Predicting Children's Reading Ability using Evaluator-Informed Features

Matthew Black, Joseph Tepperman, Sungbok Lee, and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, CA

{matthepb,tepperma,sungbokl}@usc.edu, shri@sipi.usc.edu

Abstract

Automatic reading assessment software has the difficult task of trying to model human-based observations, which have both objective and subjective components. In this paper, we mimic the grading patterns of a "ground-truth" (average) evaluator in order to produce models that agree with many people's judgments. We examine one particular reading task, where children read a list of words aloud, and evaluators rate the children's overall reading ability on a scale from one to seven. We first extract various features correlated with the specific cues that evaluators said they used. We then compare various supervised learning methods that mapped the most relevant features to the ground-truth evaluator scores. Our final system predicted these scores with 0.91 correlation, higher than the average inter-evaluator agreement.

Index Terms: children's speech, reading assessment, feature selection, pronunciation verification, disfluency detection

1. Introduction

Reading assessment is a vital component of children's early education and second language learning. This task is rooted in objective rule-based observation but has an important human subjective component as well. For example, humans might listen for specific phones within a word pronunciation when evaluating whether it was read correctly or not (a decision largely based on objective observation) but may use some combination of more subjective cues, such as fluency and speaking rate, when evaluating the pronunciation quality of a read sentence. Here, we use the term subjectivity to indicate the level of personal judgment that goes into the decision. If two or more evaluators judge the same material, this level can be measured quantitatively using, for example, Pearson's correlation (i.e. the lower the correlation between the evaluators, the higher the level of subjectivity).

Automatic reading assessment software could offer numerous advantages, including saving human evaluators time and offering consistent evaluations across different subjects and over time. One way in which to design automatic reading assessment software is to train it to grade like a typical human evaluator or bank of evaluators. This can be viewed as a challenging signal processing and machine learning problem. First, we need to extract features that correlate with cues humans use. Second, the relevant features must be mapped to the evaluators' judgments. This two-step problem is further complicated when we do not know exactly how evaluators are making their judgments and/or in cases where evaluators are deriving them in different ways. However, if we are able to cover the spectrum of cues humans use and have some training data of actual human evaluations, then we can use supervised learning methods that incorporate the most relevant features.

In this paper, we assess children's overall reading ability based on their performance on an isolated word reading task. We concentrate on predicting the *ground-truth* evaluator's scores (scores averaged across the evaluators) in order to train models that mimic their common grading patterns. Future research will explore predicting the more subjective individual evaluator's scores. In our previous work, we showed we could successfully predict ground-truth evaluator's scores using two features: one correlated with pronunciation correctness and the other correlated with the fluency of the children's speech [1]. This paper expands upon our previous work in a number of ways: 1) we have evaluators rate 42 children, more than three times as many, 2) we do not have evaluators rate the pronunciation quality of each word individually but rather allow them to listen to the entire list of words and then make their judgment using their own criteria, 3) we do not assume the two aforementioned features are sufficient but instead over-generate features based on cues the evaluators said they used, and 4) we try several supervised learning techniques that make use of this pool of features to predict the ground-truth evaluator's scores of children's reading ability.

2. Corpus

We used speech from the Tball Corpus [2], collected in kindergarten, first, and second grade classrooms in the Los Angeles area from both native English and Spanish speaking children. We randomly selected 42 children who were tested on an isolated English word reading task. Each word was displayed on a computer screen for a maximum of five seconds before the next word was shown. The word list and order for each child was identical, beginning with simple words (e.g., map, rip) before progressing to more difficult ones (e.g., cute, rested). Some children were unable to complete the word list, so the number of words per child ranged from 10 to 23.

3. Human evaluation

Eleven engineering graduate students rated the speech from each child (we found no significant difference between student evaluators and expert linguists/teachers in [1]). We randomized the order of the children for each evaluator but maintained chronological word order within each child's speech. To simulate the reading task, we combined all the words for one child into a single audio file and inserted a short beep to indicate word boundaries. We provided the word list to the evaluators for each child, so they could track the child's progress while they listened. After listening to the audio file, each evaluator rated the child's overall reading ability on an integer scale from one ("poor") to seven ("excellent"). We purposefully did not instruct evaluators how to grade the children or offer them examples of an "excellent" versus a "poor" reader for two main reasons: 1) we did not know in advance what an "excellent" versus a "poor" reader was to *all* evaluators, and 2) we wanted evaluators to come up with their own grading criteria. Evaluators could listen to each child's audio file as many times as they felt necessary and were also encouraged to change previously assigned scores. After rating all 42 children, we asked evaluators to list the criteria they used when making these judgments.

