Graphical Abstract

Predicting Children's Perceived Reading Proficiency with Prosody Modeling[⋆]

Kamini Sabu, Preeti Rao

Highlights

Predicting Children's Perceived Reading Proficiency with Prosody Modeling

Kamini Sabu, Preeti Rao

- Data set created of oral reading of connected text passages screened for minimum word-decoding skill and rated for comprehensibility.
- Our trained model predictions of human rater scores demonstrate the complementarity of accuracy, rate and prosodic features.
- A novel implementation is presented for prosodic miscue detection that incorporates trained boundary and prominence detection models along with a new 3-level reference information structure that takes reading speed into account.

Predicting Children's Perceived Reading Proficiency with Prosody Modeling

Kamini Sabu[a,1,*], Preeti Rao[a,**]

[a]*Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, 400076 India.*

Abstract

Reading is a foundational skill and the focus of school-level education efforts across countries. The assessment of linguistic competence from oral reading has long been the subject of scientific studies linking the reader's comprehension of the text to various measures of oral reading fluency. Given the time and resource intensive nature of such assessment, it is of interest to automate the prediction of reading fluency from audio recordings using the same pedagogical rubrics. Given recent findings about the importance of prosody to the communicative purpose of reading aloud, we discuss new approaches to modeling it reliably for the automatic assessment task. We present a new data set of children's oral reading screened for minimum word decoding skill and rated for comprehensibility by two human experts. We develop a system for the automatic prediction of rater scores that also facilitates insights about the complementarity and inter-dependence of computed lexical accuracy, rate and prosodic features as corroborated by multiple performance measures. With achieved values of correlation and agreement that surpass the corresponding inter-rater measures, we also show how text-dependent prosodic features, informed by speech rate and speaking style, contribute prominently to system performance.

Keywords: comprehensibility, L2 prosody, non-native children's speech, literacy assessment

1. Introduction

Reading is a foundational skill and therefore the focus of school-level education efforts across countries. Learning to read is the path to reading to learn. According to Torgesen (1998), 75% of the children who were poor readers in the third grade remained poor readers in the ninth grade, the disability persisting into adulthood. In countries such as India, low literacy levels well beyond primary school years are the principal reason underlying large school dropout numbers. Being the target of large-scale education interventions by government and not-for-profit organizations, the reliable monitoring of progress is a crucial requirement for the effectiveness of the programs. Alleviating the need for trained human resources, automatic assessment of reading proficiency from audio recordings can play an important role in the ecosystem by providing for efficient evaluation protocols and potential for insights based on realistic data samples (Rao et al., 2016). The recent pandemic-triggered shift to online schooling has further accentuated the need for digital assessment tools while also enabling more widespread connectivity than ever before.

The ultimate goal of reading is to derive meaning from the text. This ability has been found to be correlated with efficient decoding of the written word freeing up the reader's mental capacity for the comprehension processing of the text. That is, the capacity to read passages and answer questions about those passages has been found to be strongly linked to oral reading fluency (Fuchs et al., 2001). Thus comprehension can, in principle, be assessed by listening to a student reading aloud. Apart from accuracy and speed, an important attribute of oral reading fluency is the prosody or expression (Zutell and Rasinski, 1991). In the developmental trajectory, such expressive reading is typically attained after accuracy and speed have been achieved. The NAEP oral reading fluency scale specially attends to the realization

⋆This document is the result of the research project funded by the Tata Centre for Technology and Design at IIT Bombay.

*Corresponding author

**Principal corresponding author

Email addresses: kaminisabu@ee.iitb.ac.in (Kamini Sabu), prao@ee.iitb.ac.in (Preeti Rao)

[1]Supported by Visvesvaraya PhD scheme by Ministry of Electronics and Information Technology, Government of India

of expressive reading in terms of proper phrasing (grouping of words in line with the text syntax) and emphasis (conveying the meaning) (White et al., 2021). In spite of such well-defined rubrics for trained raters, expression in oral reading has been difficult to evaluate, partly due to the high variability of even correct realizations (Fuchs et al., 2001). As a consequence, oral reading fluency continues to be widely measured solely in terms of WCPM (words read correctly per minute). This metric is easy to manage and captures both the word decoding accuracy and reading pace.

Recent scientific studies on oral reading indicate however that WCPM is insufficient to assess a child's language proficiency. Despite achieving age-appropriate decoding efficiency, children can still exhibit strongly varying comprehension levels (Benjamin and Schwanenflugel, 2010; Groen et al., 2018). There are several studies that propose that a child's *understanding* of the text is usually conveyed through the reading prosody (Schwanenflugel et al., 2004; Miller and Schwanenflugel, 2008; Breen et al., 2016). Indeed, accurate and automatic reading is the foundation for prosodic reading and should be considered an essential criterion for reading assessment. However, inferential comprehension is more linked to the reading prosody, while literal comprehension is more linked to the reading automaticity (Keskin et al., 2019). Readers who read aloud with appropriate prosody tend to have higher scores on silent reading comprehension assessments as well (Breen et al., 2016). In a recent study aimed to disentangle the contribution of decoding from that of prosody skills, Groen et al. (2018) found that the production of appropriate prosody while reading a story, as assessed in terms of speech rhythm and word boundaries, and stress placement as related to hyphenated words, discriminated the good comprehenders from poor.

Some of the above works use acoustic measures for prosody like pause structure and pitch variation (Miller and Schwanenflugel, 2008) while others use rating scales such as the Multi-Dimensional Fluency Scale (MDFS, (Zutell and Rasinski, 1991)). In most studies, a standardized reading comprehension test is employed while prosodic fluency is measured with the four MDF (multi-dimensional fluency) parameters, viz. pace (automaticity), smoothness (accuracy), phrasing, and expression (Zutell and Rasinski, 1991). Based on the understanding that decoding proficiency underlies prosodic reading, Paige et al. (2017) examined the individual contribution of these four MDF parameters in early grade readers to find that the word identification accuracy and expressive phrasing accounted for 64.9% of unique variance in comprehension. In a comprehensive review and meta-analysis of two decades of research on the relation between prosodic reading and reading comprehension, Wolters et al. (2021) found all works reporting a positive correlation but differing in its magnitude, which they claim depended to a large extent on how prosody was measured. They find that rating scales provide the best predictions and conclude that it is the multiple aspects that they take into account (rather than individual features such as pause structure or similarity to adult F0 contour).

Closely paralleling the research in oral reading is that on the evaluation of speaking proficiency of adult L2 learners. Here, achieving comprehensibility, or being easily understood by native speakers, is the critical goal. To comprehend non-native speech, native speakers are known to employ segmental, suprasegmental, and fluency (temporal) cues while the perception of 'accentedness' depends mainly on segmental accuracy. Saito et al. (2015) investigated how the linguistic correlates of comprehensibility in L2 spontaneous speech varied according to the speakers' L2 skill to find that good prosody played a specially important role at intermediate to advanced levels, together with lexico-grammatical aspects. A more recent study used manipulated and resynthesized speech to demonstrate the significant influence of varying intonation and speech rate on comprehensibility ratings (van Maastricht et al., 2017). We can conclude therefore that comprehension of oral reading of a fixed text is judged by attributes common to comprehensibility of spontaneous speech with the lexico-grammatical aspects eliminated. Speech rate, segmental accuracy and prosody are important for both.

In the present work, we are interested in the automatic prediction of a child's comprehension of read-aloud text from computed acoustic features that relate to accuracy, rate and expression. We are particularly interested in the development of effective components for prosody evaluation in the context of automatic assessment of L2 English learners. To train and evaluate our system, we use a new data set of oral reading recordings of stories by Indian school students that has been rated by human experts for comprehensibility. Short passages, as reading material, are better suited to the study of the association between prosody and comprehension, rather than isolated sentences (Benjamin and Schwanenflugel, 2010). In the next section, we discuss the relevant past work on automatic assessment and motivate our own approach to the task. Our data set is presented next, followed by a description of our system and its experimental evaluation.

2. Related work

One of the earliest systems to automate feedback on oral reading proficiency was project LISTEN by Mostow and Aist (1997) who proposed WCPM, computed using ASR, as a measure of reading comprehension. Later, Mostow and Duong (2009) addressed the prediction of expressiveness and fluency on the MDF scale by comparing computed word-level prosodic contours of mean pitch, mean intensity, duration and latency with adult reference contours for the same text. For text-independent use, Duong et al. (2011) and Sitaram and Mostow (2012) developed phoneme-level pitch, duration and intensity models by training on a variety of sentences spoken by a set of adult speakers.

FLORA, introduced by Bolaños et al. (2013a), was a web-based system that estimated an individual student's oral reading fluency. Besides the measurement of WCPM, disfluencies and variation in reading rate, FLORA had expressiveness evaluation linked to the NAEP scale (Bolaños et al., 2013b). Pause based features were used to check the pause consistency with sentence punctuation. They also considered the difference between average pitch and duration across the expected primary stress bearing syllables in the words and the other syllables. They obtained scores from each of two raters for three 20 sec chunks of a 1-min recorded utterance and combined these to get a global score for the utterance. Three linear SVM classifiers are trained in hierarchical fashion to predict 2-level and 4-level ratings. The binary classifier gives 90% accuracy, with 76% accuracy achieved for the 4-category task.

In a more recent work, Proença et al. (2017a) predict the 6-level scores by teachers for overall reading ability of 150 children across grades 1 to 4 in reading tasks including sentences and pseudo-words. Features related to reading speed, silence and disfluencies are considered, with the number of characters correctly read per minute, the rate of disfluencies and syllables per minute turning out to be most correlated with the subjective ratings. A later work (Proença et al., 2017b) found that WCPM is already a very good predictor for isolated sentence reading indicating that decoding abilities are the main indicators of performance in such contexts. Cheng (2018) predict three scores for school reading evaluation: WCPM, expression and passage retelling comprehension. All the features are derived from ASR and are related chiefly to phone likelihood, reading rate and silences. Using a neural network model, they achieve their best correlation of 0.856 with 6-level human expression scores, better than their inter-rater average of 0.740. Recognising the challenge of assessing expressivity of oral reading in the absence of any "gold standard", very recent work by Bailly et al. (2022) exploits multiple references performed by adults. They project features derived from F0 and syllabic stretching values in two dimensional space, where distances are measured between the model adult speech and the child's test utterance. These are combined with accuracy and rate features to predict the student's reading fluency.

There has also been considerable research on exploiting prosodic features in predicting human scores for delivery proficiency of spontaneous speech. Liscombe (2007) showed that distributional measurements of prosodic events such as syllabic stress and phrase boundary tones correlated well with speaking proficiency of English language test takers. The use of automatically derived prosodic events somewhat weakened the correlation. Education Testing Service (Chen et al., 2018a) has been a major contributor to automated speech scoring research via their application SpeechRater aimed at non-native speech from the TOEFL exams. They use different features related to structural events (clause boundaries and disfluencies), pronunciation, rhythm, intonation, stress, vocabulary, grammatical complexity, content, and discourse coherence. In an earlier application to non-native read speech scoring, Zechner et al. (2011) used features based on the positional match of automatically detected stress and tone with a human gold standard, i.e. the same text as read by a native speaker. The stress and tone detection was based on models trained on spontaneous speech labeled by experts using listening and visual signal/spectral representations. They found an average correlation with human proficiency scores of 0.42 with stress labels and 0.31 with tone labels. The same group used deep learning to obtain high-level prosodic features from frame-level acoustic parameter (F0, loudness, etc.) contours that were then combined with aggregated features representing grammar, pronunciation, etc. with one of linear regression or MLP (Yu et al., 2015). They rated overall speaking proficiency on 4-point scale to get a correlation of 0.727 with expert ratings using linear regression. In another work, Chen et al. (2018b) used CNN, BLSTM, and attention models to obtain utterance-level acoustic and lexical encodings from four word-level acoustic features and 300-dimensional GloVe word-embedding, respectively. In the interest of the model interpretability, critical to educational assessments, these are combined through a linear regression layer. The system, with a Pearson correlation of 0.602, outperformed their traditional model with hand-crafted features. Very recently in the context of children reading Korean, Fontan et al. (2022) compared multiple linear regression, support vector regression and random forest regression based systems with different sets of low-level and high-level speech rate and pause features derived

from segmentation and clustering for pseudo-syllables and silence breaks. Their results indicated that for the given features, there was no significant difference in performance of the different regressors in the prediction of sentence level ratings.

While the presentation above is limited to a few representative works, our extended review indicates that while prosody has been clearly acknowledged as pedagogically critical, it remains less researched for automatic fluency assessment, especially for attributes other than pause structure and duration. Oral reading assessment systems that have incorporated expression scoring, via attributes such as pitch contours, have based it on reference models derived from adult speakers of the same text. Given that the reading studies reviewed earlier have shown that reading comprehension relates best with subjective prosody rating scales, we follow the approach of modeling the rating scales in our choice of features for automatic assessment. We seek particularly to improve the effectiveness of text-informed prosodic miscue features by exploiting our recent work (Sabu and Rao, 2018, 2021) on speaker-independent prosodic event detection. Combining this with lexical accuracy and rate features in a random forest classifier, we develop and evaluate a system to predict the comprehensibility of read-aloud passages as scored by human experts. Further, given that that middle and senior grades, having attained a minimum level of word decoding skill, show the best relations to the prosody rating scales (Wolters et al., 2021), we consider speakers for our dataset who have been screened based on WCPM. This enables us to examine more clearly the effectiveness of incorporating prosodic features in automatic reading fluency assessment and also to study the interaction of the different attributes.

