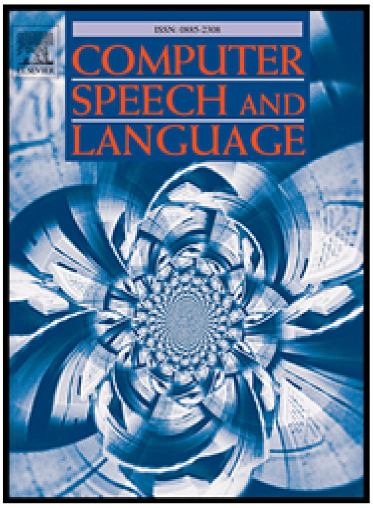


Prosodic Event Detection in Children's Read Speech

Kamini Sabu, Preeti Rao

PII: S0885-2308(21)00007-3
DOI: https://doi.org/10.1016/j.csl.2021.101200
Reference: YCSLA 101200

To appear in: *Computer Speech & Language*

Received date: 2 August 2020
Revised date: 18 November 2020
Accepted date: 31 January 2021

Highlights

- Data set of oral reading across L2 learner skill levels annotated for boundary and prominence
- Acoustic features evaluated in speaker-independent and speaker-dependent testing
- Trained model feature importances provide new insights related to speaking style

Prosodic Event Detection in Children's Read Speech[☆]

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

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

Abstract

Prosody is the supra-segmental aspect of speech that helps to convey the structure and intended meaning of lexical content unambiguously. The automatic detection of prosodic events, such as phrase boundary and word prominence, has a number of applications in discourse analysis, where a combination of syntactic and acoustic-prosodic features is typically employed. This work addresses prosodic event detection in the context of assessing oral reading skills of middle-school children. We discuss the observed characteristics of a specially created labeled data set of oral reading recordings of English stories by non-native speakers. The obtained diversity of language skills adds to the known challenges of high speaker variability in the acoustic realization of prosodic events. A combination of knowledge- and data-driven feature selection is implemented to identify a compact set of word-level features from the acoustic correlates of prosody considering different ways of incorporating the necessary temporal context. The system is benchmarked with reference to a widely known prosodic event recognition system in a speaker-independent set-up to obtain a competitive performance with greatly reduced feature dimensionality. The interpretable features enable us to use the predictor model importance scores to identify high-level speaker traits that influence the acoustic realization of prosodic events, suggesting a potential extension to systems that can extract and utilize speaker idiosyncrasies for superior prosodic event detection.

Keywords:
prosodic event, phrasing, prominence, L2 prosody, non-native children's speech, literacy assessment

1. Introduction

Pedagogical research on reading for school-age children shows that a child who reads with proper phrasing (grouping of words) and prominence (using emphasis for new or important information) is able to construct meaning of the text at the time of reading itself (Miller and Schwanenflugel, 2008). Such appropriate use of prosody is indicative of reader's understanding of the text (Bock and Mazzella, 1983; Whalley and Hansen, 2006; Paige et al., 2017; Groen et al., 2018). Prosody has therefore been recognized as a critical component of oral reading training and evaluation systems (Danne et al., 2005; Sinambela, 2017; Levis and Silpachai, 2017). Recent studies (Lochrin et al., 2015; Breen et al., 2016; Groen et al., 2018) indicate that learners with good word decoding ability can still span a wide range of prosody skills and correspondingly varying levels of comprehension. The work presented in this paper is motivated by the need for objective and scalable methods for the automatic assessment of children's oral reading skills, and more specifically, of the use of phrasing and prominence when good word decoding skills have already been attained.

