Predicting Prosodic Boundaries for Children's Texts

Mansi Dhamne[1]*, Sneha Raman[2] and Preeti Rao[2]

[1]Sardar Patel Institute Of Technology, [2]Indian Institute of Technology, Bombay
mansi.dhamne22@spit.ac.in, {sneharaman, prao}@ee.iitb.ac.in

Abstract

Reading fluency in any language requires accurate word decoding but also natural prosodic phrasing i.e the grouping of words into rhythmically and syntactically coherent units. This holds for, both, reading aloud and silent reading. While adults pause meaningfully at clause or punctuation boundaries, children aged 8-13 often insert inappropriate pauses due to limited breath control and underdeveloped prosodic awareness. We present a text-based model to predict cognitively appropriate pause locations in children's reading material. Using a curated dataset of 54 leveled English stories annotated for potential pauses, or prosodic boundaries, by 21 fluent speakers, we find that nearly 30% of pauses occur at non-punctuation locations of the text, highlighting the limitations of using only punctuation-based cues. Our model combines lexical, syntactic, and contextual features with a novel breath duration feature that captures syllable load since the last major boundary. This cognitively motivated approach can model both allowed and "forbidden" pauses. The proposed framework supports applications such as child-directed TTS and oral reading fluency assessment where the proper grouping of words is considered critical to reading comprehension.

1 Introduction

Human speech is not a continuous stream of words, but rather unfolds in syntactically and semantically organized segments. Native speakers instinctively group words into prosodic units or chunks that are marked by pauses, pitch resets, boundary tones, and durational cues (Rosenberg, 2009). These segments include Prosodic Words (PW), Intermediate Phrases (iP), and Intonational Phrases (IP) nested hierarchically to reflect underlying structure and meaning (Figure 1). The endpoints of these units are known as phrase boundaries and are central to the naturalness and comprehensibility of speech.

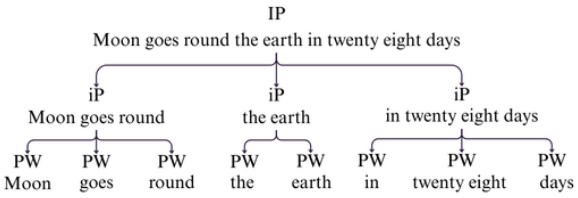

Figure 1: Example showing prosodic hierarchy within a sentence. IP: Intonational Phrase; iP: Intermediate Phrase; PW: Prosodic Word

In this work, we are concerned with the prediction of phrase boundaries from text in a manner that mimics the speech of a fluent child reader. However, Phrase or Prosodic Boundary Detection (PBD) is inherently challenging. First, prosodic phrasing often diverges from syntactic phrasing. Semantic emphasis, speaker intent, and discourse structure may override grammatical constraints, leading to syntax–prosody mismatches (Atterer and Klein, 2002). Second, most prior research on PBD targets acoustic cues such as pitch resets, pre-boundary lengthening, and silent pauses, which are absent in plain text (Jeon and Liu, 2009; Kuang et al., 2022; Ananthakrishnan and Narayanan, 2006). Third, prosodic boundaries are hierarchically structured, requiring models to distinguish between no break, minor breaks (e.g., PP-level), and major breaks (e.g., IP-level), making PBD a nuanced multi-class classification problem (Rosenberg, 2009; Yoon, 2006).

Predicting prosodic boundaries for children's texts, using only text-derived features, is thus a particularly underexplored problem. Reliably predicted boundaries can serve as references for the realised breaks in the child's oral reading. This is important in the assessment of oral reading fluency (ORF), a topic of growing interest due to its strong connection to reading comprehension (Bailly et al., 2022).

*Work carried out during an internship at IIT Bombay

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31875–31885

Acoustic cues to perceived breaks include pauses and other phrase-boundary cues based on pitch and duration. Most current approaches to ORF assessment either ignore prosody or treat it superficially. Yet recent studies show that listeners are highly sensitive to the placement of breaks (also termed 'pauses'). Unexpected disjunctures, especially within cohesive prosodic units like PWs or syntactic phrases, are perceived as unnatural and detrimental to comprehension (Sabu and Rao, 2024). In our study, we use the term *forbidden pauses* for such unexpected events.

In the context of reading fluency development, these concerns are magnified. Children differ significantly from adults in both linguistic maturity and physiological capacity. For instance, the average child between the age of 8 to 13 speaks at a rate of 2.87 $\pm$ 0.5 syllables per second (Logan et al., 2011), but their respiratory capacity limits breath groups to roughly 2 to 3.33 seconds, allowing for only 5 to 7 syllables per breath (Fleming et al., 2011). In contrast, adults can sustain breath groups with 14 to 16 syllables (Papadopoulos, 2014; Venkatagiri, 1999). This discrepancy underscores the need to model breath-driven phrasing constraints in child speech, rather than relying fully on adult prosody models.

