interspeech 2008 incorporating SST 08

Estimation of Children's Reading Ability by Fusion of Automatic Pronunciation Verification and Fluency Detection

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

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

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

Abstract

Pronunciation verification of children's reading is a difficult task in itself, but automatic reading assessment software must also detect and evaluate other phenomena that influence human evaluators. Using an isolated word-reading task, we first show that humans use both pronunciation correctness (accuracy) and fluency information in their assessment of the reading ability of kindergarten to second grade children. Next, we used disfluency-specialized grammars and trained a Bayesian Network to automatically classify the fluency and accuracy of an utterance. Finally, we used these automatically determined scores to estimate evaluators' scores of children's reading ability with a 0.91 correlation.

Index Terms: children's speech, disfluency detection, pronunciation verification, automatic assessment

1. Introduction

Formative assessment of the development of children's reading skills is important to ensure that children are learning. Assessment of early literacy for children in grades kindergarten and first grade involves testing for proficiency in phonemic awareness by reading letter-names, letter-sounds, and blending syllables. Older and more advanced children are tested on reading words and sentences, and reading comprehension [1,2]. These assessment tasks are meant to inform instruction in an ongoing and individualized fashion.

Automating aspects of the assessment of these reading tasks can have several advantages. Reading assessment takes one-on-one time, which a teacher may not always afford to give. An automatic reading assessment system can supply a teacher with information on which children need what specific help, thereby facilitating their instructional planning. It also has the benefit of standardizing the grading process, reducing subjective variability in oral production-based skill evaluations, arising for example, due to pronunciation differences. Furthermore, automatic reading assessment systems allow for self-paced practice and provide meaningful longitudinal information that can be used to track children's performance.

There are numerous engineering challenges, however, in creating automatic reading assessment tools. The first is in the design of the reading task to mimic actual assessment conditions. Next, modeling and processing the children's speech is challenging due to a range of robustness factors. For instance, language and age-specific acoustic models may have to be trained if there is a disparity between children's demographics [3,4]. Of critical importance is the ability to differentiate acceptable speaker accent variability from unacceptable pronunciations. Finally, the system must also be robust to both environmental noise and production variability, including disfluencies [5,6].

Importantly, an automatic reading assessment system should grade as an ideal evaluator would, in an unbiased yet expert-like fashion. These systems need to be carefully designed to detect all phenomena that affect an evaluator's judgment (including, when applicable, pronunciation accuracy, fluency, speaking rate, and speaker confidence). If multiple factors are relevant, the designer has a choice to make: report each separately, or find a meaningful way to fuse these factors together in providing an integrated assessment measure.

In this paper, using the case of an isolated word-reading task, we show that both the correctness (*accuracy*) of the pronunciations and the *fluency* of the speech play significant roles in the perception of the reading ability of children. This paper also describes an approach for how we can automatically estimate and fuse these two information sources together to mimic actual human evaluators rating the overall reading ability of children.

2. Corpus

The data used in this study was from the Tball Project [1], recorded in Los Angeles Kindergarten to second grade classrooms in realistic noise conditions with standard headset microphones [7]. The corpus was comprised of the speech from over 250 children from multi-lingual backgrounds; the young ages and accent variability made this corpus particularly challenging [3,4]. We focused on the data corresponding to the isolated word reading task, in which the children read 55 words aloud. One word was displayed by an animated user interface on a computer screen for a maximum of 5 seconds before the next one was shown. These transition times were automatically recorded and used to segment the resulting audio file into single word utterances, which were then used for automated assessment.

3. Subjective Human Evaluation

In order to model the overall reading ability of a child, we first needed to understand how human evaluators assigned an overall score. Toward that, we performed a subjective human evaluation administered using 14 people, five with teaching experience and seven with some linguistics education [6]. Each evaluator listened to approximately ten word utterances each from 13 children. The children's data were chosen to represent a range of reading abilities and so-called disfluency types (sounding-out, hesitations, whispering, elongated onsets, and question intonations). We chose words with similar difficulty, so we could make the assumption that each utterance was of equal significance to the evaluators.

