Towards understanding text factors in oral reading

Anastassia Loukina, Van Rynald T. Liceralde, Beata Beigman Klebanov
Educational Testing Service
Princeton, NJ, USA
{aloukina,vliceralde,bbeigmanklebanov}@ets.org

Abstract

Using a case study, we show that variation in oral reading rate across passages for professional narrators is consistent across readers and much of it can be explained using features of the texts being read. While text complexity is a poor predictor of the reading rate, a substantial share of variability can be explained by timing and story-based factors with performance reaching r=0.75 for unseen passages and narrator.

1 Introduction

Listening to and performing oral reading are activities that permeate daily life, from parents reading aloud to young children, through reading instruction in elementary school, to audiobook narrations increasingly chosen by adults as the form of book-reading that fits in a busy schedule. Oral reading is also used in assessment of language skills for children and language learners, and in professions such as teaching and news broadcasting.

Reading rate is a common metric used to control or evaluate oral reading. It is usually computed as a number of words read per minute, and is used in many applications. For example, research in second language acquisition has considered both optimal reading rates for listening materials aimed at English language learners and reading rates that ensure the highest comprehensibility of accented speech (Munro and Derwing, 1998). Speech rate is a standard feature in systems for automated scoring of second language proficiency (Higgins et al., 2011) including read aloud tasks (Zechner et al., 2012; Evanini et al., 2015). Reading rate is also one of the main measures used to assess the fluency of oral reading (Hasbrouck and Tindal, 2006).

The assumption underlying these uses is that reading rate is a property of the reader (or controlled by the reader). However, variation in reading rate across different passages for the same readers has also been reported (Foulke, 1968; Tauroza and Allison, 1990; Ardoin et al., 2005; Compton et al., 2004; Beigman Klebanov et al., 2017).

Improving the understanding of the properties of oral reading, such as reading rate, is thus an important theoretical goal. We also have a specific practical reason to study text-based variation in reading rate. We are developing an intervention for improving literacy that would encourage sustained reading by having the student read aloud multiple passages from an engaging novel-length book, taking turns with others. While it is technically easy to compute reading rate by timing the readers, if a reader's rate across different texts is not stable given his current reading skill, it is not clear that tracking the rate over time would yield a valid measurement of improvement in skill. However, if such variation is systematically dependent on the text being read, rather than a random or idiosyncratic fluctuation, we might be able to adjust the measurement to account for text effect.

In order to inform both the theoretical and the applied goals, we address the following research questions in this paper:

1. Is reading rate constant for a given reader across various texts?

2. If not, do different readers show similar patterns of variation across texts, or is variation idiosyncratic?

3. If variation exists and is systematic across readers, can we identify the properties of texts that impact reading rate?

In this paper, we study reading rates in two professional narrations of the same book-length text. By using professional narrations we are able to

Proceedings of NAACL-HLT 2018, pages 2143–2154
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

eliminate other factors that might cause variation in reading rate, such as reader fatigue or disfluencies. While these would play a role in a practical application, we seek first to answer the research questions in a setup that allows focusing on the relationship between reading rate and the passage being read, controlling for other factors.

2 Related work

2.1 Passage effects in reading

Passage effects in reading have been addressed most directly in the context of assessment of reading. Since the intention is to measure the student's reading ability, any difference in performance that is not due to reading ability confounds the measurement. In particular, since comprehension complexity of a passage is known to impact reading comprehension, it seems reasonable to assume that it would also impact other aspects of reading skill, including oral reading fluency. In fact, this assumption underlies text selection for tests of oral reading fluency such as DIBELS (Good and Kaminski, 2002) that rely on readability to select comparable passages (Francis et al., 2008).

Yet research also suggests that controlling for readability does not entirely solve the problem of text-based variation in reading fluency. Ardoin et al. (2005) examined readability formulas for their ability to predict fluency and generally found only low-to-moderate correlations ($r<0.5$). Researchers also observed that fluency measurements for the same students varied across texts even for passages of comparable readability (Ardoin et al., 2005; Compton et al., 2004; Petscher and Kim, 2011; Francis et al., 2008).

Moreover, Francis et al. (2008) found that while actual fluency scores vary across different readability-controlled passages, the relative ranking of students is only minimally different when estimated using different passages, suggesting that variation in fluency has some consistency across readers; results to a similar effect were reported by Beigman Klebanov et al. (2017).

Oral reading fluency is commonly measured using words correct per minute – a combination of reading accuracy and reading rate. It is thus not clear whether the observations above pertain more to the accuracy aspect of oral reading (not considered in the current paper) or to reading rate, although Beigman Klebanov et al. (2017) noted that consistent variation across students was observed both for reading rate and for reading fluency.

To summarize, it appears that while readability could explain some of the variation in oral reading performance, there are also indicators that it is not sufficient on its own to effectively control for variation in oral reading performance caused by the properties of the passage being read.

2.2 Factors that affect duration of segments and pauses

2.2.1 Sentence-level timing

Since oral reading involves saying the text aloud, the durations of individual segments, words and phrases as well as location and duration of silent pauses are subject to constraints that have been extensively studied in literature on phonetic timing; see White (2014); Hirschberg (2002) for a review.

Thus it has been long known that different segments have different intrinsic durations which account for a lot of variation in segmental durations (Peterson and Lehiste, 1960; Klatt, 1976; van Santen, 1992): for example, high vowels tend to be shorter than low vowels. At the syllable level, in many languages vowels tend to be shorter when followed by a voiceless consonant than when followed by a voiced consonant (House and Fairbanks, 1953; Crystal and House, 1988) while consonants within a consonant cluster tend to be shorter than single consonants (Klatt, 1976).

