# Atypical Prosodic Structure as an Indicator of Reading Level and Text Difficulty

**Julie Medero and Mari Ostendorf**
Electrical Engineering Department
University of Washington
Seattle, WA 98195 USA
{jmedero,ostendor}@uw.edu

## Abstract

Automatic assessment of reading ability builds on applying speech recognition tools to oral reading, measuring words correct per minute. This work looks at more fine-grained analysis that accounts for effects of prosodic context using a large corpus of read speech from a literacy study. Experiments show that lower-level readers tend to produce relatively more lengthening on words that are not likely to be final in a prosodic phrase, i.e. in less appropriate locations. The results have implications for automatic assessment of text difficulty in that locations of atypical prosodic lengthening are indicative of difficult lexical items and syntactic constructions.

## 1 Introduction

Fluent reading is known to be a good indicator of reading comprehension, especially for early readers (Rasinski, 2006), so oral reading is often used to evaluate a student's reading level. One method that can be automated with speech recognition technology is the number of words that a student can read correctly of a normed passage, or Words Correct Per Minute (WCPM) (Downey et al., 2011). Since WCPM depends on speaking rate as well as literacy, we are interested in identifying new measures that can be automatically computed for use in combination with WCPM to provide a better assessment of reading level. In particular, we investigate fine-grained measures that, if useful in identifying points of difficulty for readers, can lead to new approaches for assessing text difficulty.

The WCPM is reduced when a person repeats or incorrectly reads a word, but also when they introduce pauses and articulate words more slowly. Pauses and lengthened articulation can be an indicator of uncertainty for a low-level reader, but these phenomena are also used by skilled readers to mark prosodic phrase structure, facilitating comprehension in listeners. Since prosodic phrase boundaries tend to occur in locations that coincide with certain syntactic constituent boundaries, it is possible to automatically predict prosodic phrase boundary locations from part-of-speech labels and syntactic structure with fairly high reliability for read news stories (Ananthakrishnan and Narayanan, 2008). Thus, we hypothesize that we can more effectively leverage word-level articulation and pause information by focusing on words that are less likely to be associated with prosodic phrase boundaries. By comparing average statistics of articulation rate and pausing for words at boundary vs. non-boundary locations, we hope to obtain a measure that could augment reading rate for evaluating reading ability. We also hypothesize that the specific locations of hesitation phenomena (word lengthening and pausing) observed for multiple readers will be indicative of particular points of difficulty in a text, either because a word is difficult or because a syntactic construction is difficult. Detecting these regions and analyzing the associated lexical and syntactic correlates is potentially useful for automatically characterizing text difficulty.

Our study of hesitation phenomena involves empirical analysis of the oral reading data from the Fluency Addition to the National Assessment of Adult



*Proceedings of NAACL-HLT 2013*, pages 715–720,
Atlanta, Georgia, 9–14 June 2013. 

Literacy (FAN), which collected oral readings from roughly 12,000 adults, reading short (150-200 word) fourth- and eighth grade passages (Baer et al., 2009). The participants in that study were chosen to reflect the demographics of adults in the United States; thus, speakers of varying reading levels and non-native speakers were included. For our study, we had access to time alignments of automatic transcriptions, but not the original audio files.

## 2 Related Work

For low-level readers, reading rate and fluency are good indicators of reading comprehension (Miller and Schwanenflugel, 2006; Spear-Swerling, 2006). Zhang and colleagues found that features of children's oral readings, along with their interactions with an automated tutor, could predict a single student's comprehension question performance over the course of a document (2007). Using oral readings is appealing because it avoids the difficulty of separating question difficulty from passage difficulty (Ozuru et al., 2008) and of questions that can be answered through world knowledge (Keenan and Betjemann, 2006).

WCPM is generally used as a tool for assessing reading level by averaging across one or more passages. It is more noisy when comparing the readability of different texts, especially when the reading level is measured at a fine-grained (e.g. word) level. If longer words take longer to read orally, it may be merely a consequence of having more phonemes, and not of additional reading difficulty. Further, for communication reasons, pauses and slow average articulation rates tend to coincide with major phrase boundaries. In our work, we would like to account for prosodic context in using articulation rate to identify difficult words and constructions.