Accepted after peer review of full paper

For each word, the evaluator rated the binary *accuracy* of the pronunciation and also rated how *fluently* the child spoke (on a scale from 1 to 5). After listening to all the words from a particular child, the evaluator also rated the overall reading ability of the child on a scale from 1 to 7. Lastly, we asked evaluators to rate, on a scale from 1 to 5, the relative importance of accuracy versus fluency information when making this judgment.

In [6], we found that accuracy was considered more important than fluency in determining the children's overall reading ability. However, fluency significantly affected the evaluators, even in instances when the child eventually read the word correctly. We extend this previous research in section 3.1 by computing evaluator agreement statistics for the three measures: accuracy, fluency, and overall scores. In section 3.2, we demonstrate that we can predict the evaluators' overall scores by using evaluators' accuracy and fluency marks as features.

3.1. Human Agreement Statistics

We computed pairwise evaluator agreement for three different measures per child: 1) mean accuracy rating i.e., the fraction of items marked as acceptable pronunciations 2) mean fluency rating 3) overall score. We were only interested in agreement of the *mean* accuracy/fluency ratings and not agreement at the individual utterance level. We calculated the *ground-truth* scores for the three measures by averaging across all evaluators for each child; we call these "ground-truth" scores since they represent the *average* evaluator's opinion. For all three measures, we also computed agreement between each evaluator and the ground-truth scores (using leave-one-out cross-validation for these calculations).

Evaluator Domain	*Mean Accuracy*	*Mean Fluency*	*Overall Score*
Minimum Pairwise Evaluator Correlation	0.8300	0.1843	0.4324
Mean Pairwise Evaluator Correlation	0.9337	0.7498	0.7638
Minimum Evaluator Correlation with GT Labels	0.9118	0.6615	0.7201
Mean Evaluator Correlation with GT Labels	0.9637	0.8557	0.8640

Table 1: *Evaluator agreement correlations between pairwise evaluators and between evaluators and ground-truth (GT) scores for three measures: mean accuracy, mean fluency, and overall score*

As shown in Table 1, mean pairwise evaluator agreement is higher for the mean accuracy marks than for the mean fluency marks ($p=0.11$) and the overall scores ($p=0.13$). This is most probably because fluency and overall scores are more loosely defined and subjective than accuracy. The relatively low agreement statistics for the overall scores indicate that a system intended for an individual teacher should be trained by that teacher. Table 1 also demonstrates that evaluator correlation with the ground-truth scores was consistently high and always significant ($p<0.01$) for all three measures. Therefore, in circumstances where the teacher may not be qualified or does not have time to train the system, modeling based on more subjective measures needs to be trained on multiple evaluators' opinions.

3.2. Estimation of Overall Scores

We used the ground-truth overall scores as our dependent variable in these experiments. We investigated three sets of independent variables: 1) mean accuracy only 2) mean fluency only 3) mean accuracy and mean fluency. We chose to use linear regression for this estimation problem and partitioned the data using leave-one-out cross-validation. We quantified the performance of the resulting models using two metrics: 1) correlation 2) percent root-mean-squared (rms) error between the estimated overall scores and the ground-truth overall scores.

Evaluator Domain	*Indep. Vars.*	*Leave-one-out Cross-Validation*	
		Correlation	*rms error*
Worst Evaluator	*Fl.*	0.087	31.76 %
Worst Evaluator	*Acc.*	0.720	21.23 %
Worst Evaluator	*Acc+Fl*	0.849	16.00 %
Mean Evaluator	*Fl.*	0.651	21.37 %
Mean Evaluator	*Acc.*	0.815	17.48 %
Mean Evaluator	*Acc+Fl*	0.923	11.22 %
GT Evaluator	*Fl.*	0.820	17.19 %
GT Evaluator	*Acc.*	0.855	15.69 %
GT Evaluator	*Acc+Fl*	0.980	5.98 %