Further constraints are at play at word, phrase and sentence level. White (2014) summarizes these as "domain-head" and "domain-edge" lengthening effects. "Domain-head" lengthening refers to lengthening of salient elements such as syllables bearing lexical stress, words in prominent positions (Peterson and Lehiste, 1960; Crystal and House, 1988; van Santen, 1992). "Domain-edge" effects include lengthening of segments in word-initial position or sentence-final lengthening (Turk and Shattuck-Hufnagel, 2000, 2007).

Finally, these domain-head and domain-edge lengthening effects do not apply uniformly: some segments and some positions are more resistant to lengthening than others (Peterson and Lehiste, 1960; Klatt, 1976; van Santen, 1992; White, 2014). The magnitude of lengthening also depends on the number of elements within each domain: in monosyllabic words, the stressed syllable receives all of the prosodic lengthening, but in disyllabic and trisyllabic words, some of the lengthening spreads to the unstressed syllables and lengthening of the stressed syllable is attenuated (Turk and Shattuck-Hufnagel, 2000; White and Turk, 2010).

In addition to segmental lengthening, phrase and sentence boundaries are often associated with some amount of pause. The location and duration of sentence-internal pauses depends both on syntactic structure and the number of syllables in each adjacent unit: sentence-internal pauses associated with punctuation or major syntactic boundaries tend to be longer than other sentence-internal pauses, with sentence-final pauses being the longest (Pfitzinger and Reichel, 2006; Burrows et al., 2005; Bailly and Gouvernayre, 2012).

2.2.2 Content and duration

Beyond domain-head and domain-edge effects, duration of segments and pauses is also affected by other aspects of text content. Frequent words tend to have shorter duration than phonologically similar less frequent words. Words that are more predictable in a given context tend to be shorter than words with higher information load and repeated words are pronounced shorter than the first mention (see Zhao and Jurafsky (2009); Bell et al. (2009) for reviews). Note that these effects persist after one controls for "domain-head" effects described in the previous section (Bell et al., 2009).

Further factors come into play in the context of story-telling where the speaker is either reading or narrating a well-rehearsed story. Montaño and Alías (2017) review approaches used to characterize story-telling speech.

Several studies observed that the duration of pauses between sentences and paragraphs in a longer story is not uniform. In their analysis of pausing in book reading, Bailly and Gouvernayre (2012) reported that pauses between paragraphs were longer than pauses between sentences. They also found that the thematic relationships between sentences affect breathing patterns although these were not immediately related to pause duration.

Reading rate has also been shown to depend on the emotional state of the speaker, whether genuine or performed as part of a dramatic reading: for example, actors tend to speak slower when expressing anger, fear or sorrow (Williams and Stevens (1972), see Scherer (2003) for a comprehensive review). Doukhan et al. (2011) analyzed pause distribution in a corpus of tales and reported "speakers' expressive reinterpretation of sentence syntactic structure" which they attributed to expressiveness of the reader.

There is also evidence that prosody may be affected by the narrative structure. Theune et al. (2006) observed in an informal analysis that Dutch actors narrating fairy-tales reduced their speech tempo when approaching the story climax. They also noticed an increase in duration in some words that indicated extreme value of a property. Doukhan et al. (2011) analyzed prosody in a corpus of French tales using Propp's morphology of Folktale (Propp, 1968). They found that narrative structures had a significant effect on various prosodic properties. For duration, epilogues were associated with lower articulation rate (syllables/min without pauses) while refrains had the lowest pausing time percentage. Finally, several studies found that impersonation by narrator of different characters leads to clear differences in pitch, intensity and spectral quality (Doukhan et al., 2011; Wang et al., 2006).

In short, previous research suggests that multiple factors may affect phone and pause duration in a reading of a story: from the phonetic properties of individual segments to where the passages falls within the narrative structure. However, most of these studies considered durations of individual segments, words, or pauses. It is not clear which of these effects will still persist when durations are averaged over a longer text as is the case for reading rate computation. In fact, studies in phonetics talk about "emergent speech rate" that can be relatively consistent over long stretches of speech (White, 2014). Furthermore, pause duration is likely to have a substantial effect on the reading rate (Kendall, 2013) yet previous research on pausing in story-telling suggests that this can be highly idiosyncratic.

3 Data

3.1 Text

We use the "Harry Potter and the Sorcerer's Stone" by J.K. Rowling (Rowling, 2015) as the case study for this paper. The book consists of 79,508 words spread across 17 chapters. We divided the text into 313 non-overlapping passages of about 250 words each (mean = 249 words; range: 190-309).[1] Boundaries of passages were set to be the starts

[1] This is roughly the intended length of a reading turn in the turn-taking reading intervention described in the Introduction.

and ends of paragraphs, where the end of a passage consists of a paragraph whose addition brings the passage closer to 250 words than without adding the paragraph. When generating passages, we took into account chapter boundaries so that no passage spanned two chapters: the word-count for passage generation was always re-set from the beginning of the chapter and any short fragments left at the end of a chapter were not included in the analysis. We randomly assigned 156 passages to the training set and 157 passages to the test set.

3.2 Audio

We used data from two narrators. The first dataset, hereafter referred to as JD, comes from a narration by the actor Jim Dale published as an audio-book (Rowling and Dale, 2016). The book is released as 17 .mp3 files with one file per chapter.

