Automatic Detection and Classification of Disfluent Reading Miscues in Young Children's Speech for the Purpose of Assessment

Matthew Black[1], *Joseph Tepperman*[1], *Sungbok Lee*[1], *Patti Price*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, USC, Los Angeles, CA
[2]PPrice Speech and Language Technology, Menlo Park, CA
{matthepb,tepperma,sungbokl}@usc.edu, pjp@pprice.com, shri@sipi.usc.edu

Abstract

This paper explores the importance of disfluent reading miscues (sounding-out, hesitations, whispering, elongated onsets, question intonations) in automating the assessment of children's oral word reading tasks. Analysis showed that a significant portion (21%) of the speech obtained from grades K-2 children from predominantly Spanish-speaking families contained at least one disfluent reading miscue. We discovered human evaluators rated the fluency nearly as important as accuracy when judging the overall reading ability of a child. We devised a lexical method for automatically detecting the sounding-out, hesitation, and whispering disfluencies, which achieved a 14.9% missed detection and 8.9% false alarm rate. We were also able to discriminate 69.4% of the sound-outs from other disfluencies with a 28.5% false alarm rate, a promising and novel result.
Index Terms: children's speech, disfluency detection, automatic assessment

1. Introduction

Significant research has been done on automatically assessing children's read speech [1,2,3,4]. Common reading miscues, including mispronunciations, partial word utterances, substitutions, insertions, deletions, and sound-outs, hold key information regarding the reading ability of children. The automatic detection of these reading miscues is a challenging task that has been accomplished through constrained automatic speech recognition (ASR). For example, CMU's Project LISTEN predicted the occurrence of reading miscues, based on rote and extrapolative means, and was able to detect 37% of the mispronunciations, substitutions, and insertions with a 7% false alarm rate [5]. The Colorado Literacy Tutor automatically detected 16% of the partial word utterances with a 0.6% false alarm rate by splitting each pronunciation in the lexicon into syllable-like units [6]. Both of these projects concentrated on native English-speaking children already reading full sentences.

Our current speech assessment system is a crucial component of the Technology-based assessment of language and literacy (Tball) Project [7]. This project aims at automatically assessing the English literacy skills of predominantly Mexican-American children in grades K-2. The specific Tball Project tests are developmentally based and pedagogically informative: phonemic awareness, syllable blending, letter naming and sounding, isolated word reading, story reading, and listening comprehension. Of these, the production of high frequency words serves as an important building block for literacy skills, and hence is a significant element of the assessment suite. This paper focuses on the data from the isolated word reading task.

The children were recorded reading 55 words that progressively increased in difficulty. One word was displayed on a computer screen for a maximum of five 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. While it might seem trivial to design an automated system that operates at this single word level, there were a number of factors that complicated the design. The Tball Project speech was collected in realistic classroom noise conditions with standard headset microphones [8]. The children's young ages and multi-lingual backgrounds resulted in high speaker variability [9,10]. Many of the children were learning to both read *and* speak English, which caused a larger number of reading miscues than in previous studies that used different corpora [2,3,11,12]. Initially we felt *reading mistakes* (substitutions, deletions, and mispronunciations) were the only important reading miscues to detect automatically. An example of a reading mistake is pronouncing the word /r ih p/ ("rip") as /r ay p/ ("ripe"). We used recognition and alignment-based features in a decision tree classifier to detect these reading mistakes; the system agreed with human evaluators 91% of the time [4]. This paper complements our previous work by examining the importance of other reading miscues heard in our data and finding ways to automate their detection.

2. Transcription of reading miscues

We transcribed a random subset of 2800 utterances from 225 children in the Tball corpus [8]. 12% of the children were in Kindergarten; 47% were in first grade; 41% were in second grade. The majority of the data transcribed (60%) were from children who lived in Spanish-speaking households.