Much of the previous work on using automatic speech recognition (ASR) output for reading level or readability analysis has focused on assessing the reading level of children (Downey et al., 2011; Duchateau et al., 2007). Similar success has been seen in predicting fluency scores in oral reading tests for L2 learners of English (Balogh et al., 2012; Bernstein et al., 2011). Project LISTEN has a reading tutor for children that gives real-time feedback, and has used orthographic and phonemic features of individual words to predict the likelihood of real word subsitutions (Mostow et al., 2002).

## 3 FAN Literacy Scores

To examine the utility of word-level pause and articulation rate features for predicting reading level when controlled for prosodic context, we use the Basic Reading Skills (BRS) score available for each reader in the FAN data. The BRS score measures an individual's average reading rate in WCPM. Each participant read three word lists, three pseudo-word lists, one easy text passage, and one harder text passage, and the BRS is the average WCPM over the eight different readings. Specifically, the WCPM for each case is computed automatically using Ordinate's VersaReader system to transcribe the speech given the target text (Balogh et al., 2005). The system output is then automatically aligned to the target texts using the track-the-reader method of Rasmussen et al. (2011), which defines weights for regressions and skipped words and then identifies a least-cost alignment between the ASR output and a text. Automatic calculation of WCPM has high correlation (.96-1.0) with human judgment of WCPM (Balogh et al., 2012), so it has the advantage of being easy to automate.

Word Error Rate (WER) for the the ASR component in Ordinate's prototype reading tracker (Balogh et al., 2012) may be estimated to be between 6% and 10%. In a sample of 960 passage readings, where various sets of two passages were read by each of 480 adults (160 native Spanish speakers, 160 native English-speaking African Americans, and 160 other native English speakers), the Ordinate ASR system exhibited a 6.9% WER on the 595 passages that contained no spoken material that was unintelligible to human transcribers. On the complete set of 960 passages, the system exhibited a 9.9% WER, with each unintelligible length of speech contributing one or more errors to the word error count.

The greatest problem with speech recognition errors is for very low-level readers (Balogh et al., 2012). In order to have more reliable time alignments and BRS scores, approximately 15% of the FAN participants were excluded from the current analysis. This 15% were those participants whose BRS score was labeled "Below Basic" in the NAAL

reading scale. Additional participants were eliminated because of missing or incomplete (less than a few seconds) recordings. With these exclusions, the number of speakers in our study was 7587.

## 4 Prosodic Boundary Prediction

We trained a regression tree[1] on hand-annotated data from the Boston University Radio News Corpus (Ostendorf et al., 1995) to predict the locations where we expect to see prosodic boundaries. Each word in the Radio News Corpus is labeled with a prosodic boundary score from 0 (clitic, no boundary) to 6 (sentence boundary). For each word, we use features based on parse depth and structure and POS bigrams to predict the prosodic boundary value.

For evaluation, the break labels are grouped into: 0-2 (no intonational boundary marker), 3 (intermediate phrase), and 4-6 (intonational phrase boundary). Words with 0-2 breaks are considered non-boundary words; 4-6 are boundary words. We expect that, for fluent readers, lengthening and possibly pausing will be observed after boundary words but not after non-boundary words. Since the intermediate boundaries are the most difficult to classify, and may be candidates for both boundaries and non-boundaries for fluent readers, we omit them in our analyses. Our model achieves 87% accuracy in predicting $\pm$ intonational phrase boundaries and 83% accuracy in predicting $\pm$ no intonational boundary, treating intermediate phrase boundaries as negative instances in both cases.

Note that our 3-way prosodic boundary prediction is aimed at identifying locations where fluent readers are likely to place boundaries (or not), i.e., reliable locations for feature extraction, vs. acceptable locations for text-to-speech synthesis. Because of this goal and because work on prosodic boundary prediction labels varies in its treatment of intermediate phrase boundaries, our results are not directly comparable to prior studies. However, performance is in the range reported in recent studies predicting prosodic breaks from text features only. Treating intermediate phrase boundaries as positive examples, Ananthakrishnan and Narayanan (2008) achieve 88% accuracy. Treating them as negative examples, Margolis and Ostendorf (2010) achieve similar results. Both report results on a single held-out test set, while our results are based on 10-fold cross validation.