The second dataset comes from the audio-book with a female narrator, provided to us by Learning Ally.[2] We will refer to it as LA. These recordings are created by volunteers and are made available on subscription-basis to students diagnosed with disabilities that impact their ability to read print-based materials. Learning Ally recordings are subject to quality control similar to that of commercial audio-books.[3]

3.3 Calculating reading rate

We used forced alignment to automatically align the audio for JD narration for each chapter with the book text and establish the passage boundaries. We used the Kaldi toolkit (Povey et al., 2011) and publicly available acoustic models trained on the LibriSpeech corpus (Panayotov et al., 2015). The forced alignment was spot-checked manually for accuracy and found to be very accurate. The LA audio was already aligned with the book text.

The LA recordings were split across multiple audio files. To avoid any artifacts of the recording process, we only used the passages where the whole audio was in the same file. Out of the original 314 passages, 270 passages (86%) satisfied this condition, of these 134 in the training set and 136 in the test set. We used these matching training and testing passages for both narrators, in order to facilitate comparisons.

For both narrators we used the time stamps for the beginning of the first word in the passage and the end of the last word in the passage to compute the total duration of the passage, which was then divided by the number of words in the passage to yield the reading rate (words per minute, WPM).

4 RQ1: Is reading rate constant?

To answer our first question, we looked at the distribution of the reading rate across the passages in the training set.

The distribution of WPM for both narrators was close to normal. JD: mean = 164.01; SD = 12.66; min = 129.2; max = 197.7. LA: mean = 125.12; SD = 11.4; min = 86.8; max = 156.9. Based on discussions in the literature regarding syllables per second being a more stable measure of reading rate than WPM (Tauroza and Allison, 1990; Griffiths, 1991; Munro and Derwing, 1998), we calculated rate in syllables per second, and observed a similar pattern of variation (JD: mean = 3.52, SD = 0.30, LA: mean = 2.72, SD = 0.27). We also found that WPM and syllables per second were highly correlated, for each of the narrators ($r \geq 0.9$). We therefore continue with WPM, as this is the commonly used measure in the reading assessment context. The distribution of WPM for each of the narrators is shown in Figure **??**.

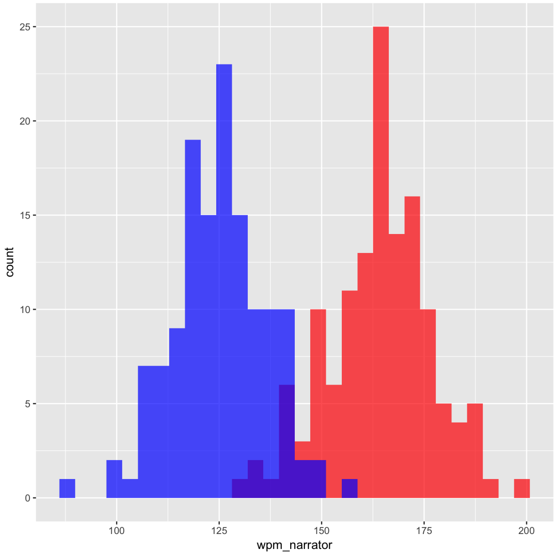

Figure 1: Distribution of WPM for LA (blue) and JD (red).

The answer to RQ1 is that while there is clearly a sense in which one narrator generally reads slower than the other, it is not the case that a narra-

[2] https://www.learningally.org/

[3] We cannot use another well-known commercially available narration, by Stephen Fry, since he narrates the British version of the book.

tor keeps the same rate of reading across different passages.

5 RQ2: Is variation in reading rate correlated across the two narrators?

We next compared the reading rates of the two narrators on the 134 training set passages. We found them to be highly correlated: Pearson's r = .81. This suggests that a substantial share of the variation across passages is systematic rather than idiosyncratic. We therefore proceed to the next question – what factors can explain this variation?

6 RQ3: What textual factors explain variation in reading rate?

6.1 Method

We use a standard model building approach to answer RQ3. We used the train partition with JD's WPM (hereafter JD-train) to identify possible textual features as well as the best learner to combine these features. We then trained separate models on the training data for the two narrators. We evaluated the two models on: (a) different passages from the test partition as read by the same narrator; (b) same passages as used for model training but read by a different narrator; (c) different passages (test partition) read by a different narrator.

6.2 Baseline: text complexity

We used text complexity as our baseline, following the practice in the reading assessment community. While we do not expect either of the narrators to experience any reading comprehension difficulties, one might reasonably assume that a skilled narrator would slow down on fragments which are harder for the listener to comprehend.

We used TextEvaluator,[4] a state-of-the-art measure of comprehension complexity of a text (Napolitano et al., 2015; Sheehan et al., 2014, 2013; Nelson et al., 2012).[5] TextEvaluator extracts a range of linguistic features and uses them to compute a complexity score on the scale of 100–2000. TextEvaluator computes three complexity scores based on the models optimized for literary, informational and mixed texts. We used the literary metric. The average complexity score for passages in the training set was 613.1, with a large variation across passages: min=240, max=1,019, SD=154.75. In other words, the selected book offers its readers much variety in the configurations of textual features across its different passages.

6.3 Features

We used the passage text to extract 107 features that capture different factors that might affect durations in oral reading. These could be grouped into four categories.

6.3.1 Sentence-level timing factors

We hypothesized that the timing effects described in section 2.2.1 are likely to be the source of at least some variation in reading rates across the text. Due to the complexity of these effects, building an accurate model that would predict segmental durations based on the text is not a trivial task.