Table 2: *Predictive power of independent variables with different evaluator domains when estimating ground-truth (GT) overall scores*

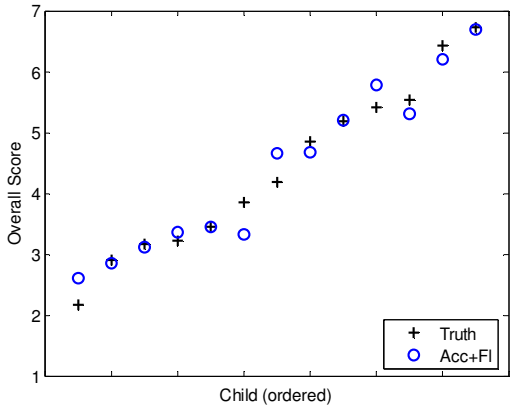

Figure 1: *Leave-one-out cross-validation linear regression results with ground-truth overall scores as dependent variable and ground-truth accuracy and fluency scores as independent variables*

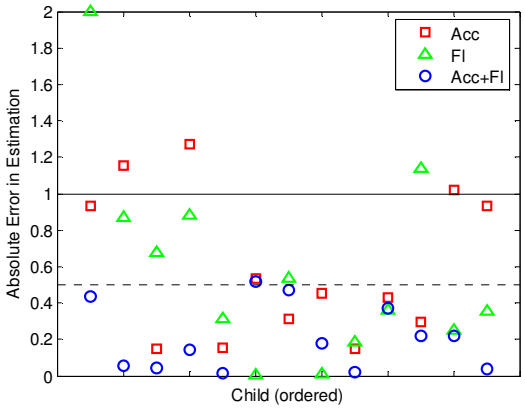

Figure 2: *Absolute error in predicting ground-truth overall scores with ground-truth mean accuracy and mean fluency scores*

Table 2 shows the results with different evaluator domains and combinations of accuracy and fluency features. Accuracy had more predictive power than fluency, but the

highest correlation and lowest errors occurred when using both features. The best linear model used the ground-truth accuracy and fluency scores as features and achieved a 0.98 correlation and 5.98 percent rms error. Figure 1 is the regression plot for this case, with Figure 2 showing the resulting absolute error plot.

4. Automatic Accuracy & Fluency Detection

Motivated by the analysis in section 3, we needed to automatically detect the accuracy and fluency of an utterance in order to estimate children's overall reading ability. We accomplished this by using our previous work in [6,8]. Sections 4.1 and 4.2 briefly explain our methods to calculate fluency and classify accuracy of an utterance, respectively.

4.1. Disfluency Detection

Since sound-outs and hesitations affected human perception of fluency the most and were also the most prevalent disfluencies in our data [6], we concentrated on detecting these partial words. We took a similar approach as in [9] by constructing disfluency-specialized grammar structures for each target word. The grammars optionally permitted the speech recognizer to detect partial words prior to a required recognition of the whole target word. This grammar-based approach successfully recognized over 88% of the sound-outs and hesitations with an 8.9% false alarm rate. Figure 3 shows the final grammar architecture, with full details in [6].

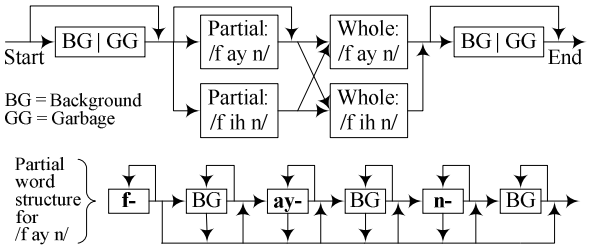

Figure 3: *Example disfluency grammar architecture for the word, "fine." The disfluency grammar is used to detect any partial-word disfluencies and to endpoint the final pronunciation*

We assigned a fluency score for each utterance heuristically from the speech recognizer output transcription. If no partial words (phones) were recognized, the utterance was deemed fluent and assigned a fluency score of 1. If two or more phones were recognized, the utterance was deemed disfluent and was given a fluency score of 0. Finally, a fluency score of 0.5 was assigned to "partially disfluent" utterances in which a single phone was recognized. This method was applied to all the utterances in the subjective human evaluation, and we calculated a mean fluency score for each child by averaging across all utterances from each child.

