7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

Predicting Clinical Evaluations of Children's Speech with Limited Data Using Exemplar Word Template References

Gary Yeung[1], *Amber Afshan*[1], *Kaan Ege Ozgun*[1], *Kantapon Kaewtip*[1], *Steven M. Lulich*[2], *Abeer Alwan*[1]

[1]Department of Electrical Engineering, University of California, Los Angeles, USA
[2]Department of Speech and Hearing Sciences, Indiana University, Bloomington, USA

garyyeung@g.ucla.edu, amberafshan@g.ucla.edu, kaanege@g.ucla.edu, kkaewtip@ucla.edu, slulich@indiana.edu, alwan@ee.ucla.edu

Abstract

The need for automated speech pathology diagnostic tools for children has increased in recent years. Such tools can help speech pathologists identify speech disorders in children at an early age. This paper introduces an approach to automated clinical evaluations of children's speech using limited data. A database of ten normally developing first-grade children administered the Goldman-Fristoe Test of Articulation, 3rd Edition (GFTA-3) was recorded. Graduate clinicians evaluated the pronunciation of the rhotic sounds by evaluating words in the GFTA-3 containing the letter 'r'. The rhotic sounds were specifically chosen due to their late acquisition in children. Experiments were performed attempting to predict the results of the clinical evaluations. Five children, judged to have proper rhotic pronunciations, were chosen as exemplar templates for the experiment. The remaining children, used for evaluation, were aligned in time to match the five templates using dynamic time warping, and the difference between a test child's 'r' and a template child's 'r' was measured using the cosine distance. Multiple linear regression on the difference scores was shown to be effective at producing predictions that were well-correlated with human clinical evaluations. Several sublists of words with rhotic sounds were used to evaluate the regression, and the sublist containing words with the most mispronunciations performed best. Further discussion includes how much each individual template contributed to the regression and how consistent the clinicians were at scoring children's speech production.

Index Terms: children's speech assessment, pronunciation evaluation, template-based

1. Introduction

In a number of clinical settings, automated assessments of children's speech production can provide clear benefits and advantages. In particular, speech-language pathologists are currently required to both administer speech evaluations for children and design treatment plans for those children diagnosed with a speech delay or disorder. Automated speech assessments for children can allow clinicians to focus on therapy instead of time-consuming examinations. Additionally, human speech pathologists do not always agree in their evaluations of children's speech. Machine assessments can maintain objectivity in judgments and assist clinicians in their evaluations.

Many systems for speech evaluations have been proposed since the 1980s. Several studies were performed by Kewley-Port and her colleagues in the 1980-90s using templates to evaluate speech [1, 2, 3, 4]. The Indiana Speech Training Aid (ISTRA) used the best recordings from a subject as templates, and new templates were recorded to replace old ones as the subject's pronunciation improved [2]. While this system was shown to be effective, it also required heavy clinician involvement since templates constantly needed to be updated.

More recently, the advancement of machine learning in automatic speech recognition (ASR) has led to a number of Hidden Markov Model (HMM) ASR systems for children's speech evaluations [5, 6, 7, 8, 9, 10]. The Speech Training, Assessment, and Remediation (STAR) system achieved an $r^2 = 0.6$ when using phoneme likelihoods in a linear regression to assess the pronunciation of the phoneme /r/ [6]. Another system achieved a 76% agreement between ASR and human listeners when measuring children's speech intelligibility [9]. A more recent study used HMMs for forced-alignment and the Mahalanobis distance to explore trade-offs caused by thresholding scores [10].

While ASR systems have improved dramatically in recent years, children's ASR is still not as well-understood as adult ASR [11, 12]. Children's HMM-ASR systems, as well as deep neural network ASR systems, generally require a sizeable amount of data to train and are highly dependent on the data used [13, 14]. However, clinical speech data (especially for child speech) are much more difficult to acquire than normal speech data, and it is impractical for clinicians to do the work of gathering enough data for such systems. More research is needed to enable the development of clinical evaluation systems that can be used with small amounts of training data.

This study proposes a method that uses a limited amount of children's speech data to train a clinical evaluation system for children. We return to a template-based approach to tackle this low-resource problem. In this paper, we examined children's pronunciation of rhotic sounds due to the late acquisition of these sounds in children [15]. The children were reported (by their parents) to have no history of speech, language, or hearing impairment in a prior screening interview. From this study, we hope to understand whether clinician perception of children's pronunciations can be modeled using a small number of exemplar pronunciations.