This problem has been extensively discussed in literature on modeling prosody for text-to-speech synthesis systems (TTS) which generally combined the insights from the phonetic studies with statistical learning in order to establish the optimal duration for each segment and pause in synthesized audio. Therefore rather than attempting to build our own model, we synthesized the audio for each passage using Apple's built-in TTS engine (OS X 10.11.6). We used the male Alex voice which in terms of overall quality and default speaking rate appeared closest to JD. According to Capes et al. (2017), linguistic features used for training this system include segment identity and segmental context, stress, part-of-speech context, prominence[6], sentence type and initial/final positional features for syllable, word, phrase and sentence;[7] in other words, features directly related to timing factors discussed in 2.2.1.

We used the generated audio to compute the WPM for each passage. The mean reading rate of TTS was close to that of JD: 157.1 vs. 164.0. There was variation across passages with WPM varying from 129.2 to 197.7 (SD = 9.13).

6.3.2 Lexical, syntactic, and discourse features

Next, we considered lexico-syntactic properties of the passages. Some of these (lexical fre-

[4] https://textevaluator.ets.org/

[5] TextEvaluator appears in the Nelson et al. (2012) benchmark as SourceRater.

[6] See for example (Hirschberg, 1993) for features used to establish prominence

[7] Note that Capes et al. (2017) describe a different engine from the one used in this study, however as noted in the paper it shares the front-end for linguistic feature extraction with other Mac OS TTS systems. The same features are also described in (Zen et al., 2009).

quency, emotion, arousal) may be associated with local changes in segment and pause durations (see 2.2.2). Many of the features are used as low-level features in readability estimation (Graesser et al., 2004; Sheehan et al., 2014), and are thus likely to capture facets of a reader's experience when reading the text.

These included: (1) Vocabulary features capturing presence as well as average score along some meaning dimension, such as concreteness, imageability, emotion, arousal, motion, academic register (Coltheart, 1981; Warriner et al., 2013; Coxhead, 2000); (2) Morphological features (e.g., count of nominalizations, count of syllables); (3) Distributional features such as average word frequency; (4) Syntactic features such as counts of different part-of-speech, as well as features based on specific constructions (relative clauses, preposed clauses etc.); (5) Discourse features that deal with paragraphing (e.g., word count of the longest paragraph, average paragraph length in sentences) and overall cohesion (e.g., average lexical overlap across adjacent sentences).

6.3.3 Story-related features

Considering previous work on prosody in storytelling, we also built features that relate to the overall story development. These included: (1) The number of occurrences of names of the main characters and other proper nouns important to the plot (*Harry*, *Hermione*, *Weasley*, *Dumbledore*, *Ollivander*, *Quidditch*), under the assumption that there might be systematic ways in which the narrators act out certain kinds of people (older vs younger, for example), as well as events that could indicate a fast-paced event, such as a commentated game of Quidditch; (2) The order in which the passage appears in the book (as a numeric continuous variable); (3) Plot arc as estimated by `syuzhet` package (Jockers, 2015). This package uses sentiment analysis to attempt to reveal the latent structure of the narrative. We used the default sentiment lexicon developed by the Nebraska Literary Lab supplied with the package.

6.3.4 Performance-related features

Finally we considered typographic features that provide clues to how the text should be performed when read aloud. These included exclamation marks (!), ellipses (. . .), words printed in all capitals and indications that a character stutters.

6.4 Learner

We used 3-fold cross-validation on JD-train to compare performance of 9 regressors available via `SKLL`[8] package including Random Forest, SVR and various regularized linear models. We used grid-search with 3-fold cross-validation within each fold to fine-tune the parameters for all learners. We found that Lasso regression achieved the highest average performance and therefore used it as the learner for subsequent evaluations.

6.5 Results

Table 1 shows the performance of all models on the four datasets in our study. Since we are interested in explaining variation across passages rather than predicting the actual reading rate of a given narrator for a given passage, we use Pearson's r as our evaluation metric, as it would capture the extent to which the predicted and the observed values deviate similarly from their respective means, and thus would not be affected by differences in absolute values between the two narrators.

Dataset	Baseline	M_{JD}	M_{LA}
JD-train	0.38	-	0.71
LA-train	0.40	0.80	-
JD-test	0.37	0.74	0.74
LA-test	0.45	0.75	0.80

Table 1: Performance (Pearson's r) of models trained using Lasso regression on JD-train (M_{JD}) or LA-train (M_{LA}). The models are evaluated on unseen data: different passages read by the same narrator, same passages read by a different narrator and different passages read by different narrator. The table also shows the correlations with baseline (text complexity score).

The correlations between the baseline (estimates of comprehension complexity) and WPM of the two narrators were r=0.37–0.45. We also note that the direction of the correlation was opposite to our expectation: more complex passages were in fact read faster.

Our models substantially outperformed the baseline with r increasing from 0.4 to 0.7–0.8. In other words, the final models explain much of the variability in reading rates. Furthermore, this level of performance holds for predicting variation in reading rate for a set of unseen passages

[8] We used v1.3 from `https://github.com/EducationalTestingService/skll`.

read by a different narrator (M_{LA} on JD-test and M_{JD} on LA-test), suggesting fairly strong generalization. Results also suggest that the prediction is somewhat easier for LA than for JD, in that evaluations on the former are in the 0.75–0.80 range, and for the latter – in the 0.71–0.74 range, no matter which narrator supplied the training data. This could be due to Jim Dale's narration being more theatrical/artistic, hence somewhat more idiosyncratic.

