# Automatic Prediction of Children's Reading Ability for High-level Literacy Assessment

Matthew P. Black*, *Student Member, IEEE*, Joseph Tepperman, *Member, IEEE*,
and Shrikanth S. Narayanan, *Fellow, IEEE*

***Abstract*—Automatic literacy assessment technology can help children acquire reading skills by providing teachers valuable feedback in a repeatable, consistent manner. Recent research efforts have concentrated on detecting mispronunciations during word-reading and sentence-reading tasks. These token-level assessments are important since they highlight specific errors made by the child. However, there is also a need for more high-level automatic assessments that capture the overall performance of the children. These high-level assessments can be viewed as an interpretive extension to token-level assessments, and may be more perceptually relevant to teachers and helpful in tracking performance over time. In this work, we model and predict the overall reading ability of young children reading a list of English words aloud. The data consist of audio recordings, collected in real kindergarten to second grade classrooms from children from native English- and Spanish-speaking households.**

**This research is broken into two main parts. The first part is a user study, in which 11 human evaluators rated the children on their overall reading ability based on the audio recordings. The evaluators were volunteers from a diverse background, seven of whom were native speakers of American English and four that were fluent speakers of English as a secondary language. While none of the evaluators were trained reading experts or licensed teachers, a subset of them were linguists and researchers with experience in automatic literacy assessment. As part of this work, we analyzed the effect of the evaluator's background on inter-evaluator agreement.**

**In the second part, we ran machine learning experiments to predict evaluators' scores using features automatically extracted from the audio. The features were human-inspired and correlated with cues human evaluators stated they used: pronunciation correctness, speaking rate, and fluency. We investigated various automated methods to verify the correctness of the word pronunciations and to detect disfluencies in the children's speech using held-out annotated data. Using linear regression techniques, we automatically predicted *individual* evaluators' high-level scores with a mean Pearson correlation coefficient of 0.828, and we predicted *average* evaluator's scores with correlation 0.946. Both these human-machine agreement statistics exceeded the mean inter-evaluator agreement statistics.**



Manuscript submitted March 27, 2010, revised and resubmitted June 25, 2010, and accepted August 25, 2010. This material is based upon work supported by the National Science Foundation under Grants No. IERI 0326228 and CAREER 0238514.

M. P. Black and S. S. Narayanan (e-mail: matthepb@usc.edu, shri@sipi.usc.edu) are with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA. J. Tepperman (e-mail: jtepperman@rosettastone.com) was with the University of Southern California and is now with Rosetta Stone Labs, Boulder, CO 80302 USA.

***Index Terms*—automatic literacy assessment, children's speech, disfluency detection, pronunciation verification**

## I. Introduction

LITERACY assessment is an important element in early education [1], helping bridge the gap between children's learning and teachers' goals [2]. These assessments can occur at different granularities (segmental or suprasegmental) depending on the intended application and reading task. For example, preliterate children are assessed on their knowledge of the letter-to-sound rules of a particular language, while more advanced students are assessed on their ability to fluently read phrases and sentences aloud [3]. Appropriate reading tasks must be designed to elicit speech that facilitates the intended assessment. One common theme among most reading assessment tasks is the use of multiple test items ("tokens") for each subject. This is done for a number of practical reasons. First, it ensures the subjects are provided enough tokens to cover many, or even possibly all, associated linguistic or category variations. Second, it allows evaluators to adjust to the speaking style of the subjects, so accent and idiosyncratic behaviors are taken into account. Third, it provides evaluators with statistically adequate evidence to make global ("high-level") assessments on the subjects' overall performance. In this paper, we are specifically interested in this final aspect: to automatically model and predict evaluators' high-level assessments for a particular reading task widely administered to young children.

There is a need for technology to be incorporated in the classroom to collaboratively assist in reading instruction [4]. We propose in this paper to use automatic computer-based literacy assessments to help teachers, allowing them to better concentrate on lesson-planning and individualized teaching. Automatic computer-based literacy assessments can have several advantages over manual human-based assessments. Manual assessments are very time consuming, requiring one-on-one time. Doing continual assessments may not be feasible in a common scenario like a classroom, where there are several students and only one teacher, and where assessment time competes with instruction. Automatic assessment systems could significantly reduce the time burden of teachers. Manual assessments are also not standardized across evaluators, dependent on factors such as the evaluator's experience, personal biases, and human limitations (e.g., fatigue). Automatic computer-based assessments can provide a more consistent assessment framework, relying on objective features extracted from the available audio-video signals. A

standardized computer-based automatic literacy assessment system could make more meaningful comparisons across children and over time. Finally, automatic literacy assessment systems can be portable and be scaled up to serve large populations of children.

There are several benefits for providing high-level overall assessments rather than (or in addition to) the more typical token-level assessments. First, having knowledge of the overall performance may be particularly useful when tracking performance over time. Second, high-level assessments provide a thumbnail view of a child's performance, which may be useful for teachers by aiding in instruction planning or designing further performance drill-down. Third, high-level assessments may model evaluators' perception better than token-level assessments. Whereas in token-level assessments, decisions are made on the goodness of that particular token, high-level assessments are directly modeling evaluators' interpretation on overall performance, which may be a multi-dimensional and/or non-linear mapping from token-level performance. Therefore, high-level assessments can be viewed as the interpretive extension to token-level assessments.

Automatic high-level literacy assessment is a difficult problem because it involves the modeling and prediction of subjective human judgments. In order to accurately make high-level assessments, the multiple cues human evaluators might use have to be automatically extracted from the available measured observations. In addition, they have to be combined in a way that accurately models the high-level assessment. People might base their assessments on different cues when forming a grading criteria, and even in cases where evaluators use the same cues, they might differ on the relative importance of each. From a signal processing viewpoint, this requires the robust extraction of perceptually relevant features, followed by an appropriate machine learning algorithm that learns the interpretation of these cues, based on individual evaluators or a bank of evaluators.

There has been significant work on reading assessment, especially in second language learning and children's reading applications. Most of the related work has involved adults or children already reading phrases and sentences. We argue that literacy assessments at an earlier age is critical, since it has been shown that early literacy proficiency is a good predictor for reading fluency and comprehension proficiency in later grades [3]-[5]. Importantly, studies have shown a significant decrease in the percentage of poor readers when interventions take place before the second grade [6]. Automatic literacy assessments targeting younger children could help catch problems earlier, and an effective intervention could give children a better chance to grow into competent readers. In addition, much of the related work has concentrated on detecting segmental and suprasegmental errors in production for various reading tasks (e.g., [7]-[14]), but overall performance is rarely estimated. Some previous work has concentrated on providing overall scores (e.g., pronunciation quality [15], fluency [16], reading level [17]), but automatic high-level reading assessments remain relatively under-researched. It should be noted that the idea of modeling global holistic human judgments is not unique to literacy assessment. For example, the computer vision community has viewed this problem in the context of reconciling human evaluations and automatic scene classification [18]-[19].

Literacy assessments can fall under a number of overlapping reading-related skills, such as decoding words, fluently reading sentences aloud, reading comprehension, and writing. In this research, we assess children in kindergarten to second grade on their *overall* ability to fluently decode a list of English words aloud. This reading task is appropriate for this age group and resulted in speech that had a high level of variability in responses, including a range of disfluencies (e.g., hesitating, sounding out the words, elongating phones). While teachers can make use of both acoustic information and visual information (e.g., mouth movement, eye gaze) when assessing children's reading skills, we only have access to one audio signal, recorded from a close-talking microphone. Both the human evaluators and the automatic methods used this single audio channel, which may have resulted in a lower baseline performance for the human evaluators, as compared to a more traditional scoring setup. Future research will incorporate both acoustic and visual information to provide a more realistic scenario to human evaluators and to enable a multimodal approach to automatic literacy assessment. The combined use of audio and video information has been shown to bring increased accuracy and robustness in the context of automatic speech recognition [20]-[21].

In this research, human evaluators listened to the children's speech and rated each on their overall reading ability on a Likert scale of 1 to 7. These human scores were the dependent variable for all our experiments and represented the high-level literacy assessment targets. There is always some level of subjectivity involved in assessment tasks, as is evident in variations across evaluators. Computers can help automate these types of judgments if they are able to make predictions that are in line with human evaluators. In this research, and in related research also involving human assessments (e.g., [12] and [22]-[24]), performance of the automatic system is measured by computing human-computer agreement. One could then view a computer as being competent if it can agree with human evaluators as much as humans agree amongst themselves. Ideally, computers would be able to adapt their grading styles to each evaluator or to a bank of evaluators.

In our previous paper [25], we showed that disfluencies have a perceptual impact on evaluators rating the overall performance of the children. We used a grammar-based automatic speech recognizer to detect disfluencies in the children's speech. In addition, we showed that by combining pronunciation correctness, disfluency features, and temporal speaking rate features, we could predict the *average* evaluator's scores with agreement that was comparable to human inter-evaluator agreement [26]-[27]. In this paper, we improve upon our pronunciation verification and disfluency detection methods and train a system using various feature selection procedures and linear regression techniques. We also extend our analysis to predict individual evaluator's scores. The final optimized system was able to learn both an *individual* evaluator's high-level scores and the *average* evaluators' scores with the same level of agreement with which evaluators agree among themselves.

This paper is organized as follows. Section II discusses the TBALL Project and the TBALL Corpus, on which this paper's work is based upon. Section III describes and analyzes the

human evaluations we administered to attain perceptual judgments. Section IV discusses the features we extracted that correlated with the cues evaluators used when making high-level judgments. Section V discusses the machine learning methods we studied to predict evaluators' high-level assessments. Section VI provides our results and discussion, and we conclude in Section VII.

## II. TBALL Project and Corpus

The Technology-Based Assessment of Language and Literacy (TBALL) Project was formed to create automatic literacy assessment technology for young children in early education from multi-lingual backgrounds [28]-[29]. The TBALL Project's main goal was not to create real-time automated literacy tutors (see [7]-[8] and [30]-[37]) but rather to provide a technological assessment framework that teachers could use to inform their teaching and track children's progress. The reading tasks were designed for and administered to children in actual kindergarten to second grade classrooms in Northern and Southern California. About half of the children were native speakers of American English, with the other half non-native or bilingual speakers of English from a Mexican-Spanish linguistic background.

The young age of the children and diverse population make this project and resulting corpus unique from other existing corpora [38]-[40]. We administered different reading tasks, compared to other automatic literacy assessment projects, to be more geared to preliterate children. These ranged from testing the production of English letter-names, the sounds corresponding to each letter ("letter-sounds"), syllable-blending tasks, to reading a list of isolated words. The resulting speech from a single close-talking headset microphone makes up the TBALL Corpus [41]. Since the reading tests were administered in actual classrooms, the background noises included typical classroom sounds, such as other children's voices and the teacher's voice. The children's demographics (gender, grade, native language) were obtained by forms filled out by assenting parents and were included as part of the corpus when available.

