for consonantal /r/ suggesting that listeners were better able to perceive differences in rhoticity in rhotic vowels than consonantal /r/.

The current study

Based on Klein et al. (2012), inexperienced listeners were better able to perceive differences in rhoticity in vocalic /r/ monophthongs than in consonantal /r/. The current study

examines acoustic properties and listener perceptions of rhoticity on the breadth of rhotic diphthong and monophthong vowels in the Western Canadian English dialect.

Listeners with training in linguistics are able to describe young children's errors in rhotic vowel production through phonetic transcription and acoustic analysis; however, not everyone has access to these descriptive tools. In addition, transcription reliability is typically much lower for rhotic vowels than for non-rhotic vowels (Pollock & Berni, 2003), suggesting that even trained listeners have difficulty determining the accuracy of rhotic vowels. Though the inexperienced listeners in Klein et al. (2012) had no clinical experience transcribing young children's speech, they had completed recent coursework in the areas of articulatory and acoustic phonetics, as well as phonology. In the current study, "inexperienced listeners" refers to listeners that do not have training or experience in phonetic transcription or acoustic analyses. "Experienced listeners" refers to listeners with said training in linguistics as well as experience in transcribing young children's speech.

The purpose of the current study was to examine the relationship between inexperienced listeners' ratings of children's production of rhotic vowels (diphthongs and monophthongs) and the acoustic properties of these vowels. The following research questions were investigated:

1. Is there a difference between listener ratings for productions of correct rhotic vowels, incorrect rhotic vowels and non-rhotic control vowels?
2. Do listener ratings distinguish between different types of errors in the production of rhotic vowels? Are some error categories rated as more r-like than other categories?
3. Is there a difference in terms of acoustic measurements (F3-F2, duration) between correct /r/ productions, incorrect /r/ productions, and control vowels?

4. What is the relationship between inexperienced listeners' ratings and acoustic traits of rhotic vowel production in children's speech? In other words, do inexperienced listeners rate vowels with a smaller distance in F3-F2 as more like rhotic vowels than vowels with a larger distance difference between F3 and F2?

As many children have difficulty with the pronunciation of rhotic vowels, their productions of target rhotic vowels are often derhoticized. This means that in lieu of the targeted rhotic vowel, the child's production does not have a rhotic component (e.g. for the word "car", the child says /kɑ/ instead of /kɑ͡ɚ/). These errors could involve: deleting the rhoticized vowel from a diphthong, substituting a rhotic vowel with a non-rhotic neutral vowel (e.g. /ə/ for /ɚ/), or substituting a rhotic vowel with a non-rhotic rounded vowel (e.g. /o/ for /ɚ/). An experienced listener can use phonetic transcription to describe the nature of derhoticization and may use an appropriate diacritic to mark the degree of derhoticization. Based on previous findings, we would expect vowels transcribed as derhoticized to have a larger difference between F3 and F2, and vowels transcribed as rhotic to have a smaller difference between F3 and F2 (Flipsen et al., 2001). For the tokens transcribed as derhoticized, the children were likely trying to approximate rhotic vowels. As such, we might expect derhoticized vowels to have a smaller difference between F3-F2 than control non-rhotic vowels.

Since rhotic vowels are present in the Western Canadian dialect of English, we would expect inexperienced listeners who are native speakers of this dialect to be able to perceive a difference between rhoticized and derhoticized rhotic vowels, given their experience and Klein et al. (2012)'s findings that listeners were more consistent in rating vocalic rather than

consonantal /r/. We might also expect inexperienced listeners to perceive a difference between target non-rhotic vowels and derhoticized rhotic vowels.

METHODS

Participants

20 participants were recruited for this study. There were 8 females and 12 males with an average age of 24.4 years (range: 19 to 28). All participants were native Western Canadian English speakers with no history of hearing impairments. 5 participants had beginner level knowledge of another language (French: 4; Italian and French: 1). 19 participants were born and raised in Western Canada (including British Columbia and Alberta). One participant was born in Ontario and moved to Western Canada when he was 5 years old (18 years in Western Canada). In terms of education, 5 participants had an undergraduate degree, 8 had a college diploma, 6 had a high school diploma and 1 participant did not have a high school diploma.

All participants were judged to be "inexperienced listeners" in that none of the participants had studied or worked in the area of linguistics for at least five years. Only one participant had experience studying linguistics more than five years ago. No participants had children, and 5 participants had experience with children through work or relatives. 4 participants had received speech and language services in the past and 3 had a relative who received speech and language services. Participants were recruited by posters and advertisements put up throughout Corbett Hall at the University of Alberta and on social networks. Participants were given a five dollar coffee gift card for their participation.