The automatic mean fluency scores were correlated with each evaluator's mean fluency scores with a mean correlation coefficient of 0.6983. The automatic mean fluency scores were correlated with the ground-truth mean fluency scores with a correlation coefficient of 0.8139. These agreement statistics are both significant ($p<0.01$). In addition to detecting disfluencies, the disfluency grammars also endpointed the final pronunciation of the target word. This portion of the utterance was then analyzed by the pronunciation verification system, which determined whether the pronunciation was acceptable or unacceptable.

4.2. Pronunciation Verification

Word-level pronunciation verification is often formulated as a hypothesis test between two competing models: the target pronunciation, and a filler model of expected mistakes [10,11]. Using standard speech acoustic models we can calculate likelihoods of the observed speech given each pronunciation model, and then set a threshold for target model acceptance on the ratio of these likelihoods. This method is inappropriate for our task for several reasons. One, it does not account for other cues to reading ability that teachers would use, such as the rate of speaking or their prior knowledge of the child's native language. Second, it assumes a closed set of well defined acceptable and unacceptable pronunciations. Many categories of expected pronunciation variants can be hypothesized, but given the variability in children's speech – and especially that of the bilingual children in our corpus – it is not always clear which of these categories should be deemed acceptable by teachers. Pronunciation production and reading comprehension, though well correlated, are not identical skills, and we would not want to penalize a Spanish-accented child's reading score just because his pronunciation is not standard.

The verification method used here is essentially a perceptual model for a teacher's unconscious "decision" to accept or reject a pronunciation, inspired by theories of competition and activation in lexical access [12]. In a Bayesian Network classifier as described in [8], we combine speech recognition- and alignment-based cues to a number of different pronunciation categories with child demographics and other prior knowledge, allowing for interaction among these cues in determining an overall qualitative reading score. Bayesian inference determines the probability of the perceptual class given all these cues, and from that we accept or reject each word based on

$$\underset{q}{\operatorname{argmax}}\, P(Q = q \mid X_1, X_2, \ldots, X_n) \tag{1}$$

where Q is the binary accept/reject variable and X_1, X_2, …, X_n are the cues for that word. With these binary item-level scores returned by the classifier, we calculated a mean accuracy score for each child by computing the fraction of words accepted for each child.

The automatic mean accuracy scores were correlated with each evaluator's mean accuracy scores with a mean correlation coefficient of 0.8295. The automatic mean accuracy scores were correlated with the ground-truth mean accuracy scores with a correlation coefficient of 0.8628. These agreement statistics are both significant ($p<0.01$).

5. Automatic Estimation of Children's Reading Ability

Having calculated automatic accuracy and fluency scores that correlate well with human evaluators, we repeated the analysis done in section 3.2 using the automatic scores as the independent variables. Results of the leave-one-out cross-validation linear regression can be found in Table 3. Figure 4 is the regression plot when using the automatic scores, with Figure 5 showing the resulting absolute error plot. Comparing the results from section 3.2 (Tables 2 and 3), we see that when using both the automatic accuracy and fluency scores, the performance approached that of the mean evaluator and outperformed five of the 14 evaluators.