For this work, we analyzed speech from an adaptation of the Beginning Phonic Skills Test (BPST) [42], an isolated word-reading task consisting of 55 pre-determined words. This word list was chosen since it evaluates children's phonemic awareness and decoding skills [43]. The difficulty of the words is steadily increased throughout the reading task, starting with monosyllabic words (e.g., map, left, cute), and ending with multisyllabic words (e.g., silent, respectfully). When administering the test, each word was displayed on a computer monitor one at a time, and the children had up to five seconds to say the word aloud before the next word was shown. The children had the option to advance to the next word before this five-second limit by pressing a button. During the data collection process, a trained research assistant listened beside the child, and if the child mispronounced three words in a row, the assistant manually stopped the session. This was done to prevent the children from getting too frustrated and is not the termination criterion from the BPST as generally administered. As a result, only 11.0% of the children read the full list of 55 words from our sample ($M$ = 21.6 words, $SD$ = 11.2 words). The transition times between words were automatically recorded, and these times were used to split each child's audio into single-word utterances.

Our test set was comprised of the speech from 42 children, each of whom completed at least the first ten words of the isolated word-reading task. These children were selected from a total of 100 children's data to ensure a wide variety of performance levels and reading styles and to be near balanced with respect to gender and native language. We chose 42 children to limit the total amount of speech to approximately 30 minutes to prevent evaluator fatigue when manually assessing the speech (described in Sec. III). To ensure the words read by each child were of comparable difficulty, we only selected words that appeared in the top 25 of the word list. In total, the test set had 770 single-word utterances, an average of 18.3 words per child ($SD$ = 5.07 words). The final demographics of the 42 children were: gender (female=21, male=21), grade (kindergarten=5, first=22, second=15), and native language (English=20, Spanish=18, bilingual=4).

We also constructed a held-out feature development set with 220 children's speech from the isolated word-reading task; this set is described in detail in Sec. IV-B. Lastly, we used 19 hours of held-out speech from word-reading and picture-naming tasks to train 33 monophone acoustic models, a word-level filler "garbage" acoustic model on all speech segments, and a background/silence acoustic model on background segments of the recordings. All acoustic models were three-state Hidden Markov Models (HMMs) with 16 Gaussian mixtures per state. For features, we extracted a 39-dimensional vector, consisting of the first 12 Mel-Frequency Cepstral Coefficients (MFCCs), log energy, and their delta and delta-delta coefficients, every 10 ms using a 25 ms Hamming window. We applied cepstral-mean subtraction across each single-word utterance to help make the features more robust to classroom noise. We used the Hidden Markov Model Toolkit (HTK) [44] for all MFCC feature extraction, acoustic model training, and decoding.

## III. Human Evaluations

### A. Evaluation 1: High-level Literacy Assessment

Evaluation 1 was administered to obtain human perceptual judgments of high-level literacy assessments for the 42 children in the test data. Eleven English-speaking volunteers rated the children on their "overall reading ability." The evaluators fit into four classes: three had worked on children's literacy research for over a year, three were linguists, four were non-native speakers of American English with an engineering background in speech-related research, and three were native English-speaking individuals with no linguistics background or experience with speech or literacy research; the evaluators belonged to only one of the four classes, except for one linguist who also worked with children's speech and a different linguist who was a non-native speaker. While none of the evaluators were licensed teachers or reading experts, we found in previous work that the inter-evaluator agreement between teachers and non-experts was not significantly different for a pronunciation verification task [45]. Analysis of the inter-evaluator agreement for the 11 evaluators in this paper will be provided in Sec. V.

The order of the children was randomized for each evaluator, but the word order within each child's session was maintained. The evaluators were provided the word list, so they could follow the children's progress. A short beeping sound was inserted between each single-word utterance, so the evaluators knew when the transitions between words took place. After listening to the speech from a child, evaluators rated her/his overall reading performance on an integer scale from 1 ("poor") to 7 ("excellent"). Examples of a "poor" reader versus an "excellent" reader were not provided to the evaluators beforehand for two reasons: 1) we did not know in advance whether all evaluators would agree on what a "poor" versus an "excellent" reader was, and 2) we wanted evaluators to come up with their own grading criteria for this reading task. Since evaluators likely needed to listen to a few children before getting comfortable with their own grading scheme, they were permitted to change previously assigned scores.

After the evaluators rated the 42 children, we asked one open-ended question to find which criteria evaluators used when grading the children. This was done to get a rough estimate of the relative importance of various cues people used for this assessment task. The evaluators' responses were grouped into three categories: pronunciation correctness (stated by 10 out of the 11 evaluators), fluency (stated by 9 of 11 evaluators), and speaking rate (stated by 9 of 11 evaluators). It should be noted that none of the evaluators specified that they based their judgment on the child's relative performance at the beginning or end of the word list or on the number of words spoken by the child. The number of spoken words was somewhat artificial for this data, since a human evaluator will not be present to stop the session if the task were administered by a computer; therefore, we do not use the number of words the child spoke as a feature for automatic high-level literacy assessment. While word order and word difficulty most likely had some effect on human evaluators, we assumed each word was equally important in this paper. Coming up with a quantitative system that takes into account a word's importance based on its location in the word list is difficult because these effects are most likely evaluator-dependent. The fact that children read a variable number of words from the word list further complicates the matter. Future work could use machine learning algorithms that take into account word list effects by weighting words differently, as was done in our previous work [12].

Based on the evaluators' responses, we concentrated on automatically extracting features/scores from the audio signal that correlated with pronunciation correctness, fluency, and speaking rate. There has been a significant amount of research on automatic pronunciation verification (accepting or rejecting the pronunciation of a target word), and we will employ some of these techniques on the development set in Sec. IV-C. Speaking rate features and other temporal correlates are also straight-forward to extract if the word pronunciations can be correctly endpointed. However, quantifying fluency is more difficult, since we did not know what made a response "fluent." We used a second human evaluation to discover this.

### *B. Evaluation 2: Perceptual Impact of Disfluencies*

Evaluation 2 explored the impact of fluency on people's perception. We noted five main "disfluencies" in the data: hesitations, sound-outs, elongations of phones, whispering, and speaking with a questioning intonation (perhaps expressing uncertainty). Here, we use the term "disfluency" to describe any speech phenomena that takes away from the natural flow of the pronunciation of the target word. Typically, the term disfluency is used in the context of spontaneous speech for events like fillers (e.g., "uh"), repetitions, repairs, and false starts [46]. However, since this is a reading task and the children are learning how to read (and some are still learning how to speak English as a second language), the types of disfluencies are different from those studied in adult spontaneous speech.

We prescribed a set of conditions necessary for each disfluency type to make the task of labeling disfluencies more objective. The types of disfluencies that occurred in the data before the target word pronunciation included *hesitations*, where the child started to pronounce the target word, paused, and then said the target word, and *sound-outs*, where the child pronounced each phone in the word, pausing between each one, and then pronounced the target word. Some children *whispered* when sounding-out and hesitating, speaking voiced phones in an unvoiced manner. The other two types of disfluencies we noted took place during the pronunciation of the target word. Some children lengthened a phone or syllable of the target word, which we call *elongations*. Lastly, some children's pitch rose at the end of a word's pronunciation, which we refer to as a *question* intonation. It should be noted that these disfluency types were not mutually exclusive within an utterance. For example, a child might hesitate at first, and then say the word with a question intonation, or a child might use a whispered voice while sounding out the word.

For this evaluation, we selected 13 children's speech from the test set which displayed varying levels of the five disfluency types. Since labeling disfluencies is partially subjective, we had two evaluators (the first and second authors) mark each utterance with the presence/absence of each disfluency type. Table I shows that the percent agreement between the evaluators was high, so we used Evaluator 1's labels as the ground-truth for the remainder of our analysis.

| *Disfluency* | *Frequency Counts (out of 146)* | | *% Agreement* |
|---|---|---|---|
| | Evaluator 1 (first author) | Evaluator 2 (second author) | |
| Sound-out | 39 | 38 | 97.95 |
| Hesitation | 27 | 29 | 97.26 |
| Whisper | 22 | 26 | 97.26 |
| Elongation | 13 | 22 | 93.84 |
| Question | 10 | 14 | 95.89 |

*Table I.* The number of utterances (out of 146) that each evaluator labeled as containing each of the five disfluency types and the percentage of utterances in which the two evaluators agreed.

We then had 16 evaluators (eight engineers with speech-related background, four with teaching experience, and four with a linguistics education) rate for each word utterance the fluency of the speech (on an integer scale from 1 to 5). The words were grouped by child, so evaluators could adjust to the speaking style of the children. The resulting fluency scores from the multiple evaluators were transformed to z-scores by subtracting the mean of each evaluator's scores and dividing by the standard deviation. This normalization was done to

allow for more meaningful comparisons of scores between evaluators. We found that the mean normalized fluency score for utterances that contained no disfluencies ($M$ = 0.637, $SD$ = 0.792) was significantly higher than the mean score for utterances that contained at least one disfluency type ($M$ = -0.484, $SD$ = 0.854), $t(2035) = 30.3$, $p < .001$. This shows that indeed utterances which were not labeled with any of the five disfluency types were considered more fluent. We also computed pairwise one-sided $t$-tests to compare the mean normalized fluency scores between disfluency types. Table II shows that the sound-out and hesitation disfluencies were considered the most disfluent, and utterances with whispers were considered more disfluent than ones with question intonations or elongations.

| *Disfluency* | *M* | *SD* | *p*-value | | | |
|---|---|---|---|---|---|---|
| | | | Hes | Wh | Qu | El |
| Sound-out | -0.648 | 0.865 | 0.154 | 0.001 | < .001 | < .001 |
| Hesitation | -0.587 | 0.804 | -- | 0.015 | < .001 | < .001 |
| Whisper | -0.397 | 0.946 | -- | -- | 0.011 | 0.012 |
| Question | -0.210 | 0.714 | -- | -- | -- | 0.271 |
| Elongation | -0.164 | 0.672 | -- | -- | -- | -- |

*Table II*. Statistics of the normalized fluency scores for each of the five disfluency types, along with the resulting $p$-values when using pairwise one-sided $t$-tests to compare the difference in mean scores.

To discover the relative contribution of each disfluency type on the perception of fluency, we also ran a regression analysis. The dependent variable was the vector of normalized fluency scores, and the independent variables were the binary ground-truth labels of the five disfluency types for each utterance. We found these independent variables were able to account for a significant portion of the variance in the fluency scores, $R^2$ = .331, $F(5, 2031) = 201.0$, $p < .001$. As shown in Table III, the coefficient magnitudes for the sound-out, hesitation, and whisper disfluencies were largest, which suggests their presence impacts evaluators' perception of fluency more than the elongation and question intonation disfluencies.