3. Dataset and Annotation

With our target application of reading assessment of children learning L2 English, we describe our dataset in terms of the choice of reading material, audio recording process and speaker selection. Our target group comprises students of Grades 5 to 8 (ages 10-14 years) studying English as a second language in schools in the state of Maharashtra in India. The students are native Marathi speakers with diverse socio-economic backgrounds. We also discuss our subjective rating scale and the human ratings process, followed by a discussion of the observed dataset characteristics.

3.1. Reading Material and Text Annotation

The reading material comprises 85 distinct short English stories chosen by the authors, and vetted by a school English teacher, for grade appropriateness. In terms of standards, the levels range from A2 (beginner) to B2 (intermediate) levels on the CEFR (Common European Framework of Reference) scale (Council of Europe, 2019). Some considerations in the selection of text included avoiding meaning ambiguities and any scope for dramatization. Most stories have a couple of sentences within direct quotes, indicating conversation.

To create the reference for the evaluation of the text-specific prosody, we require the story text to be marked up with the basic information structure in terms of the phrase boundaries and prominences (i.e. words expected to be emphasized). The prosodic events are expected to be based on the syntactic and semantic structure of the text, including the givenness and importance of the information being conveyed. Critical to intelligibility, transmitting this structure is one of the main functions of prosody in the context of oral reading. This can be understood better from the information structure marking rules as summarized by Levis and Silpachai (2017). The rules include marking the final word in each grammatical clause and phrase as a boundary candidate and the rightmost content word in the phrase as the default prominent word. If the rightmost word is the 'given' information through the earlier context, then the prominence shifts to the previous qualified word. Prominent words often introduce new or important information to the discourse, or they may bear contrastive focus. However, we expect reading speed to influence the subset of potential prosodic events that a speaker will realize with, for example, more minor boundaries skipped at higher speeds while sentence endings are realised at all speeds. Similarly, for prominence, the words conveying new and important information are typically expected to be emphasized. Other content words may or may not be stressed but prominence on function words is usually prohibited.

With our interest in deriving a reference information structure that takes reading speed into account (just as a human listener might do), we designate each word in the text with a tag for each of boundary and prominence in terms of the following three descriptors: mandatory, optional, and forbidden. A 'mandatory' event is necessarily expected to be realised on the word while an 'optional' event may or may not be realised. A 'forbidden' event is strictly not to be realised on the word in question irrespective of the reading speed. An automatic tag classifier is developed using

supervised training with a fully labeled dataset of our 85 story texts, as described next. The ground-truth labels for training were obtained with the help of a fluent Indian English adult narrator who recorded each story at two paces, a slow and a fast pace, both within acceptable limits for reading speed. The slow readings are around 125 to 165 words per minute, while the fast readings are in the range of 155 to 200 words per minute. The narrator utterances were manually marked by one of the authors based on listening for boundaries and emphases. Table 1 shows the perceived prosodic events for an example of the narrator reading a passage at slow and fast paces. We observe that the fast pace shows far fewer phrase boundaries. All the occurring boundaries are also found in the slow paced version, as expected. In the case of prominent words, the frequency of occurrence is more similar across the slow and fast paces while the actual words chosen differ. This may be attributed to cue trading between boundary and prominence at the higher speaking rates (Barhate et al., 2016). Table 2 shows the distribution of the prosodic event tags across the words in all the 85 story texts.

Finally, words associated with the same prosodic event at both reading paces are marked mandatory, and those words not marked at either pace as forbidden. The remaining words are marked optional. We use these tags based on narrator's slow and fast recordings to train a boundary and prominence tag classifier. The classifier is based on fine-tuning a pre-trained NLP language model - BERT, an approach used previously for text prominence marking by Talman et al. (2019). We obtained a word-level accuracy of 94.9+/-3.6% and 83.7+/-12.3% in detecting the 3-class phrase boundary and prominence tags respectively in the leave-one-story out cross validation mode over the 10959 words in 85 stories. We then use this automatically derived reference information structure in the comprehensibility prediction system in order to validate our automated approach to reference structure generation.

Table 1: An example showing perceived information structure annotation for a narrator recording at two different reading paces. The phrase break positions are indicated with a pipe | symbol, while the prominent words are shown in bold.

Slow Pace	Fast Pace
one **day** | **three** tortoises | **went** for a **picnic** | near a **pond** | they **put** down the **basket** of **food** they had **brought** | at the **shore** | and went **inside** the pond | to **enjoy** a **bath** | **when** they came **back** | they **found** a **fox** | **eating** their food | **on** being asked | the fox replied | **anything** lying on the **ground** | is **public property** | **so** | the basket and the food were **also public** property | and **hence** | **i ate** the food |	**one** day | **three** tortoises went for a **picnic** near a **pond** | they **put** down the **basket** of food they had **brought** at the **shore** | and went **inside** the pond to **enjoy** a **bath** | when they came **back** | they found a **fox eating** their food | on being **asked** | the fox replied | **anything lying** on the **ground** is **public property** | so the **basket** and the food were also **public** property | and **hence i** ate the **food** |

Table 2: Distribution of reference information structure tags across 10959 words from the 85 stories used in the work.

Boundary			Prominence		
Event expectation	No. of words	% of words	Event expectation	No. of words	% of words
Forbidden	8478	77.4	Forbidden	6284	57.3
Optional	673	6.1	Optional	1749	16.0
Mandatory	1808	16.5	Mandatory	2926	26.7

Each story text is split into shorter passages (45 to 70 words each) to define the recorded chunks used for the human ratings of comprehensibility. This length of text corresponds to approximately 20 s of reading at normal reading pace and has been found to be suitable for holistic judgements of reading fluency by teachers (Bolaños et al., 2011). With our stories averaging to 2 to 3 such chunks each that typically correspond to distinct paragraphs of the story, we obtain 202 unique passages of text to be used as prompts in our data collection.

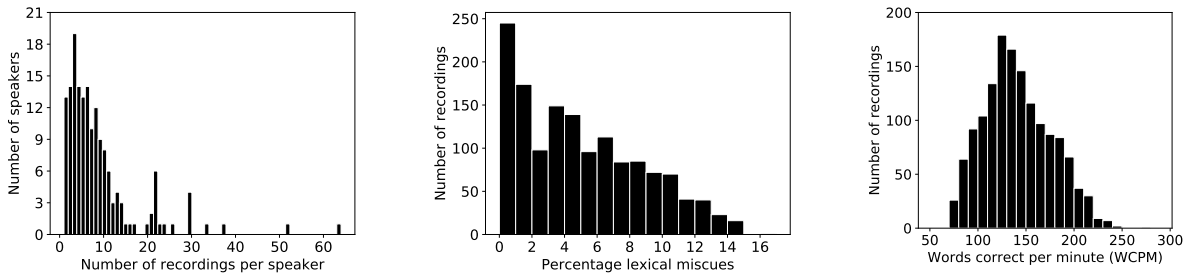
3.2. Data Acquisition and Screening

We conducted oral reading recording sessions across seven schools in two cities. Given the widespread interest in English language proficiency, the schools enthusiastically volunteered to set up the oral reading sessions during school hours. The recording is carried out via an app on an Android tablet that presents the selected passage on the

screen for the child to read aloud and record. Each child reads at least one complete story in a single session but with a single chunk of text presented on one screen. The child goes over the text silently before recording. Each passage is recorded in one take at 16 kHz sampling frequency in 16-bit PCM format. To the extent possible, the recording sessions are conducted in relatively quiet spaces such as the school library. Further, we employ headset microphones to ensure SNRs (Signal to Noise Ratio) over 20 dB (Sabu and Rao, 2021).

A pre-test process of getting the child to read a couple of sentences of grade-appropriate text was used to gauge the approximate level of the child's proficiency. Next a story of the estimated level was randomly selected from the set of stories and assigned for recording by the child. Further stories, if any, assigned to the same child were different samples drawn from the same or next higher level. The additional assignments were based on the child's perceived interest in continuing the reading exercise. We obtain a total of 2855 recordings across 330 distinct speakers and the entire set is manually transcribed at word level. The transcripts are screened for a minimum level of word decoding accuracy and rate in order to create a data set for use in the present study where one of the chief aims is to evaluate the contribution of prosodic features to comprehensibility. This is achieved with the counting of lexical miscues (i.e. word substitutions, omissions and insertions with reference to the passage text). We discard recordings with a 'lexical miscue rate' beyond 15% (or one in six words). We also compute the WCPM for each utterance based on the duration from the first word uttered to the last. WCPM values for grades 4-6 students are expected to be in the range 70 to 160 for good readers (Learning A-Z LAZEL Inc., 2020). Other reading studies such as (Hasbrouck and Tindal, 2006) and DRA scale (Pearson School, 2019) indicate that poor readers display WCPM range below 70, which also corresponds to the lowest category in the dataset of Proença et al. (2017b). Hence we use 70 WCPM as a lower bound for our dataset.

Applying the above two criteria (which, from our observations, almost completely overlap), we have 1447 recordings by 165 students reading from a pool of 148 unique passages. Based on the precise passage length and reading speed, the recording duration varies from 12 s to 56 s (mean = 25 s, standard deviation = 8 s). The total dataset duration is about 10 hrs. With no passage repeated within speaker, Figure 1a shows the distribution of the number of recordings per speaker. We see that while the majority of speakers contribute less than 10 recordings each, up to 20 speakers have read more than 20 distinct passages each. The recordings have relatively few miscues as apparent from the lexical miscues distribution in Figure 1b. Figure 1c also shows the distribution of WCPM. We note that the short speaking duration from the limited passage length promotes greater speeds and skews the WCPM to the right compared to that observed in studies using the traditional 1 minute reading task (White et al., 2021).



(a) (b) (c)

Figure 1: Distribution of (a) number of recordings per speaker across number of speakers; (b) percentage lexical miscues and (c) WCPM across 1447 recordings

3.3. Comprehensibility Rating

Subjective evaluations are typically plagued by low inter-rater agreements due to the differences in perception, understanding, experience, and leniency of each rater. Isaacs and Trofimovich (2012) studied the perceived comprehensibility ratings from three expert teachers to get pair-wise correlations of 0.81, 0.62 and 0.66. After-rating discussions revealed differences between raters in the basic interpretation of the term 'comprehensibility'. The first

two raters assumed it to be the listener's understanding of the 'message', while the third rater considered it to be the understanding of each 'individual word'. A common understanding of the rubrics across the raters is therefore important, and achievable through training sessions, discussions on the rating scales and the associated criteria.

Rating scale design is another important factor, with low resolution and ceiling and floor effects seriously limiting the representation of the diversity in the data (Hayes and Hatch, 1999). A larger scale for rating is preferred, which can later be collapsed to fewer levels if needed. Further, Bone et al. (2015) insist on sample ratings before the actual task where the raters can go through the full range of available levels to help them anchor their ratings. Discussions following the sample ratings can be very helpful for better inter-rater agreement.

In a task of Oral English Proficiency Test scoring, Yang (2010) appointed language teachers with experience in teaching English as a second language. They found that even after conducting sessions and training raters for using the same scale, final ratings varied due to the different degrees of severity or leniency. The variation depended on the familiarity with the speaker's accent, the priority of grammar over content, etc. Rater training helps in improved self-consistency of individual raters but cannot eliminate the differences in rater severity or leniency (Weigle, 1994). So for healthy ratings, the raters should not be compelled to agree but instead their scores can be used to obtain insights about the aspects important to them individually (Yang, 2010; Bone et al., 2015). In order to solve the problem of different rater strictness, Black (2012) and Bone et al. (2015) z-normalized autism ratings from each rater in their study. They found that this improved the inter-rater agreement, nullifying the individual rater's bias.

For the target predictions, we can use the scores from each individual rater separately. Instead, Bone et al. (2015) average the normalized scores from different raters for every utterance and treat these as the target labels to predict. Another approach for the target gold standard is to weigh the ratings from different raters with the corresponding mean pair-wise inter-rater correlation values (Ringeval et al., 2017). Isaacs and Trofimovich (2012) found that the average ratings from 60 naive raters are well-correlated (>0.8) with the individual expert raters. They also found that expert raters are able to articulate their criteria properly and easily, while novice raters don't specify all the possible features. The scoring criteria of expert raters also span more aspects than those of novice raters.