We focus on linguistically grounded, child-aware modeling of prosodic boundaries using rich syntactic, positional, and semantic features from text. Our motivation lies in the fact that prosodic segmentation is not merely a low-level speech feature, but a reflection of developmental, physiological, and cognitive constraints, and that modeling it accurately is key to understanding and supporting child readers. Importantly, our models operate exclusively on text-based features, without relying on audio or prosodic contours, thus making them applicable in settings like text preprocessing for text-to-speech (TTS) systems and oral reading fluency (ORF) assessment and pedagogy.

Our work makes two primary contributions. First, we propose a cognitively motivated feature based on breath duration to model prosodic boundary strength. This feature quantifies the number of syllables uttered since the last major prosodic boundary and is grounded in developmental and physiological research on children aged 8 to 13. Second, we release a dataset containing annotations and extracted text-based features from the reading materials, which was used to predict prosodic boundaries.[1].

2 Previous Work

Prosodic boundary detection (PBD) has evolved considerably over the past few decades, transitioning from rule-based systems grounded in linguistic heuristics to data-driven methods leveraging syntactic, semantic, and acoustic features.

Early approaches relied on handcrafted rules or punctuation-based alignment (Möbius, 1999), which proved insufficient for handling spontaneous or expressive speech. More advanced heuristics leveraged function-content word patterns (Ejerhed, 1988; Taylor and Black, 1998; Brierley and Atwell, 2007), but still fell short in syntactic depth. These methods primarily depended on surface-level features such as part-of-speech sequences or word categories, lacking access to the hierarchical and relational structures captured by full syntactic parses. Rule-based systems informed by syntax, including the PHI algorithm (Gee and Grosjean, 1983) and various clause-structure heuristics (Atterer and Klein, 2002; Fitzpatrick, 2001), achieved more accurate phrasing predictions, yet their performance was contingent on high-quality parses and often failed to generalize across domains.

The availability of annotated corpora such as the Boston University Radio News Corpus (BURNC) (Ostendorf et al., 1995) with TOBI-labeled prosodic annotations (Beckman and Ayers, 1997) enabled a shift toward supervised learning. CART models like Wang (1991) combined syntactic and prosodic features to achieve strong accuracy of nearly 90% on datasets like ATIS (Hemphill et al., 1990). Later studies extended classification-based methods using maximum entropy models, LightGBM, and memory-based learning. Zhang et al. (2006) proposed a multi-pass ME model with post-hoc rule refinement for Mandarin, Trang et al. (2021) demonstrated performance gains in Vietnamese PBD using PhoBERT embeddings and tree-based syntactic features.

Sequence models such as Conditional Random Fields (CRFs) became popular for capturing contextual label dependencies (Kim et al., 2009; Qian et al., 2010; Levow, 2008). CRFs outperformed decision trees in modeling sequential prosodic patterns and showed robustness on corpora like BURNC. Linear models like logistic regression and SVMs provided interpretability, with lexical

[1]Link to Dataset

and contextual features shown to be predictive of boundary strength (Rosenberg, 2009; Jeon and Liu, 2009).

Mishra et al. (2015) achieved an accuracy of 94.7% using a logistic regression-based binary classifier to identify intonational phrase breaks. Yoon (2006) explored memory-based learning that combined semantic, syntactic, and phonological cues, achieving high accuracy—93.23% for boundary classification and 88.06% for boundary strength prediction—on the BURNC corpus. Sloan (2023) further improved TTS by injecting explicit prosodic events (breaks and pitch accents) using random forest classifiers trained on BURNC, achieving an accuracy of 93.4% in identifying phrase boundaries.

Recent work has focused on deep learning, especially LSTMs and BiLSTMs, which model temporal dependencies effectively when combined with embeddings. Rendel et al. (2016) showed that continuous embeddings can predict symbolic prosodic events in TTS, with an F1-score of 0.83.

Despite these advances, many systems still rely on proxy labels—such as punctuation or TTS heuristics — that poorly reflect perceived prosody, especially in children's speech that is shaped by unique cognitive and physiological factors. To address this, we model prosodic boundary strength as a continuous variable derived from human annotations, enabling a more nuanced and perceptually grounded formulation of phrasing suited for child-directed speech applications.

3 Data and Tasks

We present our dataset of human-annotated text passages followed by a discussion of the research tasks.

3.1 Dataset

The dataset developed for this study comprises 54 passages drawn from the stories in the Reading Cards created by the Central Institute of English and Foreign Languages (CIEFL), India[2] and marked for reading level by grade. The selected stories are evenly distributed across six grade levels (Grades 3–8), with nine passages per grade, covering both narrative and expository genres. To maintain a balance of text types, each grade includes a mix of both genres. Direct speech in the original stories was converted to indirect speech, as is typically done for texts created for use in oral reading fluency assessments, to encourage a consistent reading style. The dataset consisted of 8662 words and 701 sentences.