We conjecture that whispers, hesitations, and sound-outs were considered more disfluent because they occurred *in addition* to the pronunciation of the target word, thus breaking up the flow of the speech more than disfluencies that occurred *during* the pronunciation of the target word. Based on these results, we set out to automatically detect these three perceptually relevant disfluencies directly from the audio signal. Sec. IV-D discusses our proposed methods and shows results based on experiments with the development set.

| *Disfluency* | *Coefficient* | *Std. Error* | *t*(2031) | *p*-value |
|---|---|---|---|---|
| Sound-out | -1.206 | 0.045 | -26.68 | < .001 |
| Hesitation | -1.047 | 0.052 | -19.99 | < .001 |
| Whisper | -0.718 | 0.078 | -9.224 | < .001 |
| Elongation | -0.500 | 0.072 | -6.930 | < .001 |
| Question | 0.150 | 0.057 | 2.645 | 0.008 |

*Table III*. Regression analysis of the five disfluency independent variables when estimating the evaluators' normalized fluency scores.

## IV. Feature Extraction

We learned in Evaluation 1 (Sec. III-A) that people considered pronunciation correctness, fluency, and speaking rate to be critical cues in determining the child's overall reading ability. In Evaluation 2 (Sec. III-B), we learned that the whispering, hesitation, and sound-out disfluencies were considered the most perceptually relevant. In this section, we concentrated on extracting features correlated with these cues. In Sec. IV-A, we describe the construction of a dictionary for each target word, which we will use for much of our subsequent analyses. In Sec. IV-B, we describe the development set in greater detail. In Sec. IV-C and IV-D, we use this development set to experiment with automatic pronunciation verification and disfluency detection methods, respectively. In Sec. IV-E, we apply these methods to the test data to extract features for high-level literacy assessment.

### A. Dictionary

For each target word, we constructed a dictionary with the help of an expert teacher and linguist. Acceptable and foreseeable unacceptable phonemic pronunciations were included in each target word's dictionary. These unacceptable pronunciations were made by substituting correct pronunciations with common letter-to-sound errors; for example, /k ah t/ ("cut") was augmented to the dictionary as a common reading mistake for /k y uw t/ ("cute"). Also, due to the large Mexican-American background in the corpus, we added common Spanish-speaking influenced variants to the dictionary, based on [47]. On average, each target word had 1.20 acceptable pronunciations and 3.03 foreseeable unacceptable pronunciations in its dictionary. Across all target words, 33 phonemes were used in these pronunciations. (We trained a monophone HMM for each, as described in Sec. II).

### B. Feature Development Set

To test various feature extraction methods, we used the development set, introduced in Sec. II; this speech data was not included in either the test set or the acoustic model training data. Most of the demographic information about the 220 children was unknown, since the children's parents did not provide this optional information: gender (female=25, male=43, unknown=152), grade (kindergarten=5, first=36, second=27, unknown=152), and native language (English=21, Spanish=38, bilingual=5, unknown=156).

Since we were interested in detecting mispronunciations and disfluencies as relevant features, we first needed to explicitly label these in the development set. Three evaluators manually verified the pronunciation of each target word in the development set (binary accept/reject) and labeled each single-word utterance with the five disfluency types. All utterances in which there was excessive background noise or problems during the recording (e.g., cut-off speech) were marked by the evaluators and ignored. There was no overlap in evaluations, since this manual labeling process is costly (we saved approximately 20 hours of time by using three evaluators with no overlap)[1]. In total, 2800 single-word utterances were annotated. 22.95% of the utterances had at least one disfluency type, and 2.49% had two or more types. Hesitations were marked in 8.93% of the utterances, sound-outs in 5.94%, elongations in 5.15%, whispering in 3.13%, and question intonations in 2.13%. 37.1% of the target word

[1] We found with a subset of 13 children's speech that evaluators agreed with one another an average of 93% of the time when verifying the correctness of a word's pronunciation [26].

pronunciations were rejected. If at least one disfluency was marked in the utterance, the probability the pronunciation was rejected increased to 0.578. This means that disfluent speech and mispronunciations were positively correlated events.

### C. Automatic Pronunciation Verification

The purpose of automatic pronunciation verification is to accept or reject a pronunciation. To characterize the performance of this task, we borrow metrics commonly used in detection theory and binary classification tasks: precision (1), recall (2), balanced F-score (3), false-alarm rate (4), misdetection rate (5), and Matthews correlation coefficient (6). In these equations, a true positive (TP) is correctly detecting a mispronunciation, a false positive (FP) is incorrectly detecting a mispronunciation, a true negative (TN) is correctly detecting no mispronunciation, and a false negative (FN) is incorrectly detecting no mispronunciation.

$$P \equiv \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

$$R \equiv \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$F \equiv \frac{2 \cdot P \cdot R}{P+R} \tag{3}$$

$$FA \equiv \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{4}$$

$$MD \equiv \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

$$MCC \equiv \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{6}$$

In our previous papers [26]-[27], we used a simple automatic pronunciation verification method, which acts as our baseline method for this work. We ran automatic speech recognition (ASR) with the dictionary of acceptable and foreseeable unacceptable pronunciations on each single-word utterance in the development set. We tried a number of different finite-state grammars (FSGs) to endpoint the pronunciation automatically: allowing for recognition of the background model (BG) vs. the garbage model (GG) at the start and end of the utterance vs. allowing both to be recognized; requiring the BG or GG models to be recognized at the start and end of the utterance vs. making it optional; allowing for repetitions of the BG and GG models at the start and end of the utterance vs. only allowing them to be recognized once. We found, in general, that allowing for the GG model to be recognized at the start and end of the utterance resulted in more false alignments of the target word pronunciation, probably because the GG model was trained on speech data. Fig. 1 shows an example of the FSG that attained the highest F-score. In this FSG, the BG model is recognized (with the option of multiple recognitions) at the start and end of each utterance, and there is one required forced alignment of either the background model (BG), the garbage model (GG), or one of the acceptable or unacceptable pronunciations in the dictionary for that target word. A pronunciation is accepted if and only if an acceptable pronunciation of the target word is recognized; otherwise, it is rejected. The first row of Table IV shows the performance of this method (called LEX), with respect to the metrics (1)-(6).

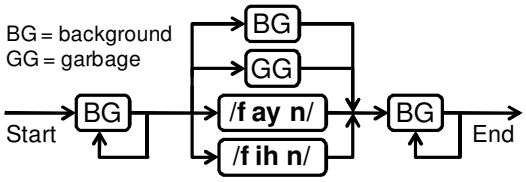


*Fig. 1.* The finite-state grammar (FSG) used for the LEX pronunciation verification method (for the sample word, "fine"). The pronunciation is accepted if and only if the correct pronunciation (/f ay n/) is recognized; otherwise, it is rejected.

The second automatic pronunciation verification method we tried was Goodness of Pronunciation (GOP) scoring [22]. In this method, a forced alignment of acceptable pronunciation(s) of the target word is first made to the utterance. The resulting output will contain the phonemes recognized and their corresponding boundaries and acoustic log-probabilities. An unconstrained phone loop is then decoded across each phone segment, and a final GOP score for each phone is computed by subtracting the acoustic log-probability of the phone loop from the log-probability of the forced-aligned phone. High GOP scores correspond to phones that are more likely to be correctly pronounced, and a GOP score threshold can be made to reject phones with GOP scores below the threshold.

We applied this technique to each utterance in the development set and got the best results, in terms of maximizing F-score, when we did not threshold on individual phones within a target word but rather thresholded on the average GOP score across the word (where each phone is counted equally). Equation (7) shows how to compute the GOP phone score (*O* is the acoustics, *p* is the phone, *PL* is the phone-loop, and *N* is the number of frames of phone *p*). Equation (8) shows how to compute the GOP word-level score, by calculating the mean of the GOP phone scores for the word. Finally, (9) shows how we thresholded the GOP word-level score to ultimately reject or accept the pronunciation. This threshold, *T*, can be chosen to attain specific performance characteristics; in this paper, we chose the *T* that maximized F-score, but other popular optimization criteria could be used (e.g., equal precision and recall, equal false-alarm and misdetection rates, maximum Matthews correlation coefficient). Table IV shows the performance of this GOP scoring method for this optimal value of *T*.

$$\mathrm{GOP}(p) \equiv \tfrac{1}{N}\log\frac{P(O|p)}{P(O|PL)} \tag{7}$$

$$\mathrm{GOP}(w) \equiv \tfrac{1}{|p\in w|}\sum_{p\in w}\mathrm{GOP}(p) \tag{8}$$

$$\mathrm{Reject}(w) \equiv \begin{cases} 1, & \mathrm{GOP}(w) \le T \\ 0, & \mathrm{GOP}(w) > T \end{cases} \tag{9}$$

We also tried combining the LEX and GOP methods. The LEX method makes use of target word knowledge and common letter-to-sound mistakes a child might make

(especially with the influences of Spanish), but this method may be unable to detect errors if the child produces an unforeseeable realization of the target word. On the other hand, the GOP method is able to detect errors made that were not foreseeable but might not be able to tease apart close pronunciations with one phone substitution. We combined the two methods by first running the LEX method and then using the GOP scoring method only on pronunciations that were *accepted* by the LEX method. Table IV shows results for all three proposed pronunciation verification methods, and Figs. 2 and 3 show performance as a function of GOP score threshold. We attained the highest F-score (0.802) and Matthews correlation coefficient (0.680) by using the combined LEX + GOP scoring method.

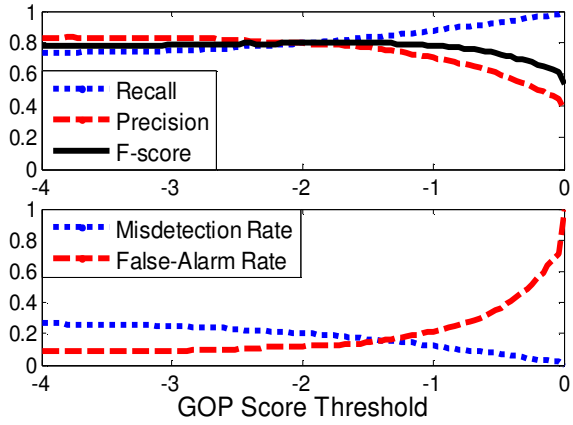


*Fig. 2.* Performance of LEX+GOP pronunciation verification method as a function of the GOP score threshold (all pronunciations with GOP scores lower than this threshold were rejected).