The rest of the paper is organized as follows. Section 2 describes the data collection and clinical evaluation process. Section 3 describes the pronunciation evaluation system. Section 4 discusses the experiments and results. Finally, Section 5 concludes the paper with a brief summary and description of future work.

10.21437/SLaTE.2017-28

Table 1: *Words containing rhotics in the GFTA-3: Sounds in Words divided into 'onset', 'coda', 'medial', and 'cluster' categories.*

Onset	Coda	Medial	Cluster
Red	Brother	Giraffe	Brother
Ring	Chair		Brushing
	Door		Crown
	Finger		Drum
	Guitar		Frog
	Hammer		Green
	Star		Princess
	Teacher		Truck
	Tiger		Zebra
	Spider		

2. Data Collection

2.1. Children's Speech Data

This paper is part of a larger study by UCLA and Indiana University which aims to improve children's speech-language pathology tools through a longitudinal analysis of children's speech. The database currently being collected consists of recordings of elementary school children. All children were screened beforehand to ensure that they did not have any speech disorders. Each child was recorded taking the Goldman-Fristoe Test of Articulation, 3rd Edition (GFTA-3) [16]. GFTA-3 is a common standardized speech test which evaluates children's pronunciation using clinically relevant utterances. Student clinicians tracked the quality of phoneme pronunciation of each child using the GFTA-3 assessment format. Each child was seated inside a double-walled sound booth with the student clinician who administered the GFTA-3. A SHURE KSM32 microphone was placed approximately 1 m in front of the child and 0.5 m to the side of the student clinician, who was facing the child. Audio was recorded at 48 kHz with 16-bit quantization.

Ten first grade children, aged between 6-7 years, were recorded for this study. All of the children made some pronunciation mistakes, but these mistakes were typical and developmentally appropriate for children of such an age. The GFTA-3 has two sections: "Sounds in Words" and "Sounds in Sentences". All tests were administered by graduate speech-language pathology clinicians. For this paper, only the "Sounds in Words" data were used, in which each child was prompted by the clinician to say specific words in a picture-naming task. For example, to cue the child to say 'table', the clinician would show a picture of a table and ask, "What is this?" There were 21 words from the GFTA-3 containing 22 'r' sounds in a variety of phonetic contexts. These words are listed in Table 1. The word 'brother' is listed twice as it has an 'r' in both onset cluster and coda positions. We will refer to the onset cluster 'r' in this word as 'brother1' and the coda 'r' as 'brother2'. These words compose the master word list for this paper. As each of the ten children said the 22 words once (counting 'brother' as two words), there were 220 total word utterances. The ten children were separated into two groups, five in a "template" group and five in a "trial" group. The children chosen for the trial group included two children who were suspected of having more errors in 'r' pronunciation than the rest of the children while still being typically developing. The remaining children were divided in such a way to ensure the template group had minimal 'r' mispronunciations. As these children had few pronunciation errors in general, the division was randomly assigned.

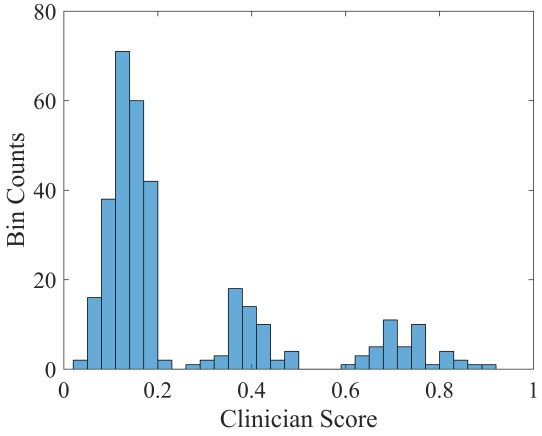


Figure 1: *Histogram of one clinician's scores on the evaluation of 330 rhotic phonemes. The scores are clearly separated into three groups. Clinician scores were in the range of 0 to 1 where 0 represented a perfect pronunciation and 1 represented a severe mispronunciation.*

2.2. Clinician Scoring

Ten graduate clinicians from Indiana University rated the quality of production of the 'r' sounds from the five trial children. Utterances were played in a random order to each clinician, and each utterance was played a total of 3 times. As such, each utterance was judged a total of 30 times, and each clinician made 330 judgments (3 judgments of 5 children saying 22 words each). Graduate clinicians rated the quality of a child's 'r' production by clicking (with a computer mouse) within a circular bullseye, displayed on a graphical user interface, with three levels: 'no impairment' as level 1 (inside), 'mild impairment' as level 2, and 'severe impairment' as level 3 (outside). Such a rating system was familiar to the graduate clinicians from their prior clinical experiences and is commonly used to explain quality of pronunciation to children.