Manual annotation by fluent adult speakers of Indian English was used to obtain the prosodic boundaries in each text. Annotators were presented with clear on-screen instructions within a custom-built web-based annotation tool, instructing them to mark the boundary words, i.e. words that they would naturally pause after, when reading aloud the presented texts to the children they were meant for. To limit the total time required of an annotator to within one hour, the 54 passages were divided into 3 batches with each batch containing 18 stories (three per grade), with a balanced representation of genres (narrative and expository). A given annotator received only one batch of passages.

A total of 21 annotators were recruited, ranging from university students to working professionals, with self-declared high proficiency in English. Each batch was assigned to a distinct group of 7 annotators, aligning with empirical findings from (Cole et al., 2017), which suggests that inter-rater agreement stabilizes around this number. Stories within each batch were randomly ordered for each annotator to mitigate order effects. Inter-annotator agreement, measured with Fleiss' Kappa scores, obtained 0.69, 0.75, and 0.61 for the 3 batches, with the overall average of 0.68 indicating substantial agreement according to the criteria of Landis and Richard (1977).

3.2 Data Insights

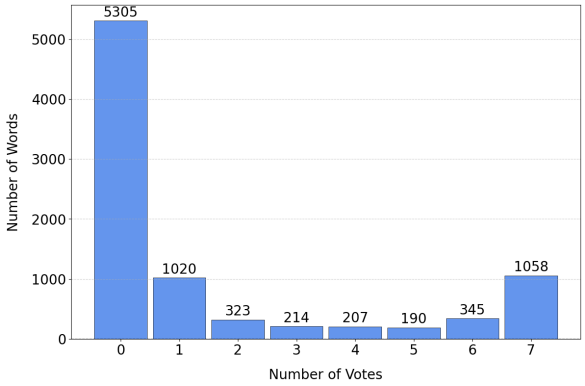

Figure 2: Distribution of the 8662 words in our dataset by number of votes for pause presence (out of a total of 7 votes).

With every word in the dataset annotated by 7 raters, we obtain a histogram of the pause votes

[2] https://www.orientblackswan.com/books?id=0&pid=0&sid=40

per word as shown in Figure 2. The distribution is skewed toward extremes indicating relatively high annotator agreement on words that receive either 0 or 7 votes. Words with 0 votes (5305, or 60%) are typically short function words in syntactically uninterrupted regions. Words receiving 1–2 votes appear within fluent, unmarked spans and reflect marginal or reader-specific pause tendencies. Those with 3–4 votes often occur near clause boundaries (e.g., after subordinators or conjunctions) and suggest variability in prosodic phrasing strategies. Words with 5–6 votes frequently occur at near-clause endings or just before breath-group transitions, determined by word and syllable counts, signaling softer boundary markers. Full agreement cases (7 votes) align with sentence or clause-final positions and punctuation, offering strong supervision cues for pause prediction. The observed correspondence of number of votes with a boundary hierarchy supports modeling of prosodic boundaries as continuous phenomena rather than strictly binary events.

Boundary Metrics: We also examine typical phrase durations in terms of number of syllables per phrase, separately for each grade given the expected dependence of text complexity on grade. Table 1 presents the average number of syllables and words per sentence, as well as between consecutive human-annotated boundaries with at least one vote. As expected, the number of syllables and words per sentence increases steadily from Grade 3 to Grade 7, reflecting longer and more syntactically complex sentence structures in higher-grade texts. Notably, the spacing between prosodic boundaries also shows a gradual increase with grade level. For example, Grade 3 stories exhibit a prosodic boundary roughly every 5.5 syllables, whereas Grade 7 texts show lower pause density with an average of 7.1 syllables between pauses. This gradual lengthening of phrase segments may reflect a shift toward more complex syntactic constructions and longer breath or thought groups in advanced-level texts.

Further, we observed that for prosodic boundaries receiving five or more annotator votes, an average of 7.0 syllables is obtained between successive boundaries, with standard deviation of 3.5. This corresponds to roughly one pause every 5–7 syllables, aligning well with estimated breath group sizes in younger children and offering physiological motivation for incorporating breath-related cues in pause prediction models.

Grade	Syl/Sent M (SD)	Wd/Sent M (SD)	Syl/Bd M (SD)	Wd/Bd M (SD)
3	11.2 (2.2)	9.2 (2.1)	5.5 (1.2)	4.5 (0.9)
4	14.9 (2.3)	11.8 (1.6)	6.0 (1.8)	4.7 (1.4)
5	16.1 (3.3)	12.5 (2.5)	6.2 (1.7)	4.8 (1.2)
6	17.6 (2.4)	13.2 (1.7)	6.4 (2.2)	5.0 (1.8)
7	22.2 (3.7)	15.8 (2.7)	7.1 (3.1)	5.3 (2.3)

Table 1: Average number of syllables and words - Mean (M), Standard deviation (SD) - between consecutive annotated boundaries, grouped by grade level. Syl: Syllables; Sent: Sentence; Wd: Word; Bd: Boundary

Role of Punctuation and Syntax: Further analysis shows that punctuation, particularly terminal markers (., ?, !), corresponds to strong prosodic boundaries. All such locations were labeled as pauses by all 7 annotators. Non terminal markers (commas, -, :, etc.), however, show more variability: out of 430 comma-marked locations, only 57% reached full agreement (Table 2).