Prosody, the suprasegmental aspect of speech, includes the manner of grouping words, stress on specific words, and the overall intonation of an utterance as related to the intended meaning, attitude, emotion, and other speaker-specific characteristics. It can be viewed as the spoken realization of textual punctuation, with added expression to convey meaning and affect. Phrasing refers to the grouping of words as related to the syntax and influenced by the

[☆]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

need for breathing pauses (Dowhower, 1987). Accordingly, the ideal chunking of an utterance involves splitting it into small meaningful groups similar to grammatical clauses (Schwanenflugel et al., 2004). Acoustically, the phrase boundaries are realized through lengthening of the final word and one or more of pauses and pitch and/or energy resets (Ericsson, 2003). Prominence, on the other hand, 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). It helps focus the listener's attention to a specific part of the utterance to mark new information or to express emotion or attitude (Tamburini et al., 2014). Prominence refers to, both, contrastive focus (correcting given information) and emphasis (marking new information), and therefore relates to the intended meaning (Maxwell, 2014). The acoustic realization of prominence is more variable than that of phrasing with its greater language and speaker dependence.

Phrasing and prominence are considered to be critical to the listener's ease of comprehension of speech (Bock and Mazzella, 1983; van Maastricht et al., 2017; Levis and Silpachai, 2017). Native English speakers are found to recognize and adjust to grammatical and pronunciation errors by non-native speakers but not so for the suprasegmental aspects, motivating increasing attention to the latter in spoken language training (Liscombe, 2007; Li et al., 2017). The present work relates to the use of prosody by children in the context of second language learning. A phrase boundary after a word or the presence of prominence on a word can be viewed as a prosodic event that is acoustically cued by local signal features. The automatic detection of prosodic events in the child's read speech can facilitate feedback and scoring of reading skills based on a comparison of the predicted locations with the expected locations of phrase boundaries and prominent words according to syntactic and semantic rules applied to the text.

Given the complex dependence of prosody perception on multiple suprasegmental attributes and linguistically defined local contexts, prosodic event detection research has been based largely on the exploration of interpretable feature representations (Rosenberg, 2009; Mishra et al., 2012; Christodoulides and Avanzi, 2014; Black et al., 2015). We present next an overview of past research in automatic prosodic event detection and place our task and approach in this perspective. Much of the well-known previous work is based on adult native speech. Further, considering the difficulty of obtaining manual prosodic annotations, the data sets are labeled by phonetic experts and typically limited to a few speakers. We describe our data set of children's read speech and the associated annotation process in Section 3. This is followed by a discussion of the proposed automatic boundary and prominent word detection system. Experimental findings of performance on test data are presented together with the useful insights obtained by relating the behavior of the acoustic-prosodic features with speaker characteristics.

2. Related Work

Prosody is realized in speech through the continuously varying low-level acoustic signal parameters such as fundamental frequency (F0), energy and segmental duration, the combinations of which give rise to psychological percepts such as phrasing (grouping), stress (prominence), and tonal movement (intonation) (Breen et al., 2010). Prosody perception is influenced not only by the low-level acoustic cues but also top-down expectations from lexico-syntactic information (Cole et al., 2010; Wagner et al., 2012). Phrase boundaries are marked at two levels - intonational/major and intermediate/minor. The former corresponds to the boundaries at sentence ending, while the latter corresponds to boundaries within sentence, e.g. at commas. Pause duration acts as the strongest cue in sentence segmentation (Shriberg et al., 2000; Rosenberg, 2009). Major boundaries are usually followed by long pauses, while intermediate phrase boundaries can be associated with the pitch reset and, optionally, short pauses (Ostendorf and Veilleux, 1994; Christodoulides et al., 2017). Word final lengthening, and pitch and energy reset are other important properties of phrase boundaries (Ericsson, 2003; Rosenberg, 2009; Christodoulides et al., 2017). Moreover, it is found that energy based features are more reliable indicators of phrase boundary than pitch based features (Kochanski et al., 2005; Rosenberg, 2009). Word prominence, on the other hand, is cued by syllable stress that is hyper-articulated through elongation, loudness, pitch accent and voice quality with the corresponding acoustic correlates being local variations in duration, intensity, F0 contour and spectral shape respectively (Li et al., 2018). The acoustic correlates of prominence have been reported for various languages. Breen et al. (2010) reported that focused words by American English speakers are produced with longer duration, higher F0 and intensity. German speakers use both intensity and pitch accents (Tamburini and Wagner, 2007). Intensity change from unstressed to stressed syllable was observed to be accompanied by a change in the spectral balance in Dutch. This was supported for American English speakers with the phenomenon being restricted to lexical stress in focused words (Campbell and Beckman, 1997). In French, prominence contributes substantially to prosodic grouping and boundary demarcation (Christodoulides and Avanzi, 2014). Focused constituents in Hindi have either extended syllables of longer length, expanded pitch range or both (Roy, 2017). Segment duration was found to be the dominant local cue to focus in Marathi, accompanied by post focal compression of F0 span (Rao et al., 2017). Speakers of languages that have flexible word order (such as Marathi or Spanish) however, also use syntactic means to cue sentence focus (Rao et al., 2016a; Cole et al., 2019).