The mean pairwise evaluator correlation was 0.825. We calculated ground-truth scores for each child by averaging across all 11 evaluators. Table 1 shows the three metrics we used to compare each evaluator's scores with the ground-truth scores (where we used leave-one-evaluator-out cross-validation for these calculations). The ground-truth scores served as the dependent variable in this paper, and the same three metrics are used later to compare our automatic results.

Metric	*Mean (std. dev.)*
Evaluator Correlation with GT Scores	**0.899** (0.038)
Mean Absolute Error with GT Scores	**0.624** (0.137)
Maximum Absolute Error with GT Scores	**2.227** (0.388)

Table 1: *Metrics comparing each evaluator's scores with the ground-truth (GT) scores (mean and standard deviation shown).*

4. Feature extraction

Analyzing the open-ended question posed at the end of the human evaluation, the three most cited cues that the 11 evaluators used when making their judgment on overall reading ability were: pronunciation correctness, fluency of speech, and speaking rate, which is consistent with other studies [3,4] and our previous work [1]. As discussed in the introduction, we feel an over-generation of potentially useful features is needed to tackle this machine learning problem, since the exact grading method of the evaluators is unknown. Therefore, we extracted many features that were based on these three main cues provided by the evaluators. Since the children were reading a list of words in isolation, it was natural to extract these features at the word-level. Sections 4.1-4.3 discuss the word-level features correlated with pronunciation correctness, fluency, and speaking rate, respectively. Section 4.4 discusses how we generate our final set of child-level features from these word-level features.

4.1. Pronunciation correctness word-level features

We applied traditional pronunciation verification methods to extract automatic scores (referred to as word-level features) correlated with pronunciation correctness. First, we created four dictionaries with various phonemic pronunciations, as described in Table 2. Since these were young children learning to read, many of the common reading mistakes are predictable (Reading Error dictionary), and since many of the children were from Mexican-American families, common Spanish-speaking confusions were also to be expected (Spanish Confusion dictionary) [5]. The Recognition dictionary was the union of the Acceptable, Reading Error, and Spanish Confusion dictionaries; note that the Reading Error and Spanish Confusion dictionaries were not disjoint.

Dictionary Name	*Avg. #*	*Entries for word, "map"*
Acceptable	1.18	/m ae p/
Reading Error	2.13	/m ey p/
Spanish Confusion	1.09	/m aa p/
Recognition	3.71	/m ae p/, /m ey p/, /m aa p/

Table 2: *Average number of entries per target word and entries for the word, "map," for the four dictionaries.*

We used the same acoustic models as in our previous work, trained on 12 hours of held-out children's speech: three-state 16 Gaussian monophone HMMs, a background HMM, and a word-level "garbage" HMM [1,6]. We extracted the first three pronunciation correctness features in Table 3 by running speech recognition with the Recognition dictionary using a grammar that allowed for one forced alignment of the target word and optional silence/garbage before and/or after this forced alignment. We then ran forced alignment over the portion of the utterance endpointed as the target word with the Acceptable, Reading Error, and Spanish Confusion dictionaries and with the garbage model to attain the log-likelihoods of the pronunciation being an acceptable pronunciation, a reading error, a Spanish-confusion related error, and garbage. These log-likelihoods were used to compute the remaining word-level features listed in Table 3.

Description	*Domain*
Was an "acceptable" pronunciation recognized?	{N=0, Y=1}
Was a "reading error" recognized?	{N=0, Y=1}
Was a "Spanish-related error" recognized?	{N=0, Y=1}
Log-likelihood of acceptable pronunciation (LL_{acc})	Continuous
Log-likelihood of reading error (LL_{read})	Continuous
Log-likelihood of Spanish-related error ($LL_{Spanish}$)	Continuous
$LL_{acc} - LL_{read}$	Continuous
$LL_{acc} - LL_{Spanish}$	Continuous
$LL_{acc} - LL_{garbage}$	Continuous
$LL_{acc} - \max\{LL_{read}, LL_{Spanish}\}$	Continuous
$LL_{acc} - \max\{LL_{read}, LL_{Spanish}, LL_{garbage}\}$	Continuous