Informed by the above discussion, we chose raters who were experienced English language teachers in Indian schools and took similar steps to obtain usable targets for our automatic predictions. The teachers also happen to have undergone a certificate course in ELT offered by the British Council in India. We had an initial discussion with potential raters in order to decide the nature of the scale and the interpretation of the levels. Our single attribute to be judged by the listener was comprehensibility, serving as a holistic measure of reading ability in contrast to using multiple individual dimensions that have been found to result in low inter-rater correlations (Mostow and Duong, 2009). We arrived at a 6-point rating scale, closely related to the typical stages in reading acquisition and similar to the standard NAEP and MDF scales. It is simpler and permits more subjectivity compared to NAEP scale where the levels are prescribed based on the precise fraction of utterance read correctly. Our scale levels and interpretation appear in Table 3. As reviewed in Section 1, beginning readers typically first achieve word decoding ability, followed sequentially by smoothness, pace consistency, proper phrasing and the use of emphasis. Since we have screened our dataset for recordings with low %miscues, the rating scale practically starts with phrasing deficiencies moving further towards the proper use of emphasis, all of which are important for high comprehensibility. An online rating interface was used for the randomized presentation of passage recordings with a play button, the text expected to be read, and a pull-down menu for the comprehensibility scoring as shown in Figure 2.

Table 3: Comprehensibility rating scale interpretations

Rating level	Interpretation
0	Random grouping, pauses in wrong places, incomprehensible to the listener
1	Poor grouping by and large with glimpse of good grouping in one or two places. The underlying reason could be word difficulty.
2	Better grouping but still not perfect in all places.
3	Grouping is good but stresses are wrong.
4	Basically a good reader, some local deviations that may be attributed to first-time reading of the given text. This level can be treated as the benchmark.
5	Exceptional (could be a practiced child who does speech/elocution training).

In order to make raters aware of the data variability and help them anchor their scales, we asked them to first go through 100 sample recordings drawn across different students. This was followed by a round of discussion around the scale interpretations. Following an approach from Isaacs and Trofimovich (2012), we asked raters to introspect on the different aspects that might influence their ratings. We provided check-boxes for items that emerged during our discussions and also invited free style comments as seen in Figure 2. While there is no limit on the number of

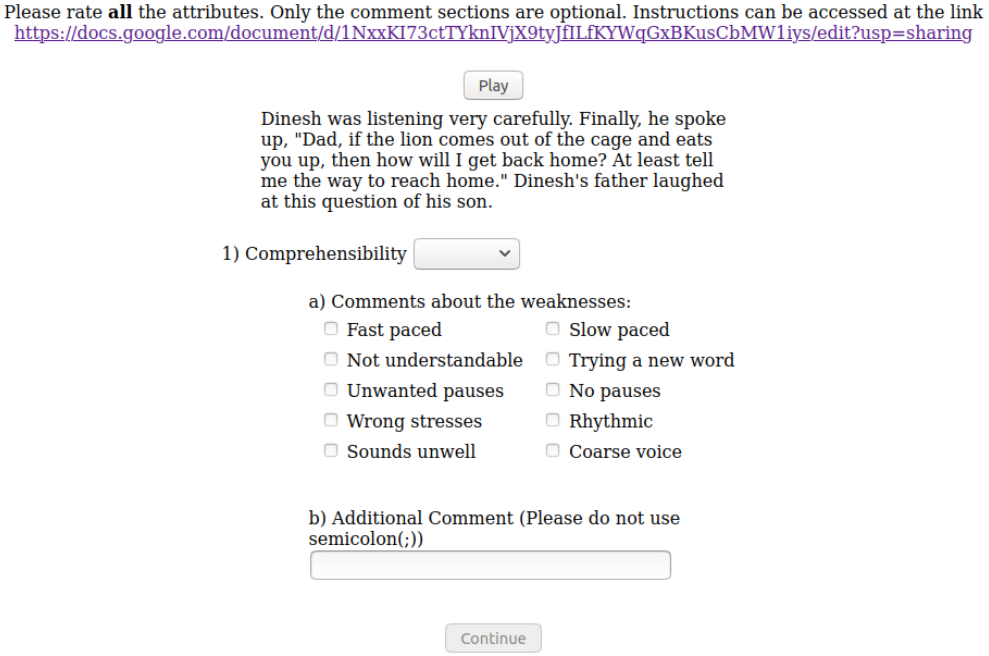
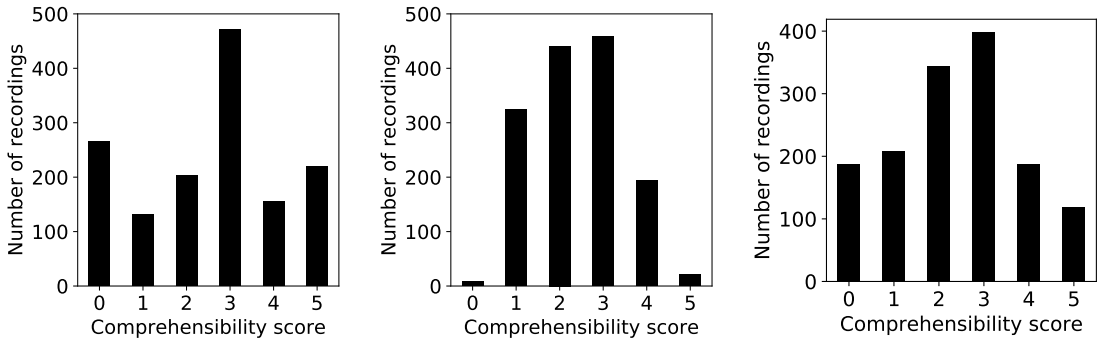


Figure 2: GUI for comprehensibility rating

times a rater can listen to the recording, we find that the raters listen to each recording typically once or twice. On an average, comprehensibility scoring for a recording takes about 1 minute. In order to avoid the raters' scale shift due to continuous exposure to poor or good student's recordings, we advise the raters to limit their sessions to not more than 1 hour and ensure a sufficient gap between two consecutive sessions. Further, they are asked to revisit sample recordings if they feel such a shift taking place. The inter-rater correlation, as measured by the Pearson correlation coefficient computed across the 1447 recordings, was found to be 0.61. We then went through the ratings, and wherever the difference in ratings from the two raters was more than 1 point, we asked the raters to do a round of re-rating without any further instructions. The setting was identical to that of first-time rating with raters having no access to their initial ratings. In many cases, but not all, the difference between the raters reduced to attain the eventual inter-rater Pearson correlation value of 0.76 (and Spearman's correlation of 0.77).

3.4. Dataset Characteristics

The final distribution of scores for each of the raters is shown in Figure 3. We note that the recordings are relatively evenly spread across low, medium, and high scores for rater 1. Rater 2, on the other hand, prefers the middle scores of 2 and 3. Table 4 presents the final confusion matrix accompanied with a visualization of the same in Figure 4. We see that the confusions are mainly around the diagonal but for the extreme ends of the scale. The subsequent discussion with the raters indicated that the differences in rater perception were due to the different extents of familiarity with the specific L1 accents as also differences in leniency, as mentioned by Yang (2010). Other sources of disagreement stemmed from individual preferences related to the reading style. While one of the raters found over-emphasized or enacted speech a sign of enthusiasm, the other found it jarring and not easily comprehended. For speakers perceived to be bored, the scores by one rater, who attributed this to low motivation rather than low comprehension, were quite good, while the other rater indicated that the comprehensibility was drastically affected. Another phenomenon



(a) Rater 1 (b) Rater 2 (c) Combined

Figure 3: Distribution of comprehensibility scores by the two expert raters and the combined target ratings across 1447 recordings

flagged by both raters, but to different extents, was a certain rhythmic or sing-song style of reading the text with periodic accents unrelated to text content. In some cases, the reading speed affected the perception of the two raters differently. Although both the raters agreed that very slow and very fast reading pace hampers comprehensibility, the choice of the optimum speed differed across the two.

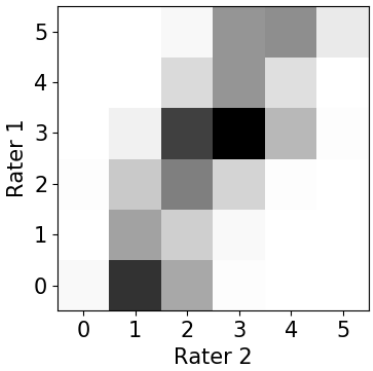


Figure 4: Distribution of co-occurring rater scores across the dataset.

As mentioned earlier, the final inter-rater agreement across the two raters is 0.757, surpassing the correlation of 0.59 obtained by Tao et al. (2016) in a similar task and 0.70 recommended by Williamson et al. (2012). This is similar to the average inter-rater correlation of 0.740 obtained by Cheng (2018) for 5-point oral reading expression ratings by 3 raters. We find that the correlation drops to 0.670 if the ratings are grouped into three categories ('poor': 0, 1, 2; 'fair': 3; and 'good': 4, 5). Such reduction in correlation points to the high confusions near the group boundaries. Other measures of inter-rater agreement are computed in Table 5. The table also indicates the inter-rater agreement for 2-class ratings (0, 1, 2 vs 3, 4, 5).

We consider the average rating from the two raters as the final target score. Since the scores by Rater 1 are more spread out and those from Rater 2 are crowded in the middle, we z-score normalize the ratings before averaging, similar to Bone et al. (2015). The final target scores are 24-point continuous values ranging between -2 to 2. For the final target ratings to be interpretable on the same lines as the original 6-point scale, we also perform a reverse normalization similar to Proença et al. (2017b). The original discrete ratings from both raters are combined across the dataset utterances to get a mean value and the standard deviation (s.d.). The overall combined score for each utterance is then computed as follows where $R1$ and $R2$ are the two rater-level z-score normalized scores for the utterance.

$$Combined\ score = 0.5 * (R1 + R2) * s.d. + mean \tag{1}$$

Table 4: Confusion matrix of comprehensibility scores by two raters

Rater 1 \ Rater 2	0	1	2	3	4	5
0	6	0	2	0	0	0
1	181	82	48	13	0	0
2	77	43	113	168	33	7
3	2	6	38	225	94	94
4	0	0	2	63	29	100
5	0	0	0	2	0	19

Table 5: Inter-rater agreement for comprehensibility scores by the two raters across 1447 recordings

Agreement type	Number of Classes		
	6	3 [0,1,2 ‖ 3 ‖ 4,5]	2 [0,1,2 ‖ 3,4,5]
Exact agreement (%)	32.8	63.9	81.4
Adjacent agreement (%)	83.5	97.1	NA
Cohen's kappa	0.176	0.433	0.632

The combined score (also called the target rating) for an utterance is a continuous value largely within (0, 5) but can have values slightly lower than 0 or greater than 5 due to the nature of the calculation. In Figure 3c we show the distribution of the combined score after rounding to the discrete levels in 0-5 in order to compare it more easily with the individual rater distributions. We note that the combination results in a more balanced distribution across the scale levels.

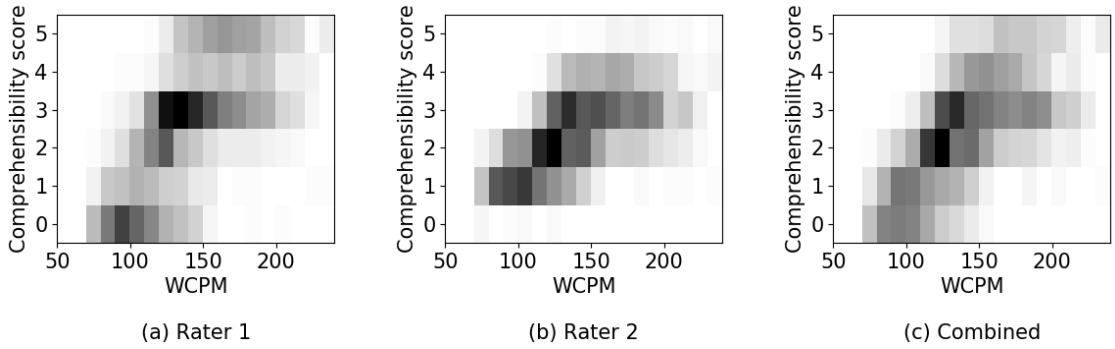
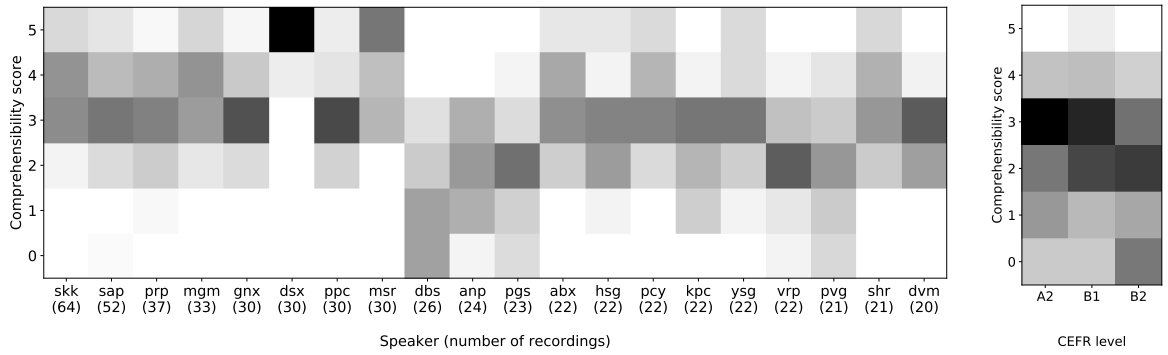


Figure 5: Distribution of comprehensibility scores across WCPM metric for each rater and for the combined rating.