Recorded Stimuli

Segments were taken from recorded samples of both male and female children's speech that were elicited as part of a larger study on the acquisition of rhotic vowels in children with and without speech sound disorders led by Drs. Karen Pollock and Hyunju Chung. Samples chosen for the current study were taken from real word productions by children between the ages of 3 and 5 years, including children who had typically developing speech (TD) and those with speech sound disorders (SSD). Audio recordings of children's productions of the target and control words were used (Table 1). Words were chosen from the recorded samples to include diphthong and monophthong rhotic vowels (targets) and a corresponding set of homorganic non-rhotic vowels (controls). All rhotic vowels were in open syllables in word final position, and control vowels were in closed syllables of CVC word structure, but were controlled for preceding consonant place of articulation (e.g., bilabial /p/ in "zipper" and /m/ in "mud"; alveolar /d/ in "door" and /t/ in "toys"). An experienced listener (a speech-language pathology student) who had training and experience transcribing young children's speech completed narrow phonetic transcriptions of all word productions for all child participants. A second listener (also a speech-language pathology student experienced in transcribing children's speech) independently transcribed 22% of the samples. Inter-transcriber reliability for overall vowel transcription was 98%. For rhotic vowels, percent agreement averaged 81% for broad transcription and 77% for narrow transcription.

Table 1: Words from which tokens were extracted.

Rhotic vowels	Non-rhotic control vowels
Ear	Sock
Tiger	Pig
Zipper	Bed
Bear	Mud
Door	Toys
Star	
Store	
Her	

Target vowels were segmented out from the single word recordings based on their spectrograms and corresponding speech waveforms using the computer program Praat (Boersma & Weenink, 2012) by undergraduate volunteers studying Linguistics. The start of each target vowel was marked at the beginning of a clear glottal pulse in the spectrogram and the upswing of the corresponding waveform. The end of the vowel was marked at the point in the spectrogram where F2 began to fade (i.e. white space was prominent in the spectrogram) and the downswing of the corresponding waveform. The segments of the audio clips corresponding to these target rhotic vowel or control non-rhotic vowel segments were extracted. The duration of the tokens was recorded after they were segmented.

The segmented vowels were used in isolation, as opposed to within the whole word, as it was suspected that if the listeners knew the stimulus word, it could affect their rhoticity ratings. For example, a listener might recognize that the target of /kɑ/ was likely 'car' /kɑ͡ɚ/ as /kɑ/ is not a word in English. It was suspected that if listeners could recognize the target word as a word with a rhotic vowel, they would potentially be biased to rate this vowel as higher in rhoticity than if they did not know whether a rhotic vowel was present in the target word or

not. The opposite was also thought possible in that listeners might rate such a token as lower in rhoticity as they would recognize it as an error.

Tokens were selected to include fully rhoticized, partially derhoticized, and non-rhoticized productions of target vowels, based on the transcription with reference to a spectrogram in Praat (Boersma & Weenink, 2008). Vowels coded as partially derhoticized were further categorized based on the rhotic vowel error categories proposed in Pollock (2002) and Pollock (2012) and used in the larger study by Chung, Farr, and Pollock (2014). A total of 198 tokens were segmented and categorized by error type (Table 2).

Table 2: Codes to differentiate children's approximations of rhotic vowels

Code	Description	Examples
DerhotMCV	Derhoticized to mid central vowel	/ɪ͡ɚ/ → [ɪ͡ə], /hɝ/ → [hʌ] /zɪpɚ/ → [zɪpə]
DerhotBRV	Derhoticized to back rounded vowel	/tá͡ɪgɚ/ → [tá͡ɪgo], /hɝ/ → [hɔ], /bɛ́͡ɚ/ → [bɛ́͡ʊ]
PartDerhot	Partially derhoticized	/tá͡ɪgɚ/ → [tá͡ɪgɚ], /ɪ͡ɚ/ → [ɪg̬]
OtherDerhot	Derhoticization to another vowel (i.e. not mid central or back rounded)	/tá͡ɪgɚ/ →[tá͡ɪgɪ], /stɑ́͡ɚ/ → [stɑ́͡ɛ]
RDR	Rhotic diphthong reduction – rhotic element deleted but pre-rhotic vowel target maintained	/stɑ́͡ɚ/ → [stɑ], /bɛ́͡ɚ/ → [bɛ]
RDR+	Rhotic diphthong reduction with vowel change – rhotic element deleted and pre-rhotic vowel target changed	/stɔ́͡ɚ/ → [sta]

Listener judgment procedures

Experimental Environment. The perception experiment was programmed using Praat (Boersma & Weenink, 2008). Listeners sat at a computer in a quiet room with noise cancelling

headphones. Selections were made by clicking a mouse on one of four options also displayed in Praat. Listeners first did a training block where they were asked to judge 10 tokens and were able to ask questions to clarify that they understood the task. These 10 training tokens included a mix of correct, incorrect, and control vowels. Once the training block was completed and the participants displayed understanding of the task, the experimental block was initiated.

Training Block (10 tokens). During the training block, listeners were made aware of what a rhotic vowel is and were instructed that they would hear a series of vowels that may or may not include a rhotic component. Ten tokens were provided in the training phase to ensure that participants understood the task. Participants only heard each token once but were able to ask questions between tokens. They were asked to listen to the token and rate the sound that they heard using the following four-point scale:

Figure 1. Inexperienced Listener Rating Scale

Experimental Block (198 Tokens). After completion of the training block, participants listened to 198 tokens consisting of 80 correct, 78 incorrect, and 40 control vowels presented in a randomized order. The participants were instructed to rate the level of rhoticity on the same four point scale as in the training phase. Participants heard each token once and were not able to ask questions. An optional two minute break was offered after 99 tokens.