We note that both types of prosodic events draw from the same set of underlying signal features. Phrase boundary detection can be considered as the binary classification of each word position in the utterance as phrase final or not with both accuracy and F-score used to quantify performance. The same applies to detection of prominence on a word. It is not surprising therefore that prosodic event detection in most works has relied on the stages of feature extraction and classification based on the statistical modeling of labeled data corresponding to the specific event. Apart from features computed from the word region itself, information related to the surrounding context plays a critical role for it is the acoustic changes that signal the prosodic event (Levow, 2005). This is typically taken into account by computing word-level functionals of low-level acoustic contours of energy, F0 and spectral shape and further applying normalization or differencing within selected temporal windows. Duration properties are brought in via the measured durations of pause, word and sub-word segments with further similar considerations about local context.

Since the time of pioneering research by Ostendorf and colleagues (Wightman and Ostendorf, 1994), a particularly detailed and influential work on prosodic event detection is by Rosenberg (2009, 2010) on BURNC, a Standard American English (SAE) corpus of seven broadcast radio announcers that is ToBI (Tones and Break Indices) (Beckman and Elam, 1997) labeled by experts. They also used the similarly labeled Boston Directions Corpus (BDC) containing read and spontaneous speech by 4 SAE speakers. While prosodic event classification to multiple ToBI labels was also investigated, we review the boundary and prominence detection results here. The use of pause features alone (silence before and after the word) gave 88.22% accuracy for intonational phrase boundaries, while a small (1.5%) improvement was seen with the addition of pitch and energy features in the form of differences between adjacent words to capture resets. Pre-boundary lengthening (length of final rhyme i.e. phones after last vowel) further improved the performance to 90.73% accuracy with 0.736 F-score. Including grammatical structure based cues like punctuation, phrase length, part-of-speech tags further improved the performance by approximately 1%. The intermediate boundary prediction was also studied separately reporting an accuracy of 89.5% with considerably lower F-score of 0.394. Detection of prominence, also termed pitch accent, employs duration, F0, energy and spectral shape features for the word followed by surrounding context normalization in up to eight distinct windows based on the included neighboring words. An accuracy of 82.9% is reported when trained on BDC read speech and tested on the spontaneous speech. Logistic regression classifiers were employed in the work which evolved into the AuToBI tool for ToBI labeling of speech corpora (Rosenberg, 2010). At the same time, a number of works showed how adding lexical and syntactic features at the word-level such as detailed part-of-speech tags, word frequency and position of word in sentence significantly improved detection accuracy for the BURNC corpus over the use of acoustic-prosodic features alone by accounting for top-down expectations in the prosodic event labeling tasks (Chen et al., 2004; Hasegawa-Johnson et al., 2005; Sridhar et al., 2008).

It is interesting to note that the contribution of the different attributes varies over languages, but the set itself is universal (Koreman et al., 2009; Endress and Hauser, 2010; Andreeva et al., 2014; Evin et al., 2018). AuToBI has also been extended to four other languages in Rosenberg et al. (2012) and cross-language performance has been studied. As expected, it has been observed that training or adaptation on the target language helps system performance. Further, Moniz et al. (2014) applied it on spontaneous European Portuguese speech, to find that prominence estimation performance was excellent for adult speech, but much lower for children's speech.