We noted two types of reading miscues in the transcribed data: reading mistakes (the focus of [4]) and disfluencies. Disfluencies included anything spoken by the child that was disruptive to the normal flow of the target word pronunciation. We labeled five types of disfluencies in the subset of data analyzed. Disfluencies that occurred 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. The other two types of disfluencies heard were noticeable during the final pronunciation of the target word. Some children lengthened the first phone or syllable of the word, which we call *stalling*. Lastly, some students said the target word with a *questioning* intonation. It should be noted that 2.4% of the utterances had more than one type of disfluency, meaning that the disfluency types are not mutually exclusive within a particular utterance.

Reading Miscues	*# of Files*	*% of Files*
Reading Mistakes	1042	**38.4**
Disfluencies (at least one)	597	**21.3**
Disfluency: Hesitations	241	**8.6**
Disfluency: Sound-outs	160	**5.7**
Disfluency: Stalling	120	**4.3**
Disfluency: Whispering	85	**3.0**
Disfluency: Questions	55	**2.0**

Table 1: *Number and percentage of utterances containing the observed reading miscues (in the 2800 transcribed files)*

Analyzing language background effects, we found children from Spanish-speaking families were more likely to sound-out during an utterance (a statistically significant difference in proportions with $p<0.001$). When comparing statistics between the three grade levels, we found Kindergartners stalled the most, first graders sounded-out the most, and second graders most often had disfluency-free utterances (all $p<0.01$). All other differences in proportions were not significant at the 95% confidence level.

Reading mistakes are arguably the most important reading miscue to detect for speech assessment tasks since they were the most prevalent error in our data. Disfluencies were positively correlated with reading mistakes; given that the speech was disfluent, the probability of a reading mistake increased from 37.2% to 57.8% ($p<0.001$). This indicated that disfluency detection could aid in the prediction of literacy assessment. We hypothesized that disfluencies affected human assessors, giving them insights towards the child's competence in letter-to-sound rules and overall confidence in reading. We first sought to determine the importance and perception of the disfluency types in the context of reading assessment by means of a subjective human evaluation.

3. Evaluation on the effect of disfluencies

We selected approximately 11 words each from 13 different children (148 utterances total), picking examples with varying levels of all disfluency types heard in our data. The audio samples were grouped by child to allow the evaluators to adapt their judgment to the children's speech, which made the assessment process more representative of realistic testing conditions.

The 16 evaluators rated both the accuracy and the fluency of each utterance. When assessing accuracy, the evaluators were instructed to concentrate solely on the final pronunciation of the target word (ignoring any disfluencies) and rate it as acceptable or unacceptable. When assessing fluency, the evaluators were instructed to rate, on a scale of 1 to 5, how fluently the child spoke throughout the utterance. After listening to all the utterances from one child, the evaluators then provided a score, on a scale of 1 to 7, for the overall reading ability of that child. Lastly, to gain insight on the importance of accuracy compared to fluency, the evaluators rated which was more important, on a scale of 1 to 5, when rating the overall reading ability of the child.

We analyzed the utterance-level fluency scores to determine how each disfluency type was perceived. To account for evaluators using the range (1 to 5) in the fluency rating scheme differently, the standardized fluency score z_{ij} of each evaluated utterance x_{ij} was calculated:

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \quad (1)$$

for each evaluator i and each utterance j, where $\bar{x}_i$ and s_i are the mean fluency rating and standard deviation, respectively, across all utterances rated by evaluator i.

The type of disfluency present in each utterance of the disfluency evaluation was already known from the transcriptions. Figure 1 shows the distribution of all standardized fluency scores for each disfluency type. The mean fluency score of all five disfluency types was lower than the mean score of disfluency-free utterances (all $p<0.001$). Thus, as expected, utterances without disfluencies were considered the most fluent. We also performed pairwise hypothesis tests on the difference in mean fluency scores between disfluency types. While the difference in means between the whisper, stalling, and question phenomena were not statistically significant at the 95% confidence level, sound-outs were considered the most disfluent, and hesitations were considered the second most disfluent (all $p<0.01$); these two disfluencies were also the most prevalent in the data (see Table 1).