We used 3-fold cross-validation on JD-train and LA-train to further consider how much of the variation can be explained by different groups of features discussed in Section 6.3. The results are shown in Table 2. For both narrators, models based on all groups of features outperformed models based on individual groups of features, but all groups of features were effective in explaining at least some variance in reading rate across passages. Timing as modeled by TTS was the highest performing feature followed by lexico-syntactic features and story-based features.

Dataset	JD-train	LA-train
Baseline	0.37	0.39
All features	0.69	0.77
Sentence-level timing	0.63	0.75
Lexical/syntactic/discourse	0.55	0.65
Story-related	0.47	0.47
Performance-related	0.35	0.35

Table 2: Performance of the four groups of features described in Section 6.3 on the training passages for each of the narrators. The table shows average Pearson's r for 3-fold cross-validation on JD-train and LA-train. All models use Lasso regression.

To summarize, we found text complexity to be a poor predictor of passage-to-passage variability in reading rates of adult narrators. These findings are consistent with recent work in the oral reading fluency community which found variation in children's reading fluency across passages after controlling for grade level (see Section 2.1).

We found that textual factors that explain a substantial share of passage-to-passage variability in reading rates include sentence-level timing factors such as distribution of segments, stressed syllables, sentences, and pauses as well as features related to passage vocabulary and syntax, story and performance. Given the good generalization of our results to both a new narrator and to new passages, we believe they hold promise for explaining some of the unaccounted-for variation in reading rates observed in the oral reading fluency studies; more research is necessary to explore this direction.

7 Discussion

Out of 107 original features, 17 features had non-zero coefficients in M_{JD} and 14 in M_{LA}, with 6 features in the overlap: timing, ellipsis, number of verbs in past tense, preposition count, *Weasley*, and *Dumbledore*. Additional features selected in only one of the two models included various vocabulary features (such as age of acquisition, imageability for M_{JD}), syntax (average word count before main verb, contractions for M_{JD}), discourse (average lexical overlap in adjacent sentences), as well as story features (syuzhet and *Dudley* for M_{LA}, *Ollivander*, *quidditch* for M_{JD}).

Some of these features lend themselves to an easy explanation. Thus in our study, a strong predictor of narrator slowdown was occurrence of ellipsis (...), a mark of hesitation or thoughtfulness; these were not modeled as such by TTS.

Similarly, the positive weight alloted to the average lexical overlap in adjacent sentences is consistent with the expectation that repeat instances would be read faster.

Effective character features included *Ollivander* and *Dumbledore*; mentions of both of these indicate a slowdown in narration. One possible explanation is that passages with multiple mentions of these characters are likely to be those where they speak. Both of these characters are elderly; acting them out could yield a slower rate of speech.[9]

In other cases the interpretation of the feature was less straightforward. Thus the feature with the second highest coefficient after timing for both narrators was that which counted occurrences of members of the *Weasley* family. Why?

Figure 1 plots standardized reading rates of JD (blue), LA (orange), and TTS (black) as a function of the location in the book. It is clear from the plot that in addition to passage-by-passage variation there is a global pattern in narrator WPM: the narrators slow down over the first few chapters,

[9] Barbara Roseblatt, an audiobook narration coach, explicitly advises to slow down when reading the contribution of the old character in a coversation with a young one: https://www.youtube.com/watch?v=MVmywsM9-h4, 5:17. Jim Dale himself describes his image of Dumbledore as *hesitant, wheezy old man*: https://www.youtube.com/watch?v=whzhEIB9Qkg: 2:45.

then speed up, and slow down again in the last third. It is also apparent that the TTS curve is flatter, suggesting that some of the slowdown and especially the speedup are not due to sentence-level timing factors.

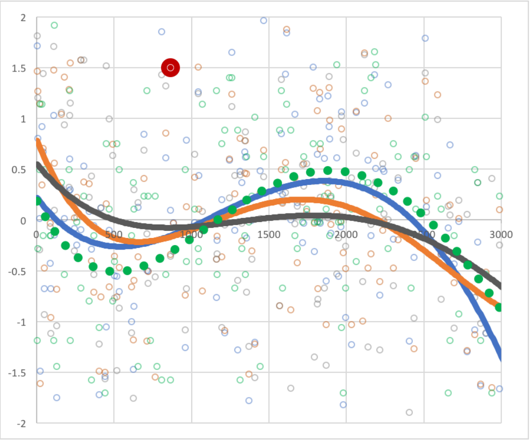

Figure 2: Standardized reading rates of JD (blue), LA (orange), TTS (black) as a function of the location of the passage in the novel. Trend lines are 4th degree polynomial approximations. The y-axis shows standardized readings rates (JD, LA, TTS). The large red dot and the green dotted line will be explained later in the text.

This book-level trend can help explain the strong performance of *Weasley*. This feature covers a number of characters that are prominent in the magic world as experienced by Harry (Ron, his brothers, sister, mother); they play no role at all in the first part of the story that is based in the Muggle world. The large red dot in Figure 1 indicates the first passage with a non-zero count for *Weasley*. This is very close to the onset of the speedup that is not captured by TTS. Apparently, the speedup coincides with an important plot transition (see Behr (2005) on plot transitions in Harry Potter), which is, in turn, indicated by a character mention pattern.