Acoustic Measures

The distance between F3 and F2 was measured by two authors using Praat. For each token, the formant settings, with 16 LPC coefficient and 25 ms window, were adjusted to best

align with the spectral energy on the spectrogram (3-5 formant tracings displayed). Values were taken from areas that had energy in the spectrogram. The authors examined the spectrogram of each token and chose the point corresponding to the lowest F3 value from the Praat generated formant tracing in the area that corresponded to the rhotic portion of the target. F3 values were taken from stable points in the spectrogram to reduce the risk of choosing outliers. If the lowest point in the tracing appeared to be an outlier, an adjacent point with a better fit was selected. The corresponding F2 value was also recorded. A total of 190 tokens were analyzed. Nine tokens were discarded due to poor recording quality.

The F2 and F3 values for 30% of the tokens (60 tokens) were rated by both raters measured by both authors. The authors came to a consensus about tokens that differed by 100 Hz or more by re-analyzing the tokens together or deciding to discard them. 4 tokens were discarded due to poor recording quality. 80% of the kept tokens (45/56) reached the reliability criterion (i.e. F3-F2 within 100 Hz). The authors were able to reach consensus on the 11 remaining items by re-measuring F3 and F2 values together. These values were included in the final analyses.

RESULTS

Listener Ratings

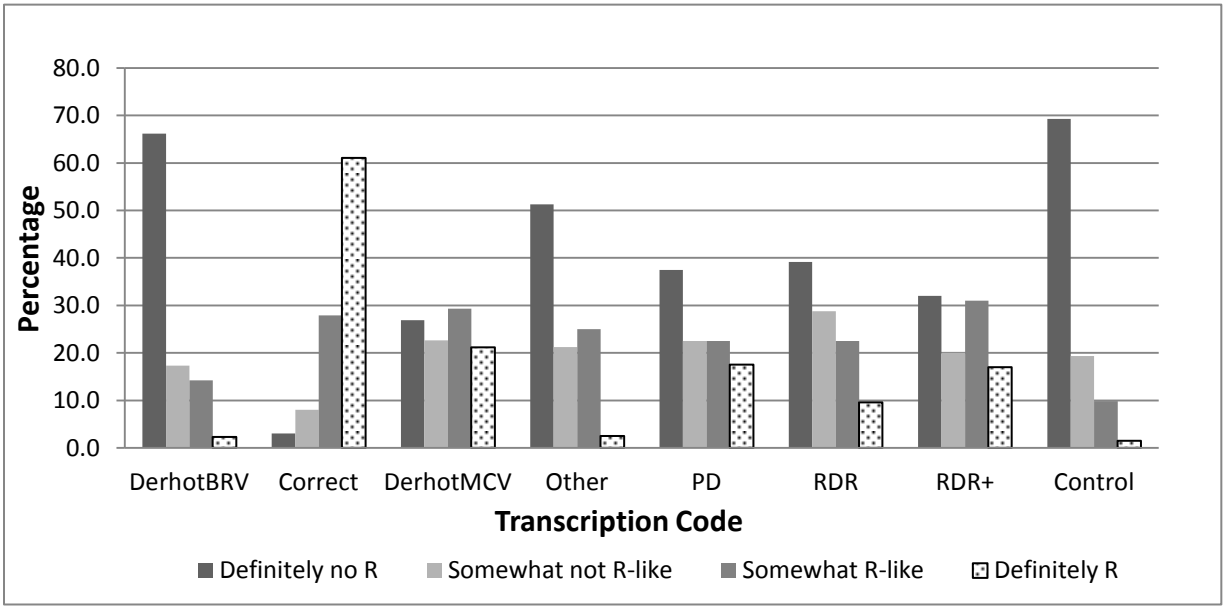
Correct vs. Incorrect Tokens. Listener ratings were translated to a numerical scale as follows for the purposes of statistical analysis:

Table 3: Perceptual listener ratings and corresponding numerical rating

"Definitely no 'r'"	1	"Somewhat not 'r-like"	2	"Somewhat 'r-like'"	3	"Definitely 'r'"	4