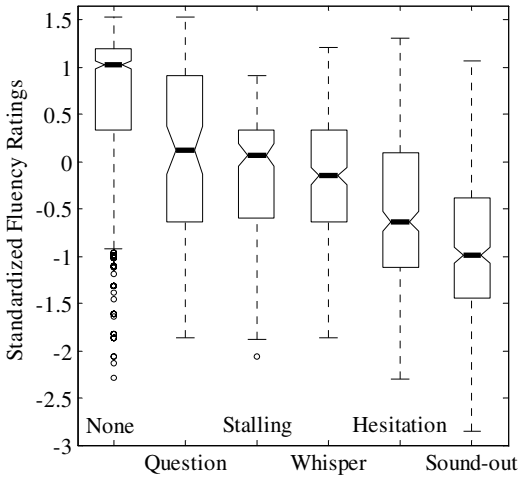

Figure 1: *Distribution of standardized fluency scores for all disfluency types. Disfluency-free files were the most fluent, while sound-outs and hesitations were the most disfluent.*

We also extracted the scores comparing the importance of fluency (which is correlated to the presence of disfluencies) to accuracy (which is correlated to the presence of reading mistakes) when rating the overall reading ability of a child. Table 2 shows the evaluator responses, where "1" meant fluency was far more important than accuracy and "5" meant accuracy was far more important than fluency. The mean score was 3.53, with a standard deviation of 0.77. Therefore, while reading mistakes were rated the most important reading miscues, disfluencies significantly affected the judgment of assessors when evaluating the reading performance of a child.

Rating	1	2	3	4	5
Meaning	F»A	F›A	F=A	F‹A	F«A
# Responses	0	18	74	90	16
% Responses	**0.0**	**9.1**	**37.4**	**45.4**	**8.1**

Table 2: *Evaluator responses on importance between fluency (F) vs. accuracy (A). The mean score was 3.53.*

4. Automatic disfluency detection

Motivated by the results of the disfluency evaluation, we next sought to incorporate disfluency detection into the automated assessment framework. Questions were the least prevalent disfluency in our data and had a higher mean fluency rating than the other disfluencies, so this paper does not concentrate on detecting questions. Since stalling is based on speaking rate, an inexact measure in isolated word productions, we did not concentrate on detecting stalling either. Rather, we focused on automatically detecting hesitations, sound-outs, and whispering disfluency phenomena.

4.1. Exploiting the lexical constraints

Since this is a reading assessment task, the target words are known ahead of time. Furthermore, nearly all of the disfluencies were partial word manifestations or singular phones of some pronunciation variant of the current target word. This facilitated our grammar-based lexical approach for disfluency detection. With the help of expert linguists and educators, we constructed a dictionary that included a phonemic breakdown of all foreseeable acceptable and unacceptable pronunciations of the target words. These predictable reading mistakes 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 of /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 [8,10].

4.2. Disfluency grammar structure

Using the dictionary for each target word, a simple way to analyze the speech would be to run forced-alignment with each of the pronunciation variants to discover which one was most probably spoken. The traditional grammar structure also allows the recognition of background noise (silence) to make the alignment more robust. Unfortunately, this would result in disfluent speech either going undetected or being detected as noise. Simply allowing repetitions of the target word to be recognized would still not facilitate the detection of partial word utterances or sound-outs with great success.