Plotting the distribution of scores with WCPM in Figure 5, we see that the higher comprehensibility is associated with WCPM > 100 words read correctly per minute, and both raters tend to improve their scores with increasing WCPM confirming that correct word decoding and adequate reading speed are essential predictors of comprehensibility. However, for any chosen WCPM, we see a significant dispersion in the comprehensibility scores (over a range of 3 to 4 levels), suggesting that prosodic features are expected to play a critical role in explaining the rater scores.

To describe the gross characteristics of our dataset, we show the distribution of scores across speakers and across passages. As seen in Figure 6a, although there are 20 speakers with greater than 20 recordings, they all (but one, who was relatively advanced) have scores distributed across 3 or more levels for either rater due to the variety of passages they read, with text difficulty playing a role in reading prosody for an individual child (Benjamin and Schwanenflugel, 2010). Further, for any given passage, we have a good distribution of ratings across levels since each passage is read by an average of 10 distinct speakers who seem to be spread out in the reading development.This is shown across 3 groups of passages where the grouping is based on CEFR passage difficulty as shown in Figure 6b.



(a) (b)

Figure 6: Distribution of target scores for (a) different highly recorded speakers, (b) for different story difficulty levels

4. Comprehensibility Prediction System

We wish to score the comprehensibility of a recorded passage as a human expert would do, using the scale discussed in Section 3.3. In the interest of obtaining interpretable scores, we take the approach of modeling the different pedagogical rubrics with the relevant acoustic features. Further using feature selection in a powerful classification framework can provide insights around the factors that underlie the human judgements. Features related to word decoding accuracy and rate are obtained with the automatic speech recognition of the utterance. With focus on the prosody functions of conveying the text syntax and semantics, we investigate computable features that capture pause structure, prosodic event realization and attributes related to reading styles that potentially influence raters' perceptions.

Figure 7 shows the different modules of the automatic scoring system. Apart from the recorded speech, the prescribed text (which we term the 'canonical text') is provided. The word-level information structure is computed from the latter as was discussed in Section 3.1. The audio signal is processed for, both, ASR input features and the acoustic parameter contours, sampled at 10 ms intervals, corresponding to F0, intensity and various spectral band energies across the utterance. The extraction of the acoustic parameters at 10 ms intervals follows the methods presented by Sabu and Rao (2021). Eventually, the high-level feature aggregates associated with lexical accuracy, speech rate and prosody, as computed from the ASR decoder hypothesis and the acoustic contours, are combined in the trained classifier to obtain the estimated comprehensibility score for the given utterance. We describe the feature extraction and score prediction methodology in this section.

4.1. ASR and Post-processing

The utterance recording is passed through an automatic speech recognition system that provides the decoded text hypothesis with the corresponding word- and phone-level alignments. The ASR has a 13-layer TDNN for the acoustic model (developed with the Kaldi toolkit by Povey et al. (2011) and trained on 80 hours of Indian English read speech by 598 adult speakers (Speech Processing Lab IIT Madras, 2020)). We employ data augmentation via VTLN and further adapt the models on another transcribed dataset previously created by us comprising of 30 hours of children's read speech with 425 speakers, not overlapping with the comprehensibility-rated dataset of the present study. In this larger dataset, we have a cross-section of speakers, including those of lower proficiency, reading diverse passages. The language model for our system is a trigram trained on the entire set of story texts. Common substitutions are entered in a zero-gram garbage model in parallel with the trigram. The garbage model thus contains common mispronunciations of the story words as well as word fragments representing false starts or hesitations. On the dataset in the current study, the ASR system has a measured word error rate (WER) of 5.55% with respect to the manual transcript.

It is necessary to obtain a word-level alignment between the ASR decoder output and the canonical text in order to determine the lexical miscues (i.e. words that have been substituted, deleted or inserted by the child) and also to detect

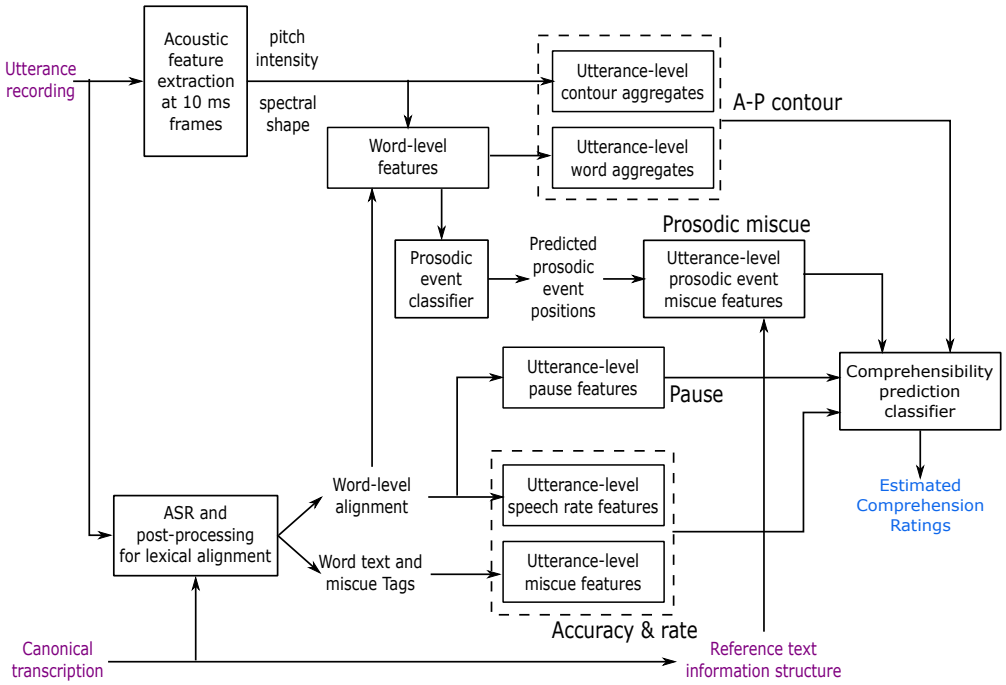


Figure 7: Proposed system block diagram

word-level prosody mismatches with the reference information structure. While the classical dynamic programming algorithm used in speech recognition evaluation simply minimizes the sum of substitutions, deletions and insertions, it often indicates substitutions that sound very different and therefore unlikely. We employ instead a phonological similarity based distance to identify the substitutions (Fisher and Fiscus, 2004). Further, it is observed that the ASR output frequently comprises of two or more short words corresponding to a single longer word in the manual transcript possibly due to the presence of slight disfluencies in uttering the word (e.g. water -> 'what are'; obesity -> 'a be city'). A post-processing stage identifies such splits based on phone sequence similarity and carries out the needed merges. The miscue type is then a single word substitution and the reference prosody assigned to the word is that of the corresponding canonical word (unless a pause>200 ms is detected within the decoded segment, in which case the most phonetically similar sub-segment is treated as the prosodic counterpart). We also obtain syllable-level alignments based on word-syllable-phone mappings.

Our task needs the accurate detection of the lexical miscue tag for each canonical word rather than the identity of the uttered words. This is achieved by the above mapping of the post-processed decoder output words to the canonical text words, and detecting lexical miscues, if any, in terms of word substitution, deletion and insertion. We evaluate the post-processed ASR output via the achieved precision and recall in the detection of the miscue tags with reference to those derived from the manual transcript, as presented in Table 6 for our dataset. The low recall for substitutions is found to arise from the mostly minor pronunciation variations noted by the transcriber but not detected by the ASR due to the limited lexicon. Such small pronunciation deviations are likely to be ignored by the human raters. Further, the slightly reduced precision in substitution detections comes from the language model where the same word occurs in multiple contexts in the story and sometimes across stories. Further, we also observe lower precision-recall values for deletions and insertions due to the dominance of the LM. These are mainly due to small words like the article 'a'. We discuss the impact of ASR error on the overall performance of the comprehensibility prediction system in a later section.

4.2. Accuracy & Rate Features

The previously detected miscues and time alignments facilitate the computation of the lexical accuracy & rate features for the utterance. We discuss next the precise features tested in this work, as also summarised in Table 7.

Table 6: ASR performance evaluation after post-processing

Miscue tag	Precision	Recall
substitution	0.63	0.23
deletion	0.82	0.74
1-word insertion	0.81	0.67
multi-word insertion	0.72	0.77
correct	0.97	0.99

4.2.1. Lexical Miscue Features

Lexical miscues are identified after aligning the ASR decoder output with the canonical text using the post-processing discussed in Section 4.1. Word decoding ability is measured by the number of words uttered correctly '#correct'. We compute the lexical miscue features '#miscues' and '%miscues' (total number of miscues and their percentage with respect to the total number of canonical text words respectively). Depending on the type of miscue, comprehensibility is affected in different ways. For example, self-corrections restrict the fluency but also signal comprehension. Therefore, we use the individual counts of different lexical miscues (substitution, omission and insertion) across the utterance as features denoted by #deletions, #insertions, #substitutions. Further, we also count the number of words which are preceded by at least one insertion (#hesitations) and the number of words which are preceded by more than one insertion (#mult_hesitations). These features may be indicative of the extent of hesitation in the course of uttering the word, related to greater difficulty in word decoding. Finally, the WCPM (number of words correctly read per minute) is computed by dividing the number of text words uttered correctly divided by the duration of the recording between the start of the first word and the end of the last word. The WCPM is a combined measure of word decoding accuracy and speaking rate.

4.2.2. Speech Rate Features

Reading speed is an important determiner of the comprehensibility of delivered speech. Students who read rapidly and smoothly are more likely to be rated highly comprehensible. Students who read at a very slow or swift pace are both difficult to comprehend (Munro and Derwing, 2001). There are a number of measures that represent speaking rate depending on the lexical unit chosen and whether or not inter-word pauses are considered in the measured duration. We compute the speech rate across the recording segment in terms of the number of syllables per second (SR), articulation rate (AR) and the number of phones per second (PR). Here the count of words or phones in the recording segment comes from the ASR decoder hypothesis, while the recording segment duration is the duration from the start of the first word to the end of the last word in the segment. Articulation rate is a variation of syllable rate, where the segment duration used refers to the speech duration excluding the non-speech regions. Apart from the above averages of the different speech rate measures, we estimate the speech rate variation (SRvar) across the utterance as the standard deviation of local syllable rate computed across 10-word segments with three-word shifts as a feature representing speech rate consistency.

4.3. Pause Features

The occurrences and durations of pauses are important for reading prosody. In normal reading, pauses are indicative of phrase boundaries. In the case of poor readers, pauses can be manifestations of hesitation or reading difficulty. More frequent pauses and a wider range of pause durations across the utterance are associated with low comprehension (Miller and Schwanenflugel, 2008). Pauses shorter than 250 ms are usually irrelevant when calculating measures of L2 fluency in the context of discourse (Jong and Bosker, 2013). In our case of oral reading, however, we found it useful to lower the threshold to 200 ms. The word alignments obtained from the ASR decoder are used to detect the inter-word gaps. The measured non-speech duration between two successive words is termed a 'silence' if it is longer than 200 ms and is further classified into pauses (< 500 ms) and long pauses (≥ 500 ms) (Liscombe, 2007).

The segment-level features computed in the pause category include the mean and standard deviation (s.d.) of silence durations (meanSil, sdSil), pause durations (meanPause, sdPause) and long pause durations (meanlongPause), and the percentage of the total silence duration (%Sil) in the recording excluding the extreme silences. Further, we

Table 7: List of features extracted across the five feature groups

Feature group (No. of features)	Utterance-level features
Accuracy (9)	#correct, #deletions, #insertions, #substitutions, #miscues, %miscues, #hesitations, #mult_hesitations, WCPM
Speech rate (4)	AR, SR, PR, SRvar
Pause (12)	%sil, Silpersec, Pausepersec, longPausepersec, Silperword, Pauseperword, longPauseperword, meanSil, sdSil, meanPause, sdPause, meanlongPause
Prosodic miscues (50)	%events (boundary and prominence) detected corresponding to optional and mandatory events; avgdist between consecutive event detections
Events-based	%TP, %TN, %FP, %FN, Acc, TPR, FPR, precision, and F-score for boundary and prominence w.r.t. optional and mandatory events, and mandatory alone
Consistency	unexpectedSilratio, expectedSilratio, mostexpectedSilratio stressedDurdiff, moststressedDurdiff, stressedF0diff, moststressedF0diff
A-P contour (69)	mean and s.d. of: raw (Hz), z-normalized and semitone F0, and raw and normalized intensity, 4 band-intensity contours;
Contour based aggregates	mean and s.d. of means of 1s chunks of normalized: F0, intensity; mean-mode difference of raw F0 (Hz); mean-to-s.d. ratio of raw (Hz) and normalized F0; F0changeVar1, F0changeVar2, F0_zcr; F0 periodicity features: mean and s.d. of F0_acfpeak and F0_acflag across 1s and 2s chunks and corresponding #aperiodicF0
Word-based aggregates	mean and s.d. of word-level mean, and s.d. of word-level s.d. of: raw F0 (Hz), intensity, spectral band intensity, and HNR contours; %positiveF0slope, %risingF0, %fallingF0, %peakyF0, %dippingF0, %positiveEnd, %negativeEnd, %flatEnd; #words with F0 (Hz) reset value more than mean F0 (Hz)

also compute the number of silences, pauses and long pause regions per unit of time (Silpersec, Pausepersec, longPausepersec) (Bolaños et al., 2013b) and also the same normalized by number of words (Silperword, Pauseperword, longPauseperword).