The distance between the center of the bullseye and the clinician's selected point was chosen as the clinician's score where the radius of the bullseye was normalized to 1. A histogram of one clinician's scores for the 330 judgments is shown in Figure 1. All graduate clinicians displayed similar Gaussian-like behavior in scoring around three central points as that shown in Figure 1. A continuous scoring method was kept over a discrete scoring method of 1, 2, or 3 because no specific instructions were given to the clinicians about whether to use the bullseye in a continuous or discrete way. As such, it was possible that some clinicians used the bullseye in a more continuous way. The final score of each utterance was chosen as the average of the 30 corresponding scores.

2.3. Word Lists

Various word lists were used in these experiments. Six different word lists were chosen as follows:

1. All words
2. brother2, chair, door, finger, guitar, hammer, spider, star, teacher, tiger
3. brother2, finger, hammer, spider, teacher, tiger

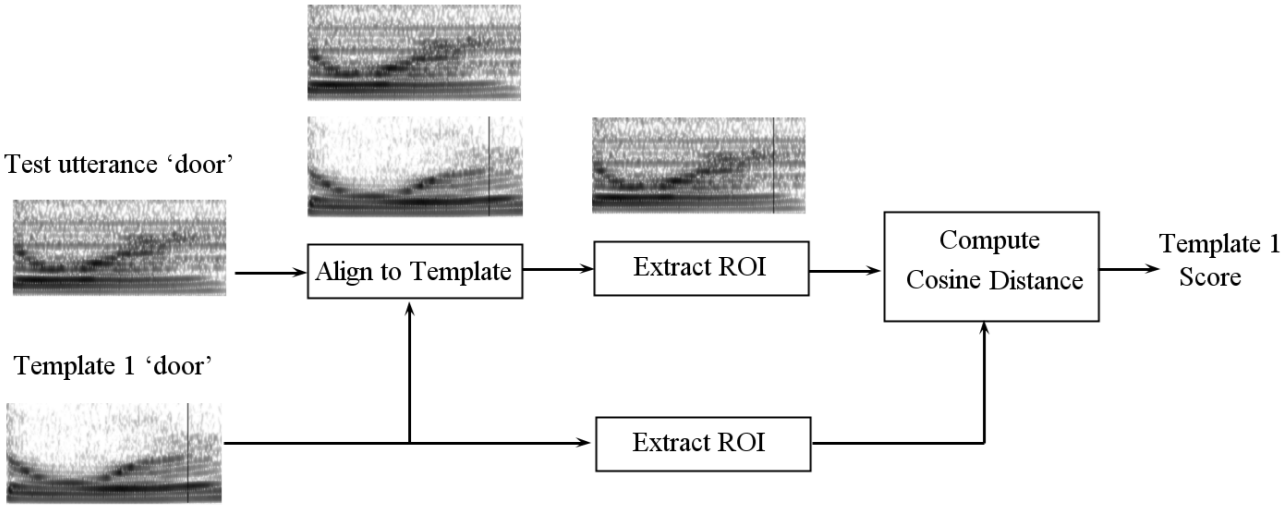


Figure 2: *Block diagram of the similarity score extraction using the example word 'door' from template child 1.*

4. brother1, brushing, chair, crown, door, drum, frog, giraffe, green, guitar, princess, red, ring, star, truck, zebra
5. brother2, drum, guitar, hammer, princess, ring, tiger, zebra
6. guitar, hammer, princess, ring, tiger

List 2 consists of words with 'r' in a syllable coda. List 3 consists of words with syllabic final 'r', while List 4 consists of the complementary set of words with non-syllabic 'r'. List 5 consists of words where at least one child had more than 30% of clinician scores in the outer section of the bullseye. Finally, List 6 consists of words where at least one child had more than 50% of clinician scores in the outer section of the bullseye.