To detect the desirable disfluencies, we took a similar approach as in [6] by constructing a disfluency-specialized grammar structure, capable of recognizing partial words. We created the disfluency grammar solely from the dictionary and took advantage of the constrained reading task. The recognition of one of the pronunciations listed in the dictionary was still required as in the traditional grammar structure. However, prior to this whole word recognition, the speech recognizer was also optionally permitted to recognize individual phones in any pronunciation variant of the target word in the order of the pronunciation. We experimented with numerous partial word structures and found we got the best results when the first phone of each pronunciation variant was required to be recognized. If this first phone was recognized, the rest of the phones in the pronunciation could be detected or skipped and single phones could be repeated. Since most children paused between phones when sounding out words, background noise (silence) was allowed to be detected between phones. Figure 2 shows the disfluency grammar structure for the specific word, "rub."

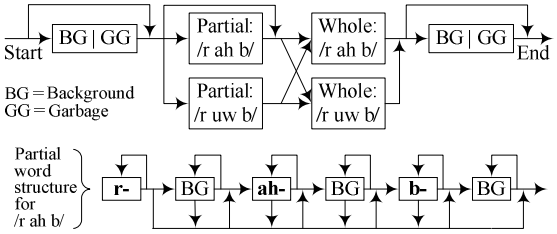

Figure 2: *Disfluency grammar architecture for the word, "rub," which has 2 pronunciation variants in the dictionary*

One caveat about the grammar structure was that if a child only hesitated or sounded-out and did not make an attempt to say the target word afterwards, the grammar structure force aligned the disfluency as the final pronunciation. However, this situation happened rarely in the subset of data we heard, and in these cases, the disfluent speech could be considered a reading mistake.

4.3. Acoustic models

With about 19 hours of classroom recordings taken from both native and nonnative speakers, we trained monophone models from the Tball corpus [8] using HTK [13]. Annotated on the word level, we used canonical pronunciation expansions and the Baum-Welch algorithm to estimate three-state HMM parameters with 16 Gaussian mixtures per state. Generic models for the background classroom noise were trained by cutting background segments out of the recordings and estimating their parameters separately. For these background models, 256 mixtures per state proved necessary for adequate target word endpointing performance. Using all the speech data, we also trained a generalized word-level filler (garbage) model with three states and 16 mixtures per state. We did not use any other special acoustic models, such as for modeling whispered speech.

Using HTK, we extracted the first 13 MFCCs and their delta and delta-delta coefficients to serve as the 39-dimensional input feature vector for ASR. Section 5 describes how we used the ASR output transcription to classify an utterance as disfluent.

5. Classification algorithm

We created a classification algorithm to determine whether an utterance was fluent or disfluent entirely based on the recognition of partial words (individual phones). If an utterance was classified as disfluent, we further used the ASR transcription to separate sound-outs from other detected disfluencies. Figure 3 shows the disfluency classification algorithm. First, we counted the total number of partial words in the ASR transcription. If there were none, we classified the utterance as fluent. If one partial word was detected, further analysis was done to rule out false detections as follows. We force-aligned the garbage (GG) model to the segment of speech that contained the partial word to attain the GG model log-likelihood score. If the GG log-likelihood score was higher than the partial word log-likelihood score (which we attained from the ASR transcription), we classified the speech as fluent; otherwise, we classified it as disfluent. If there were two or more partial words in the ASR transcription, we classified the utterance as disfluent, and if background/silence (BG) was recognized between any two partial words, we further classified the disfluency as a sound-out.