4.4. Prosodic Miscue Features

Given that the comprehensibility of the read aloud text is linked to the correct placement of phrase boundaries and word prominence, we define prosodic miscues in terms of word-level mismatches with the expected prosodic events. The latter are determined by the given text syntax and semantics. The acoustic realization of prosodic events is known to be speaker-dependent. While pauses are commonly used to cue phrase boundaries, pitch and/or energy resets as well as final word lengthening also contribute to the listener's perception of a boundary (Ericsson, 2003). Given the possibility for cue-trading, the precise choice of acoustic cues is speaker-dependent and further depends on the reading speed. The same holds for prominence with potentially even more variability. Prominence is defined as "the phenomenon of a linguistic unit (syllable, word, or even larger stretch of speech) being perceived as standing out of its environment" (Christodoulides and Avanzi, 2014) with the change occurring in one or more of the acoustic parameters of fundamental frequency, energy and segmental duration. We investigate two approaches here, (i) checking for consistency by observing directly the variations in specific attributes across unexpected and expected prosodic event locations, and (ii) detecting, and later validating, the realized event locations using available boundary and prominent word classifiers trained with multiple suprasegmental features on a perceptually labeled dataset of children's oral reading (Sabu and Rao, 2021). The second method is expected to be more robust to speaker variations in prosodic event realization. In all cases, we employ the 3-level reference information structure to better account for different reading speeds.

4.4.1. Prosodic Consistency Cues

Bolaños et al. (2013b) propose features examining the consistency of the pause structure with the expected event locations, i.e. the text punctuation marks. Based on the information structure labels from Section 3.1, we compute the features 'expectedSilratio' and 'unexpectedSilratio' as the ratio of the duration of silence made at expected locations (or at unexpected locations) with respect to the total duration of silence. Further, we add one more feature 'mostexpectedSilratio', which is the ratio of silence duration made at mandatory boundary locations only to the total duration of silence, to benefit from the 3-level information structure, this being similar to the within-sentence and sentence-final punctuation conditions of Bolaños et al. (2011). The silence durations are computed across the utterance as in Section 4.3 with non-speech regions shorter than 200 ms ignored.

Bolaños et al. (2013b) also evaluate reading fluency and expressiveness in terms of correct realization of lexical stress. They estimate the increase in F0 and duration in the expected stressed syllables with reference to the remaining syllables. Lexical stress is less important and more variable, both in placement and realization, across speakers of Indian English (Fuchs and Maxwell, 2015). We instead distinguish expected prominent and non-prominent word acoustics with the similar features. The expected prominent words are determined by the reference information strcture (Section 3.1) to get the features 'stressedDurdiff' and 'stressedF0diff'. The 'stressedDurdiff' is the difference between the duration of the estimated longest syllable in the word averaged across words expected to be emphasized and words expected not to be emphasized. Similarly, 'stressedF0diff' is the difference between the word-level mean F0 value averaged across words expected to be emphasized and words expected to be not emphasized. Again, we also extend this approach to utilize the 3-level information structure through the features 'moststressedDurdiff' and 'moststressedF0diff' where only the words that are marked to bear mandatory prominence (i.e. assigned level 2) are considered.

4.4.2. Prosodic Event-based Miscues

This set of prosodic features is obtained by explicitly detecting the realised prosodic events in the utterance and comparing these with the expected event locations to estimate prosodic 'miscues'. Using multiple word-level features computed from across the different supra-segmental parameters, each word in the utterance is evaluated for the degree of each of boundary and prominence via the prosodic event detection (PED) system of Sabu and Rao (2021). It is assigned a value in [0,7] representing the number of votes received from the 7 annotators marking perceived events for the training data. The continuous value is thresholded to detect event presence on a word. The performance of the models measured in terms of F score was 0.80 and 0.76 for boundary and prominence detection from the acoustics alone when the reference label corresponds to $\geq$2 votes out of 7, while $\geq$3 votes resulted in correspondingly lower F-scores of 0.79 and 0.63 (Sabu and Rao, 2021).

We have the reference information structure marked for every word as discussed in Section 3.1. We transfer this to the target (spoken) word based on the automatic alignment of the transcript with the text as discussed in Section 4.1. Inserted words are left unmarked. Further, the canonical information structure is compared with the detected prosodic event positions to get the prosodic miscue tag for each word. These include the four tags: correctly realized event (true positive, TP), correctly recognised non-event (true negative, TN), missed realizing an event (false negative, FN), and falsely realized an event (false positive, FP). For the comparison of the predicted event locations with the expected event locations, we use the thresholds discussed above to convert the predicted event degree to the binary event presence decision. The threshold for ($\geq$2 votes) is used when comparing with the mandatory and optional event positions, while the threshold for ($\geq$3 votes) case is used when comparing with the only-mandatory event positions. All the features are computed separately for the set of mandatory event positions and the combined set of event positions.

For the utterance-level features related to the prosodic miscues, we compute the percentage of realized event words of the total number of text words (%events) and the average distance (avgdist) between two consecutive event realisations in terms of the number of words. From the prosodic miscue tags obtained by comparing the detected and expected event positions, we compute the fraction of true positives (%TP), true negatives (%TN), false positives (%FP) and false negatives (%FN) by normalizing the corresponding counts with respect to the total number of words in the canonical text. We also compute the accuracy, true positive rate (TPR), false positive rate (FPR), precision, and F-score for each of the events in both these conditions.

4.5. Acoustic-Prosodic Contour Features

Fluent speech can be recognised by its prosody and the associated global variations of the suprasegmental parameters independent of the specific sequence of words uttered. As noted from the literature and from our data including the human rater comments, speakers who do not vary their pitch or loudness across the utterance are not easily comprehended. On the other hand, those with unnatural, undulating variations are unintelligible too. These observations motivate the study of features which can potentially capture both natural and other characteristic speaking styles such as monotonous, rhythmic, drawling, etc. by way of various functionals computed across contours of the frame-level acoustic-prosodic (A-P) parameters. Given the high speaker dependence of the parameter ranges, it is necessary to consider also features computed from speaker-normalized acoustic contours. As in (Sabu and Rao, 2021), we use utterance level (i.e. across the single speaker-passage recording) z-normalization to obtain the corresponding speaker-normalized (SN) contours.

Pitch and loudness are the common prosodic parameters influencing human perception, with fundamental frequency and sound intensity being the respective acoustic correlates. The percept of loudness is associated with both intensity and spectral shape. Variations in voice quality, especially brightness, as represented by relative intensities of various spectral band energies, and also by the harmonics-to-noise ratio, were found useful in the prediction of perceived confidence from recordings of oral reading (Sabu and Rao, 2020). We investigate the potential utility of the different acoustic parameter contours in the prediction of comprehensibility by extracting features as presented below. The features comprise two kinds of utterance-level aggregates: those computed directly from the 10 ms-sampled contours across the utterance and those obtained from intermediate word level features. The latter are computed from word-aligned segments of the contours with word boundaries obtained from the ASR post-processing step. The list of aggregate features, as presented in Table 7, is described next.

4.5.1. Contour Aggregated Features

The contour features are computed directly on the frame-level estimates of each A-P parameter. In order to determine the loudness variation across the recording segment, we compute the mean and standard deviation (s.d.) of the intensity across the non-silence regions of the utterance. Similarly, the mean and s.d. across the utterance are obtained for each of the relative spectral intensity and HNR, restricted to the speech regions. We also represent the corresponding dynamics via the s.d. across the sequence of mean and s.d. values computed over short (1 s duration with 500 ms hop) segments for the normalized intensity contour.

In order to represent the pitch dynamics across the utterance, we compute the mean and s.d. of the samples of the normalized pitch contour (raw values in Hz, as well as semitone values) across the voiced region of the segment. The ratio of mean and standard deviation of the pitch (in Hz) helps overcome the speaker-specific pitch range effect. The difference between the arithmetic mean and the most frequently occurring value as represented by the mode (computed over quantized values of the raw F0 in Hz) is used as an estimate of the variation. It was also found useful to summarise the variation in the local dynamics via the s.d. of the variances across short segments (1 s and 2 s durations) of the pitch contour taken every 500 ms. These are the features $F0changeVar1$ and $F0changeVar2$ in Table 7.

An acquired reading style that is sometimes observed is the regular variation in pitch that is unrelated to the text. This 'sing-song' rendition of the text is easily recognised by the human raters who consider it at odds with the comprehensibility of the oral reading. It manifests itself in a pitch contour that exhibits regular oscillations about its mean. In order to capture this, we compute the zero-crossing rate across the utterance of the z-normalized F0 contour ($F0\_zcr$). We also compute the windowed autocorrelation function over 1 s (and 2 s) segments of the pitch contour to get the corresponding autocorrelation peak ($F0\_acfpeak$) and lag ($F0\_acflag$). The mean and standard deviation of the peak and lag values across the utterance are then computed to get an overall estimate of the periodicity. The number of 1 s (and 2 s) segments for which the autocorrelation function peak lies below a threshold are counted to get $\#aperiodicF0$ feature.

4.5.2. Aggregation of Word-level Features

With the word being a prominent prosodic unit, the perceived prosodic variation across an utterance may be associated with the aggregation of word-level estimates. We obtain the sequence of word-level means and word-level s.d.s for each A-P parameter contour across the utterance. In order to capture the variation across words, we compute

mean of word-level mean sequence. Next, to know the extent of variability across words, as listed in Table 7, we compute the s.d. of the above mean and s.d. sequence to use in the feature set.

Apart from the mean pitch, the actual shape of the pitch contour across a word can be distinctive (Sridhar et al., 2008). Some students utter the text in a list-like fashion, with every word perceived as a phrase boundary, obscuring the structure of the text. Sometimes, this can even be a type of upspeak, where each word has a rising pitch. In order to capture this, we compute the slope for every word segment's pitch contour through line fitting. The fraction of words with a positive slope ($\%positiveF0slope$) forms the corresponding feature, expected to be high for list-like reading. The number of words with pitch reset value (absolute difference between the average pitch value across last 100 ms of a word and across the initial 100 ms of the next word) more than the average word level mean pitch are also considered. Another set of features to predict this is the fraction of words ending (last 100 ms of voiced region) with falling ($\%negativeEnd$), rising ($\%positiveEnd$) or flat ($\%flatEnd$) pitch contour shape. On the similar lines, we also compute the fraction of words with rising ($\%risingF0$) or falling ($\%fallingF0$) contours indicative of phrase break type, and peak-like ($\%peakyF0$) or valley-like ($\%dippingF0$) contours indicative of stressed and unstressed words. Only those words with the F0 contour highly correlating (> 0.5) with the corresponding ideal piece-wise contour (Tamburini, 2003) are considered in this case. The ideal contour is computed from measured key points in the word-level F0 contour.

4.6. Comprehensibility Prediction

The set of 144 distinct features, computed from a speaker's oral reading recording of a passage, are summarised in Table 7 in four feature groups according to the underlying suprasegmental attributes.The broader categories, corresponding to the pedagogic rubrics discussed in Section 1, are as follows.

1. Accuracy & rate features: lexical miscue aggregates, WCPM, and other speech rate based features across the recording.
2. Prosody features: pause features, prosodic event consistency and miscue aggregates, and aggregates of the acoustic-prosodic contour and the word-level features.

As depicted in Figure 7, the features across the 4 groups are used as inputs to the trained random forest regressor to predict the comprehensibility rating score. The target scores used for training are obtained by combining the individual rater's scores as in Section 3.4, Eq. 1. The target score is therefore continuous-valued in the range 0-5 and the regressor is trained to minimise the mean squared error between predicted and target scores. We employ feature selection to reduce redundancies with minimal loss of predictive power. Feature selection is known to help system performance while also improving the interpretability of the model. We use the process of recursive feature elimination (RFE) to obtain the optimal feature set (Scikit-learn documentation, 2020). Here, the least important features are removed from the feature set recursively until the removal causes the R^2 (coefficient of determination) score to drop more than 0.1%. The feature selection outcomes are therefore valuable in terms of providing an understanding of the precise acoustic cues that explain the raters' choices.

First, each feature group is processed individually based on cross-validation prediction performance on the training set to obtain the optimal set of features representing that group. Next, we apply feature selection on the combined feature set drawn from across the individually reduced feature groups to discover and eliminate strongly covarying components that do not necessarily add value together. Such a staged approach helps remove the correlated features within a group without losing on possibly less important but very distinct features from a different group (Sabu and Rao, 2021). Finally, we use ablation experiments, evaluating predictions separately with each individual feature group as well as with specific combinations, to compare the contributions of the different high-level attributes in explaining a given rater's scores. In the next section, we present our experiments starting with a presentation of the speaker-independent data set splits and the chosen evaluation measures.