[1] Our approach differs slightly from previous work in the use of a regression (vs. classification) model; this gave a small performance gain.

## 5 Experiments with Prosodic Context

### 5.1 Word-level Rate Features

We looked at two acoustic cues related to hesitation or uncertainty: pause duration and word lengthening. While pause duration is straightforward to extract (and not typically normalized), various methods have been used for word lengthening. We explore two measures of word lengthening: i) the longest normalized vowel, and ii) the average normalized length of word-final phones (the last vowel and all following consonants). Word-final lengthening is known to be a correlate of fluent prosodic phrase boundaries (Wightman et al., 1992), and we hypothesized that the longest normalized vowel might be useful for hesitations though it can also indicate prosodic prominence.

For word-level measures of lengthening, it is standard to normalize to account for inherent phoneme durations. We use a z-score: measured duration minus phone mean divided by phone standard deviation. In addition, Wightman et al. (1992) found it useful to account for speaking rate in normalizing phone duration. We adopt the same model, which assumes that phone durations can be characterized by a Gamma distribution and that speaker variability is characterized by a linear scaling of the phone-dependent mean parameters, where the scaling term is shared by all phones. The linear scale factor $\alpha$ for a speaker is estimated as:

$$\alpha = \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{\mu_{p(i)}} \quad (1)$$

where $d_i$ is the duration of the $i$-th phone which has label $p(i)$ and where $\mu_p$ is the speaker-independent mean of phone $p$. Here, we use a speaker-independent phone mean computed from the TIMIT Corpus,[2] which has hand-marked phonetic labels and times. We make use of the speaking rate model

[2] Available from the Linguistic Data Consortium.

to adjust the speaker-independent TIMIT phone durations to the speakers in the FAN corpus by calculating the linear scale factor $\alpha$ for each speaker. Thus, the phone mean and standard deviation used in the z-score normalization is $\alpha\mu_{p_i}$ and $\alpha\sigma_{p_i}$, respectively.

From the many readings of the eight passages, we identified roughly 777K spoken word instances at predicted phrase boundaries and 2.0M spoken words at predicted non-boundaries. For each uttered word, we calculated three features: the length of the following pause, the length of the longest normalized vowel, and the averaged normed length of all phones from the last vowel to the end of the word, as described above. The word-level features can be averaged across instances from a speaker for assessing reading level or across instances of a particular word in a text uttered by many speakers to assess local text difficulty.

The phone and pause durations are based on recognizer output, so they will be somewhat noisy. The fact that the recognizer is biased towards the intended word sequence and the omission of the lowest-level readers from this study together contribute to reducing the error rate ($< 10\%$) and increasing the reliability of the features. In addition, noise is reduced by averaging over multiple words or multiple speakers.

### 5.2 Reading Level Analysis

To assess the potential for prosodic context to improve the utility of word-level features for assessing reading difficulty, we looked at duration lengthening and pauses at boundary and non-boundary locations, where the boundary labels are predicted using the text-based algorithm and 3-class grouping described in section 4.

First, for each speaker, we averaged each feature across all boundary words read by that person and across all non-boundary words read by that person. We hypothesized that skilled readers would have shorter averages for all three features at non-boundary words compared to at boundary words, while the differences for lower-level readers would be smaller because of lengthening due to uncertainty at non-boundary words. The difference between the boundary and non-boudnary word averages for normalized duration of end-of-word phones is plotted in Figure 1 as a function of reading level. As expected, the difference increases with reading skill, as measured by BRS. A similar trend is observed for the longest normalized vowel in the word.