Votes	IP (700)	ip (430)
$\leq$4	0	12 (3%)
5–6	0	173 (40%)
7	700 (100%)	245 (57%)

Table 2: Distribution of phrase boundaries based on punctuation and annotator agreement. IP: Intonational Phrase boundary marked by terminal punctuation; ip: Intermediate Phrase boundary marked by non-terminal punctuation.

Despite being a punctuation cue, commas function as a soft constraint and show significant variability in annotator-rated boundary strength. Figure 3 illustrates the distinct ways in which commas can be interpreted prosodically.

Long,| long ago, | the peacock had a lovely voice | and beautiful feathers|.

When you eat a banana, | you see small, | soft, | black seeds | inside the fruit|.

Figure 3: Examples where red | denotes locations with $\leq$4 votes (low agreement) and green | shows pause locations with $\geq$ 5 annotator agreement.

Syntactic Cues: Even after punctuation-based boundaries are removed, the inter-annotator agreement remains moderately high (κ = 0.45), indicating that prosodic phrasing is not solely governed by surface punctuation cues but also by deeper linguistic structures such as syntax and breath planning. Notably, clause boundaries — especially those

marked by coordinating conjunctions like "and" or "but" — show strong alignment with pause placement. If we consider locations that receive 5 or more pause votes as perceived pauses, we find that out of 323 instances of coordinating conjunctions, 255 (nearly 79%) were immediately preceded by a perceived pause. This consistent pattern reinforces the role of syntactic segmentation in pause prediction, suggesting that readers instinctively place boundaries at points of structural or cognitive reset.

3.3 Tasks

We define three automatic prediction tasks of interest in downstream applications of this work. We intend to evaluate models based on text-derived features on these tasks.

- **Prediction of perceived boundary strength by regression:** The number of annotator votes per word (ranging from 0 to 7) serves as a continuous measure of how strongly a word is perceived as a natural pause point. This formulation allows for nuanced modeling beyond binary decisions. For analysis purposes, we group the words into three ranges to reflect different levels of perceived boundary strength (see Table 3): no pause (0 votes), low to moderate agreement (1–4 votes), and high agreement (5–7 votes).

Votes	Number of Words
0	5305 (61.24%)
1–4	1764 (20.36%)
5–7	1593 (18.40%)

Table 3: Grouping of vote counts for the 8662 words into three classes for analysis and visualization.

- **Binary classification for presence/absence of prosodic boundaries:** Words receiving votes from at least 5 out of 7 annotators indicate strong consensus and are labeled as *boundary*. The remaining words are labeled *no boundary*.

- **Binary classification for presence/absence of forbidden pauses:** Words unanimously marked with 0 votes (i.e., all annotators agreed that no pause should occur) are categorized as *forbidden pauses*. These represent positions in the text where pausing may be particularly disruptive to the comprehension of the text. All other words (with 1–7 votes) are considered *allowed pauses*. This formulation supports applications that aim to discourage disfluent or unnatural pausing behavior during reading as in ORF assessment.

4 Feature Extraction

Informed by the literature reviewed previously, we extracted a comprehensive set of lexical, syntactic, semantic, positional, and physiological features for each word in the corpus to support accurate modeling of prosodic boundary prediction, as presented here.

Lexical Features: At the surface level, we included the word form and its lemmatized version, along with orthographic and typographic characteristics such as whether the word is capitalized, whether it is the last word in the sentence, and whether it is followed by a punctuation mark. Additional features include word length (in characters), raw word frequency (computed from the corpus statistics), and a binary indicator for whether the word is a function word. To capture lexical predictability, we also incorporated the word's score from pre-trained BERT language model (Devlin et al., 2018).

Semantic Features via Word Embeddings: To represent semantic similarity without relying on raw high-dimensional embeddings, we extracted contextualized word vectors using Google's `bert-base-uncased` (Devlin et al., 2018). Since embeddings are not directly usable as features in classification tasks, we reduced their dimensionality via PCA and then applied K-Means clustering to group semantically similar embeddings. Each word was thus assigned a discrete cluster ID. Notably, these clusters showed strong alignment with coarse-grained part-of-speech categories—for example, one cluster predominantly represented pronouns, another grouped nouns and proper nouns, and so on. In addition to each word's cluster ID, we included the cluster IDs of its preceding and succeeding words to provide lightweight semantic context.