3. Pronunciation Evaluation System

3.1. Feature Extraction

Feature sets investigated included Mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) coefficients, and linear predictive cepstral coefficients (LPCC). For all feature sets, a window size of 25 ms, a window shift of 10 ms, a pre-emphasis filter with coefficient 0.97, and a sinusoidal lifter with coefficient 22 were used. A filter bank with 23 filters was used for the MFCC features. A 12^{th} order linear predictive coding (LPC) polynomial was used for both the PLP and LPCC features. For all feature sets, 13 coefficients were extracted, and the 0^{th} (energy) coefficient was removed for a total of 12 dimensions per frame. Utterances were downsampled to 8 kHz before feature extraction.

3.2. Template Setup

Of the ten first-grade children recorded, utterances from 5 children who were judged to have no rhotic pronunciation errors from the original GFTA-3 assessment, and few errors in general, were chosen to serve as templates. These children are referred to as "template children" for the remainder of the paper. All 22 words containing 'r' were used for each of the 5 template children for a total of 110 templates, 5 templates per word. For each template, 3 consecutive frames from the corresponding feature sets were chosen manually at the center of the rhotic sound as a region of interest (ROI). The ROI for each word utterance can be thought of as an exemplar pronunciation of 'r'.

3.3. Evaluation Procedure

The 5 children not chosen to be templates were used to model clinician scores. We will refer to these children as "trial children" for the remainder of the paper.

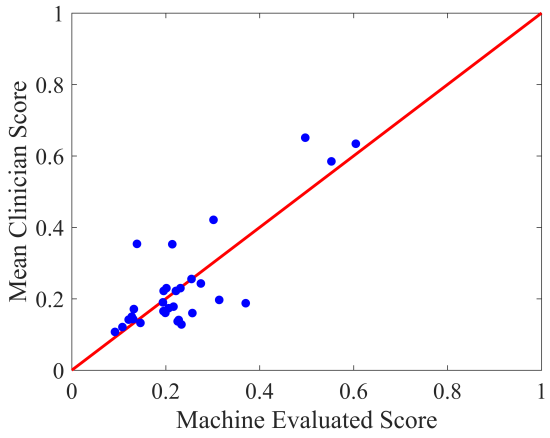
The feature set of each word utterance from a trial child was aligned to the corresponding word from a template child using dynamic time warping (DTW) with Euclidean distance as a metric. For example, the utterance 'door' spoken by a trial child was aligned in time to match the utterance 'door' spoken by a template child. After time alignment, the cosine distance between the ROI of the template and the corresponding frames in the aligned trial utterance was calculated, averaged over 3 frames. The resulting distance served as a similarity measure of the trial word's 'r' and the template's 'r'.

For each word from a trial child, the above procedure was repeated 5 times, once for each template child. As a result, each trial child's word utterance had 5 different scores representing the similarity of the trial child's 'r' and each template child's 'r'. Figure 2 illustrates this procedure. For subsets of the word list, these 5 similarity ratings were used as inputs to a multiple linear regression with the mean clinician score as the prediction.

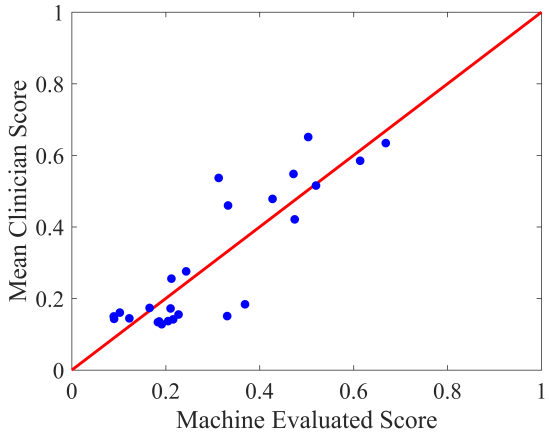
An alternative procedure considered using all words from a subset of the word list to create a single template to represent each template child. In this case, all ROIs from the chosen word list spoken by a single template child were averaged to create a single exemplar 'r', one for each template child. The DTW alignment and choice of ROI for the trial utterance was identical to the first procedure. However, the cosine distance was computed between the ROI in the aligned trial utterance and the mean exemplar 'r' of the corresponding template child. This procedure can be illustrated with Figure 2 by simply replacing the single template ROI with the mean ROI as the input of the cosine distance. As with the first procedure, the 5 resulting scores were used in a multiple linear regression to predict the mean clinician score for each trial child's word utterance. However, the procedure using mean templates did not perform as well and will not be reported.