Independent Variable	*Leave-one-out Cross-Validation*	
	Correlation	*rms error*
Fluency only	0.197	30.86 %
Accuracy only	0.710	22.02 %
Accuracy + Fluency	0.910	12.38 %

Table 3: *Predictive power of automatic accuracy and fluency scores when estimating ground-truth (GT) overall scores*

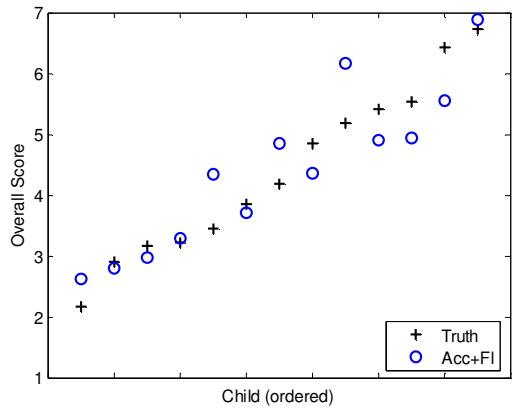

Figure 4: *Leave-one-out cross-validation linear regression results with ground-truth overall scores as dependent variable and automatic mean accuracy/fluency scores as independent variables*

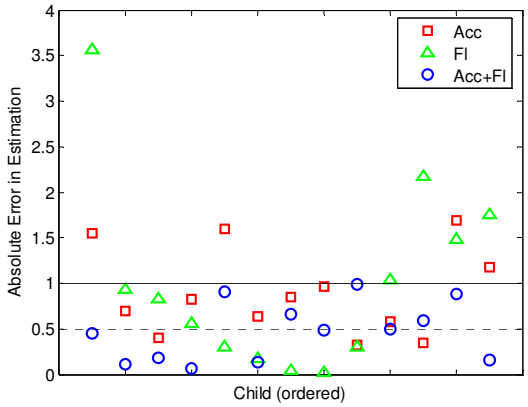

Figure 5: *Absolute error in predicting ground-truth overall scores with automatic mean accuracy and fluency scores*

6. Conclusion

We showed that automatically determined accuracy and fluency scores are capable of capturing the information needed to rate overall reading ability of young children for an isolated-word reading task. We detected fluency using a grammar-based approach and standard automatic speech recognition techniques and trained a Bayes Net that was capable of differentiating acceptable and unacceptable pronunciations. We demonstrated that having knowledge of both fluency and accuracy information significantly improves prediction of the overall reading ability of children. The *automatic* fluency and accuracy scores *combined* proved to be better features than the *manual* accuracy or fluency scores when used in isolation, an impressive result.

Our final automatic system was able to predict children's overall reading ability with a 0.910 correlation with the ground-truth scores (12.4% rms error), approaching the performance when using the mean accuracy and fluency manual scores from human evaluators (0.923 correlation and 11.2% rms error). By training our system on multiple people's perception (ground-truth overall scores), we were able to combat the problem of low pairwise evaluator agreement and reduce individual evaluator bias effects. The final system implicitly models the relative importance between accuracy and fluency information.

In the future, we plan to do a second subjective human evaluation with actual elementary education teachers and more children's data to ensure statistically significant results. In this study, we hope to further demonstrate our ability to train an automatic reading assessment system to mimic the grading of teachers in a supervised fashion. We also want to extend this research to sentence-level data, where other factors such as speaking rate may play larger roles in evaluators' perception of reading ability.

7. Acknowledgments

This work was supported in part by the National Science Foundation. Special thanks to the entire Tball Project team and all participants of the subjective human evaluation.

8. References

[1] Tball. http://diana.icsl.ucla.edu/Tball/assess_frame.html

[2] 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, S. Wang, "A system for technology based assessment of language and literacy in young children: the role of multiple information sources," Proc. MMSP, Greece, 2007.

[3] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. of Acoust. Soc. Am., 105:1455-1468, Mar. 1999.

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

[5] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," Proc. ASRU, St. Thomas, Virgin Islands, 2003.

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

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

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

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

[10] 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," Proc. ICSLP, Pittsburgh, PA, 2006.

[11] D. Willet, A. Worm, C. Neukirchen, and G. Rigoll, "Confidence measures for HMM-based speech recognition," Proc. ICSLP, Sydney, Australia, 1998.

[12] P. A. Luce and C. T. McLennan, "Spoken word recognition: the challenge of variation," *The Handbook of Speech Perception*. Ed. D. B. Pisoni and R. E. Remez. Oxford: Blackwell, 2005.