Word-based features obtained from frame-level acoustic contours have been used with several different supervised classification schemes including decision trees, support vector machines and neural networks (Ananthakrishnan and Narayanan, 2008; González-Ferreras et al., 2012; Escudero-Mancebo et al., 2014). Speech utterances labeled for boundary and prominence at the word level comprise the training and testing data. The representation of temporal context has been considered important and achieved either explicitly within the input feature computation by considering the surrounding regions, or implicitly via sequential models such as HMM or RNN (Rosenberg et al., 2015). More recently, Stehwien et al. (2020) employed a CNN to learn higher level feature representations from low-level frame-based acoustic features corresponding to the demarcated context region. They found it helpful to add extra position features indicating the current word in the input context matrix. For prominence detection, an unsupervised method exploiting surprise or unpredictability in acoustic properties over a region was attempted by Kakouros and Räsänen (2016). We see therefore that the deep learning methods applied so far to prosodic event detection have used well-motivated frame-based acoustic features at the input while seeking to automatically learn higher, word-level features together with the crucial context information. Making the task complex is the fact that prosodic information lies at multiple time scales in the signal.

The current work draws from the large body of previous research on acoustic features for prosodic event detection. With no known reported studies on children's read speech in the L2 learner context, we justify some of our choices for the proposed automatic system based on the characteristics of our data. Languages are known to have variations in the prosodic structures and L2 learners tend to retain some of the speech attributes of their L1 (Levow, 2009; Black et al., 2015). More importantly, given the possibly lower comprehension and fluency levels of our speakers, we do not expect the adherence to text syntax and semantics, and must account for this in, both, the instructions for data annotation and the choice of computational features. We build on our previous work for prominence detection (Sabu and Rao, 2018a,b) by greatly expanding the labeled data set (from the earlier six speakers) and extending the study to phrase boundary detection. As opposed to the expert annotations employed for data sets in the surveyed works, we employ a recently proposed and more practical approach to the perceptual labeling of the prosodic events that involves a large number of naive raters (Cole et al., 2017). We investigate available and new acoustic-prosodic features in speaker-independent testing of our data set to obtain performance surpassing that of the chosen baseline system while also providing useful insights about speaker-specific behaviors.

3. Data Set Description

Our data source comprises children from grades 5-7 (age group 10-14 years) in schools where English is the language of instruction. The speakers are exposed to the language from Grade 1 onwards although they rarely engage in spoken communication outside school. The oral reading sessions conducted for the data collection were enthusiastically facilitated by the school administration given the high interest in spoken English proficiency in India. We collected audio recordings at different times over a one year period from three different schools in and around the city of Mumbai in Maharashtra. The students were native Marathi speakers. In this section we present the characteristics of our data and the annotation process, which eventually provides the data set for this study, comprising speech recordings that are aligned to the transcription at the word level and manually labeled for prosodic events.

3.1. Materials and Recording Procedure

The reading material comprised of 34 distinct short English stories (∼ 100 words each). The story context facilitates expressive reading and can serve to showcase reading comprehension better over the reading of isolated sentences (Breen et al., 2016; Paige et al., 2017). The difficulty level, based on the vocabulary, and sentence length and structure ranged from A2 to B2 levels on the CEFR scale (McDowell and Settles, 2019). Most stories have a couple of sentences as quotatives indicating conversation. Half of the stories also have one Wh-question inside quotation marks. Meaning ambiguity is avoided as far as possible so that there is a single reference prosody for a sentence, given that our data is expected to serve also our larger goal of overall fluency assessment across skill levels (Rao et al., 2016b).

The data collection was facilitated by a specially created Android application which also allowed for metadata entry for each speaker. The story texts were presented in plain Roman script with the normally expected punctuation. Each story was split into paragraphs of about 50-70 words with the text presented on the screen, one paragraph at a time, for the speaker to read aloud with a headset microphone. The speaker was given the opportunity to go over the text silently before recording. Each story is recorded in one take. Recordings were at 16 kHz sampling frequency in 16-bit PCM format. To the extent possible, the recording sessions were conducted in off-class hours in relatively quiet spaces such as the school library to ensure SNRs (Signal to Noise Ratio) over 20 dB.