5. Experiments and Results

We use cross-validation testing with a relatively large number of non-speaker-overlapping folds. The dataset of 1447 recordings discussed in Section 3.4 is divided into six cross-validation folds with no speaker overlap. Each fold has close to 240 recordings by about 28 speakers. Each of the 6 folds serves once as test split with the remaining

5 as the corresponding train split. With the training set being relatively large in each case (5/6 fraction of the full dataset), we obtain a similar distribution of ratings across each of the 6 training splits, that also matches that of the overall dataset distribution as seen in Figure 3c. It was not possible to achieve the separation of the text passages across the folds simultaneously with the speaker separation; however this does not compromise our performance evaluation given that all recordings of a given passage correspond to distinct speakers and therefore obtain a range of ground-truth scores as discussed previously with reference to Figure 6b. For both, the feature selection and the random forest hyperparameter tuning, we use 5-fold cross-validation within the training split corresponding to each test fold. The RF model hyperparameters are the number of trees selected from [500, 750, 1000] and the maximum tree depth selected from [5, 8, 10]. The final model is obtained by fitting the entire train split using the selected features and hyperparameters, and predictions are made on the test set. The reported system performance is then the mean performance achieved across the six test folds and hence speaker-independent.

We evaluate the predictions of the automatic system with reference to the target scores obtained from the human judges as discussed in Section 3.4. In order to summarize the relevant aspects of the assessment system, we employ five distinct measures that are commonly used in similar tasks, viz. Pearson correlation (r), R-squared (R^2), RMSE, concordance correlation coefficient (CCC) and adjacent agreement (adjAgree). The Pearson correlation coefficient measures the strength of the linear relationship between the predicted and the target scores as an indicator of consistency. However it is insensitive to data translation and scaling, and consequently uninformative as far as absolute agreement is concerned. Given that the scores on our rating scale (as presented in Table 3) have clear interpretations in terms of the stages of reading development, the extent of deviation of the predicted score from the target score is important to the system's effectiveness. This is summarized by the root of the mean-squared error, with lower being better. Root mean squared error (RMSE) and r have been used previously to evaluate automatic predictions in oral reading tasks (Proença et al., 2017b). Chicco et al. (2021) showed that the coefficient of determination R^2 is more informative than RMSE in regression analysis evaluation since it takes into account the variance of ground truth labels. We therefore report R^2 along with RMSE, with the former computed with the'r2_score' function in the Python scikit-learn package (version 0.22.1). We also include the percentage of predicted scores, after rounding to the scale levels, that lie within +/-1 of the target score (%adjAgree) which has been found useful in educational assessments where the number of levels on the scale is more than a few (White et al., 2021). Finally, we bring in a measure that balances the functions of correlation and agreement, known as the concordance correlation coefficient (Lin, 1989), that has been popular in evaluations of emotion recognition systems (Ringeval et al., 2015).

5.1. Feature Contribution to System Performance

Table 8 indicates the number of features obtained from the feature selection exercise applied separately to each group and the corresponding automatic prediction performance on the test set. The individual feature groups are further clustered into the two broad groups 'accuracy & rate' and 'prosody'. The first is the combination of the lexical miscue and speech rate features. We see that its performance is superior on all measures compared to either one of the components. The superiority of the lexical miscue features is explained by its single most important feature, the WCPM, which implicitly brings in the speech rate as well. The prosody group comprises the pause, A-P contour and the prosodic miscue features. Within this group, prosodic miscue-based features play the most crucial role, followed by the pause and A-P contour features. The former are the only text-dependent prosodic features, taking into account the reference information structure and closely modeling the pedagogic rubrics suggested by the NAEP scale for oral reading assessment. The prosody features obtain consistently better performance than the accuracy & rate features, underlining the importance of prosody for our dataset of above-threshold word decoders. This is consistent with the observations by Benjamin and Schwanenflugel (2010) indicating the importance of prosody for difficult text comprehension.

Finally, applying a round of feature selection to the entire set of features across the individual groups in Table 8 gives us a subset of 21 features, which together represent all the groups, in the penultimate row. This row shows performance parameters that are significantly better than those of the rows above it, confirming the complementarity of information across the feature groups. In the final row, we show the inter-rater agreement across all the same measures for comparison. We observe better correlation and agreement between the automatically predicted and combined-rater scores compared to that obtained across the scores of the two human raters. In all our experiments, we obtain p-values for the Pearson correlation coefficient that are all less than 0.001 indicating that the reported differences are significant.

Table 8: Performance of individual and combined feature groups in comprehensibility prediction.

Feature group (#selected features/#total features)	r	R^2	RMSE	CCC	% adjAgree
Lexical miscue (8/9)	0.69	0.496	0.902	0.651	83.8
Speech rate (4/4)	0.67	0.459	0.937	0.624	81.3
Accuracy & rate	0.71	0.522	0.879	0.674	84.9
Pause (12/12)	0.66	0.453	0.944	0.589	82.2
A-P contour (28/69)	0.65	0.409	0.983	0.517	79.4
Prosodic miscue (15/50)	0.71	0.510	0.890	0.664	85.3
Prosody	0.75	0.566	0.839	0.697	87.1
All (21/144)	0.77	0.603	0.800	0.736	88.3
Inter-rater	0.76	-	-	0.681	83.6

Table 9: Performance of different prosodic feature combinations in comprehensibility prediction.

Feature group	r	R^2	RMSE	CCC	% adjAgree
Pause & A-P contour	0.70	0.493	0.910	0.617	82.4
Prosodic miscue	0.71	0.510	0.890	0.664	85.3
Add SR	0.76	0.577	0.826	0.716	87.0
Add A-P contour	0.77	0.591	0.814	0.722	87.7

5.2. Analysis of Prosody Feature Interactions

In this section, we take a closer look at the prosody features, the main focus of this work. We study the individual contributions to system performance and the interactions within and across the feature groups. We also evaluate the specific benefits, if any, of the newly proposed 3-level reference information structure to the prosodic miscue features. Given that the accuracy of prosodic event detection is critical to the effectiveness of prosodic miscues, we examine the potential of PED that has been improved with additional word tags such as parts-of-speech (PoS).

5.2.1. Feature Group Interactions

To better understand the roles and the interactions of the component feature groups under the prosody category, we separately evaluate the non-text-dependent features (pause and A-P contour). We note from Table 9 that while the combination is much better than either of the two in isolation as presented in Table 8, it is still not as effective as the prosodic miscue feature group by itself. This underlines the significant role of text-dependent prosodic features in reading assessment. To test our hypothesis that a human rater's judgment of appropriate prosody in the reading of a text is influenced by the perceived reading speed, we also examine the effect of adding speech rate (SR) features to the prosodic miscue features and note a large improvement in all the measures. Finally, we observe that further bringing in the A-P contour features registers another boost, even if smaller, possibly indicating the effectiveness of including cues to reading style while judging prosodic accuracy.

5.2.2. 3-level Information Structure

The prosodic miscue features were seen to bring significant value to the system performance in Table 8. Apart from the text-dependent prosodic features adapted from Bolaños et al. (2013b), we have the features computed from explicitly detected prosodic events in the utterance as discussed in Section 4.4.2. Table 10 compares the individual contributions of the two feature categories. We also show the effects of 3-level reference information structure in each case where every word in the text is tagged for expected boundary and prominence as forbidden, optional or mandatory as discussed in Section 3.1. We observe that this enhancement of the reference information structure, over the usual 2-level (event present/absent) actually improves the performance of the system for, both, the Bolaños et al. (2013b) features and the proposed PED based features. The PED features show the better performance but the combination of the two sets is best overall suggesting that the inaccuracies in PED are compensated to an extent by the implicit prosodic consistency cues, as discussed further in Section 6. The slight inaccuracies in the automatic

extraction of reference information structure by the BERT-embedding based model (Section 3.1) were not found to impact the performance in any measurable way (Sabu, 2022).

Table 10: Performance of different prosodic miscue-based feature groups in comprehensibility prediction

Feature group	r	R^2	RMSE	CCC	% adjAgree
Bolaños et al. (2013b) 2-level	0.64	0.414	0.975	0.587	80.4
Bolaños et al. (2013b) 3-level	0.65	0.423	0.968	0.595	81.5
Prosodic event 2-level	0.65	0.436	0.958	0.596	81.9
Prosodic event 3-level	0.68	0.468	0.930	0.622	83.3
Combined prosodic miscue 3-level	0.71	0.510	0.890	0.664	85.3

5.2.3. Improving Prosodic Event Detection

In the experimental results reported in Section 5.1, we employed the prosodic event detection system of Sabu and Rao (2021) based on purely acoustic features in a trained classifier. As discussed in Section 4.4, prominence detection achieves modest accuracies while boundary detection is considerably better. The accuracy of prosodic event detection with respect to perceived labels is known to benefit however from added lexical information given the top-down mechanisms at work in perceptual rating (Sridhar et al., 2008; Stehwien et al., 2018). The concatenation of PoS tag, number of syllables in the word and the information structure label, all together termed 'word tags', with the acoustic features were found to boost the detection performance for boundary and prominence significantly (Sabu, 2022). The PED performance with the inclusion of the word tags improved the detection accuracies for both boundary and prominence by up to 6%. However additional experiments with the improved PED system with word tags incorporated showed, contrary to expectation, a small drop in performance of the prosody feature group across measures. This may be explained by the differing natures of the two kinds of ratings, viz. that by the prosodic event annotators who carried out word level rating of the oral reading recordings influenced by top-down expectations from their knowledge of the text, and the other by the teachers who were holistically scoring each utterance for comprehensibility by paying attention to the child's actual realization of the passage prosody. The latter were driven by the scale interpretations that prompt them to attend to prosodic miscues based on acoustic realizations.

6. Discussion

We discuss different aspects of the experimental results reported in the previous section to derive further insights about the working of the automatic assessment system. The composition of the compact set of features that resulted from feature selection can tell us about the acoustic and lexical cues that influence the rater scores. Any systematic differences between the two raters' perspectives can further be revealed with the comparison of features from the separate modeling of each rater's scores as the training target. We also study the class confusions and the performance obtained with the reduction of the number of classes via meaningful clustering. Finally, we examine the extent to which ASR word decoding errors affect system performance.

6.1. Combined and Rater-specific Feature Preferences

With redundant and highly correlated features removed by the process of recursive feature elimination, we have a relatively small subset drawn from across the feature groups originally listed in Table 7. Further, having seen that the raters are differently inclined from the distributions in Figure 3, we also model each rater separately to learn what cues they individually paid most attention to. The selected features computed for training targets corresponding to the combined ratings and to each of the two raters appear in Table 11.

We note that the prediction performance on the combined ratings is better than that on the individual rater scores, matching the observations of Black (2012) who attribute this to the denoising achieved by the process of normalizing and combining individual rater scores and the consequent improved consistency. Further, the combined scores take on continuous values in the range and can be considered to be particularly informative whenever the two rater scores differ. Comparing the performance metrics across the two raters, we see that the system trained for Rater 1 shows

Table 11: System performance for individual and combined target scores. The final set of selected features are shown in order of decreasing feature importance values. The features common across the three scenarios appear in bold.

Target (# features)	r	R^2	RMSE	CCC	%adjAgree	Features in decreasing order of importance
Combined (21)	0.77	0.603	0.800	0.736	88.3	**AR, WCPM, unexpectedSilratio, PR, sdmeanpitch, mostexpectedSilratio, bndAcc1, bndFPR1, bndAcc2**, band3full_mean, **SR, sdpitchsemitone**, stressedDurdiff, meanmeanband2, **bnd%FP1, bndPR1**, bnd%TN1, sdsdinten, sdmeanband4, meanmodepitch, sdmeaninten
Rater 1 (23)	0.75	0.578	1.055	0.711	85.4	**AR, WCPM, PR, unexpectedSilratio, bndFPR1, bndAcc2, sdpitchsemitone, bndAcc1, mostexpectedSilratio, SR**, stressedDurdiff, **sdmeanpitch**, meanmeanband3, bnd%TN1, Silpersec, Pauseperword, **bnd%FP1**, SRvar, sdmeanhnr, meanpitch, **bndPR1**, Silperword, sdsdband4
Rater 2 (29)	0.70	0.498	0.731	0.643	95.6	**AR, WCPM, unexpectedSilratio, PR, mostexpectedSilratio, sdmeanpitch, bndAcc1**, band3full_mean, **bndAcc2, SR**, band2full_mean, sdmeanband4, **bndFPR1**, F0_zcr, **sdpitchsemitone**, %miscue, Pauseperword, meanmodepitch, sdmeanband1, bnd%TP2, meanpitch, meanmeanband2, **bndPR1**, meanmeaninten1s, band2full_sd, **bnd%FP1**, SRvar, sdmeaninten, sdsdhnr

the higher correlation between the predicted and target score but that the RMSE is worse. The latter is explained by the predicted values deviating considerably from those of her targets that appear at either 0 or 5. Rater 2, instead, displays superior RMSE and adjacent agreement values due to her targets being largely confined to a narrow range as seen in the ratings distribution of Figure 3b. However, the CCC is relatively low for Rater 2 due to the reduced Pearson correlation. The R^2 values here cannot be compared directly due to the differences in, both, ground-truth and predicted labels across the rows.