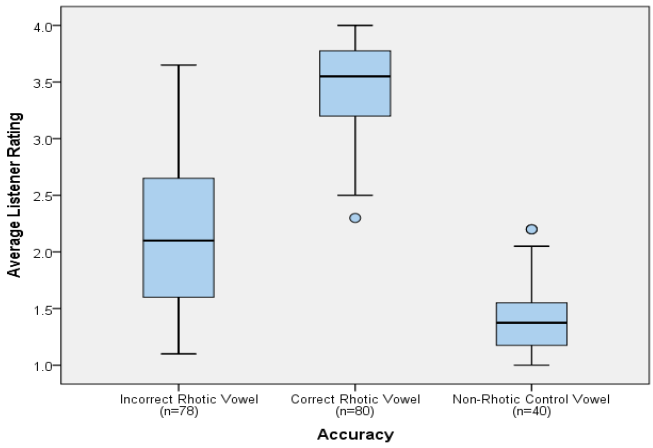
Listener Rating Distribution. Out of 198 tokens x 20 inexperienced listeners' judgments, "definitely 'r'" was selected 30.9% of the time. "Definitely no 'r'" was selected 29.9% of the time, "Somewhat 'r-like'" was selected 23.3% of the time, and "somewhat not 'r-like'" was selected 16.0% of the time. Neither "somewhat not 'r-like'" nor "somewhat 'r-like" received the highest percentage of listener ratings for any transcription code; however, when combined as an "in-between" category, they were chosen 39.3% of the time (the most often). Across the 20 participants, "definitely 'r'" was selected 61.1% of the time out of the 80 correct tokens, "definitely 'no-r'" was selected 69.3% of the time out of the 40 control tokens. The percentage of inexperienced listener ratings by error code for incorrect tokens can be seen in Figure 2 below. For all error codes except for *DerhotMCV*, "definitely no 'r'" was selected most of the time, whereas for DerhotMCV, "somewhat 'r-like'" was the most frequent response, chose 23.9% of the time across all participants out of 42 tokens.

Figure 2. Percentage of Listener Rating by Error Code



The distribution of listener ratings by stimulus type is presented in Figure 3 (see below). There is no overlap between correct rhotic productions and non-rhotic controls. The range of ratings for incorrect rhotic productions is extensive, but the majority of ratings fall between 1.51 and 2.86 (+/- 1 SD). A between-subjects univariate analysis of variance (ANOVA) was conducted to see the effect of stimulus type (correct rhotic, incorrect rhotic, or non-rhotic control) on average listener ratings. The effect was significant at the $p \leq 0.05$ level ($F=246.136$, $p<0.0005$). Tamhane's post hoc test of multiple comparisons revealed that there were significant differences between the average rating of control tokens and both the correct and incorrect rhotic vowel tokens ($p<0.0005$ for all comparisons).

Figure 3. Average listener rating by stimulus type

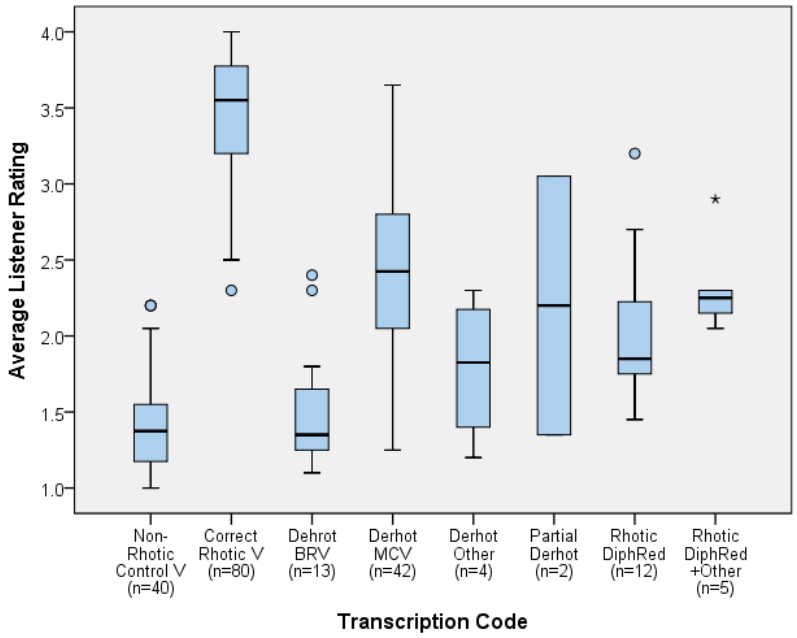


Note: The horizontal line represents the median, the box represents the interquartile range, the upper and lower whiskers represent the 95th and 5th percentiles, and circles represent outliers. The Y-axis represents the average listener rating. For each token, the average rating was calculated by taking the mean rating of 20 participants. The average ratings for each token were then used to determine the average listener rating per group (correct, incorrect, and control).

Listener ratings by categories. The distribution of listener ratings for each transcription code is included in Figure 4. The transcription code *Correct* had the highest average rating (3.5) and the transcription code *Control* had the lowest average listener rating (1.4). All error

categories had average ratings that lie somewhere between *Correct* and *Control*. The category of *DerhotMCV* had the greatest variability.

Figure 4. Average listener ratings by transcription code



Note: The horizontal line represents the median, the box represents the interquartile range, the upper and lower whiskers represent the 95th and 5th percentiles, and circles represent outliers.