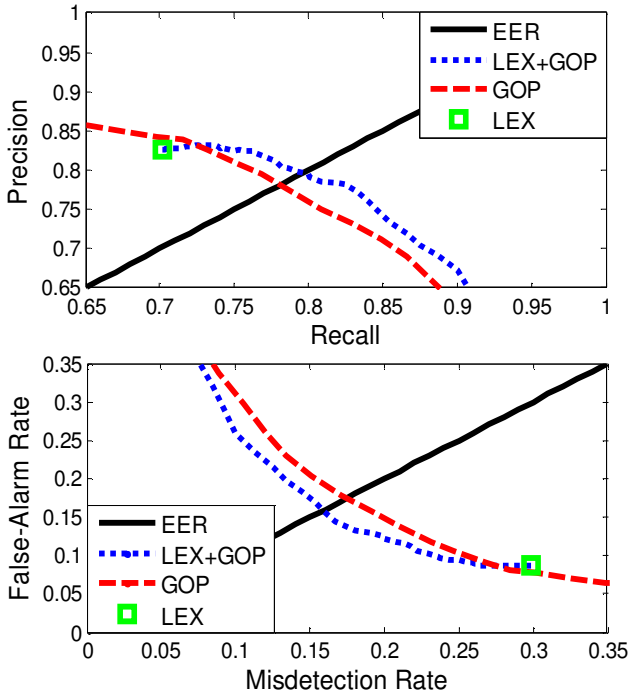


*Fig. 3.* Performance of the three proposed pronunciation verification methods (LEX, GOP, LEX+GOP). The GOP method performances are shown as the GOP score threshold is varied from -10 to 0. EER is the equal error rate for the displayed metrics.

| *System Type* | *R* | *P* | *F* | *MD* | *FA* | *MCC* |
|---|---|---|---|---|---|---|
| LEX | 0.702 | **0.826** | 0.759 | 0.298 | **0.087** | 0.639 |
| GOP* | 0.785 | 0.785 | 0.785 | 0.216 | 0.127 | 0.657 |
| LEX+GOP* | **0.832** | 0.775 | **0.802** | **0.168** | 0.143 | **0.680** |

* GOP score threshold chosen to maximize F-score

*Table IV.* Performance of the pronunciation verification methods: LEX, GOP, and the combination LEX+GOP, in terms of (1)-(6). The LEX+GOP method attained the highest F-score and MCC.

### *D. Automatic Disfluency Detection*

Since this is a reading assessment task, the target words are known ahead of time. Furthermore, the sounding-out, hesitation, and whispering disfluencies were partial word manifestations of some pronunciation variant of the current target word. This facilitated the use of automatic speech recognition using finite-state grammars (FSGs) to detect disfluent speech. We first developed two simple baseline FSGs. The first baseline (Base1) allowed for repetitions of the target word with optional silence decoded in between. If two or more target words were recognized, the utterance was deemed disfluent; otherwise, it was deemed fluent. This baseline was chosen since the disfluencies usually consisted of phonemes that were present in the target word. The second baseline (Base2) inserted a phone loop (again with optional silence decoded between phones) prior to a required forced alignment of the target word. If one or more phones were recognized, the utterance was deemed disfluent; otherwise, it was deemed fluent. This second baseline was chosen since oftentimes the full target word was not spoken during a disfluency, so a phone loop allowed for partial words to be recognized. Table V shows the performance of these two baselines, in terms of the same six metrics we used before (1)-(6). Here, a "true positive" is the correct detection of a disfluency. As shown in Table V, Base1 suffered from low recall (high misdetection rate), since the grammar was unable to recognize partial words, while Base2 suffered from low precision (high false-alarm rate), since its unconstrained phone loop resulted in a high number of false alarms.

To improve upon these baselines, we created a two-stage procedure for detecting disfluencies that combined both baselines, allowing for partial words to be recognized using only phones present in the target word. In the first stage, we designed a disfluency-specialized FSG to ensure a low misdetection rate (high recall). In the second stage, we rejected some of these detections to reduce the false-alarm rate. The first stage in the disfluency detection was introduced in [25] and based on work in [48]-[50]. We created target-word specific FSGs to recognize partial words. Since most disfluencies were partial word manifestations of the target word (or a partial word manifestation of a common mispronunciation of the target word), we created constrained FSGs that only allowed phones in the target word to be recognized and only in the order they appear in the dictionary. We experimented with many FSG designs: an unconstrained phone-loop consisting only of phones within the target word pronunciation(s) vs. requiring phones to be recognized in the order they appear in the target word pronunciation(s); allowing for repetitions and skipping of phones; requiring the first phone to be recognized vs. allowing it to be skipped; and allowing for optional repetitions of the BG model to be

recognized between phones. All the FSG designs had high recall statistics above 0.94, so we chose to use the FSG shown in Fig. 4, since it had the highest precision statistic (Table V).

Analyzing the errors made in stage 1, we noticed that many of the false-alarms were due to the recognition of unvoiced phones like stops (/k/, /p/) and fricatives (/f/, /s/). These "noise-like" phones were similar to the classroom noise, and therefore, more susceptible to false alarms than vowels and other voiced phones. We tried a number of methods to reject some of these false alarms while still maintaining a low misdetection rate: 1) rejecting utterances below a minimum number of partial words recognized, 2) rejecting partial words that were below a minimum length in time, 3) rejecting partial words that were below a minimum acoustic model log-likelihood, 4) rejecting partial words that were below a minimum GOP phone-level score (7). We got the best results, in terms of maximizing F-score, by rejecting recognized partial words that were shorter than a minimum time threshold. Figs. 5 and 6 show how these performance metrics vary as a function of the threshold, and Table V shows the performance of the proposed two-stage disfluency detector when using the threshold that maximized F-score.

Compared with the two baseline methods, we attained the highest F-score (0.783) and Matthews correlation coefficient (0.737) with this two-stage FSG method. Further examining the performance of the two-stage FSG method when choosing the threshold that maximizes the F-score, 94.35% of the hesitations and 93.94% of the sound-outs were successfully detected. It most likely was unable to detect as many instances of whispering (58.62%) because of acoustic mismatches with the non-disfluent speech we used to train the acoustic models. In addition, whispered speech is more likely to be dominated by background noise.

| *System Type* | *R* | *P* | *F* | *MD* | *FA* | *MCC* |
|---|---|---|---|---|---|---|
| Base1: Word Reps | 0.175 | **0.965** | 0.297 | 0.825 | **0.001** | 0.376 |
| Base2: Phone Loop | **0.989** | 0.273 | 0.428 | **0.011** | 0.568 | 0.336 |
| FSG: Stage 1 | 0.942 | 0.611 | 0.741 | 0.058 | 0.129 | 0.697 |
| FSG: Stage 2* | 0.885 | 0.702 | **0.783** | 0.115 | 0.081 | **0.737** |

* Stage 2 threshold of 0.125 s chosen to maximize F-score

*Table V.* Performance of the multiple disfluency detection methods: baseline 1 (Base1), baseline 2 (Base2), and the 2 stages of the target word-specific finite-state grammar (FSG) procedure. The proposed 2-stage FSG method achieved the highest F-score and MCC.

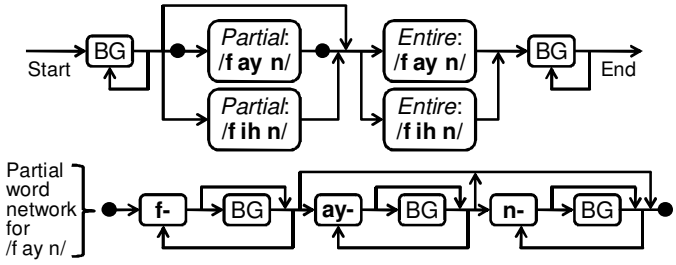


*Fig. 4.* The stage 1 disfluency detection finite-state grammar (FSG) for the sample word, "fine," which has two entries in the dictionary (/f ay n/, /f ih n/). The FSG allows partial word manifestations of the target word to be recognized before a required forced-alignment of the entire target word. (BG is the background acoustic model.)

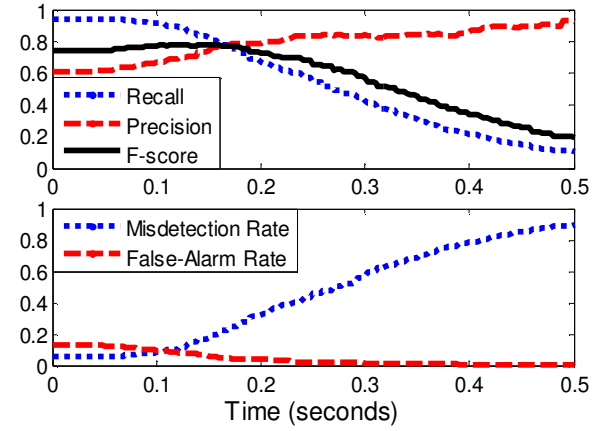


*Fig. 5.* The performance of the stage 2 finite-state grammar (FSG) method as a function of the partial word length threshold (below which all partial words were rejected).

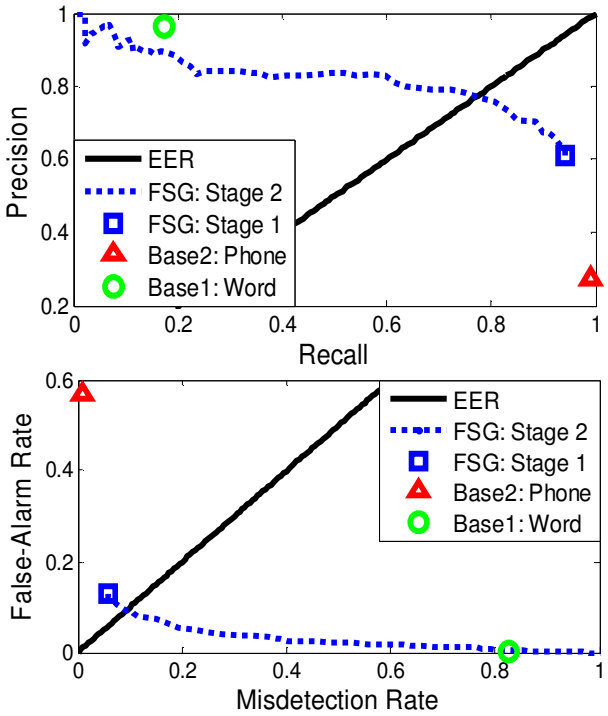


*Fig. 6.* Performance of the two baseline systems (Base1 and Base2) and target word-specific finite-state grammar (FSG) procedure (stages 1 and 2). The FSG stage 2 performance is shown as the minimum partial word length threshold is varied from 0 to 2 seconds. EER is the equal error rate for the displayed metrics.