With the features common to all 3 target choices marked in bold font in Table 11, we also observe a number of non-overlapping features across the raters. The final selected features appear in the order of decreasing importance. We observe that all the three types of speech rate estimates, viz. articulation rate (AR), phones per sec (PR) and syllables per sec (SR), serve as important features. This is indicative of differences in perception of syllable versus phone rates, and whether the non-speech region is included in the estimation (Pfitzinger, 1998). Further, the variation in speech rate also seems to affect the ratings by individual raters. *WCPM*, a measure that represents both word-level accuracy and speech rate, is a prominent feature for both raters and the combined target. The other lexical miscue features are conspicuously absent from the combined ratings feature set. We explain this later with reference to ASR decoding errors.

Once the accuracy and rate features are accounted for, the next most important features are related to prosodic miscues. We see that the boundary miscues dominate with both the mandatory (suffixed by '2' as in *bndAcc2*) as well as combined mandatory and optional boundaries (suffixed by '1') featuring in the selected set. Although the correct realization of the mandatory boundaries is important, even more important is 'not realizing' phrase boundaries in forbidden positions as is indicated by the false positive rate feature $bndFPR1$, the event realization precision features

*bndPR*1, and miscue realization accuracy features *bndAcc*1 and *bndAcc*2. The feature *bnd%FP*1 also appears across the raters with Rater 1 also adding *bnd%TN*1. The importance of correct boundary realization and especially the importance of avoiding pauses at unexpected boundary positions is indicated not only through these features, but also through the features indicating the text structure consistency of silence durations (*unexpectedSilratio* has higher importance than *mostexpectedSilratio* across the models). We note that the text-independent pause features (i.e. all the pause group features) are not among the top features except for *Pauseperword*. Rater 1 further has other features like *Silpersec* and *Silperword* in the selected set. Prominence miscues are represented in the combined ratings prediction by only the *stressedDurdiff* feature which does not depend on the explicit detection of the event. The lower reliability of prominence detection as well as the more complicated prominence perception could be the causes for this. This is similar to the observations by Liscombe (2007) in the context of speech delivery rating, where the syllable stress based rhythm features are rather low performing compared to the syllable tone based intonation features.

Among the A-P-contour features, the energy content in band 3 (2000 to 5000 Hz) plays an important role, pointing to voice quality effects on comprehensibility. Pitch and intensity dynamics flagging monotonicity (standard deviation of the word level mean pitch and intensity values) form part of the important acoustic features. In particular, the standard deviations of the word-level mean pitch values (*sdmeanpitch*) and the semitone F0 contour values (*sdpitch-semitone*) are prominent features highlighting the place of pitch variation in the perception of expressiveness. Rater 2 has *F0_zcr* as an important feature endorsing her perception of the rhythmic or sing-song style as associated with poor comprehension. A greater number of pitch and intensity variation related features appear in the selected features for Rater 2.

The selected features in each target setting are assigned to the appropriate feature group as discussed in Section 5.1. The feature importance values of all features in each group are then summed up to get the group importance as plotted in Figure 8 for the three target settings. We see that the relatively large number of A-P contour features chosen by Rater 2 lead to the larger representation of this feature group indicating greater attention to speaking style by Rater 2.

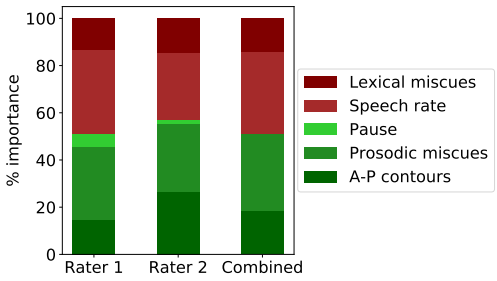


Figure 8: Feature group-wise importance (%) across target settings

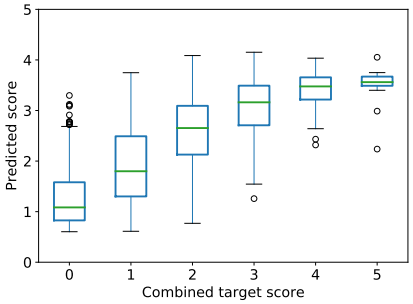


Figure 9: Distribution of predicted scores with discrete target scores

6.2. Analysis of Class Confusions

We take a closer look at the prediction errors and their possible dependence on the rating scale levels. Figure 9 shows the distribution of the continuous valued prediction for each of the combined target rating levels obtained via rounding to the nearest scale level of Table 3. We see that the predictions corresponding to scale levels 2 and 3 are the most widely dispersed. Further, we also observe that the system does not ever predict a score above 4.0 and also projects the 0-level ground-truth to a prediction region around 1.0. In Table 12, we show the confusion matrix resulting from predictions rounded to the scale levels. The poorer performance (as seen from the purely non-diagonal entries) at the extreme ends of the scale may be attributed to the lower representation of training target scores at the two ends.

The confusion matrix also facilitates a study of system performance for alternate interpretations of the scale levels corresponding to a reduced number of classes. A useful categorization is that with 3 levels, interpreted as follows: poor (levels 0, 1, 2, i.e., where correct chunking is yet to be achieved), fair (level 3, i.e., phrasing is proper but the realization of prominence or expression is inadequate) and good (levels 4, 5). We also consider a binary categorization into poor (levels 0, 1, 2) and good (levels 3, 4, 5, i.e., where at least one of phrasing and prominence realization has

Table 12: Confusion matrix of target and predicted scores after rounding both to nearest integers

Target \ Pred	0	1	2	3	4	5
0	0	137	44	6	0	0
1	0	115	61	32	1	0
2	0	58	136	135	16	0
3	0	5	78	240	76	0
4	0	0	13	118	57	0
5	0	0	2	51	66	0

been attained). The agreement measures, as reported in Table 13, indicate that performance is well above chance levels.

Table 13: Performance metrics comparing the target and predictions after rounding

Agreement type	Number of classes		
	6	3 [0,1,2 ‖ 3 ‖ 4,5]	2 [0,1,2 ‖ 3,4,5]
Exact agreement (%)	37.7	63.0	80.0
Adjacent agreement (%)	88.3	97.7	NA
Cohen's kappa	0.204	0.410	0.600

We derive from Table 12 that there are 32 recordings where the target and predicted are separated by greater than one level in the 3-class categorization. A closer inspection of these recordings, including the human raters' comments accompanying these, points to the inability of the system to penalize certain speaking styles that strongly affect the raters' perceptions. For example, one of the speakers speaks in a list-like manner, but is still able to realize phrase break in an acoustically acceptable manner. Similarly, some speakers display a noticeably rhythmic style of speaking that is apparently not captured well enough by the acoustic-prosodic features that survive feature elimination as seen in Table 11. With binary classification, we have 80.0% accuracy as indicated in Table 13. This is below the 90% accuracy achieved by Bolaños et al. (2013b), with the difference that 25% of the samples in their dataset lie in the <70 WCPM range.

It is of interest to compare the agreement values of Table 13 with the corresponding inter-rater agreements presented in Table 5. We see that in the 6-category ratings, the automatic predictions agree with the target ratings much better than do the two human raters with each other. For the more coarse 3- and 2-category clusters, we observe comparable values across the two tables except for the slightly superior Cohen's kappa in the inter-rater 2-class case. From-scratch retraining of the system with the reduced target classes is likely to improve the system performance further.

6.3. Performance Degradation due to ASR Errors

The word-level features computed for the prediction of comprehensiblity are derived from the utterance using the ASR decoder hypothesis word sequence and word boundaries. As indicated in Section 4.1, the ASR has a word-error rate of 5.55% with reference to the manually transcribed ground truth on the dataset used in this work. Given the seemingly critical dependence of the lexical and prosodic feature extraction on ASR accuracy, it is of interest to understand the impact of the ASR errors on system performance. We achieve this by simulating an ASR error-free situation with the direct use of the manual transcript for utterance segmentation via forced alignment. We then report the system performance on two representative measures in Table 14 so as to facilitate a direct comparison with the previously reported feature-group performances with ASR in Table 8.

We observe that the use of the manual transcript leads to fewer selected features (21 versus 19) and an overall improvement in both the Pearson correlation and the agreement as indicated by lowered RMSE. However, the individual feature groups show varying behaviour. The lexical miscue features, based as they are on the accurate identification of the speaker's word decoding errors, show a marked improvement when ASR errors are eliminated. This, as expected,

translates to significant improvement for the accuracy & rate features. We note a worsening of performance for the Pause and A-P contour features. Since these features depend more on word alignment rather than word identity, a possible explanation is the difficulty in achieving the accurate alignment of the acoustic signal with text that is not necessarily the best match in terms of the acoustic models. The prosodic miscue computations require, both, good word boundaries and the accurate mapping to reference text words. Overall, performance with prosodic features is not affected much, indicating their relative robustness to ASR errors.

In Table 15, we compare the set of selected features in the two cases. While WCPM and the speech rate features appear in both sets (with the former dominating when there are no ASR word errors), a notable difference is the inclusion of *%miscues* as a prominent feature with the manual transcript. This is absent from the feature set selected with the ASR hypothesis with ASR errors rendering it less reliable. We see instead the prosodic miscue feature *stressedDurdiff* and also a few A-P contour features corresponding to pitch and intensity although these are lower in importance. These observations demonstrate the diversity of cues associated with oral reading fluency and how they can contribute to the automatic prediction to different extents based on the reliability of the associated feature extraction.

Table 14: Prediction performances compared with ASR decoder and manual transcript based forced alignment

Feature group (#features selected)	ASR hypothesis			Manual transcript		
	r	R^2	RMSE	r	R^2	RMSE
Lexical miscues (8,5)	0.69	0.496	0.902	0.73	0.551	0.853
Speech rate (4,4)	0.67	0.459	0.937	0.66	0.447	0.947
Accuracy & rate	0.71	0.522	0.879	0.76	0.584	0.821
Pause (12,12)	0.66	0.453	0.944	0.66	0.446	0.951
A-P contour (28,30)	0.65	0.409	0.983	0.63	0.374	1.013
Prosodic miscues (15,18)	0.71	0.510	0.890	0.72	0.523	0.879
Prosody	0.75	0.566	0.839	0.76	0.570	0.836
All (21,19)	0.77	0.603	0.800	0.79	0.636	0.767

Table 15: The selected features in order of decreasing feature importance values for the two scenarios: ASR decoder alignment and manual transcript based forced alignment. The common features across the two cases are indicated in bold.

ASR hypothesis (21 features)	Manual transcript (19 features)
AR, WCPM, unexpectedSilratio, PR, sdmeanpitch, mostexpectedSilratio, bndAcc1, bndFPR1, bndAcc2, band3full_mean, SR, sdpitchsemitone, stressedDurdiff, **meanmeanband2, bnd%FP1, bndPR1, bnd%TN1**, sdsdinten, **sdmeanband4**, meanmodepitch, sdmeaninten	**WCPM, AR, PR, unexpectedSilratio**, %miscue, **bndAcc1, mostexpectedSilratio, bndFPR1, bndAcc2, sdpitchsemitone, band3full_mean, bnd%FP1, bndPR1, sdmeanpitch, SR**, band2full_sd, **sdmeanband4, meanmeanband2, bnd%TN1**

7. Conclusion

A chief goal of this work was to contribute to the automatic assessment of expressiveness in oral reading, a strong determiner of reading comprehension in the middle and higher ranges of reading ability where acceptable word decoding accuracy and rate have been already achieved. We presented a dataset of audio recordings of the oral reading of short passages of connected text that was screened for a minimum level of word-decoding ability and rate by lower limiting the WCPM and the percentage of lexical miscues as determined from the manual transcription. The recordings were rated for comprehensibility by each of two expert raters on a scale that closely followed the NAEP scale for reading expressiveness linked to conveying structure and meaning, achieving an inter-rater correlation of 0.76. While the prompt passages were selected to span a range of difficulty levels, any given passage also obtained a range of scores across speaker recordings.