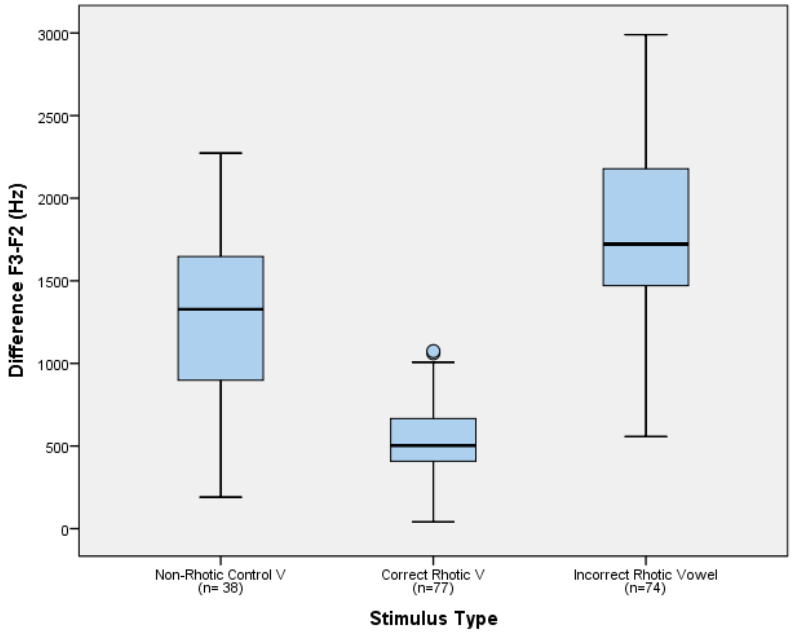
A between-subjects univariate analysis of variance (ANOVA) was also conducted to see if untrained listeners perceived differences in rhoticity for different types of errors based on transcription code (C*ontrol, Correct, DerhotBRV, DerhotMCV, OtherDerhot, PartDerhot, RDR, RDR+*). The effect of transcription code on average listener rating was significant at the $p \leq 0.05$ level ($F=90.815$, $p < 0.0005$). Tamhane's post hoc test of multiple comparisons showed that the average listener rating for tokens with the following transcription codes were significantly different from tokens transcribed as *Correct* at the $p \leq 0.05$ level: *Control* ($p<0.0005$), *DerhotBRV* ($p<0.0005$), *DerhotMCV* ($p<0.0005$), *RDR* ($p<0.0005$), *RDR+* ($p=0.027$). The categories of *OtherDerhot* and *PartDerhot* did not reach significance.

The only error category for which the average listener rating was significantly different than the average listener rating for controls was *DerhotMCV* ($p<0.0005$) though *RDR* and *RDR+* approached significance ($p=0.057$, $p=0.062$). *DerhotBRV* was not significantly different ($p=0.999$) from the control group. The categories *OtherDerhot* ($p=1.000$) and *PartDerhot* ($p=1.000$) were also not significantly different than controls.

Acoustic Measures

F3-F2. There was more overlap between the stimulus categories of correct and incorrect rhotic vowels and non-rhotic control vowels (see Figure 5). A between-subjects univariate analysis of variance (ANOVA) was conducted to see the effect of stimulus type (correct rhotic, incorrect rhotic, or non-rhotic control) on acoustic measures. The effect was significant at the $p \leq 0.05$ level ($F=188.56$, $p<0.0005$). Posthoc comparisons revealed that there were significant differences between all comparisons ($p<0.0005$).

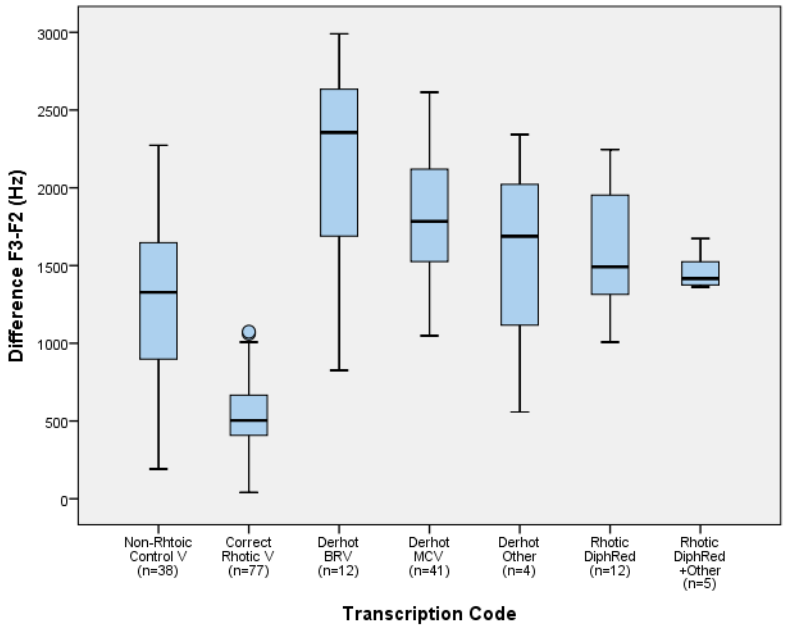
Figure 5. Distribution of F3-F2 values by stimulus type



Note: The horizontal line represents the median, the box represents the interquartile range, the upper and lower whiskers represent the 95th and 5th percentiles, and circles represent outliers.