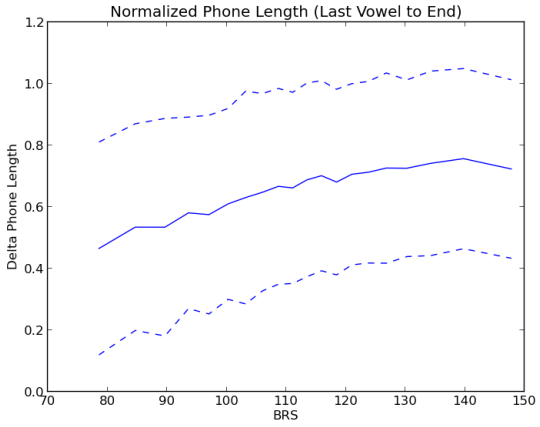


Figure 1: Mean end-of-word normalized phone duration (+/- standard deviation) as a function of BRS score

We also looked at pause duration, finding that the average pause duration decreases as reading skill increases for both boundary and non-boundary words. Since pauses are not always present at intonational phrase boundaries, but are more likely at sentence boundaries, we investigated dividing the cases by punctuation rather than prosodic context. Table 1 shows that for both the top 20% of readers and the bottom 20% of readers, sentence boundaries had much longer pauses on average, followed by comma boundaries, and unpunctuated word boundaries. The drop in both pause frequency and average pause duration is much greater for the more skilled readers.

Looking at all speakers, the unpunctuated words had an average pause duration that scaled with the speaking rate estimate for that passage, with high correlation (0.94). The correlation was much lower for sentence boundaries (0.44). Thus, we conclude that the length of pauses at non-boundary locations is related to the speaker's reading ability.

### 5.3 Identifying Difficult Texts

Instead of averaging over multiple words in a passage, we can average over multiple readings of a particular word. We identified difficult regions in texts by sorting all tokens by the average normalized length of their end-of-word phones for the lowest

| | Top 20% | | Bottom 20% | |
|---|---|---|---|---|
| | Pause Rate | Avg. Pause Duration | Pause Rate | Avg. Pause Duration |
| Sentence-final | 81.0% | 177 ms | 84.7% | 283 ms |
| Comma | 26.1% | 94 ms | 47.0% | 168 ms |
| No punctuation | 4.6% | 77 ms | 16.6% | 139 ms |

Table 1: Frequency of occurrence and average duration of pauses at sentence boundaries, comma boundaries, and unpunctuated word boundaries for the top and bottom 20% of all readers, as sorted by BRS score

20% of readers. The examples suggest that lengthening may coincide with reading difficulty caused by syntactic ambiguity. Two sentences, with the lengthened word in bold, illustrate representative ambiguities:

- She was there for **me** the whole time my grandfather was in the hospital.
- Since dogs are **gentler** when raised by a family the dogs are given to children when the dogs are about fourteen months old.

In the first example, "me" could be the end of the sentence, while in the second example, readers may expect "gentler" to be the end of the subordinate clause started by "since". The lengthening on these words is much smaller for the top 20% of readers, suggesting that the extra lengthening is associated with points of difficulty for the less skilled readers.

Similarly, we identified sentences with non-boundary locations where readers commonly paused, with the word after the pause in bold:

- We have always been able to share our **escapades** and humor with our friends.
- Check with your doctor first if you are a man over forty or a woman over fifty **and** you plan to do vigorous activity instead of moderate activity.

We observe a wider variety of potential difficulties here. Some are associated with difficult words, as in the first example, while others involve syntactic ambiguities similar to the ones seen in the lengthening cases.

## 6 Summary

We have shown that duration lengthening and pause cues align with expected prosodic structure (predicted from syntactic features) more for skilled readers than for low-level readers, which we hope may lead to a richer assessment of individual reading difficulties. In addition, we have proposed a method of characterizing text difficulty at a fine grain based on these features using multiple oral readings. In order to better understand the information provided by the different features, we are conducting eye tracking experiments on these passages, and future work will include an analysis of readers' gaze during reading of these constructions that have been categorized in terms of their likely prosodic context.

In this work, where the original recordings were not available, the study was restricted to duration features. However, other work has suggested that other prosodic cues, particularly pitch and energy features, are useful for detecting speaker uncertainty (Litman et al., 2009; Litman et al., 2012; Pon-Barry and Shieber, 2011). Incorporating these cues may increase the reliability of detecting points of reading difficulty and/or offer complementary information for characterizing text difficulties.