Syntactic and POS Features: Syntactic features were derived from both dependency and constituency parses. Each word is tagged with its part-of-speech (POS), fine-grained morphological tag, and dependency relation. To provide broader syntactic context, we included the POS tags and dependency labels of the three preceding and three

succeeding words in the context window (Mishra et al., 2015). These were extracted using the spaCy NLP toolkit (Honnibal et al., 2024)

From constituency parses generated using the Berkeley Neural Parser (Kitaev et al., 2019; Kitaev and Klein, 2018), we extracted deeper syntactic cues such as the depth and label of the smallest constituent containing the word. We computed the forward and backward positional indices of the word within this constituent. To quantify syntactic distance, we constructed a minimal spanning tree between the current word and its neighboring word, from which we derived both the tree's label and its length (Sloan et al., 2022). We also calculated the word's distance from the phrase root and the phrase head, both in terms of syntactic depth and position within the parse structure. These features together capture a rich representation of the phrase-structural boundaries and hierarchical grammatical cues that are known to influence prosodic phrasing.

Positional Features: These features encode the word's position in its sentence, both as an absolute index and as a relative value normalized by the sentence length. Sentence length itself is also included as a feature.

Physiological Features: To model breathing and planning constraints in speech, we included prosody-motivated physiological features. For each word, we recorded its number of syllables, the cumulative syllables since the last sentence ending, and the number of syllables remaining in that breath group duration (taken as 7 syllables) . These features are grounded in cognitive-linguistic theory, which suggests that prosodic boundaries often align with respiratory planning units (Wang et al., 2010).

5 Model Evaluation

To assess our models' generalization ability across grade levels and discourse types, we adopt a grade-stratified, style-balanced evaluation protocol. For each grade level (Grades 3–8), one narrative and one expository story are held out as test data, comprising 12 stories in all. This setup ensures independence between training and test splits while maintaining lexical, syntactic, and stylistic diversity in each split. The model was trained on features and targets of the training set texts.

All models were trained using the LightGBM framework (Ke et al., 2017), an efficient version of the popular gradient boosting decision trees. For regression tasks, we used the LGBMRegressor, while for classification tasks, we employed the LGBMClassifier. Features described in Section 4 were extracted for each word token w_i in the dataset, excluding punctuation tokens. However, punctuations were included in the context window for part-of-speech and dependency features. Feature ablation studies were conducted to assess the contribution of different linguistic and prosodic cues.

5.1 Results

Model evaluation results are reported on the held out test set of 6 passages spanning all the grades.

Regression task: Table 4 shows the regression performance across different feature combinations. The full model achieved an R^2 of 0.87, RMSE of 0.94, and MAE of 0.57, outperforming outperforming all cumulative ablation variants. Models restricted to subsets such as lexical, punctuation, or syntactic features showed consistently lower performance, highlighting the complementary contribution of each feature group.

Feature Set	R^2	RMSE	MAE
Punctuation	0.67	1.48	1.05
+ Lexical	0.70	1.44	0.95
+ Semantic	0.76	1.27	0.85
+ Syntactic and POS	0.79	1.16	0.75
+ Positional	0.80	1.14	0.73
+ Physiological	**0.87**	**0.94**	**0.57**

Table 4: Regression performance across cumulative feature subsets. Bold values indicate the best performance. Feature categories include punctuation (1), lexical (7), semantic (3), syntactic (12), POS and dependencies (15), positional (2), and physiological (3). Numbers in parentheses denote the feature count within each category.

To better understand model behavior across the spectrum of annotator agreement, we analyzed regression error by vote count. As shown in Table 5, the model performs best at the extremes: MAEs are lowest when there is unanimous agreement on no-pause (0 votes) or strong consensus on a pause (5–7 votes). Errors peak in the ambiguous mid-range (1–4 votes), where inter-annotator variability and prosodic ambiguity are highest.

Classification tasks: We report performance on two classification tasks: Prosodic Boundary Detection (PBD) and Forbidden Pause Detection (FPD). Table 6 summarizes the obtained precision, recall,

Votes	Meaning	MAE
0	No Pause	0.38
1–4	Low to Moderate Agreement	0.67–1.77
5–7	High Agreement	0.37–1.53

Table 5: Mean Absolute Error (MAE) by annotator agreement level (vote count).

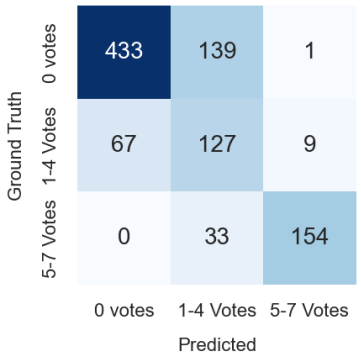

Figure 4: Confusion matrix between annotation number of votes and rounded regression outputs, both grouped as 0 votes: No Pause, 1-4 votes: Low to Moderate Agreement, 5-7 votes: High Agreement

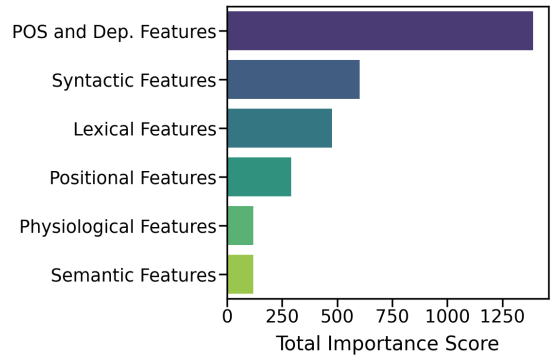

Figure 5: Importance of features grouped by their category as referred in Section 4

F1-score, and accuracies. Overall, we note performances that are comparable to similar tasks in the reviewed literature.