Next, we looked closely at one of the vocabulary features, specifically, imageability, calculated as the number of word tokens in a passage that belong to the MRC Imageability list (Coltheart, 1981). This feature has a partial correlation of -0.186 with JD controlling for TTS. In an attempt to identify the subset of the 1,194 words on the list that drive the correlation, we removed stopwords, all words that appeared in only one training passage, as well as short words (2-3 letters) and long words (7 letters or more). The partial correlation remained virtually the same (-0.178, $p < 0.05$).

These manipulations left us with 573 non-stop reasonably frequent 4-6 letter words. These words tend to name common everyday objects and properties (henceforth, **everyday** list), such as body parts (*knee, skin, neck, hair, nose, face, teeth*), colors (*blue, gray, green, white, black, orange, yellow*), family (*aunt, uncle, mother, father, sister, wife*), elements (*fire, water, wind, rain*) and materials (*silver, gold, stone, metal, glass, paper, silk*), eating (*cake, wine, dinner, hungry, eating*), common properties of objects (*warm, cold, broad, narrow, soft, hard, tall, short, long, clean, dirty*) and humans (*kind, evil, rude, polite, eager, proud, stupid, famous*), standard house interior (*chair, table, mirror, door, wall, room, clock*), feelings and emotions (*fear, hurt, hate, pain, anger, gloom, tired, panic, safe, boring, afraid, relief*), as well as numbers (*first, nine, half, dozen*), directional (*inside, back, front, behind, bottom*) and time expressions (*soon, hour, late, week, month, early, minute, moment*). These words "carry" the story, so to speak, in that on average about one third of all nonstop words in each passage belong to this list, albeit with substantial variation (min = 0.20; max = 0.49).

If the effect of the feature was simply due to higher incidence of the short high frequency everyday words, we would expect a *positive* correlation with the reading rate; in fact, the correlation is negative, suggesting that perhaps the feature is useful as an indirect indicator of something else, rather than for the phonological properties of the words on the list.

Variation across passages in the use of everyday words appears non-random. In particular, the first third of the book averages 41.4 matches per passages; the rest of the book averages 37.3. Given the above observations with *Weasley*, this is easy to explain in reference to the story line – the first part of the book mainly happens in Muggle ("normal") world, while the rest of the book happens in an alternative world of Hogwarts that is familiar enough (and so references to human feelings, bodies, and character still draw on the culturally familiar stock) yet different enough to drive a 10% average decline in the use of stock vocabulary, where special foods, special money units, the special game of quidditch, special subjects on the school's curiculum remain off the common list.

Since overlap with the everyday list has a negative correlation with reading rate, we flip the sign of the standardized everyday token counts, and overlay the plot with that of reading rates; see the green dotted line in Figure 1. It is apparent that the global pattern of JD WPM is closely traced by the feature, especially in the middle area where JD is speeding up and then slowing down again.

The observed global slowdown, speedup, and slowdown appear to align with the traditional three-part narrative structure (exposition, complication, and resolution) (Chandler and Munday, 2016). One of our features (`syuzhet`) was based on the plot arc. While this feature was selected in one of the models, its partial correlation with JD after controlling for TTS was not significant. Our results suggest that important plot transitions can sometimes be captured indirectly by tracing patterns of word usage for other specific classes of words such as characters or everyday words and that for skilled readers these transitions can be associated with systematic changes in reading rate.

8 Conclusions

The main contributions of this paper are as follows. First, we demonstrate using a case study that variation in reading rate across passages for professional narrators is consistent across readers and much of it can be explained using features of the texts being read. These findings suggest that it is possible to estimate the expected variation in durations of oral reading across texts. In the assessment context, this has a potential of providing a powerful control mechanism for selecting comparable passages for parallel forms of a test of oral reading; in a context when one cannot adjust the materials (such as a reading intervention using a particular book), it might be possible to adjust the measurement of reading rate to compensate for the effects of the text on the observed performance.

Secondly, we found that timing is a very powerful feature, yet not a perfect predictor of reading rate (the two narrators are still highly correlated controlling for timing, partial r=0.64). This opens up a possibility for a sophisticated assessment of oral reading using both TTS and human benchmark to separate reading that adheres to basic timing constraints of English speech (which constitutes a demonstrably big part of fluent reading) from a more nuanced expressive reading that TTS is not currently doing, but good human readers are. Thus beyond assessment context, our findings can also inform work on text-to-speech synthesis for book-length texts.

Extending and validating the results reported here using additional types of text and separating the effect of text factors on the two components of reading rate, articulation rate and pausing, is an important next step to get a more comprehensive picture of the impact of text on oral reading.

Acknowledgments

We thank John Sabatini and Tenaha O'Reilly for many comments and suggestions that inspired this work; Binod Gyawali and Patrick Lange for preprocessing the book; Diane Napolitano for help with extracting TextEvaluator features; Learning Ally for allowing us to use their version of the narration; Nitin Madnani, Michael Flor, Keelan Evanini, and three anonymous NAACL reviewers for their comments and suggestions.

References

Scott P. Ardoin, Shannon M. Suldo, Joseph Witt, Seth Aldrich, and Erin McDonald. 2005. Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly* 20(1):1–22. https://doi.org/http://dx.doi.org/10.1016/j.jsp.2012.09.004.

Gérard Bailly and Cécilia Gouvernayre. 2012. Pauses and respiratory markers of the structure of book reading. In *Proceedings of Interspeech 2012, Portland, OR*. pages 67–70. https://isca-speech.org/archive/interspeech_2012/i12_2218.html.