### *E. Feature Extraction on the Test Data*

We next applied these pronunciation verification and disfluency detection methods on the test data to extract scores correlated with evaluators' perception of the children's reading ability. Since this was an isolated word-reading task, we extracted all features at the word-level. Table VI shows the 48 scores extracted for each word. There are 10 scores based on the pronunciation verification methods, 12 scores based on the disfluency detection methods, and 26 speaking rate and other temporal scores based on both methods. When applying the pronunciation verification and disfluency detection methods discussed in Sec. IV-C and IV-D, we used all threshold and parameter values that maximized the F-score

on the development set. Note that we extracted the square root of all temporal features as an additional feature. This was done since the temporal features oftentimes had distributions that were skewed because of a small percentage of long times. The square root helped push the distributions towards a more bell-shaped distribution, which better fit the distributions assumed in the linear models we applied in Sec. V. We found this square root transformation performed empirically well in our previous work [27]; future work could find a more optimal transform by choosing the root that makes the distribution most "normal."

We extracted our final set of features for each child by computing 12 statistics across each word-level score for all the words spoken by the child: mean, standard deviation, skewness, minimum, minimum location (normalized by number of words spoken by child), maximum, maximum location (normalized), range, lower quartile, median, upper quartile, interquartile range. This produced our final feature set of 576 features per child. The next section will discuss how we used feature selection and supervised learning algorithms to properly deal with this over-generation of potentially useful features.

## V. Prediction of Children's Reading Ability

Sec. IV explained our feature extraction, which resulted in 576 child-level features. In this section, we used this feature set to predict children's reading ability, as rated by the 11 evaluators (see Sec. III-A). Since there were 11 evaluators, there were many ways to pose this learning problem. We first analyzed the inter-evaluator agreement of the evaluators using Pearson's correlation coefficient. Equation (10) is Pearson's correlation between two vectors of scores, $y_1$ and $y_2$, where $y_j = \begin{bmatrix} y_j^1 & \dots & y_j^{42} \end{bmatrix}^T$, and $\mu_{y_j}$ is the mean score for $y_j$. Note that the "42" in this equation refers to the total number of children we are assessing.

$$Corr(y_1, y_2) \equiv \frac{\sum_{i=1}^{42}(y_1^i - \mu_{y_1})(y_2^i - \mu_{y_2})}{\sqrt{\sum_{i=1}^{42}(y_1^i - \mu_{y_1})^2 \sum_{i=1}^{42}(y_2^i - \mu_{y_2})^2}} \qquad (10)$$

Table VII shows the pairwise inter-evaluator agreement using (10) and also displays four sets of average agreement for each evaluator. All 11 evaluators' scores had higher correlations with *ground-truth* scores (computed by averaging the other evaluators' scores), as compared to the mean pairwise correlation with the other evaluators. This means that the ground-truth scores are representative of the "average" evaluators' perception. In addition, for 9 of the 11 evaluators, agreement was higher when using all evaluators to compute ground-truth scores, as compared to using just evaluators within the evaluators' background(s). While Table VII shows that the "experts" had higher average correlations, none of the correlation coefficients were significantly different (all $p > 0.1$), using a difference in correlation coefficients test that transformed the coefficients with the Fisher Z-transform. As a result, we considered all evaluators in this paper.

| *Name* | *Description* | *Domain* |
|---|---|---|
| $VER_1$^ | Was unacceptable pronunciation recognized? | {0, 1} |
| $VER_2$^ | Was common reading error recognized? | {0, 1} |
| $VER_3$^ | Was Spanish-related error recognized? | {0, 1} |
| $VER_4$^ | Was garbage (GG) recognized? | {0, 1} |
| $VER_5$^ | Was background/silence (BG) recognized? | {0, 1} |
| $VER_6$* | Log-likelihood of *acceptable* pronunciation | (-∞, 0] |
| $VER_7$ | GOP(*w*) – see (7) | (-∞, 0] |
| $VER_8$ | Reject(*w*) – see (8) | {0, 1} |
| $VER_9$^* | 2-stage verification method – see Sec. IV-C | {0, 1} |
| $VER_{10}$ | $VER_1$ + $VER_8$ | {0, 1, 2} |
| $FL_1$# | Number of recognized partial words | {0, 1, …} |
| $FL_2$# | Was at least one partial word recognized? | {0, 1} |
| $FL_3$# | Length of recognized partial words [s] | [0, 5) |
| $FL_4$# | Length of silence *between* partial words [s] | [0, 5) |
| $FL_5$# | Length of *all* silence recognized [s] | [0, 5) |
| $FL_6$ | $FL_3$ + $FL_4$ | [0, 5) |
| $FL_7$ | $FL_3$ + $FL_5$ | [0, 5) |
| $FL_8$ : $FL_{12}$ | Square root of $FL_3$ through $FL_7$ | $[0, 5^{1/2})$ |
| $SR_1$ | Utterance length [s] | (0, 5] |
| $SR_2$^ | Target word start time [s] | [0, 5) |
| $SR_3$^ | Target word end time [s] | (0, 5] |
| $SR_4$^ | Number of syllables spoken / ($SR_3$ – $SR_2$) | (0, ∞) |
| $SR_5$^ | ($SR_3$ - $SR_2$) / Number of syllables spoken | (0, 5] |
| $SR_6$^ | Number of phones spoken / ($SR_3$ – $SR_2$) | (0, ∞) |
| $SR_7$^ | ($SR_3$ - $SR_2$) / Number of phones spoken | (0, 5] |
| $SR_8$# | Speech start time (partial word or target word) | [0, 5) |
| $SR_9$# | Speech end time | (0, 5] |
| $SR_{10}$# | Number of syllables spoken / ($SR_9$ – $SR_8$) | (0, ∞) |
| $SR_{11}$# | ($SR_9$ – $SR_8$) / Number of syllables spoken | (0, 5] |
| $SR_{12}$# | Number of phones spoken / ($SR_9$ – $SR_8$) | (0, ∞) |
| $SR_{13}$# | ($SR_9$ – $SR_8$) / Number of phones spoken | (0, 5] |
| $SR_{14}$ : $SR_{26}$ | Square root of $SR_1$ through $SR_{13}$ | -- |

^ using a finite-state grammar as depicted in Fig. 1
* using forced alignment of acceptable pronunciations of target word
# using a finite-state grammar as depicted in Fig. 4

*Table VI.* Features extracted for each word in the test data (VER = verification, FL = fluency, SR = speaking rate). The temporal features have an upper bound of 5 seconds since this was the maximum time allotted per word. All GOP scores in this study were finite, since all phone probabilities were non-zero.

We chose three different learning problems, meant to show how well the system could do in three typical scenarios. In all scenarios, we trained and tested the system using leave-one-child-out cross-validation, i.e., trained the system on 41 children and tested it on the held-out child, and repeated this process for all 42 children. In the first scenario, we trained the system on an individual evaluator's scores and tested on the same evaluator's held-out score. Scenario 1 is a test for how well the system can predict a single evaluator's scores if trained on that evaluator. In scenario two, we predicted individual evaluator's scores using ground-truth scores to train the system. In this scenario, we computed a ground-truth score for each child by taking the mean score across the 10 held-out evaluators. Scenario 2 is a test for how well the system can predict single evaluator's scores if trained on a bank of held-out evaluators; scenario 2 is analogous to testing how much an evaluator agrees with "off-the-shelf" assessment tools trained on a group of different evaluators. In the third scenario (and the only one we did in our previous work [26]-[27]), we predicted ground-truth scores using these ground-truth scores to train the system. Therefore, scenario 3 is a test for how well the system can predict a bank of evaluators if that same bank of evaluators trains the system.

| *Evaluator (Background)* | *Pairwise Evaluator Correlation* | | | | | | | | | | *Avg. Correlations* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | mean | | ground-truth | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | intra | all | intra | all |
| 1 (Naïve) | | | | | | | | | | | 0.776 | 0.770 | 0.810 | 0.833 |
| 2 (Naïve) | 0.70 | | | | | | | | | | 0.767 | 0.803 | 0.808 | 0.874 |
| 3 (Naïve) | 0.85 | 0.83 | | | | | | | | | 0.843 | 0.860 | 0.909 | 0.940 |
| 4 (Non-native) | 0.72 | 0.70 | 0.84 | | | | | | | | 0.813 | 0.780 | 0.850 | 0.844 |
| 5 (Non-native) | 0.76 | 0.85 | 0.86 | 0.84 | | | | | | | 0.857 | 0.848 | 0.913 | 0.928 |
| 6 (Non-native) | 0.82 | 0.84 | 0.89 | 0.86 | 0.91 | | | | | | 0.880 | 0.868 | 0.944 | 0.949 |
| 7 (Non-nat., Ling.) | 0.82 | 0.79 | 0.88 | 0.74 | 0.82 | 0.87 | | | | | 0.810 | 0.816 | 0.866 | 0.886 |
| 8 (Linguist) | 0.69 | 0.86 | 0.84 | 0.73 | 0.87 | 0.86 | 0.73 | | | | 0.777 | 0.814 | 0.801 | 0.888 |
| 9 (Linguist, Expert) | 0.79 | 0.82 | 0.88 | 0.76 | 0.83 | 0.83 | 0.88 | 0.83 | | | 0.860 | 0.840 | 0.923 | 0.916 |
| 10 (Expert) | 0.77 | 0.80 | 0.86 | 0.79 | 0.86 | 0.86 | 0.81 | 0.87 | 0.86 | | 0.857 | 0.837 | 0.886 | 0.913 |
| 11 (Expert) | 0.78 | 0.84 | 0.87 | 0.82 | 0.88 | 0.88 | 0.82 | 0.86 | 0.87 | 0.86 | 0.863 | 0.844 | 0.895 | 0.922 |
| | | | | | | | | | | ***Avg:*** | **0.828** | **0.827** | **0.873** | **0.899** |

*Table VII*. Pairwise evaluator correlations between the 11 evaluators (*Naïve* = native English speakers with no background in linguistics or children's literacy, *Non-Native* = non-native English speakers with an engineering background in speech-related research, *Linguist* = taken at least two graduate-level linguistics courses, *Experts* = more than a year working on children's literacy research). Average correlations were computed two different ways ("mean" and "ground-truth") and across two different groupings of evaluators ("intra" and "all"). "Mean" is the average pairwise evaluator correlation, and "ground-truth" is the correlation between an evaluator's scores and the averaged scores of the other evaluators. "Intra" calculations compare evaluators with the same background(s), while "all" calculations compare all evaluators' scores.

To validate our results, we chose three metrics. Pearson's correlation coefficient (10) is the primary metric. Equation (11) is the mean absolute error between vectors of scores, $y_1$ and $y_2$. Equation (12) is the maximum absolute error between the two vectors of scores, $y_1$ and $y_2$.

$$E_{mean}(y_1, y_2) \equiv \frac{1}{42}\sum_{i=1}^{42}\left|y_1^i - y_2^i\right| \qquad (11)$$

$$E_{max}(y_1, y_2) \equiv \max\left(|y_1^1 - y_2^1|, \ldots, |y_1^{42} - y_2^{42}|\right) \qquad (12)$$

Before running experiments, we calculated human agreement statistics for all three metrics. Table VIII shows the human agreement statistics between the 11 evaluators, calculated in two ways: 1) using pairwise comparisons between individual evaluators and 2) comparing individual evaluators to the ground-truth scores of the other 10 evaluators. The pairwise comparisons had lower agreement than the ground-truth comparisons for all three metrics (lower correlation, higher mean absolute error, and higher maximum absolute error).

| *Evaluator Domain* | Mean (Standard Deviation) | | |
|---|---|---|---|
| | *Corr* | $E_{mean}$ | $E_{max}$ |
| Pairwise | **0.827** (0.032) | **0.810** (0.180) | **2.800** (0.701) |
| Ground-Truth | **0.899** (0.038) | **0.624** (0.137) | **2.227** (0.388) |

*Table VIII*: Human agreement statistics for the 3 metrics (10)-(12).