Table 2: *Results of the multiple linear regressions using similarity scores between trials and templates to predict clinical evaluations of rhotic phonemes for all six word lists. Both r^2 and adjusted r^2 are shown.*

Word List	MFCC		PLP		LPCC	
	r^2	adjusted r^2	r^2	adjusted r^2	r^2	adjusted r^2
1	0.267	0.231	0.220	0.183	0.103	0.060
2	0.473	0.413	0.420	0.354	0.370	0.299
3	0.691	0.627	0.537	0.440	0.615	0.535
4	0.107	0.047	0.119	0.059	0.108	0.048
5	0.574	0.511	0.431	0.347	0.310	0.208
6	0.762	0.699	0.586	0.478	0.570	0.457



(a) *Word List 3*



(b) *Word List 6*

Figure 3: *Clinician evaluation scores vs. scores predicted from the regression model using Word List 3 (top) and World List 6 (bottom) with MFCCs. The line represents an ideal regression.*

4. Experiments and Results

4.1. Regression Results

Table 2 shows the r^2 and adjusted r^2 results of the clinical evaluation regression models for the various word lists. In general, MFCCs performed the best out of all the feature sets. Figures 3a and 3b show the results of the regression scores plotted against clinician scores using MFCCs on Word Lists 3 and 6, respectively, which gave the two best regression results. Word List 6 with MFCCs gave the best regression results, modeling over 76% of the variance of the clinician scores. This is likely due to the fact that Word List 6 better represented mispronounced 'r' phonemes (higher clinician scores) while the other word lists may have overrepresented words with proper pronunciation (lower clinician scores). Additionally, Word List 3 gave decent regression results when used with MFCCs, modeling almost 70% of the variance of the clinician scores. This suggests that using the specific subset of syllabic 'r' sounds can improve the evaluation procedure. One noticeable issue is that Word List 3 did not have many words that were judged as severely impaired by the clinicians. As seen in Figure 3a, only a small number of points represented higher scores in the linear regression. As such, the results from Word List 3 may be questionable.

Interestingly, the regression results in most cases indicated that some of the template children contributed to the model significantly more than others. Table 3 shows the significance of contribution for the five template children from the regression using MFCCs and Word List 6. Only template child 1 and 4 contributed significantly in predicting the clinician scores. As such, the remaining templates were likely not reliable exemplars of properly pronounced 'r' phonemes as judged by clinicians. Recomputing the regression using only the two significant template children resulted in an $r^2 = 0.721$ and adjusted $r^2 = 0.695$, representing only a small decrement in performance.

Table 3: *Significance of contribution of the five individual template children for the clinical evaluation regression model using Word List 6 with MFCCs.*

Template Child ID	**p-value**
1	0.006
2	0.220
3	0.314
4	0.002
5	0.789

In an attempt to improve the results, vocal tract length normalization (VTLN) was tested at the feature extraction step to improve alignment and template scoring. Various numbers of filters for MFCCs, LPC orders for PLP and LPCC, and window sizes were tested as well. Additionally, the Euclidean and Mahalanobis distances were also tested for similarity scoring between templates and trials. However, these approaches did not reveal any notable improvements in most cases and were thus discarded.