Performance Metrics	PBD	FPD
Precision	0.92	0.83
Recall	0.85	0.91
F1-score	0.89	0.87
Accuracy	0.96	0.84

Table 6: Performance on the two classification tasks. Precision and recall are computed for the positive class (i.e., presence of boundary or forbidden pause).

Prosodic Boundaries Detection: Despite high overall performance in prosodic boundary detection, we note a recall of 85%. The model fails on a small but important subset of boundaries, particularly in complex syntactic or prosodically nuanced contexts. A common pattern among false negatives involves clause-internal boundaries that are not marked by punctuation but are clearly perceived as prosodic breaks by human annotators. For instance, the model often misses pauses after verbs introducing embedded clauses (e.g., *"Nobody really knows | how they find their way..."*) or after noun phrases ending in rare or meaningful content words (*e.g., "the thick and rough khadi | doesn't make me suffer"*). It also struggles with coordinated or contrastive structures, such as between adjectives (*"a thick coat of short | soft hair"*) or content-heavy noun phrases used for emphasis or affect (*"I was living alone | in an empty flat..."*). These cases suggest that while the model captures many syntactic patterns, it underestimates boundaries that depend on context prominence or subtle shifts in information structure. The examples are presented in full below.

GT: Nobody really knows | how they find their way on these...
Model: Nobody really knows how they find their way | on these...

GT: ... the thick and rough khadi | doesn't make me suffer. |
Model: ... the thick and rough khadi doesn't make me suffer. |

GT: It is a thick coat of short | soft hair. |
Model: It is a thick coat of short soft hair. |

GT: I was living alone | in an empty flat | in a quiet Moscow street. |
Model: I was living alone in an empty flat | in a quiet Moscow street. |

Forbidden Pause Detection: While the model exhibits strong recall (0.91) in detecting true forbidden pauses i.e instances with zero annotator votes, it also mislabels a notable number of words as forbidden despite evidence of pausing, resulting in a lower precision of 0.83. Most of these false positives involve words with exactly one vote, typically function words or mid-clause verbs like *"fly"* in *"they fly to warmer lands."* Such cases reflect reader-specific phrasing preferences, where occasional pauses occur for emphasis, hesitation, or rhythm, but are not consistently marked across readers. We also observe a few misclassifications on words with

3 or more votes where pauses are more reliably perceived. These include punctuation-induced emphasis (e.g., *"...and say just one word: | Money"*), coordinated structures (*"to swear | and to smoke"*), and discourse-driven boundaries (*"That is why | ..."* or *"She thought | we had all gone mad"*). Such errors indicate that the model sometimes overlooks expressive or intonational pauses that lie beyond strict syntactic rules or punctuation. Examples in full appear below.

GT: So | when winter sets in, | they fly | to warmer lands. |
Model: So | when winter sets in, | they fly to warmer lands. |

GT: She thought | we had all gone mad. |
Model: She thought we had all gone mad. |

GT: The street taught me | to swear | and to smoke | and to keep...
Model: The street taught me | to swear and to smoke | and to keep...

GT: That is why | even the tiger | is afraid of it. |
Model: That is why even the tiger | is afraid of it. |

5.2 Discussion

This work aimed to predict the expected prosodic boundaries in children's oral reading using only the corresponding prompt text. We focused on pause patterns driven by linguistic structure and developmental reading behavior. Our findings highlight that prosodic phrasing in child-directed texts is shaped not just by surface cues like punctuation, but by a complex interaction of syntax, semantics, and physiological factors such as breath control.

To reflect real-world educational and storytelling contexts adult annotators were explicitly instructed to mark pauses as they would naturally insert them when reading aloud to children, rather than to adult audiences. Analysis of the annotated data (Table 1) revealed that the distribution of syllables between marked boundaries closely aligns with known breath group sizes for children. This suggests that annotators implicitly adopted pausing patterns consistent with children's physiological and cognitive constraints.

Nearly 30% of human-marked pauses occurred at non-punctuation locations, highlighting that punctuation alone is an insufficient marker for prosody in children's reading. While terminal punctuation reliably aligned with high-agreement boundaries, commas showed notable variability—only 57% of comma-marked pauses had unanimous annotator consensus—suggesting punctuation is a probabilistic cue.

Further, the pausing patterns across six grade levels revealed clear developmental trends in prosodic segmentation. Even though annotators were not explicitly informed of the grade level of the stories they labeled, we observed longer stretches between pauses as grade level increased (see Table 1). This suggests that prosodic phrasing naturally aligns with the growing syntactic complexity, cognitive load, and breath capacity of older children. These findings underscore the importance of modeling prosody in a grade-sensitive manner, as the same features may signal different prosodic behavior across reading proficiency levels.