For all three scenarios, we chose to use linear regression techniques because of their simplicity and interpretability. The choice of function estimation methods made particular sense for scenarios 2 and 3, where the trained dependent variable was quasi-continuous. We also chose to use regression techniques for scenario 1, even though the dependent variable is ordinal, in order to ensure the results across the three scenarios are comparable. We did not z-normalize the dependent variable in any of the three scenarios since it had no impact on performance and since knowledge of the mean and standard deviation of the evaluator's scores in a real-life scenario is not always practical to attain.

For all experiments, we used leave-one-child-out cross-validation to separate train and test sets. Optimal learning parameters and feature subsets (when applicable) were computed on each cross-validation *train* set separately by using leave-one-child-out cross-validation; we chose the parameter settings (feature subsets) that maximized correlation between the automatic predictions and the evaluators' scores. This cross-validation approach effectively made use of all labeled data and simultaneously ensured that we were testing the true predictive power of our features/methods.

We developed two baseline systems for this paper, based on token-level pronunciation assessment research, where pronunciation correctness is often solely considered. Both baselines use simple linear regression with single features. The first uses the mean of feature $VER_1$, and the second uses the mean of feature $VER_8$ (Table VI). These two features represent the fraction of words mispronounced by the child, as determined by the LEX and GOP pronunciation verification methods, respectively (Sec. IV-C). Therefore, the baseline methods test whether one-dimensional token-level assessments can be extended to high-level assessments by simply computing an average over the token-level assessments.

A logical extension to these baseline systems would be to use multiple linear regression with the full set of 576 child-level features. Equation (13) shows this linear model, where $\bar{y}$ is the centered (mean subtracted) vector of human scores, $X$ is the matrix of child-level features, $w$ is the vector of coefficient weights, and $\epsilon$ is a zero mean Gaussian random variable. The objective function $J$ in this case is (14), and (15) is the analytical solution which minimizes $J$.

$$\bar{y} = Xw + \epsilon \qquad (13)$$

$$J = \|\bar{y} - Xw\|^2 \equiv (\bar{y} - Xw)^T(\bar{y} - Xw) \qquad (14)$$

$$w = (X^TX)^{-1}X^T\bar{y} \qquad (15)$$

Due to multicollinearity in the feature set, the solution to the inverse in (15) would be numerically unstable. We addressed this problem by trying various feature selection methods that model the dependent variable as a linear combination of a sparse set of independent variables. Choosing a subset of the features implicitly filters out redundant, irrelevant, and/or

noisy features and makes the model easier to interpret. To show the relative merits of each feature, we ran simple linear regression (SLR) with each child-level feature individually.

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) [51]. Forward selection iteratively adds features that optimize Pearson's correlation coefficient (10). 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 $\lambda$-weighted $L_1$ regularization term to the objective function, as shown in (16). This penalizes solutions with large weight coefficients (which often occurs when features are correlated) and promotes sparse models. Thus, many of the weight coefficients will be identically zero. We implemented the lasso using the least angle regression (LARS) algorithm, since there is no analytical solution to the lasso objective function [52]-[53]. Note that we must standardize the features to ensure the regularization term is applied equally to all features. We accomplished this by centering the feature matrix $X$ and dividing by the standard deviation of each feature; this normalization is denoted in (16) as $\bar{\bar{X}}$.

$$J = \left\|\bar{y} - \bar{\bar{X}}w\right\|^2 + \lambda\|w\| \qquad (16)$$

## VI. Results & Discussion

Table IX shows the performance for the two aforementioned baseline methods, the performance of the best SLR features for each of the three feature types, and the performance for the three feature selection methods. Table X provides coefficient statistics and lists which features were selected in at least 20% of the 42 cross-validations for the best performing feature selection method in each of the three train/test scenarios. We see from these results that scenario 1 (training and testing on individual scores) is the hardest, followed by scenario 2 (training on ground-truth scores and testing on a held-out evaluator), followed by scenario 3 (training and testing on ground-truth scores). We can explain the relative difficulty of the three scenarios using the following high-level description. Individual evaluators' scores can be viewed as "noisy," due to the subjective nature of the assessment task. Averaging the evaluators' scores can be seen as a method to "de-noise" individual evaluators' scores. We get the best results in scenario 3, where we train and test on ground-truth ("de-noised") scores and the worst results when we train and test on individual ("noisy") evaluators' scores.

In Table X, we see that the baseline methods (that used the means of $VER_1$ and $VER_8$), did not use the best features, since the mean of $VER_{10}$ proved to be a better predictor of the children's overall reading ability in all three learning scenarios. $VER_{10}$ combines $VER_1$ and $VER_8$ into one trinary verification feature (Table VI). When limited to one feature, this single verification feature achieved the best results in terms of all three metrics and for all three scenarios, compared with using a single fluency or speaking rate feature (Table IX).

| *Scenario:Method* | Mean (Standard Deviation when applicable) | | |
|---|---|---|---|
| | *Corr* | $E_{mean}$ | $E_{max}$ |
| **1**: Base1 ($VER_1$) | **0.734** (0.062) | **0.914** (0.106) | **2.880** (0.358) |
| **1**: Base2 ($VER_8$) | **0.746** (0.048) | **0.930** (0.121) | **2.682** (0.475) |
| **1**: SLR (best *VER*) | **0.769** (0.065) | **0.882** (0.072) | **2.610** (0.632) |
| **1**: SLR (best *FL*) | **0.748** (0.054) | **0.895** (0.139) | **3.041** (0.480) |
| **1**: SLR (best *SR*) | **0.705** (0.105) | **0.924** (0.197) | **3.385** (0.799) |
| **1**: Forward LR | **0.792** (0.074) | **0.815** (0.160) | **2.659** (0.700) |
| **1**: Stepwise LR | **0.805** (0.055) | **0.786** (0.143) | **2.852** (0.722) |
| **1**: Lasso | **0.807** (0.087) | **0.814** (0.223) | **2.467** (0.565) |
| **1**: Lasso, then LR | **0.828** (0.070) | **0.721** (0.153) | **2.549** (0.560) |
| **2**: Base1 ($VER_1$) | **0.741** (0.053) | **0.968** (0.111) | **3.044** (0.376) |
| **2**: Base2 ($VER_8$) | **0.756** (0.044) | **0.970** (0.107) | **2.763** (0.687) |
| **2**: SLR (best *VER*) | **0.812** (0.041) | **0.856** (0.084) | **2.510** (0.643) |
| **2**: SLR (best *FL*) | **0.731** (0.051) | **0.979** (0.137) | **3.345** (0.505) |
| **2**: SLR (best *SR*) | **0.724** (0.062) | **0.975** (0.175) | **3.374** (0.554) |
| **2**: Forward LR | **0.869** (0.038) | **0.712** (0.138) | **2.407** (0.520) |
| **2**: Stepwise LR | **0.861** (0.035) | **0.730** (0.133) | **2.589** (0.703) |
| **2**: Lasso | **0.851** (0.041) | **0.846** (0.139) | **2.544** (0.552) |
| **2**: Lasso, then LR | **0.854** (0.037) | **0.753** (0.125) | **2.526** (0.495) |
| **3**: Base1 ($VER_1$) | **0.809** | **0.735** | **2.405** |
| **3**: Base2 ($VER_8$) | **0.822** | **0.743** | **1.909** |
| **3**: SLR (best *VER*) | **0.888** | **0.596** | **1.601** |
| **3**: SLR (best *FL*) | **0.799** | **0.759** | **2.762** |
| **3**: SLR (best *SR*) | **0.783** | **0.789** | **2.858** |
| **3**: Forward LR | **0.946** | **0.365** | **1.594** |
| **3**: Stepwise LR | **0.946** | **0.365** | **1.594** |
| **3**: Lasso | **0.925** | **0.535** | **1.837** |
| **3**: Lasso, then LR | **0.940** | **0.414** | **1.636** |

*Table IX*. Automatic performance for the three scenarios described in Sec. V. The methods above the dotted line use single features, and the ones below use multiple features. The numbers in red are the best performance achieved for the three scenarios.

| *Scenario:Method* | *Feature* | *% folds selected* | Coefficient stats | |
|---|---|---|---|---|
| | | | *M* | *SD* |
| **1**: Base1 ($VER_1$) | Mean($VER_1$) | -- | -0.755 | 0.051 |
| **1**: Base2 ($VER_8$) | Mean($VER_8$) | -- | -0.771 | 0.042 |
| **1**: SLR (best VER) | Mean($VER_{10}$) | -- | -0.851 | 0.012 |
| **1**: SLR (best FL) | Uquart($FL_{12}$) | -- | -0.801 | 0.036 |
| **1**: SLR (best SR) | Uquart($SR_{14}$) | -- | -0.771 | 0.055 |
| **1**: Lasso, then LR | Range($VER_7$) | 50.9 | 0.140 | 0.129 |
| | Mean($VER_7$) | 44.4 | 0.258 | 0.267 |
| | Iquart($SR_2$) | 38.1 | -0.314 | 0.203 |
| | Uquart($FL_{12}$) | 31.2 | -0.142 | 0.116 |
| | Mean($VER_6$) | 27.9 | 0.199 | 0.149 |
| | Lquart($FL_2$) | 21.3 | -0.276 | 0.147 |
| **2**: Base1 ($VER_1$) | Mean($VER_1$) | -- | -0.824 | 0.009 |
| **2**: Base2 ($VER_8$) | Mean($VER_8$) | -- | -0.839 | 0.007 |
| **2**: SLR (best VER) | Mean($VER_{10}$) | -- | -0.898 | 0.005 |
| **2**: SLR (best FL) | Uquart($FL_{12}$) | -- | -0.852 | 0.006 |
| **2**: SLR (best SR) | Uquart($SR_{14}$) | -- | -0.829 | 0.009 |
| **2**: Forward LR | Mean($VER_{10}$) | 99.1 | -0.604 | 0.017 |
| | Uquart($FL_{12}$) | 97.0 | -0.442 | 0.019 |
| **3**: Base1 ($VER_1$) | Mean($VER_1$) | -- | -0.825 | 0.007 |
| **3**: Base2 ($VER_8$) | Mean($VER_8$) | -- | -0.840 | 0.006 |
| **3**: SLR (best VER) | Mean($VER_{10}$) | -- | 0.899 | 0.004 |
| **3**: SLR (best FL) | Uquart($FL_{12}$) | -- | -0.852 | 0.005 |
| **3**: SLR (best SR) | Uquart($SR_{14}$) | -- | -0.829 | 0.006 |
| **3**: Forward LR | Mean($VER_{10}$) | 100.0 | -0.605 | 0.012 |
| | Uquart($FL_{12}$) | 97.6 | -0.442 | 0.013 |

*Table X*. Statistics of the standardized coefficients for the baseline, single feature, and best performing feature selection methods.