Table 3: *Pronunciation correctness word-level features.*

4.2. Fluency word-level features

To extract features correlated with the fluency of the speech, we re-ran the speech recognition with the Recognition dictionary, except this time we used a grammar that allowed for individual phones within the various pronunciations of the dictionary to be recognized. Thus, we were able to recognize "partial word" pronunciations of the target word, which helped detect when a child made more than one attempt at reading the word (e.g., hesitations, sounding-out the word, repetitions) [6,7]. We found in our previous studies that the presence of these partial words was significantly negatively correlated with people's perception of fluency [1,6]. Table 4 shows the fluency features we extracted at the word-level from the output of the speech recognizer when using the disfluency-specialized grammars. We considered voiced partial words separately because we noticed a higher false alarm rate for unvoiced phones and because voiced phones may have a larger impact on the perception of fluency. For all temporal features, we also included the square root of the feature, since this transformation resulted in a less-skewed distribution.

Description	*Domain*
Number of recognized partial words	{0, 1, 2, …}
Number of unique recognized partial words	{0, 1, 2, …}
Was # of recognized partial words ≥ 1?	{N=0, Y=1}
Number of recognized voiced partial words	{0, 1, 2, …}
Number of unique recognized voiced partial words	{0, 1, 2, …}
Was # of recognized voiced partial words ≥ 1?	{N=0, Y=1}
Time of all recognized partial words [ms]	Continuous
Time of all recognized voiced partial words [ms]	Continuous
Time of silent regions between partial words [ms]	Continuous
Time of all partial words and silent regions [ms]	Continuous
Square root of all temporal features [$ms^{1/2}$]	Continuous

Table 4: *Fluency word-level features.*

4.3. Speaking rate word-level features

We extracted a number of word-level temporal features (Table 5) that are correlated with speaking rate, based on the same speech recognition output described in Section 4.1 with the Recognition dictionary. These features were all continuous.

Target word start time (relative to when word first displayed) [ms]
Target word total length in time (L_{target}) [ms]
Number of syllables spoken / L_{target} [syllables / ms]
L_{target} / Number of syllables spoken [ms / syllable]
Number of phones spoken / L_{target} [phones / ms]
L_{target} / Number of phones spoken [ms / phone]
Square root of all temporal features listed above

Table 5: *Temporal and speaking rate word-level features.*

4.4. Child-level features

Since evaluators based their ratings after listening to *all* words read by each child, we needed to map the word-level features described in Sections 4.1-4.3 to "child-level" features. We accomplished this by calculating the following descriptive statistics across the words of each child for all features: mean, standard deviation, skewness, kurtosis, minimum, minimum location (normalized by the number of words for the child), maximum, maximum location (normalized), range, lower quartile, median, upper quartile, and interquartile. This produced our final set of 481 child-level features (13 statistics for each of the 37 word-level features). While many of these features will be redundant and highly correlated, it is not obvious which ones are best. For example, the median may be better than the mean in cases where the feature is susceptible to outliers. Rather than address these robustness issues for each feature, we leave it up to the machine learning algorithm to eliminate irrelevant, noisy, and/or redundant features.

5. Learning Methods

Section 4 explained our over-generative approach to feature extraction, which resulted in 481 child-level features. Since the children's ground-truth overall reading ability scores were quasi-continuous values between one and seven, we chose to use supervised regression techniques. We used linear methods for simplicity and interpretability and because applying nonlinear techniques on such a small dataset (42 data points) may be prohibitive. We used leave-one-out cross-validation to separate train and test sets. All learning parameters for the various methods we tried were optimized on each cross-validation train set using leave-one-out cross-validation.

An obvious baseline method is multiple linear regression, which finds the linear weight coefficients of the features that minimize the square of the residual. The objective function J in this case is equation 1, where equation 2 is the analytical solution which minimizes J. Here, X is the matrix of child-level features, w is the vector of coefficient weights, and $\bar{y}$ is the centered (mean subtracted) vector of ground-truth scores.