Kate Behr. 2005. "Same-as-difference": Narrative transformations and intersecting cultures in Harry Potter. *Journal of Narrative Theory* 35(1):112–132. http://www.jstor.org/stable/30224622.

Beata Beigman Klebanov, Anastassia Loukina, John Sabatini, and Tenaha O'Reilly. 2017. Continuous fluency tracking and the challenges of varying text complexity. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Copenhagen, Denmark, pages 22–32. http://aclweb.org/anthology/W/W17/W17-5003.pdf.

Alan Bell, Jason M. Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1):92–111. https://doi.org/10.1016/j.jml.2008.06.003.

Tina Burrows, Peter Jackson, Katherine Knill, and Dmitry Sityaev. 2005. Combining Models of Prosodic Phrasing and Pausing. In *Proceedings of Interspeech 2005, Lisbon, Portugal*. pages 1829–1832. https://isca-speech.org/archive/interspeech_2005/i05_1829.html.

Tim Capes, Paul Coles, Alistair Conkie, Ladan Golipour, Abie Hadjitarkhani, Qiong Hu, Nancy Huddleston, Melvyn Hunt, Jiangchuan Li, Matthias Neeracher, Kishore Prahallad, Tuomo Raitio, Ramya Rasipuram, Greg Townsend, Becci Williamson, David Winarsky, Zhizheng Wu, and Hepeng Zhang. 2017. Siri on-device deep learning-guided unit selection text-to-speech system. In *Proceedings of Interspeech-2017*. ISCA, ISCA, pages 4011–4015. https://doi.org/10.21437/Interspeech.2017-1798.

Daniel Chandler and Rod Munday. 2016. Narrative structure. In *A Dictionary of Media and Communications*, Oxford University Press. https://doi.org/10.1093/acref/9780191800986.001.0001.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33(4):497–505. https://doi.org/10.1080/14640748108400805.

Donald L. Compton, Amanda C. Appleton, and Michelle K. Hosp. 2004. Exploring the Relationship Between Text-Leveling Systems and Reading Accuracy and Fluency in Second-Grade Students Who Are Average and Poor Decoders. *Learning Disabilities Research & Practice* 19(3):176–184. https://doi.org/10.1111/j.1540-5826.2004.00102.x.

Averil Coxhead. 2000. A new academic word list. *TESOL Quarterly* 34(2):213–238. https://doi.org/10.2307/3587951.

Thomas H. Crystal and Arthur S. House. 1988. Segmental durations in connected speech signals: Syllabic stress. *The Journal of the Acoustical Society of America* 83(4):1574–1585. https://doi.org/10.1121/1.395912.

David Doukhan, Albert Rilliard, Sophie Rosset, Martine Adda-Decker, and Christophe D'Alessandro. 2011. Prosodic analysis of a corpus of tales. In *Proceedings of Interspeech 2011*. pages 3129–3132. https://isca-speech.org/archive/interspeech_2011/i11_3129.html.

Keelan Evanini, Michael Heilman, Xinhao Wang, and Daniel Blanchard. 2015. Automated Scoring for the TOEFL Junior ® Comprehensive Writing and Speaking Test. *ETS Research Report Series* 2015(1):1–11. https://doi.org/10.1002/ets2.12052.

Emerson Foulke. 1968. Listening Comprehension as a Function of Word Rate. *Journal of Communication* 18(3):198–206. https://doi.org/10.1111/j.1460-2466.1968.tb00070.x.

David J. Francis, Kristi L. Santi, Christopher Barr, Jack M. Fletcher, Al Varisco, and Barbara R. Foorman. 2008. Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of school psychology* 46(3):315–42. https://doi.org/10.1016/j.jsp.2007.06.003.

Ronald H. Good and Ruth A. Kaminski. 2002. DIBELS oral reading fluency passages for first through third grades. *Technical Report* 10. Eugene, OR: University of Oregon. https://dibels.uoregon.edu/docs/techreports/shaw_csap_technical_report.pdf.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2):193–202. https://doi.org/10.3758/BF03195564.

Roger Griffiths. 1991. Pausological research in an L2 context: A rationale, and review of selected studies. *Applied Linguistics* 12(4):345–364. https://doi.org/10.1093/applin/12.4.345.

Jan Hasbrouck and Gerald A. Tindal. 2006. Oral Reading Fluency Norms: A Valuable Assessment Tool for Reading Teachers. *The Reading Teacher* 59(7):636–644. https://doi.org/10.1598/RT.59.7.3.

Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language* 25(2):282–306. https://doi.org/10.1016/j.csl.2010.06.001.

Julia Hirschberg. 1993. Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence* 63(1):305 – 340. https://doi.org/https://doi.org/10.1016/0004-3702(93)90020-C.

Julia Hirschberg. 2002. Communication and prosody: Functional aspects of prosody. *Speech Communication* 36(1-2):31–43. https://doi.org/10.1016/S0167-6393(01)00024-3.

Arthur S. House and Grant Fairbanks. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America* 25(1):105–113. https://doi.org/10.1121/1.1906982.

Matthew L. Jockers. 2015. *Syuzhet: Extract Sentiment and Plot Arcs from Text*. https://CRAN.R-project.org/package=syuzhet.

T. Kendall. 2013. *Speech Rate, Pause and Sociolinguistic Variation: Studies in Corpus Sociophonetics*. Palgrave Macmillan UK.

Dennis H. Klatt. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America* 59(5):1208. `https://doi.org/10.1121/1.380986`.