Table X shows that the best performing speaking rate feature was the upper quartile of $SR_{14}$, which is simply the duration of the utterance. The best fluency feature was the upper quartile of $FL_{12}$, which is the square root of the total duration of silence and disfluencies. $FL_{12}$ can be viewed as a

hybrid fluency and speaking rate feature; it is a fluency feature since more disfluencies will increase its value, and it is a speaking rate feature since slower speaking rates (longer periods of silence between words) will also increase its value. Table X shows that the signs of the trained coefficients for these features ($VER_1$, $VER_8$, $VER_{10}$, $SR_{14}$, and $FL_{12}$) were all negative, which means lower ratings of overall reading ability would be predicted for children with many mispronunciations, long periods of disfluent speech or silence, and longer (slower) responses. These interpretations agree with intuition.

Within each scenario, the automatic methods that used multiple features outperformed the single feature methods (including the two baselines) for all three metrics. For scenario 1, we achieved the best results in terms of correlation (10) and mean absolute error (11) using the lasso regression as a pre-processing feature selection algorithm and then training the coefficient weights using multiple linear regression; we achieved the best results in terms of maximum absolute error (12) using the lasso method to select features and train the weights. For scenario 2, we achieved the best results for all three metrics using forward feature selection. For scenario 3, we got equally good results with both the forward selection and stepwise linear regression methods. Forward linear regression most likely achieved the best results for Scenarios 2 and 3 because the resulting feature set included only two features, so a greedy forward selection process was sufficient and outperformed more complicated feature selection methods. On the other hand, for Scenario 1, the lasso algorithm provided a more robust objective function for the more difficult learning problem, and the average number of features selected at each cross-validation was much higher at 5.6. Thus, in this case, the forward selection algorithm was unable to robustly select this higher number of features. The stepwise linear regression method can be viewed as the middle ground, which explains why its performance generally fell between that of the forward selection and the lasso. Table X also shows that for scenarios 2 and 3, the forward selection algorithm chose the top performing verification and fluency features for almost all of the cross-validation folds. However, for scenario 1, the lasso algorithm selected a variety of features, depending on the evaluator.

Scenario 3 was the only one in which we achieved a significantly higher correlation coefficient, compared to the best baseline system ($z = 2.78$, $p = .005$). Fig. 7 shows performance (in terms of correlation) of the different automatic feature selection methods for all three learning scenarios, compared to the human agreement statistics computed earlier. For the human agreement in this plot, we show the pairwise inter-evaluator correlations in scenario 1, and the ground-truth correlations in scenarios 2 and 3. We see from this plot that we were able to achieve a comparable level of human agreement for scenario 1 with the lasso and linear regression learning method. The mean automatic performance correlation of 0.828 was actually higher than the average pairwise human evaluator correlation of 0.827, although this difference was not significant ($z = 0.014$, $p = 0.989$). This means that the system trained on a particular evaluator will agree with that evaluator about as much as other evaluators will agree with that evaluator. In scenario 2, the automatic performance improved, benefiting from being trained on the perceptions of multiple evaluators, but its average performance was less than human agreement in this scenario, since the scores being predicted were from a held-out evaluator (resulting in a mismatched train/test condition). For scenario 2, the human evaluators' scores were correlated with ground-truth scores with 0.899 correlation, which was not significantly higher than automatic correlation of 0.869 ($z = 0.609$, $p = 0.542$). In scenario 3, the automatic performance is greater than average human agreement, although not significantly ($z = 1.44$, $p = 0.151$). In this scenario, the automatic system had the benefit of having multiple evaluators to train the system and also a matched test set composed of the same evaluators.

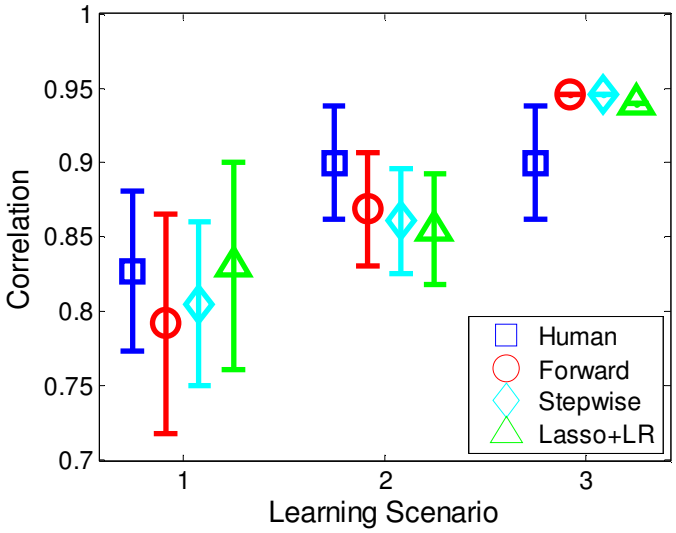


*Fig. 7.* Mean and standard deviation of human evaluator agreement compared to the automatic performance for the three feature selection methods: forward selection, stepwise regression, and the lasso followed by linear regression.

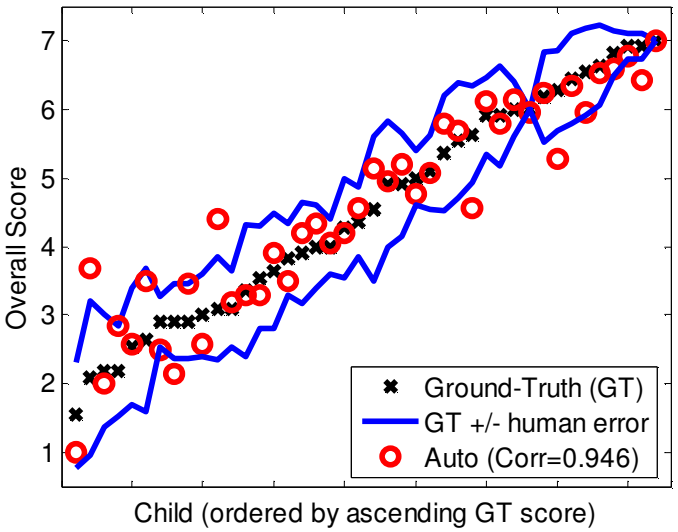


*Fig. 8.* Linear regression results when using features selected using forward selection for scenario 3. "Human error" is the mean absolute difference from the ground-truth (GT) to held-out evaluators' scores.

Fig. 8 shows the automatic predictions for the best automatic system in scenario 3. The automatic predictions were inside the mean human errors for 34 out of 42 (81%) of the children. We ran a final experiment by re-running scenario 3 using random subsets of evaluators (ranging from 2 to 10 evaluators). Fig. 9 shows these results when using the forward selection and lasso/linear regression methods. Again, for this plot, we also show agreement between the human evaluators (comparing individual evaluators to the ground-truth scores of the other selected evaluators). We chose 10 random subsets of evaluators for each value of the number of evaluators chosen. We see from this plot that human agreement and automatic performance both improve as a function of the number of evaluators. More importantly, we see that automatic performance is relatively high, even when using multiple evaluators with just two evaluators. This shows that the system benefits from the joint modeling of evaluators with as few as two evaluators.

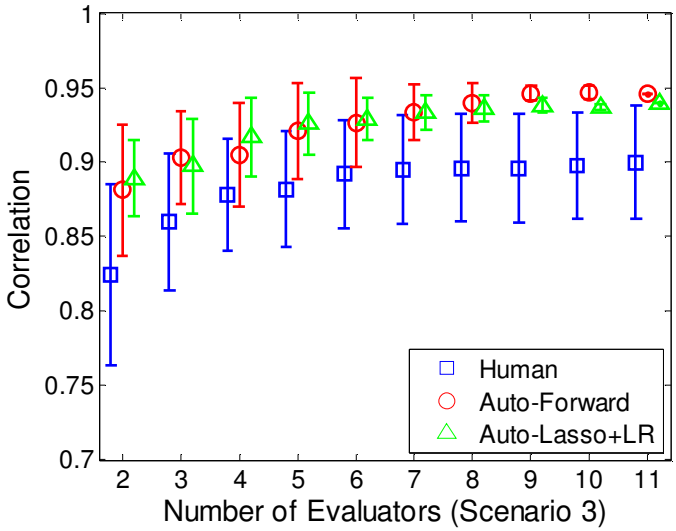


*Fig. 9.* Correlation between predictions and evaluators' scores for learning scenario 3 as a function of the number of evaluators used to compute the ground-truth scores. It shows that both human agreement and automatic performance increase as the number of evaluators increases. Automatic performance with nine or more evaluators is significantly higher than with two evaluators ($z = 1.94$, $p = 0.048$).

## VII. Conclusion

This paper addresses the need for automatic literacy assessments by predicting high-level ratings of children's overall reading ability, based on their performance reading a list of words aloud. We chose to use a modeling scheme that linearly combined a sparse set of features that spanned the ones actual human evaluators said they used (pronunciation correctness, fluency, and speaking rate). The resulting multi-dimensional models implicitly weight the importance of the selected features and offer a more interpretive assessment than the more common token-level assessments. As part of this work, we developed methods to automatically detect mispronunciations and disfluencies on a development training set, using grammar-based automatic speech recognition.

The automatic models performed best when trained on a bank of evaluators and when the train and test set were matched. This type of automatic processing could be especially useful in a classroom environment, where the teacher or a number of teachers could train the system to mimic their grading trends. High-level assessments could then be used by teachers to ensure the children are learning at an appropriate rate and to help inform their lessons. This type of collaboration between technology and teachers could transform the classroom.

In the future, we would like to incorporate both audio and video information for a more realistic scoring scenario. We would also like to extend this high-level literacy assessment to other reading tasks. We imagine applying it within a framework that examines children's skills across various reading tasks, so as to provide teachers with analysis on areas in which a child might be excelling versus an area in which he/she may need more practice or instruction.

## Acknowledgment

Special thanks to the entire TBALL Project team and to Professor Fei Sha for his suggestions on appropriate machine learning algorithms.

## References

[1] P. Black and D. Wiliam, "Assessment and classroom learning," *Assessment in Education: Principles, Policy, & Practice,* vol. 5, no. 1, pp. 7-74, Mar. 1998.