$$J = \left\| Xw - \bar{y} \right\|^2 \equiv (Xw - \bar{y})^T (Xw - \bar{y}) \qquad (1)$$

$$w = (X^T X)^{-1} X^T \bar{y} \qquad (2)$$

However, due to the redundancy in the proposed features, the solution to the inverse in equation 2 would be numerically unstable. We addressed this problem by trying various feature selection methods. As a baseline, we ran simple linear regression with each child-level feature individually. Table 6 shows the performance for the best features for each of the three feature types. We next tried three feature selection methods within the linear regression framework: a forward selection method, stepwise linear regression, and the "lasso" (least absolute shrinkage and selection operator) [8]. Forward selection iteratively adds features that optimize the correlation. Stepwise regression is less greedy in that it can remove entered features if their coefficient's p-values become too large. The lasso algorithm finds a solution to the least-squares error minimization when adding a λ-weighted L_1 regularization term to the objective function:

$$J = \left\| \tilde{X}w - \bar{y} \right\|^2 + \lambda \left\| w \right\| \qquad (3)$$

This penalizes solutions with large weight coefficients (which often occurs when features are correlated) and promotes sparse models i.e. many of the trained weight coefficients are zero [8]. There is no analytical solution to the lasso objective function, but we used a modification of the LARS (least angle regression) algorithm to efficiently implement the lasso [9,10]. Note that we must standardize the features to ensure the regularization term is applied equally to all features. We accomplished this by centering X and dividing by the standard deviation of each feature (denoted by $\tilde{X}$ in equation 3).

6. Results & Discussion

Table 6 shows the performance for all the aforementioned feature selection linear regression methods. We obtained our best results, in terms of all three metrics, when we used the lasso algorithm and then re-trained the non-zero weights using linear regression without the regularization term (equation 2) i.e. multiple linear regression with the features selected by the lasso. Figure 1 is the regression plot for this case, with Figure 2 showing the resulting absolute error plot. This model performed better than the mean evaluator in all three metrics, although this difference was not significant. The final automatic model's correlation with the ground-truth scores was significantly higher than the three baseline systems' correlations when using single features (all p<0.05).

Method (Features)	*R*	*Mean* $\lvert\varepsilon\rvert$	*Max* $\lvert\varepsilon\rvert$
LR (best correctness)	0.783	0.746	2.852
LR (best fluency)	0.586	1.077	3.270
LR (best temporal)	0.757	0.832	2.669
LR (forward 2 features)	0.876	0.616	2.107
LR (forward 3 features)	0.860	0.651	2.187
LR (forward 4 features)	0.827	0.712	2.275
Stepwise LR	0.880	0.604	2.107
Lasso	0.886	0.815	2.326
Lasso, then *LR*	**0.906**	**0.502**	**1.971**
Mean human evaluator	**0.899**	**0.624**	**2.227**

Table 6: *Performance of various learning methods (LR = linear regression) compared to human evaluators for the 3 metrics.*

The lasso and linear regression combination method selected an average of 3.29 features at each cross-validation (primarily from four features). Two features it always selected were the *mean of the binary acceptable pronunciation* i.e. the fraction of words recognized as being an acceptable pronunciation, and the *upper quartile of the square root of the target word start time*. The former pronunciation correctness feature was positively correlated with reading ability (mean regularized weight coefficient = 0.803), while the latter temporal feature was negatively correlated (mean regularized weight coefficient = -0.508). This agrees with intuition, since we would expect children who read the words correctly and wait less time to start speaking to receive better scores. The other two features selected disjointly in about half of the 42 cross-validations each were fluency features: the *maximum square root of time recognized as voiced partial words* and the *upper quartile of* *the square root of time recognized as any partial words*. Both of these features' trained weights were negative (mean regularized coefficients of -0.379 and -0.340, respectively), which agreed with our previous findings that disfluent speech negatively impacts evaluators' perception of reading ability.

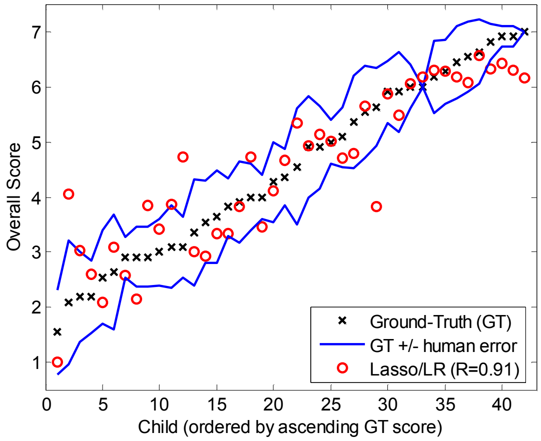

Figure 1: *Linear regression results when using features selected by the lasso (human error = mean human absolute error).*