Raúl Montaño and Francesc Alías. 2017. The role of prosody and voice quality in indirect storytelling speech: A cross-narrator perspective in four European languages. *Speech Communication* 88:1–16. `https://doi.org/10.1016/j.specom.2017.01.007`.

Murray J. Munro and Tracey M. Derwing. 1998. The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning* 48(2):159–182. `https://doi.org/10.1111/1467-9922.00038`.

Diane Napolitano, Kathleen M. Sheehan, and Robert Mundkowsky. 2015. Online Readability and Text Complexity Analysis with TextEvaluator. In *Proceedings of NAACL-HLT 2015, Denver, Colorado, May 31 - June 5, 2015*. pages 96–100. `http://www.aclweb.org/anthology/N15-3020`.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. In *Technical Report to the Gates Foundation*. `http://achievethecore.org/content/upload/\nelson_perfetti_liben_measures_of_text_difficulty_\research_ela.pdf`.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 5206–5210. `https://doi.org/10.1109/ICASSP.2015.7178964`.

Gordon E. Peterson and Ilse Lehiste. 1960. Duration of Syllable Nuclei in English. *The Journal of the Acoustical Society of America* 32(6):693–703. `https://doi.org/10.1121/1.1908183`.

Yaacov Petscher and Young-Suk Kim. 2011. The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of school psychology* 49(1):107–29. `https://doi.org/10.1016/j.jsp.2010.09.004`.

Hartmut R. Pfitzinger and Uwe D. Reichel. 2006. Text-based and Signal-based Prediction of Break Indices and Pause Durations. In *Proceedings of Speech Prosody 2006*. Dresden, Germany, page 269. `https://www.isca-speech.org/archive/sp2006/sp06_269.html`.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas Press.

J.K. Rowling. 2015. *Harry Potter and the Sorcerer's Stone*. Published as EPUB file by Pottermore Ltd.

J.K. Rowling and Jim Dale. 2016. *Harry Potter and the sorcerer's stone*. Listening Library/Penguin Random House.

Klaus R. Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40(1-2):227–256. `https://doi.org/10.1016/S0167-6393(02)00084-5`.

Kathleen M Sheehan, Michael Flor, and Diane Napolitano. 2013. A Two-Stage Approach for Generating Unbiased Estimates of Text Complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*. June, pages 49–58. `http://www.aclweb.org/anthology/W13-1506`.

Kathleen M. Sheehan, Irene Kostin, Diane Napolitano, and Michael Flor. 2014. The TextEvaluator Tool: Helping teachers and test developers select texts for use in instruction and assessment. *Elementary School Journal* 115(2):184–209. `https://doi.org/10.1086/678294`.

Steve Tauroza and Desmond Allison. 1990. Speech rates in British English. *Applied Linguistics* 11(1):90–105. `https://doi.org/10.1093/applin/11.1.90`.

Mariët Theune, Koen Meijs, Dirk Heylen, and Roeland Ordelman. 2006. Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech and Language Processing* 14(4):1137–1144. `https://doi.org/10.1109/TASL.2006.876129`.

Alice E. Turk and Stefanie Shattuck-Hufnagel. 2000. Word-boundary-related duration patterns in English. *Journal of Phonetics* 28(4):397 – 440. `https://doi.org/https://doi.org/10.1006/jpho.2000.0123`.

Alice E. Turk and Stefanie Shattuck-Hufnagel. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 35(4):445 – 472. `https://doi.org/https://doi.org/10.1016/j.wocn.2006.12.001`.

Jan P.H. van Santen. 1992. Contextual effects on vowel duration. *Speech Communication* 11(6):513–546. https://doi.org/10.1016/0167-6393(92)90027-5.

Lijuan Wang, Yong Zhao, Min Chu, Yining Chen, Frank K. Soong, and Zhigang Cao. 2006. Exploring expressive speech space in an audio-book. In *In Proceedings of Speech Prosody 2006*. Dresden, Germany, page 182. https://www.isca-speech.org/archive/sp2006/sp06{_}182.html.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45(4):1191–1207. https://doi.org/10.3758/s13428-012-0314-x.

Laurence White. 2014. Communicative function and prosodic form in speech timing. *Speech Communication* 63-64:38–54. https://doi.org/10.1016/j.specom.2014.04.003.

Laurence White and Alice E. Turk. 2010. English words on the procrustean bed: Polysyllabic shortening reconsidered. *Journal of Phonetics* 38(3):459 – 471. https://doi.org/https://doi.org/10.1016/j.wocn.2010.05.002.

Carl E. Williams and Kenneth N. Stevens. 1972. Emotions and Speech: Some Acoustical Correlates. *The Journal of the Acoustical Society of America* 52(4B):1238–1250. https://doi.org/10.1121/1.1913238.

Klaus Zechner, Keelan Evanini, and Cara Laitusis. 2012. Using Automatic Speech Recognition to Assess the Reading Proficiency of a Diverse Sample of Middle School Students. In *Proceedings of the Third Workshop on Child, Computer Interaction (WOCCI 2012), Portland, OR, USA*. International Speech Communications Association., Portland, OR, pages 45–52. http://www.isca-speech.org/archive/wocci_2012/wc12_045.html.

Heiga Zen, Keiichi Tokuda, and Alan W. Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51(11):1039–1064. https://doi.org/10.1016/j.specom.2009.04.004.

Yuan Zhao and Dan Jurafsky. 2009. The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *Journal of Phonetics* 37(2):231–247. https://doi.org/10.1016/j.wocn.2009.03.002.