The distribution of F3-F2 values by different transcription error codes is illustrated in Figure 6. The *Control* group had the lowest average F3-F2 value (534.99 Hz) and the *DerhotBRV* group had the highest average F3-F2 value (2186.83 Hz). The average F3-F2 value for all tokens was 1182.50 Hz (standard deviation=697.98 Hz). When F3-F2 was compared between transcription categories, a significant difference was found between *Control* and *Correct* ($p<0.0005$), *Control* and *DerhotBRV* ($p=0.008$), *Control* and *DerhotMCV* ($p<0.0005$), *Correct* and *DerhotBRV* ($p<0.0005$), *Correct* and *DerhotBRV* ($p<0.0005$), *Corrects* and *RDR* ($p<0.0005$), *Correct* and *RDR+* ($p<0.0005$) and *DerhotMCV* and *RDR+* ($p=0.019$). There was no significant difference between *OtherDerhot* and any other category.

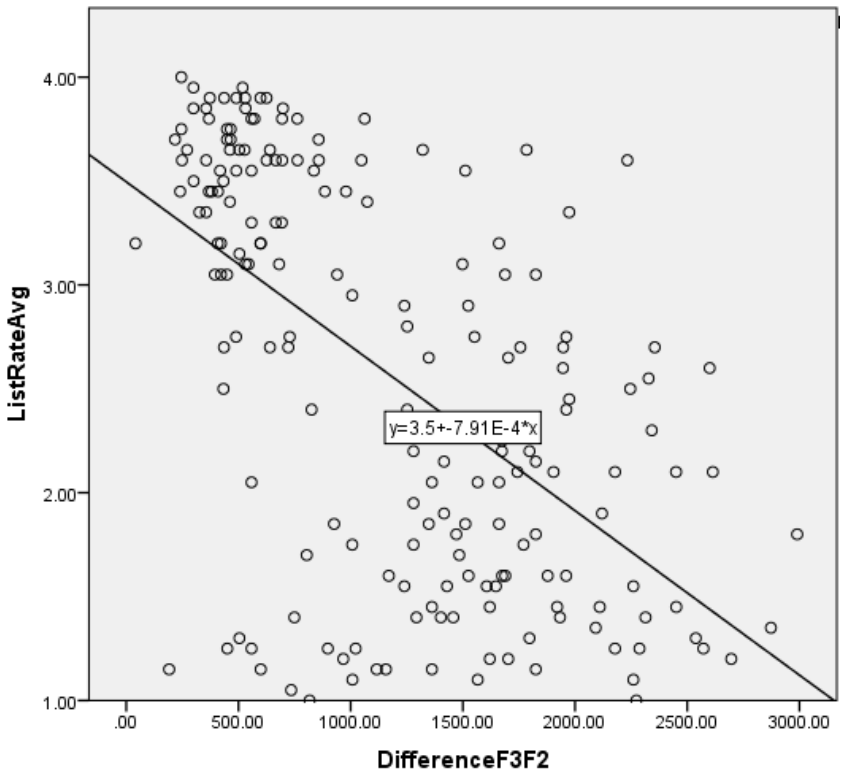
Figure 6. Transcription codes and difference between F2 and F3



Note: The horizontal line represents the median, the box represents the interquartile range, the upper and lower whiskers represent the 95th and 5th percentiles, and circles represent outliers.

There was a strong negative correlation between the average listener rating and the difference between F3-F2 ($r(187)=-0.578$, $p<0.0005$), as shown in Figure 7.

Figure 7. Relationship between the averaged listener ratings and acoustic measures (F3-F2 in Hz).



Note: Each circle represents the average listener rating and corresponding acoustic measure (F3-F2) for a single token.

Duration. The tokens used in this study were all of relatively short duration, and included stressed rhotic vowels, unstressed rhotic vowels, and diphthong rhotic vowels. As such, it was suspected that the duration of the tokens would vary depending on vowel type and in turn could affect listener's ratings. A between subjects ANOVA compared the average duration of tokens in the transcription categories. There was a significant difference between the

duration of *Correct* tokens and the duration of *RDR* tokens (p=0.009). There was no significant difference between any other categories. When the duration of the *Control* and *Correct* groups was compared to all *Incorrect* tokens, there was a significant difference between the *Incorrect* group and the *Correct* group (p=0.001). Pearson's r was calculated to determine if a correlation was present between listener ratings and token duration and no significant correlation was found (r(196)=0.045, p=0.529).