### Acknowledgments

We are grateful to the anonymous reviewers for their feedback, and to our colleagues at Pearson Knowledge Technologies for their insights and data processing assistance. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0718124 and by the National Science Foundation under Grant No. IIS-0916951. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## References

S. Ananthakrishnan and S.S. Narayanan. 2008. Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. *IEEE Trans. Audio, Speech, and Language Processing*, 16(1):216–228.

J. Baer, M. Kutner, J. Sabatini, and S. White. 2009. Basic Reading Skills and the Literacy of Americas Least Literate Adults: Results from the 2003 National Assessment of Adult Literacy (NAAL) Supplemental Studies. Technical report, NCES.

J. Balogh, J. Bernstein, J. Cheng, and B. Townshend. 2005. Final Report Ordinates Scoring of FAN NAAL Phase III: Accuracy Analysis. Technical report, Ordinate.

J. Balogh, J. Bernstein, J. Cheng, A. Van Moere, B. Townshend, and M. Suzuki. 2012. Validation of Automated Scoring of Oral Reading. *Educational and Psychological Measurement*, 72:435–452.

J. Bernstein, J. Cheng, and M. Suzuki. 2011. Fluency Changes with General Progress in L2 Proficiency. In *Proc. Interspeech*, number August, pages 877–880.

R. Downey, D. Rubin, J. Cheng, and J. Bernstein. 2011. Performance of Automated Scoring for Children's Oral Reading. *Proc. Workshop on Innovative Use of NLP for Building Educational Applications*, (June):46–55, June.

J. Duchateau, L. Cleuren, H. Van, and P. Ghesqui. 2007. Automatic Assessment of Childrens Reading Level. *Proc. Interspeech*, pages 1210–1213.

J.M. Keenan and R. Betjemann. 2006. Comprehending the Gray Oral Reading Test Without Reading It: Why Comprehension Tests Should Not Include Passage-Independent Items. *Scientific Studies of Reading*, 10(4):363–380.

D. Litman, M. Rotaru, and G. Nicholas. 2009. Classifying turn-level uncertainty using word-level prosody. In *Proc. Interspeech*.

D. Litman, H. Friedberg, and K. Forbes-Riley. 2012. Prosodic cues to disengagement and uncertainty in physics tutorial dialogues. In *Proc. Interspeech*.

A. Margolis, M. Ostendorf, and K. Livescu. 2010. Cross-genre training for automatic prosody classification. In *Proc. Speech Prosody Conference*.

J. Miller and P.J. Schwanenflugel. 2006. Prosody of Syntactically Complex Sentences in the Oral Reading of Young Children. *Journal of Educational Psychology*, 98(4):839–843.

J. Mostow, J. Beck, S. Winter, S. Wang, and B. Tobin. 2002. Predicting Oral Reading Miscues. In *Proc. ICSLP*.

M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. 1995. The Boston University Radio News Corpus. Technical report, Boston University, March.

Y. Ozuru, M. Rowe, T. O'Reilly, and D.S. McNamara. 2008. Where's the difficulty in standardized reading tests: the passage or the question? *Behavior Research Methods*, 40(4):1001–1015.

H. Pon-Barry and S.M. Shieber. 2011. Recognizing uncertainty in speech. *CoRR*, abs/1103.1898.

T. Rasinski. 2006. Reading fluency instruction: Moving beyond accuracy, automaticity, and prosody. *The Reading Teacher*, 59(7):704–706, April.

M.H. Rasmussen, J. Mostow, Z. Tan, B. Lindberg, and Y. Li. 2011. Evaluating Tracking Accuracy of an Automatic Reading Tutor. In *Proc. Speech and Language Technology in Education Workshop*.

L. Spear-Swerling. 2006. Childrens Reading Comprehension and Oral Reading Fluency in Easy Text. *Reading and Writing*, 19(2):199–220.

C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P.J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3):1707—-1717.

X.N. Zhang, J. Mostow, and J.E. Beck. 2007. Can a Computer Listen for Fluctuations in Reading Comprehension? *Artificial Intelligence in Education*, 158:495–502.