Considering our observations, our model integrates lexical, syntactic, and contextual embeddings with a novel breath-duration feature that captures syllable load since the last major pause. This cognitively motivated feature draws on our finding that prosodic boundaries typically appear every 5–7 syllables—consistent with known breath group limits in children (see Section 1).

The regression-based formulation, which modeled annotator agreement as a continuous measure of prosodic boundary strength, enabled a more graded interpretation of pause likelihood. As observed in Figure 4, the model performs reliably at the extremes with intermediate words (1–4 votes) exhibiting significant confusion with adjacent classes. This reflects the inherent ambiguity of mid-strength boundaries, where even human raters show limited consensus.

In the classification tasks, our models performed well across both prosodic boundary detection and forbidden pause identification. As discussed in Section 5.1, the errors observed in prosodic boundary detection involve syntactically complex or semantic emphasis contexts where explicit textual cues are limited. The model's reliance on structural features may thus limit its sensitivity to these more nuanced forms of pausing, which can also be listener or discourse driven rather than strictly rule-governed.

On the task of forbidden pause detection, the model achieved high recall (0.91), effectively iden-

tifying most positions where no annotator indicated a pause which is essential for flagging disfluent or unnatural phrasing in ORF assessment. This suggests potential for supporting educators in diagnosing phrasing-related stemming from the lack of comprehension and guiding targeted reading interventions for learners.

The observed errors, discussed in Section 5.1, point to a core limitation that expressive boundaries often depend on discourse-level and pragmatic cues—such as contrast or elaboration which are not reliably encoded in surface syntax or local lexical context. Given that the model's features are primarily local and structurally driven, it may lack the semantic flexibility needed to capture the full range of phrasing decisions that readers make, especially those involving interpretive or stylistic variation.

Analyzing the feature importances grouped by their linguistic roles (Figure 5) provides deeper understanding of how the model makes predictions. Features related to syntax and dependency structures carry the most weight, followed by lexical information and prosodic indicators such as syllable counts and breath group length. This pattern supports our hypothesis that children's phrasing choices are primarily guided by grammatical structure while also being influenced by cognitive and physiological factors like breath capacity. Instead of depending on a single source of information, the model leverages a combination of diverse features, reflecting the complex, multifaceted nature of reading with expression.

6 Conclusion

Our findings reveal that a combination of syntactic, lexical, and prosodic features enables accurate prediction from text of prosodic boundaries and forbidden pauses for children's oral reading. This approach holds significant promise for applications in education such as learning to read with comprehension for children, and for second-language learners by enhancing TTS systems with natural prosodic phrasing information. Since the applicable features can be extracted with standard NLP tools and are fully interpretable linguistically, this methodology has strong potential for adaptation to other languages with similar syntactic and prosodic structures. Specifically, we plan to extend this work to Hindi and Marathi children's texts where the use of punctuation, such as comma, in printed text tends to be relatively low. By expanding datasets, future research can further improve model accuracy and practical usability in multilingual educational settings. A future impactful validation of our model can involve comparison of its predictions with data on prosodic boundaries actually realised by fluent children in oral reading assessments as detected by acoustic measurements.

Limitations

Our work assumes that text-based features alone can adequately capture prosodic patterns, but prosody is influenced by additional factors such as speaker intent, emphasis, and pragmatic context, which are not represented in our model. This simplification may limit robustness across diverse speech styles or domains. Our work assumes prosodic patterns generalize across readers and linguistic backgrounds; however, variability in dialects and individual reading habits may affect performance. Human annotations introduce subjectivity and potential inconsistencies, adding noise that can impact accuracy. Furthermore, the relatively small dataset size restricts exposure to varied prosodic phenomena, limiting generalizability. Finally, our approach does not explicitly model multilevel prosodic features such as boundary tones or pitch accents, which could enhance the naturalness and realism of predicted prosody.

Acknowledgments

We sincerely thank the human annotators for their valuable efforts in marking the pauses, which were instrumental in the creation of the dataset used in this study. We also gratefully acknowledge the support of the Industrial Research and Consultancy Centre (IRCC), IIT Bombay, for providing the first author with the opportunity to pursue this research through an internship program.

References

Sankaranarayanan Ananthakrishnan and Shrikanth S Narayanan. 2006. Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling. In *INTERSPEECH*.

Michaela Atterer and Ewan Klein. 2002. Integrating linguistic and performance-based constraints for assigning phrase breaks. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Gérard Bailly, Erika Godde, Anne-Laure Piat-Marchand, and Marie-Line Bosse. 2022. Automatic

assessment of oral readings of young pupils. *Speech Communication*, 138:67–79.

Mary E Beckman and Gayle Ayers. 1997. Guidelines for tobi labelling. *The OSU Research Foundation*, 3(30):255–309.