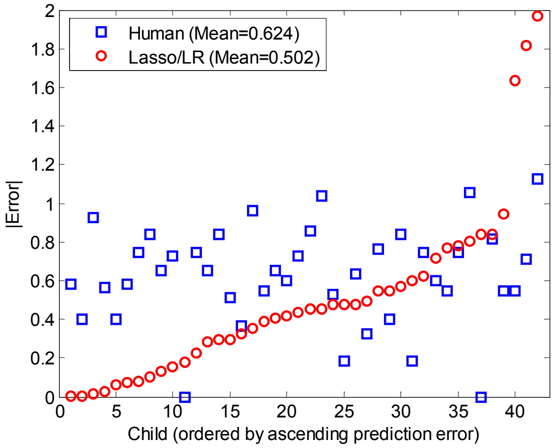

Figure 2: *Comparison of the mean human evaluator errors and the prediction errors made by our best automatic model.*

Figures 1 and 2 show that two-thirds of the predictions made by the linear regression model with the features selected by the lasso had prediction errors less than the mean human evaluator errors. However, three predictions made by the best automatic model had relatively large errors. We noticed in all three cases that the feature with the largest regularized weight coefficient, the *mean of the binary acceptable pronunciation*, was not accurate. Two of the children were making unexpected reading errors; since these pronunciations were not included in any of the dictionaries, some of their mispronunciations were wrongly accepted. We might gain from more complex pronunciation verification methods to improve our results, such as those used in [11]. For the third child with poor automatic prediction, our system rejected many acceptable pronunciations. We feel this was due to acoustic model mismatches, a problem that could be avoided in the future if a small amount of training data for each child was used for acoustic model adaptation. Even with these three outliers, the maximum absolute error made by our best automatic model (1.971) was still less than the mean maximum error made by the 11 evaluators (2.227); the high human error can be attributed to the subjective nature of the assessment and/or noise factors (e.g., evaluator fatigue).

7. Conclusions & Future Work

We found we could accurately predict judgments about children's overall reading ability for one specific reading assessment task. We first extracted many features correlated with the cues evaluators said they used: pronunciation correctness, fluency, and speaking rate. We then used the lasso algorithm to select the most relevant features and applied linear regression to learn a ground-truth evaluator's grading trends. This model: 1) chose, on average, one feature from each of the three feature classes, 2) significantly beat baseline methods using single features, and 3) predicted scores within the mean human error for 28 out of the 42 children.

In some cases, it is important to grade like a single expert evaluator. For example, in a classroom setting, automatic reading assessment software should grade like the children's specific classroom teacher. Future work will attempt to model the subjective judgments of individual expert evaluators.

8. Acknowledgements

This research was funded in part by the National Science Foundation. Special thanks to Professor Fei Sha for his suggestions on appropriate machine learning algorithms.

9. References

[1] M. Black, J. Tepperman, S. Lee, and S. Narayanan, "Estimation of children's reading ability by fusion of automatic pronunciation verification and fluency detection," Proc. Interspeech, Brisbane, Australia, 2008.

[2] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "Tball data collection: the making of a young children's speech corpus," Proc. Eurospeech, Lisbon, Portugal, 2005.

[3] C. Cucchiarini, H. Strik, D. Binnenpoorte, and L. Boves. "Pronunciation evaluation in read and spontaneous speech: a comparison between human ratings and automatic scores," Proc. of the New Sounds, 2000.

[4] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," Computer Speech and Language, 23 (1): pp. 65-88, January, 2009.

[5] H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan, "Pronunciation variations of Spanish-accented English spoken by young children," Proc. Eurospeech, Lisbon, Portugal, 2005.

[6] M. Black, J. Tepperman, S. Lee, P. Price, and S. Narayanan, "Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment," Proc. InterSpeech, Antwerp, Belgium, 2007.

[7] A. Hagen and B. Pellom, "A multi-layered lexical-tree based recognition of subword speech units," Proc. of L&TC, Poznan, Poland, 2005.

[8] R. J. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of Royal Statistical Society, 58:267-288, 1996.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least Angle Regression," Annals of Statistics 32 (2): pp. 407-499.

[10] SparseLab 2.0. Matlab toolbox. http://sparselab.stanford.edu.

[11] J. Tepperman, M. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. Narayanan, "A Bayesian Network classifier for word-level literacy assessment," Proc. Interspeech, Antwerp, Belgium, 2007.