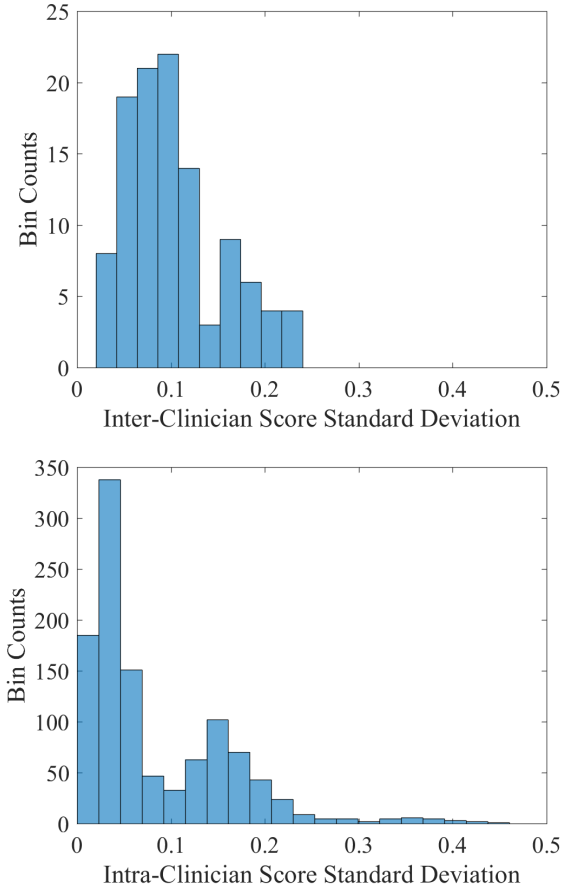


Figure 4: *Histogram of inter-clinician (top) and intra-clinician (bottom) score standard deviations across all words.*

4.2. Discussion

One major point of interest is how well DTW was able to align trial utterances to templates to ensure that the ROIs corresponded to 'r' sounds. A manual inspection of the trial-to-template alignments revealed that only 2 of the trial-to-template alignments were misaligned out of 550 (5 trial children aligned to 5 template children for 22 words) when using MFCCs. In general, the alignment was successful. One of the cases of misalignment had an obvious mispronunciation in the trial word 'ring' as [w i ŋ]. As a misalignment would likely cause the system to score the phoneme as mispronounced, we believe that these alignment mistakes are acceptable for identifying difficult words for children. The other case was in the word 'brother2' with no obvious pronunciation error. In this worst case scenario, the system would mistakenly classify this phoneme as mispronounced, which may be preferable to mistakenly classifying a phoneme as well-pronounced in some applications (e.g. a screening tool with high sensitivity).

Another point of interest is how difficult clinicians found the task of scoring the 'r' quality to be. Inter-clinician score standard deviations, defined as the standard deviation across the mean judgments for each of the ten clinicians, and intra-clinician score standard deviations, defined as the standard deviation across the three judgments from a single clinician, were computed for each word utterance. The average inter-clinician score standard deviation across words was 0.103, or 10.3% of the total range of possible scores. The maximum inter-clinician score standard deviation was 0.237 (23.7% of the range of possible scores) for a questionable pronunciation of the word 'door' by one particular child, indicating that some words and pronunciations were much less agreed upon across clinicians than others. The average intra-clinician score standard deviation across words and clinicians was 0.084 suggesting that clinicians were more consistent with themselves than with other clinicians. However, the maximum intra-clinician score standard deviation was 0.450 (45% of the range of possible scores) for a questionable pronunciation of the word 'truck' by one particular child, indicating that some pronunciations presented consistency issues for clinicians. Figure 4 shows histograms of inter-clinician and intra-clinician score standard deviations, which have most of their density in the lower standard deviations. The bimodal nature of these histograms indicates that most scores for a particular word either were within a single bullseye level or spanned two adjacent levels. The low peak between 0.3 and 0.5 in the intra-clinician score standard deviation histogram indicates that a small number of words was scored across all three levels. Table 4 shows the mean inter-clinician and intra-clinician score standard deviations for each word in the master word list. It is clear that some words, such as 'princess', caused more difficulty in scoring than others, indicated by a high mean score standard deviation in Table 4. In general, a large standard deviation in scoring an utterance was not correlated with the ability of the regression model to predict the mean clinician score.

Finally, we note that one fundamental difficulty in this study was the usage of only five trial children to predict clinician scores. Most of the poorly pronounced 'r' sounds were due to only two of the trial children, although the remaining three chil-

Table 4: *Average values of inter-clinician and intra-clinician score standard deviations for rhotic sounds.*

Word	Standard Deviation	
	Inter-Clinician	Intra-Clinician
brother1	0.130	0.100
brother2	0.135	0.104
brushing	0.132	0.100
chair	0.079	0.085
crown	0.083	0.071
door	0.085	0.085
drum	0.135	0.099
finger	0.090	0.075
frog	0.080	0.087
giraffe	0.088	0.098
green	0.090	0.082
guitar	0.100	0.057
hammer	0.103	0.073
princess	0.141	0.114
red	0.087	0.073
ring	0.143	0.078
spider	0.097	0.073
star	0.097	0.082
teacher	0.092	0.088
tiger	0.083	0.071
truck	0.067	0.069
zebra	0.123	0.078

dren contributed a few poor productions as well. Additionally, since this study was performed using children recorded in the state of Indiana alone, the ability to generalize to other dialect areas is uncertain.