[2] M. Heritage, "Knowing what to do next: The hard part of formative assessment," in *Proc. Association of Educational Assessment*, Valletta, Malta, Nov. 2009.

[3] J. R. Paratore and R. L. McCormack, *Classroom Literacy Assessment: Making Sense of What Students Know and Do*. New York, NY: The Guilford Press, 2007.

[4] National Reading Panel, "Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implication for reading instruction," Tech. Rep. 00-4769, National Institute for Child Health and Human Development, National Institute of Health, Washington, DC, 2000.

[5] A. DeBruin-Parecki, "Evaluating early literacy skills and providing instruction in a meaningful context," *High/Scope Resource*, vol. 23, no. 3, pp. 5-10, 2004.

[6] S. Otaiba and J. Torgesen, "Effects from intensive standardized kindergarten and first-grade interventions for the prevention of reading difficulties," *Handbook of response to intervention*, Springer Publishers, pp. 212-222, 2007.

[7] K. Lee, A. Hagen, N. Romanyshyn, S. Martin, and B. Pellom, "Analysis and detection of reading miscues for interactive literacy tutors," in *Proc. Int. Conf. on Computational Linguistics*, Geneva, Switzerland, Aug. 2004.

[8] J. Mostow, A. G. Hauptmann, L. L. Chase, and S. Roth, "Towards a reading coach that listens: Automated detection of oral reading errors," in *Proc. Nat. Conf. on Artificial Intelligence*, Washington DC, USA, July 1993.

[9] M. Black, J. Tepperman, A. Kazemzadeh, S. Lee, and S. Narayanan, "Automatic pronunciation verification of English letter-names for early literacy assessment of preliterate children." in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009.

[10] M. Black, J. Tepperman, A. Kazemzadeh, S. Lee, and S. Narayanan, "Pronunciation verification of English letter-sounds in preliterate children," in *Proc. Interspeech*, Brisbane, Australia, Sept. 2008.

[11] J. Tepperman, M. Gerosa, and S. Narayanan. "A generative model for scoring children's reading comprehension," in *Proc. Workshop on Child, Computer and Interaction*, Chania, Crete, Greece, Oct. 2008.

[12] J. Tepperman, S. Lee, A. Alwan and S. Narayanan, "A generative student model for scoring word reading skills," *IEEE Transactions on Audio, Speech, and Language Processing*, Accepted 2010.

[13] 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 reading assessment," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.

[14] S. Wang, P. Price, M. Heritage, and A. Alwan, "Automatic evaluation of children's performance on an English syllable blending task." In *Proc. of SLaTE Workshop*, Farmington, PA, USA, Oct. 2007.
[15] 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 & Language*, vol. 23, no. 1, pp. 65-88, Jan. 2009.
[16] C. Cucchiarini, H. Strik, and L. Boves. "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *J. of the Acoustical Society of America*, vol. 111, no. 6, pp. 2862-2873, 2002.
[17] J. Duchateau, L. Cleuren, H. Van Hamme, P. Ghesquière, "Automatic assessment of children's reading level," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.
[18] M. R. Greene and A. Oliva, "Recognition of natural scenes from global properties: seeing the forest without representing the trees," *Cognitive Psychology*, vol. 58, no. 2, pp. 137-176, Mar. 2009.
[19] A. Oliva, and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal in Computer Vision*, vol. 42, no. 3, pp. 145-175, May 2001.
[20] S. Chu and T. S. Huang, "Bimodal speech recognition using coupled hidden Markov models," Proc. ICSLP, Beijing, 2000.
[21] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, A. Kantor, P. Lal, L. Yung, A. Bezman, S. Dawson-Haggerty, and B. Woods, "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop," Proc. ICASSP, Hawaii, 2007.
[22] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95-108, 2000.
[23] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimations of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787-800, 2007.
[24] S. Tuchschmid, M. Bajka, and M. Harders, "Comparing automatic simulator assessment with expert assessment of virtual surgical procedures," *Lecture Notes in Computer Science*, F. Bello and S. Cotin (Eds.), Springer-Verlag, pp. 181-191, 2010.
[25] 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," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.
[26] M. Black, J. Tepperman, S. Lee, and S. Narayanan, "Estimation of children's reading ability by fusion of automatic pronunciation verification and fluency detection," in *Proc. Interspeech*, Brisbane, Australia, Sept. 2008.
[27] M. Black, J. Tepperman, S. Lee, and S. Narayanan, "Predicting children's reading ability using evaluator-informed features," in *Proc. Interspeech*, Brighton, U.K., Sept. 2009.
[28] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: the role of multiple information sources," in *Proc. Int. Workshop on Multimedia Signal Processing*, Chania, Crete, Greece, Oct. 2007.
[29] P. Price, J. Tepperman, M. Iseli, T. Duong, M. Black, S. Wang, C. K. Boscardin, M. Heritage, P. D. Pearson, S. Narayanan, and A. Alwan, "Assessment of emerging reading skills in young native speakers and language learners." *Speech Communication*, vol. 51, no. 10, pp. 968-984, Oct. 2009.
[30] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, Virgin Islands, Dec. 2003.
[31] A. Hagen, B. Pellom, S. Van Vuuren, and R. Cole, "Advances in children's speech recognition within an interactive literacy tutor," in *Proc. Human Language Technology Conference*, Boston, MA, USA, May 2004.
[32] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proc. Nat. Conf. on Artificial Intelligence*, Seattle, WA, USA, Aug. 1994.
[33] J. Mostow and J. Beck, "When the rubber meets the road: Lessons from the in-school adventures of an automated reading tutor that listens," *Scale-up in education*, Vol. 2, B. Schneider & S.-K. McDonald (Eds.), Rowman & Littlefield Publishers, Lanham, MD, pp. 183-200, 2007.
[34] J. Mostow, G. Aist, C. Huang, B. Junker, R. Kennedy, H. Lan, D. Latimer, R. O'Connor, R. Tassone, B. Tobin, and A. Wierman, "4-month evaluation of a learner-controlled reading tutor that listens," *The path of speech technologies in computer assisted language learning: From research toward practice*, V. M. Holland and F. P. Fisher (Eds.), New York: Routledge, pp. 201-219, 2008.
[35] S. M. Williams, D. Nix, and P. Fairweather, "Using speech recognition technology to enhance literacy instruction for emerging readers," in *Proc. Int. Conf. of Learning Sciences*, Mahwah, NJ, USA, June 2000.
[36] P. Cosi and B. Pellom, "Italian children's speech recognition for advanced interactive literacy tutors," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005.
[37] J. Duchateau, M. Wigham, K. Demuynck, and H. Van Hamme, "A flexible recogniser architecture in a reading tutor for children," in *Proc. Speech Recognition & Intrinsic Variation*, Toulouse, France, May 2006.
[38] M. Eskenazi, J. Mostow, and D. Graff, *The CMU Kids Corpus*, published Linguistic Data Consortium, Philadelphia, PA, USA, 1997.
[39] K. Shobaki, J. P. Hosom, and R. A. Cole, "The OGI kids' speech recognizers and corpus," in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000.
[40] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005.
[41] 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," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005.
[42] J. Shefelbine. *BPST – Beginning Phonics Skills Test*, 1996.
[43] S. Wren, "Descriptions of early reading assessments," Southwest Educational Development Laboratory, Nov. 2004.
[44] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK Version 3.4)*, University of Cambridge, Cambridge, U.K., http://htk.eng.cam.ac.uk, Dec. 2006.
[45] 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. Interspeech*, Pittsburgh, PA, USA, Sept. 2006.
[46] E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California, Berkeley, CA, USA, 1994.
[47] H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan, "Pronunciation variations of Spanish-accented English spoken by young children," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005.
[48] A. Hagen, and B. Pellom, "A multi-layered lexical-tree based token passing architecture for efficient recognition of subword speech units," in *Proc. Language and Technology Conf.*, Poznań, Poland, Apr. 2005.
[49] A. Hagen, B. Pellom, "Data driven subword unit modeling for speech recognition and its application to interactive reading tutors" in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005.
[50] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861-873, Dec. 2007.
[51] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of Royal Statistical Society*, vol. 58, no. 1, pp. 267-288, 1996.
[52] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407-499, Apr. 2004.
[53] D. Donoho, V. Stodden, and Y. Tsaig, "About SparseLab," Stanford University, http://sparselab.stanford.edu, Version 2.0, Mar. 2007.

**Matthew Black** (S'07) received the B.S. degree with highest distinction and honors in electrical engineering from the Pennsylvania State University, University Park, in 2005, and the M.S. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2007. He was a graduate-level research intern at the IBM T.J. Watson Research Center, Yorktown, in summer 2007 and is currently a Ph.D. candidate in the Signal Analysis and Interpretation Laboratory (SAIL) at USC.

His research interests are in behavioral signal processing, specifically in the automatic quantification and emulation of human observational processes to describe human behavior. This includes the development of engineering tools and solutions for societally-significant domain applications in education, family studies, and health.

Mr. Black is a member of the IEEE Signal Processing Society. He is a recipient of the Alfred E. Mann Innovation in Engineering Doctoral Fellowship 2010-2011 and the Simon Ramo Scholarship 2009-2010 at USC.

**Joseph Tepperman** received his Ph.D. degree in Electrical Engineering from the University of Southern California in 2009.

His thesis work was on automatic pronunciation evaluation over multiple time-scales, designed for applications in literacy and second-language instruction. Prosodic cues, articulatory representations, subjective judgments, and pedagogical applications continue to be recurring interests throughout his research work. He also uses speech technology to make art installations and music.

Some awards he has received include a USC President's Fellowship (2003-2009) and an ISCA Grant (2007). He has served as a peer reviewer for the journals Speech Communication, Bilingualism: Language and Cognition, and IEEE Transactions on Audio, Speech, and Language Processing. Currently he is a Speech Researcher with Rosetta Stone Labs in Boulder, Colorado.

**Shrikanth (Shri) Narayanan** is the Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics and Psychology and as the founding director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered information processing and communication technologies with a special emphasis on behavioral signal processing and informatics.

Shri Narayanan is a Fellow of IEEE, the Acoustical Society of America, and the American Association for the Advancement of Science (AAAS) and a member of Tau-Beta-Pi, Phi Kappa Phi and Eta-Kappa-Nu. Shri Narayanan is also an Editor for the Computer Speech and Language Journal and an Associate Editor for the IEEE Transactions on Multimedia, IEEE Transactions on Affective Computing, and the Journal of the Acoustical Society of America. He was also previously an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000-04) and the IEEE Signal Processing Magazine (2005-2008).

Shri Narayanan is a recipient of a number of honors including Best Paper awards from the IEEE Signal Processing society in 2005 (with Alex Potamianos) and in 2009 (with Chul Min Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010-11. He has published over 400 papers and has eight granted U.S. patents.