Our dataset screened for a minimum WCPM obtained a wide range of comprehensibility scores for given WCPM clearly indicating the need for additional features for the automatic prediction. Correspondingly, our experiments showed the significant role of prosody features in the system performance. The pedagogical rubrics of accuracy, rate and expression were modeled with computed signal features drawn from across lexical and acoustic-prosodic analyses and aggregated across the utterance. In the interest of interpretability, and potentially new insights, we used random forest regression with recursive feature elimination to determine the optimal set of features within and across feature groups. Lexical miscues, speech rate and pause structure were estimated from ASR hypotheses. While the acoustic-prosodic features used in this work are largely drawn from available speech prosody studies, an important new contribution is the computation of the text-dependent prosodic miscue features. The use of 3-level information structure, motivated by the observed dependence of prosody realization on reading speed, was found to improve the prediction performance of the prosodic miscue features. Within prosodic miscues, the boundary miscues emerge as important with the speaker's realization of boundaries on unexpected words being a prominent negative marker and conforming with the interpretations of the ratings scale levels. The system achieved a Pearson correlation of 0.77 with the combined rater scores, slightly superior to the observed inter-rater correlation, and also surpassing inter-rater agreement in the 6-level classification. This is similar to the observations of Cheng (2018) and Fontan et al. (2022) who also obtained automatic scores that exceeded inter-rater metrics on their datasets of 94 and 13 speakers respectively. While these two works found speech rate and pause features sufficient in terms of non-phonemic cues, we find that prosodic event based features also contribute to system performance on our dataset of 165 speakers screened for high WCPM. Finally, while WCPM emerged as the single most important feature when ASR errors are eliminated, the prosody feature group achieves a performance comparable to that of lexical accuracy and rate features, and is also resilient to the inevitable ASR errors for children's speech. Not yet modeled very well, are the aspects of reading style found to influence the human raters' judgments in examples with otherwise similar features.

Our feature-engineering approach made the prediction model explainable in terms of pedagogical rubrics. Tree models, such as ours, lend themselves also to local interpretability, i.e. how a particular example receives its score (Lundberg and Lee, 2017). On the other hand, given the large search space for the low-level features as well as the types of across-utterance aggregation needed to predict the top-down scores of the human raters, we could have missed some potentially important features. Deep learning applied to the modeling of individual high-level attributes with simple interpretation-preserving fusion schemes is a promising future direction (Chen et al., 2018a,b). A future application of interest is the extension of expression scoring to Indian language contexts given the well-known language and culture dependence of prosody (Rao et al., 2017).

References

Bailly, G., Godde, E., Piat-Marchand, A., and Bosse, M. (2022). Automatic assessment of oral readings of young pupils. Speech Communication, 125:128–141.

Barhate, S., Kshirsagar, S., Sanghvi, N., Sabu, K., Rao, P., and Bondale, N. (2016). Prosodic features of Marathi news reading style. In Proceedings of TENCON, pages 2215–2218, Singapore.

Benjamin, R. G. and Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. Reading Research Quarterly, 45(4):388–404.

Black, M. P. (2012). Automatic Quantification and Prediction of Human Subjective Judgements in Behavioral Signal Processing. PhD thesis, University of Southern California.

Bolaños, D., Cole, R. A., Ward, W., Borts, E., and Svirsky, E. (2011). FLORA: Fluent oral reading assessment of children's speech. ACM Transactions on Speech Language Processing, 7(4):16:1–16:19.

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., and Schwanenflugel, P. J. (2013a). Human and automated assessment of oral reading fluency. Journal of Education Psychology, 105(4):1142–1177.

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Schwanenflugel, P. J., and Kuhn, M. R. (2013b). Automatic assessment of expressive oral reading. Speech Communication, 55(2):221–236.

Bone, D., Black, M. P., Ramakrishna, A., Grossman, R., and Narayanan, S. (2015). Acoustic-prosodic correlates of 'awkward' prosody in story retellings from adolescents with autism. In Proceedings of INTERSPEECH, pages 1616–1620, Dresden, Germany.

Breen, M., Kaswer, L., Dyke, J. A. V., Krivokapić, J., and Landi, N. (2016). Imitated prosodic fluency predicts reading comprehension ability in good and poor high school readers. Frontiers in Psychology, 7(1026):1–17.

Chen, L., Davis, L., Zechner, K., Lee, C. M., Yoon, S., Ma, M., Evanini, K., Mundkowsky, R., Wang, X., Lu, C., Loukina, A., Leong, C. W., Tao, J., and Gyawali, B. (2018a). Automated scoring of nonnative speech using the SpeechRater v. 5.0 engine. Technical report, Education Testing Service.

Chen, L., Taoz, J., Ghaffarzadegan, S., and Qian, Y. (2018b). End-to-end neural network based automated speech scoring. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 6234–6238, Calgary, AB, Canada.

Cheng, J. (2018). Real-time scoring of an oral reading assessment on mobile devices. In Proceedings of INTERSPEECH, pages 1621–1625, Hyderabad, India.

Chicco, D., Warrens, M. J., and Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peer J. Computer Science, 7:e623.

Christodoulides, G. and Avanzi, M. (2014). An evaluation of machine learning methods for prominence detection in French. In Proceedings of INTERSPEECH, pages 116–119, Singapore.

Council of Europe (2019). Assessment grid - English - displaydctmcontent. `https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168045bb52`.

Duong, M., Mostow, J., and Sitaram, S. (2011). Two methods for assessing oral reading prosody. ACM Transactions on Speech Language Processing, 7(4):14.1–14.22.

Ericsson, C. (2003). Predicting prosodic phrase boundaries for speech synthesis. Master's thesis, Kungiga Tekniska Hogskolan, Stockholm.

Fisher, W. M. and Fiscus, J. G. (2004). Better alignment procedures for speech recognition evaluation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 59–62, Minneapolis, USA.

Fontan, L., Kim, S., Fino, V. D., and Detey, S. (2022). Predicting speech fluency in children using automatic acoustic features. In Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pages 1085–1090, Chiang Mai, Thailand.

Fuchs, L. S., Fuchs, D., Hosp, M. K., and Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. Scientifc Studies of Reading, 5(3):239–256.

Fuchs, R. and Maxwell, O. (2015). The placement and acoustic realisation of primary and secondary stress in Indian English. In Proceedings of International Congress of Phonetic Sciences, Glasgow, Scotland, UK.

Groen, M. A., Veenendaal, N. J., and Verhoeven, L. (2018). The role of prosody in reading comprehension: evidence from poor comprehenders. Journal of Research in Reading, 42(1):37–57.

Hasbrouck, J. and Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. The Reading Teacher, 59(7):636–644.

Hayes, J. R. and Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. Written Communication, 16(3):354–367.

Isaacs, T. and Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. Studies in Second Language Acquisition, 34:475–505.

Keskin, H. K., Ari, G., and Baştuğ, M. (2019). Role of prosodic reading in listening comprehension. International Journal of Education & Literacy Studies, 7:59–65.

Learning A-Z LAZEL Inc. (2020). Fluency standards table. `https://www.readinga-z.com/fluency/fluency-standards-table/`.

Levis, J. M. and Silpachai, A. O. (2017). Prominence and information structure in pronunciation teaching materials. In Proceedings of the Pronunciation in Second Language Learning and Teaching conference, pages 216–229, Ames, IA, USA.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1):255–268.

Liscombe, J. (2007). Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency. PhD thesis, Columbia University.

Lundberg, S. M. and Lee, S. (2017). A unified approach to interpreting model predictions. In Proceedings of Neural Information Processing Systems, pages 4768–4777, Long Beach, CA, USA.

Miller, J. and Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. Reading Research Quarterly, 43(4):336–354.

Mostow, J. and Aist, G. (1997). The sounds of silence: Towards automated evaluation of student learning in a reading tutor that listens. In Proceedings of the National Conference on Artificial Intelligence, pages 355–361, Rhode Island, US.

Mostow, J. and Duong, M. (2009). Automated assessment of oral reading prosody. In Proceedings of the Conference on Artificial Intelligence in Education, pages 189–196, Brighton, UK.

Munro, M. J. and Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. Studies in Second Language Acquisition, 23:451–468.

Paige, D. D., Rupley, W. H., Smith, G. S., Rasinski, T. V., Nichols, W., and Magpuri-Lavell, T. (2017). Is prosodic reading a strategy for comprehension? Journal for Educational Research, 141(4):245–275.

Pearson School (2019). DRA2+ app broucher. `http://assets.pearsonschool.com/asset_mgr/current/201316/ReaBro121705DRA2_lo.pdf`.

Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. In Proceedings of International Conference on Spoken Language Processing, Sydney, Australia.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., NagendraGoel, Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In IEEE Workshop on Automatic Speech Recognition and Understanding, Hawaii, US.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., and Perdigão, F. (2017a). Automatic evaluation of children reading aloud on sentences and pseudowords. In Proceedings of INTERSPEECH, pages 2749–2753, Stockholm, Sweden.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., and Perdigão, F. (2017b). Automatic evaluation of reading aloud performance in children. Speech Communication, 94:1–14.

Rao, P., Sanghavi, N., Mixdorff, H., and Sabu, K. (2017). Acoustic correlates of focus in Marathi: production and perception. Journal of Phonetics, 65:110–125.

Rao, P., Swarup, P., Pasad, A., Tulsiani, H., and Das, G. (2016). Automatic assessment of reading with speech recognition technology. In Proceedings of International Conference on Computers in Education, Mumbai, India.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., and Pantic, M. (2017). AVEC 2017 - Real-life depression, and affect recognition workshop and challenge. In Proceedings of 7th International Workshop on Audio/Visual Emotion Challenge, pages 3–9, Mountain View, United Statesy.

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., and Pantic, M. (2015). AV+EC 2015 - The first affect

recognition challenge bridging across audio, video, and physiological data. In Proceedings of International Workshop on Audio/Visual Emotion Challenge, pages 3–8, Brisbane, Australia.

Sabu, K. (2022). Automatic Assessment of Fluency in Children's Oral Reading using Prosody Modeling. PhD thesis, Indian Institute of Technology Bombay, Mumbai, India.

Sabu, K. and Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. In CSI Transactions on ICT, volume 6, pages 221–225.

Sabu, K. and Rao, P. (2020). Automatic prediction of confidence level from children's oral reading recordings. In Proceedings of INTERSPEECH, pages 3141–3145, Shanghai, China.

Sabu, K. and Rao, P. (2021). Prosodic event detection in children's read speech. Computer Speech and Language, 68:1–19.

Saito, K., Trofimovich, P., and Isaacs, T. (2015). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. Applied Psycholinguistics, pages 1–24.

Schwanenflugel, P. J., Hamilton, A. M., Wisenbaker, J. M., Kuhn, M. R., and Stahl, S. A. (2004). Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. Journal of Educational Psychology, 96(1):119–129.

Scikit-learn documentation (2020). Feature selection – scikit-learn. `http://scikit-learn.org/stable/modules/feature\_selection.html`.

Sitaram, S. and Mostow, J. (2012). Mining data from project LISTEN's Reading Tutor to analyze development of children's oral reading prosody. In Proceedings of International Florida Artificial Intelligence Research Society Conference, pages 478–483, Marco Island, Florida.

Speech Processing Lab IIT Madras (2020). Hindi ASR challenge. `https://sites.google.com/view/asr-challenge`.

Sridhar, V. K. R., Bangalore, S., and Narayanan, S. S. (2008). Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. IEEE Transactions on Audio, Speech and Language Processing, 16(4):797–811.

Stehwien, S., Vu, N. T., and Schweitzer, A. (2018). Effects of word embeddings on neural network-based pitch accent detection. In Proceedings of Speech Prosody, pages 314–318, Poznan, Poland.

Talman, A., Suni, A., Celikkanat, H., Kakouros, S., Tiedemann, J., and Vainio, M. (2019). Predicting prosodic prominence from text with pre-trained contextualized word representations. In Proceedings of Nordic Conference on Computational Linguistics, pages 281–290, Turku, Finland.

Tamburini, F. (2003). Prosodic prominence detection in speech. In Proceedings of International Symposium on Signal Processing and Its Applications, pages 385–388, Paris, France.

Tao, J., Ghaffarzadegan, S., Chen, L., and Zechner, K. (2016). Exploring deep learning architectures for automatically grading non-native spontaneous speech. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 6140–6144, Shanghai, China.

Torgesen, J. K. (1998). Catch them before they fall - Identification and assessment to prevent reading failure in young children. American Educator/American Federation of Teachers, 22(1-2):32–39.

van Maastricht, L., Zee, T., Krahmer, E., and Swerts, M. (2017). L1 perceptions of L2 prosody: The interplay between intonation, rhythm, and speech rate and their contribution to accentedness and comprehensibility. In Proceedings of INTERSPEECH, pages 364–368, Stockholm, Sweden.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. Language Training, 11(2):197–223.

White, S., Sabatini, J., Park, B. J., Chen, J., Bernstein, J., and Li, M. (2021). The 2018 NAEP Oral Reading Fluency Study (NCES 2021-025). Technical report, Institute of Education Sciences, National Center for Education Statistics, U.S. Department of Education, Washington, DC.

Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31:2–13.

Wolters, A. P., Kim, Y. G., and Szura, J. W. (2021). Is reading prosody related to reading comprehension? A meta-analysis. Scientific Studies of Reading, 26(1):1–20.

Yang, R. (2010). A many-facet Rasch analysis of rater effects on an Oral English Proficiency Test. PhD thesis, Purdue University.

Yu, Z., Ramanarayanan, V., Suendermann-Oeft, D., Wang, X., Zechner, K., Chen, L., Tao, J., Ivanou, A., and Qian, Y. (2015). Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pages 338–345, Scottsdale, AZ, USA.

Zechner, K., Xi, X., and Chen, L. (2011). Evaluating prosodic features for automatic scoring of non-native read speech. In Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pages 461–466, Hawaii, USA.

Zutell, J. and Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. Theory into Practice, 30(3):211–217.