Table 4: Duration Measurements

		Average Duration (s)	Standard Deviation (s)
Non rhotic (control) vs. rhotic target (correct and incorrect)	Non-rhotic target (40 tokens)	0.428172s	0.373751s
	Rhotic Target (158 tokens)	0.279921s	0.334290 s
Monophthong vs. diphthong for rhotic targets (correct and incorrect)	Monophthong (58 tokens)	0.373751s	0.279921 s
	Diphthong (100 tokens)	0.334290s	0.428172 s
Average		0.365778992 s	0.365778992 s

DISCUSSION

Listener Perceptions

It was expected that inexperienced listeners would be able to perceive differences in vowel rhoticity given that rhotic vowels are present in the Western Canadian dialect of English. The results showed that there was a significant difference between the average ratings of correct rhotic productions and incorrect rhotic productions. Overall, these results suggest that listeners were able to complete task and were able to accurately perceive differences in rhoticity when listening to segmented vowels. Vowels transcribed as rhotic vowels had an

average listener rating closer to "definitely r" and vowels transcribed as derhoticized had an average listener rating closer to "definitely no 'r'". Furthermore, the results showed that inexperienced listeners were able to differentiate between correct rhotic vowel productions, incorrect productions, and control vowels. In other words, they perceived errors in rhotic vowel productions by young children as different than non-rhotic target vowels even without word context.

When each error code was compared to correct productions, each error group was rated as significantly different from correct productions of rhotic vowels except for *OtherDerhot* and *PartDerhot*. This was likely due to the small sample size of these two groups (n= 4, and n=2 respectively). These results suggest that listeners were able to identify accurate productions of rhotic vowels compared to each different error type.

All error categories had average ratings between those of *Correct* and *Control* groups suggesting that inexperienced listeners perceived tokens in these categories as in between correct rhotic vowels and non-rhotic target vowels. The category of *DerhotMCV* had the highest standard deviation (SD=0.663) suggesting that listeners had the most difficult time assigning a rating for these errors and that this error type may best reflect an "in-between" perceptual category. *DerhotMCV* was also the only error category significantly different than controls. This is further evidence that vowels derhoticized to a mid-central vowel are perceived as in-between correct rhotic vowels and control vowels. This suggests that there are elements of MCV that are perceived as 'r-like' which is not entirely unexpected given the similar place of articulation for both types of sounds and the fact that mid-central vowels are common substitutions for rhotic vowels in young children's speech. A possible explanation could be the tendency of some

dialects of English to substitute a mid-central vowel for rhotic vowels in some contexts resulting in some types of rhotic vowel substitutions being perceived as more acceptable or closer to 'correct' rhotic vowels. The categories *OtherDerhot* and *PartDerhot* were also not significantly different from *Controls*, however it is likely that this was again due to the small sample size of these groups.

The results overall show that inexperienced listeners were able to a use a descriptive scale effectively to describe different levels of rhoticity perceived. The scale utilized in this study consisted of four points ranging from "definitely no 'r'" to "definitely 'r'". Though significant differences were found in inexperienced listeners' ratings of control, incorrect and correct productions of rhotic vowels, neither in-between category ("somewhat not 'r-like'" and "somewhat 'r-like") was selected the most often to rate rhoticity. Inexperienced listeners were least likely to rate a token as "somewhat not 'r-like'" suggesting this was not a very meaningful category to them. However, when the two middle categories, "somewhat not 'r-like'" and "somewhat 'r-like", were collapsed, this "in-between" category was chosen the most often. Overall, these results suggest that inexperienced listeners perceive an intermediate category of vowels between rhotic vowels and non-rhotic target vowels (controls); however they do not distinguish between two different in-between categories. This also suggests the possibility that children are not just substituting non-rhotic vowels for rhotic vowels and that their attempts at rhotic targets are different from their non-rhotic correct productions. To further investigate this possibility, future researchers could examine listener ratings and acoustic properties of different non-rhotic vowels and compare them to incorrect r productions that were transcribed as the same vowel (e.g. [ɛ] in "bed" vs. [bɛ] for "bear"). If a difference existed, it would suggest

that the child has separate representations or categories for the incorrect vocalic r productions and the non-rhotic vowels, but is not yet able to produce the rhotic quality.

Acoustics

As expected, the *Correct* category had the lowest F3-F2 values due to the decrease in F3 in the /r/ part of the rhotic vowel. The *Control* category had the second lowest average values. This low average in the *Control* category could be caused by to the /ɔɪ/ diphthong in "toy" which has lower F3-F2 values due to an increase in the F2 value at the end of the diphthong. The *Incorrect* category had the highest F3-F2 values. Significant differences were noted when comparing *Correct* vs. *Control* as well as *Correct* vs. *Incorrect* with the exception of 'other'. No significant difference was found between the *OtherDerhot* category and any other category. This may be due to the small sample size or the contents of the category may fit more appropriately in other categories rather than in a category of their own.

The possibility of a medial category corresponding to values between *Correct* and *Control* (i.e. somewhat r-like; partially derhoticized) was considered. There was a significant difference between *Control* and *DerhotBRV* and *DerhotMCV*, both of which had higher F3-F2 values. Further exploration in this area could examine possible causes of this difference in greater detail than the current study. Also interesting to note is a significant difference found between *DerhotMCV* and *RDR+*, where *DerhotMCV* had higher average F3-F2 values.