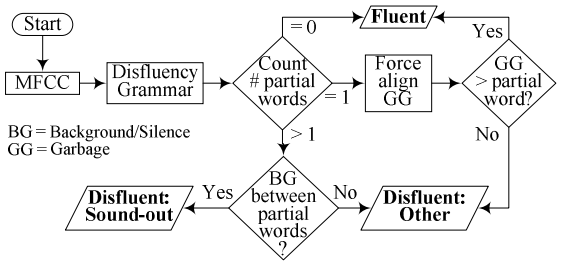

Figure 3: *Disfluency classification algorithm*

6. Results and discussion

Table 3 shows the missed detection statistics for the three disfluency types. Both sound-outs and hesitations were correctly detected as disfluent more than 88% of the time. The high whispering missed detection rate (35.3%) was likely caused by the inherent production differences, and suggests we need a separate whispering detection method. Overall, we detected 85.1% of the desired disfluencies, with an 8.9% false alarm rate. In comparison, University of Colorado researchers detected only 16% of the partial words with a low false alarm rate of 0.6% [6]. Thus, we detected a higher percentage of the partial words but suffered from a higher false alarm rate. This comparison is not entirely fair, though, since our research was on isolated word assessment, and we did not bias the ASR with prior probabilities of disfluencies. Sentence-level data was used in [6], and n-gram language models were trained to estimate partial word probabilities.

Disfluency Type	*Missed Detections*	
Sound-outs	16 / 160	**10.0** %
Hesitations	28 / 241	**11.6** %
Whispering	30 / 85	**35.3** %

Table 3: *Missed detection statistics for each disfluency type*

Of the 90% of utterances containing sound-outs we detected as being disfluent, we correctly classified 69.4% as being a sound-out. Unfortunately, this aspect of the classification algorithm suffered from a high false alarm rate, misclassifying 28.5% of the other utterances deemed disfluent as sound-outs when they were not transcribed as such. These results are promising, and to our knowledge, no previous work has attempted to classify speech as being sounded-out.

In analyzing the errors made by the proposed classification algorithm, we found that the difference in proportions of classifier errors for native and nonnative speakers was not statistically significant at the 95% confidence level. This suggests that our classifier is not biased against either category of student. However, the proportion of errors for Kindergartners was statistically higher ($p<0.05$) than those in the other grades, suggesting that age-dependent modeling might be necessary.

7. Conclusion

By administering a subjective human evaluation on a subset of isolated word utterances containing various reading miscues, we discovered disfluencies were nearly as important as reading mistakes when assessing the reading ability of a child. Furthermore, we found evaluators rated hesitations and sound-outs the most disfluent. We automatically detected hesitations, sound-outs, and whispering phenomenon by adding a target word constrained phone recognizer to the traditional ASR grammar. Using just the ASR output transcription, we achieved a low 14.9% missed detection rate and an 8.9% false alarm rate. Since sound-outs were considered the most disfluent, we further used the ASR transcription to successfully classify 69.4% of the detected disfluencies as sound-outs, with a 28.5% false alarm rate.

In the future, we will incorporate other features (prosodic, articulatory, child demographics) using a decision tree or Bayesian Network structure to improve the detection and classification of disfluencies [14]. We also plan to fuse this disfluency work with our automatic target word verification system, combining both the accuracy of the target word pronunciation and the fluency of the speech for an overall score for each child on the word-reading task.

8. Acknowledgements

This work was supported in part by the National Science Foundation. Special thanks also to Shizhen Wang for help with the design of the disfluency evaluation.

9. References

[1] J. Mostow, S. Roth, A.G. Hauptmann, and M. Kane, "A prototype reading coach that listens", Proc. *AAAI*, Seattle, WA, 1994.

[2] S.M. Williams, D. Nix, and P. Fairweather, "Using speech recognition technology to enhance literacy instruction for emerging readers," Proc. *ICLS*, Mahwah, NJ, 2000.

[3] 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.

[4] 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.

[5] S. Banerjee, J.E. Beck, and J. Mostow, "Evaluating the effect of predicting oral reading miscues," Proc. *Eurospeech*, Geneva, Switzerland, 2003.

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

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

[8] 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.

[9] 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.

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

[11] M. Eskenazi, J. Mostow, and D. Graff, "The CMU kids speech corpus," published Linguistic Data Consortium, 1997.

[12] K. Shobaki, J.P. Hosom, R. Cole, "The OGI kids' speech corpus and recognizers," Proc. *ICSLP*, Beijing, China, 2000.

[13] Cambridge University, HTK 3.2, htk.eng.cam.ac.uk.

[14] 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," *Eurospeech*, Antwerp, Belgium, 2007.