3.2. Data Annotation

In the course of informal listening, we observed large differences in reading skills across children in the same group (school and grade). While word decoding accuracy is good in many cases, this does not necessarily hold for prosody. The latter ranges from list-style or monotonous utterances to the fluent and expressive rendering of the text.

Several speakers use phrase breaks and emphasis with no relation to the text indicating the absence of comprehension while reading. Given that prosodic skills are typically acquired by readers after they attain good word decoding ability (Danne et al., 2005), prosody evaluation clearly makes sense in this context only. We use a two-stage semi-automatic transcription process to select recordings with an acceptable level of word decoding accuracy (i.e. the words are largely uttered correctly). Paragraph recordings (each termed an 'utterance'), numbering 921 in all, were collected from across 35 students of the three grades. Each utterance is of duration between 15 s and 50 s depending on the paragraph length and reading speed. This is in accordance with the literature, where 18 s duration of reading at normal pace is found to be sufficient for making good reading fluency evaluation decisions (Bolaños et al., 2013; Roy et al., 2017). In the rest of this section, we describe the semi-automatic transcription and prosodic event annotation processes implemented by us to obtain the word-level segmented data necessary for the subsequent training and evaluation of the automatic prosodic event detection system.

3.2.1. Word-level Transcription

Each utterance, corresponding to a single speaker-paragraph was first passed through an available automatic speech recognition (ASR) system. The hybrid DNN-HMM acoustic models had been trained on 5 hours of speech from bilingual children reading a variety of English and Hindi texts unrelated to the story texts of this study with 47 acoustic models of the phones and silence fillers expected in the English speech of native Hindi or Marathi speakers (Sabu et al., 2017). A trigram language model is trained on the corresponding paragraph text with a parallel zero gram garbage model containing commonly used substitutions (Li et al., 2007). An Indian English pronunciation lexicon is used. From the ASR decoded hypothesis, we computed the number of lexical miscues in terms of word substitutions, insertions and deletions with reference to the text. Recordings with a miscue rate of less than 20% were selected for the next stage of manual verification and possibly correction of the automatic transcription.

Transcribers are appointed for the word level transcription of the collected audio recordings. Long silences in the recording before and after the paragraph utterance were eliminated. Sometimes, saturating noise was observed in the recording due to mishandling of microphone during recording. Such noise was removed from the audio during transcription. Incorrect words were transcribed to whatever English word they sounded close to. For intelligible but invalid English words, Devanagari script was used to represent the perceived phone sequence (due to its grapheme-to-phoneme mapping that was familiar to the transcribers). The transcription was carried out by one transcriber per recording and followed with a quality check and possible correction by another.