There was a strong negative correlation between F3-F2 and the average listener rating. This implies that as F3-F2 increased, therefore becoming acoustically less r-like, listeners were more likely to rate it as not r-like. Figure 7 shows that listeners had more agreement in their ratings for lower F3-F2 values. As F3-F2 increases, the data points become more dispersed,

illustrating more inconsistency between listener ratings. There was no significant correlation between average listener ratings and duration of the tokens, suggesting that vowel duration may not have had a significant impact on listeners' rating choice.

CONCLUSION

The current study set out to investigate inexperienced listeners' perception of children's productions of rhotic vowels and the acoustic properties of these vowels. Results indicated that naïve listeners can perceive the difference between correct and incorrect rhotic vowel productions and that these categories are perceptually different from control (non-rhotic) vowels. This is consistent with the finding of the study by Peterson and Barney (1952), where the researchers found that listeners were able to easily distinguish between the rhotic vowel /ɝ/ and other vowel sounds. However, it suggests that untrained listeners have the ability to distinguish a middle category as well. Future studies could examine listeners' perceptions of different error types in greater detail to further define this zone between 'correct' productions of rhotic and non-rhotic vowels.

The study examined the relationship between coded transcription of rhotic vowels and the measured distance between F3-F2 and found that F3-F2 was greater for non-rhotic productions (both incorrect and control tokens) than for correct productions. This finding is consistent with previous findings (Flipsen et al., 2001); however, contrary to what was expected, the control vowels did not have a higher F3-F2 value than the incorrect productions. This could possibly be due to the increase of F2 in the /ɔ͡ɪ/ diphthong in "toy". In future studies, the trajectory of F3 could be examined to avoid the effect of the changing F2 on the F3-F2

value. The methods of the current study involved looking at inexperienced listeners' perception in an unnaturalistic environment. A more naturalistic way would involve testing listeners' perceptions of the targeted phonemes in the context of whole words; however, this could lead to bias as it is possible that the knowledge of the word would influence some listeners to rate vowels as more 'r-like' or less 'r-like' based on what they expect to hear. This would be more similar to how listeners perceive sounds in everyday contexts. Future studies could examine the extent and direction of this bias and what impacts it might have on understanding children's speech sound errors. Future studies could also examine effects of vowel duration, vowel type (monophthong vs. diphthong) and vowel characteristics (front vs. back, high vs. low and rounded vs. unrounded) on the perception of rhotic vowels.

Some limiting factors include: small sample size, poor recording quality of some tokens, and varying loudness in the samples. As the tokens were selected from previously recorded words from a larger study, the control words did not correspond directly to the target words. The program chosen for acoustic analysis was limited in terms of selection of F3 and F2 values. A different process of obtaining F3 and F2 values could be used as comparison. Following the methodology in Klein et al. (2012), the current study selected the lowest point of F3 and the corresponding F2 value. An alternative process could be to extract F3 and F2 values from the point at which they are closest together. Furthermore, the current study selected single points in the spectrogram to obtain F3 and F2 values. Future studies could use multiple points along the spectrogram to obtain average values for F3 and F2, or to focus on the trajectory of F3.

REFERENCES

Boersma,P. & Weenink, D. (2012). Praat: doing phonetics by computer [Computer program]. Version 5.3.16, http://www.praat.org.

Chung, H., Farr, K., & Pollock, K. E. (2014). Transcription-based and acoustic analyses of rhotic vowels produced by children with and without speech sound disorders: Further analyses from the Memphis Vowel Project. *Clinical Linguistics & Phonetics, 28,* 297-315.

Flipsen, Jr., P., Shriberg, L. D., Weismer, G., Karlsson, H. B., & McSweeny, J. L. (2001). Acoustic phenotypes for speech-genetics studies: reference data for residual /ɚ/ distortions. *Clinical Linguistics and Phonetics, 15,* 603-630.

Guenther, F. H., Espy-Wilson, C. Y., Boyce, S. E., Matthies, M. L., Zandipour, M.,& Perkell, J. S. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *Journal of the Acoustical Society of America, 105,* 2854-2865.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97,* 3099-3111.

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

McGowan, R. S., Nittrouer, S., & Manning, C. J. (2004). Development of /r/ in young, Midwestern, American children. *The Journal of the Acoustical Society of America, 115,* 971-884.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184.

Pollock, K. E. (2013). The Memphis Vowel Project: Vowel errors in children with and without phonological disorders.. *The Handbook of Vowels and Vowel Disorders* (). New York: Psychology Press.

Pollock, K. E. & Berni, M. C. (2003). Incidence of non-rhotic vowel errors in children: data from the Memphis Vowel Project. *Clinical Linguistics & Phonetics, 17,* 393-401.

Pollock, K. & Keiser, N. (1990). An examination of vowel errors in phonologically disordered children. *Clinical Linguistics and Phonetics, 4,* 161-178.

Potter, R. K., & Peterson, G. E. (1948). The representation of vowels and their movements. *Journal of the Acoustical Society of America, 20,* 528-535.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: extensions and reliability data. *Journal of Speech, Language & Hearing Research, 40,* 708-722.