5. Conclusion

This study proposed a framework to predict clinician scores of 'r' sounds produced by children. A database of ten first grade children was used. Speech utterances from five children were chosen as exemplar templates of 'r' production. The remaining five children were scored by clinicians and used to model clinician responses. The template 'r' and trial 'r' sounds were aligned with DTW, and the cosine distance between the phonemes was used as an automated scoring metric. The five scores from each trial word were used in a linear regression to predict mean clinician scores. Various word lists were used for regression, and it was found that the regression performed best when poorly pronounced phonemes were well-represented.

Future work will include expanding clinical scoring of children's speech with more data, as well as taking into account various ages, dialects, and speech disorders.

6. Acknowledgements

The research was supported in part by the NSF.

7. References

[1] D. Kewley-Port, C. S. Watson, D. Maki, and D. Reed, "Speaker-Dependent Speech Recognition as the Basis for a Speech Training Aid," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 12, 1987, pp. 372–375.

[2] D. Kewley-Port, C. S. Watson, M. Elbert, D. Maki, and D. Reed, "The Indiana Speech Training Aid (ISTRA) II: Training curriculum and selected case studies," *Clinical Linguistics and Phonetics*, vol. 5, no. 1, pp. 13–38, 1991.

[3] S. Anderson and D. Kewley-Port, "Evaluation of Speech Recognizers for Speech Training Applications," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 229–241, 1995.

[4] J. Dalby and D. Kewley-Port, "Explicit Pronunciation Training Using Automatic Speech Recognition Technology," *CALICO Journal*, vol. 16, no. 3, pp. 425–445, 1999.

[5] M. Russell, C. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, "Applications of automatic speech recognition to speech and language development in young children," in *Proc. of the Fourth International Conference on Spoken Language Processing (ICSLP)*, 1996, pp. 176–179.

[6] H. T. Bunnell, D. M. Yarrington, and J. B. Polikoff, "STAR: Articulation Training for Young Children," in *Proc. of the Sixth International Conference on Spoken Language Processing (ICSLP)*, 2000, pp. 85–88.

[7] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation Verification of Children's Speech for Automatic Literacy Assessment," in *Proc. of INTERSPEECH*, 2006, pp. 845–848.

[8] O. Saz, E. Lleida, and W.-R. Rodríguez, "Avoiding Speaker Variability in Pronunciation Verification of Children's Disordered Speech," in *Proc. of the 2nd Workshop on Child, Computer and Interaction (WOCCI)*, 2009, pp. 11.1–11.5.

[9] J. Lilley, S. Nittrouer, and H. T. Bunnell, "Automating an Objective Measure of Pediatric Speech Intelligibility," in *Proc. of INTERSPEECH*, 2014, pp. 1578–1582.

[10] R. Sadeghian and S. A. Zahorian, "Towards an Automated Screening Tool for Pediatric Speech Delay," in *Proc. of INTERSPEECH*, 2015, pp. 1650–1654.

[11] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A Review of ASR Technologies for Children's Speech," in *Proc. of the 2nd Workshop on Child, Computer and Interaction (WOCCI)*, 2009, pp. 7.1–7.8.

[12] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child Speech Recognition in Human-Robot Interaction : Evaluations and Recommendations," in *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 82–90.

[13] A. Potamianos, S. Narayanan, and S. Lee, "Automatic Speech Recognition for Children," in *Proc. of EUROSPEECH*, 1997, pp. 2371–2374.

[14] A. Potamianos and S. Narayanan, "Robust Recognition of Children's Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[15] A. B. Smit, L. Hand, J. J. Freilinger, J. E. Bernthal, and A. Bird, "The Iowa Articulation Norms Project and its Nebraska replication," *Journal of Speech and Hearing Disorders*, vol. 55, no. 4, pp. 779–798, 1990.

[16] R. Goldman and M. Fristoe, "Goldman-fristoe test of articulation: Third edition," *Pearson*, 2015.