Claire Brierley and Eric Atwell. 2007. Prosodic phrase break prediction: problems in the evaluation of models against a gold standard. *TAL Journal: Traitement Automatique des Langues*, 48(1):187–206.

Jennifer Cole, Timothy Mahrt, and Joseph Roy. 2017. Crowd-sourcing prosodic annotation. *Computer Speech & Language*, 45:300–325.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Eva Ejerhed. 1988. Finding clauses in unrestricted text by finitary and stochastic methods. In *Second conference on applied natural language processing*, pages 219–227.

Eileen Fitzpatrick. 2001. The prosodic phrasing of clause-final prepositional phrases. *Language*, 77(3):544–561.

Susannah Fleming, Matthew Thompson, Richard Stevens, Carl Heneghan, Annette Plüddemann, Ian Maconochie, Lionel Tarassenko, and David Mant. 2011. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. *The Lancet*, 377(9770):1011–1018.

James Paul Gee and François Grosjean. 1983. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive psychology*, 15(4):411–458.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The atis spoken language systems pilot corpus. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '90, page 96–101, USA. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2024. spaCy: Industrial-strength natural language processing in python.

Je Hun Jeon and Yang Liu. 2009. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4565–4568.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Minho Kim, Youngim Jung, and Hyuk-Chul Kwon. 2009. Prediction of korean prosodic phrase boundary by efficient feature selection in machine learning. pages 323–327.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Jianjing Kuang, May Pik Yu Chan, Nari Rhee, Mark Liberman, and Hongwei Ding. 2022. The mapping between syntactic and prosodic phrasing in english and mandarin. In *INTERSPEECH*, pages 3443–3447.

J Landis and Gary G Richard. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Gina-Anne Levow. 2008. Automatic prosodic labeling with conditional random fields and rich acoustic features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Kenneth J Logan, Courtney T Byrd, Elizabeth M Mazzocchi, and Ronald B Gillam. 2011. Speaking rate characteristics of elementary-school-aged children who do and do not stutter. *Journal of Communication Disorders*, 44(1):130–147.

Taniya Mishra, Yeon-jun Kim, and Srinivas Bangalore. 2015. Intonational phrase break prediction for text-to-speech synthesis using dependency relations. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4919–4923. IEEE.

Bernd Möbius. 1999. The bell labs german text-to-speech system. *Computer Speech Language*, 13:319–358.

Mari Ostendorf, Patti J Price, and Stefanie Shattuck-Hufnagel. 1995. The boston university radio news corpus. *Linguistic Data Consortium*, pages 1–19.

Georgina Paulette Maria Papadopoulos. 2014. Speech breathing and prosody during statement and question productions in seven-year-old children.

Yao Qian, Zhizheng Wu, Xuezhe Ma, and Frank Soong. 2010. Automatic prosody prediction and detection with conditional random field (crf) models. In *2010 7th International Symposium on Chinese Spoken Language Processing*, pages 135–138. IEEE.

Asaf Rendel, Raul Fernandez, Ron Hoory, and Bhuvana Ramabhadran. 2016. Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5655–5659.

Andrew Rosenberg. 2009. *Automatic detection and classification of prosodic events*. Columbia University.

Kamini Sabu and Preeti Rao. 2024. Predicting children's perceived reading proficiency with prosody modeling. *Computer Speech & Language*, 84:101557.

Rose Sloan. 2023. *Using Linguistic Features to Improve Prosody for Text-to-Speech*. Columbia University.

Rose Sloan, Adaeze Adigwe, Sahana Mohandoss, and Julia Hirschberg. 2022. Incorporating prosodic events in text-to-speech synthesis. In *Proceedings of the Conference on Speech Pro sody 2022*, pages 287–291.

Paul Taylor and Alan W Black. 1998. Assigning phrase breaks from part-of-speech sequences. *Computer Speech & Language*, 12(2):99–117.

Nguyen Thi Thu Trang, Nguyen Hoang Ky, Albert Rilliard, and Christophe d'Alessandro. 2021. Prosodic boundary prediction model for vietnamese text-to-speech. In *Interspeech 2021*, pages 3885–3889. ISCA.

H.S Venkatagiri. 1999. Clinical measurement of rate of reading and discourse in young adults. *Journal of Fluency Disorders*, 24(3):209–226.

Michelle Q Wang. 1991. Predicting intonational phrasing from text. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 285–292.

Yu-Tsai Wang, Jordan R Green, Ignatius SB Nip, Ray D Kent, and Jane Finley Kent. 2010. Breath group analysis for reading and spontaneous speech in healthy adults. *Folia Phoniatrica et Logopaedica*, 62(6):297–302.

Tae-Jin Yoon. 2006. Predicting prosodic phrasing using linguistic features. In *Speech Prosody*.

Xiaonan Zhang, Jun Xu, and Lianhong Cai. 2006. Prosodic boundary prediction based on maximum entropy model with error-driven modification. In *International Symposium on Chinese Spoken Language Processing*, pages 149–160. Springer.