3.2.2. Prosodic Event Annotation

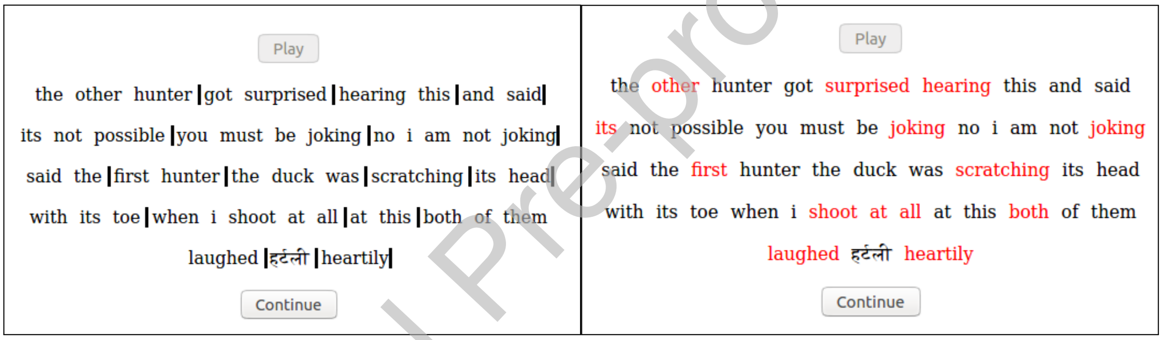
Most research on prosody has been based on annotations of speech recordings produced by trained annotators using listening, sometimes supplemented by visual information from spectrograms and pitch tracks. In an investigation looking to simplify the annotation process, Cole et al. (2017) found that crowd-sourcing manual labeling can help achieve reliable ground-truth provided a relatively large number of raters is employed. They found that the agreement rates for untrained raters stabilize at a cohort size of about 7 raters for both boundary and prominence labeling. We employ their method of Rapid Prosody Transcription (RPT), where several untrained raters label each word for boundary and prominence in separate tasks based on auditory cues alone. For our data, we confirmed that the standard deviation in the inter-rate agreement (Fleiss' kappa value) reduced by an amount below 0.05 as we went from 6 raters to 7 raters. We enlisted in all 26 university students who were fluent English speakers to label the prosodic events. They were compensated for their time with gift vouchers. None of the raters had been exposed to prosody labeling before. They were provided each story recital with the corresponding text transcriptions without any punctuation marks or capital letters which otherwise could have caused an expectancy bias. The raters were instructed to rely on their auditory impression rather than meaning or any prior expectation of prosodic event locations based on their knowledge of the text. The raters used the LMEDS web interface by Cole et al. (2017). Screenshots of the same are shown in Figure 1. Raters were then asked to mark the perceived phrase boundaries and, in a separate task, to highlight words perceived as emphasized by the speaker. Repeated listening was permitted, but not more than three times.

After the annotation by a given rater, the decisions were binary coded, where presence of event is marked as '1' and absence of event as '0'. Each recording was rated by 7 different raters and the decisions were combined to give a number per word which corresponds to the number of raters agreeing on the event presence for the word. For this study, we have 41326 words from 791 recordings by 35 speakers for phrase boundary detection and 42138 words

Table 1: Distribution of words in terms of number of votes for event present out of a total of 7 votes

No. of Votes	No. of Words (%) Marked Phrase Final	No. of Words (%) Marked Prominent
0	25071 (60.7%)	16283 (38.6%)
1	4104 (9.9%)	8406 (19.9%)
2	2115 (5.1%)	6840 (16.2%)
3	1809 (4.4%)	5221 (12.4%)
4	1554 (3.8%)	2719 (6.4%)
5	2031 (4.9%)	1513 (3.6%)
6	2089 (5.0%)	819 (1.9%)
7	2553 (6.2%)	337 (0.8%)
Total	41326	42138

from 808 recordings by 35 speakers for the prominent word detection task. Each word is assigned a ground-truth label between 0 and 7 for each of the prosodic events depending on the number of votes it obtained for event present.



the other hunter | got surprised | hearing this | and said |
its not possible | you must be joking | no i am not joking |
said the | first hunter | the duck was | scratching | its head |
with its toe | when i shoot at all | at this | both of them
laughed | हटंली | heartily |

(a)

the other hunter got surprised hearing this and said
its not possible you must be joking no i am not joking
said the first hunter the duck was scratching its head
with its toe when i shoot at all at this both of them
laughed हटंली heartily

(b)

Figure 1: Screenshots of the LMEDS GUI (Cole et al., 2017) with rater markings for (a) phrase break (vertical lines after word), and (b) prominent word (red font). The speaker repeats the word 'heartily', first with incorrect pronunciation.

3.3. Data Characteristics and Inter-Rater Agreement

The distribution of number of votes for each of the prosodic events across the data set appears in Table 1. We observe complete agreement (i.e. either 0 votes or 7 votes) for 39.4% of the total number of words in the prominence task, while 66.9% of the words show complete agreement in the phrase boundary task. As in the work of Roy et al. (2017), we consider the markings by each annotator for each attempt as independent of the other and find that 20.9% of the words are phrase-final and 21.5% words are prominent. This may be compared with data sets of native English speech with perceived phrase breaks reported to be 14.4% and prominent words to be 26.1% of the total number of words (Roy et al., 2017). We note that the differences in the numbers can arise from both the speaker productions and rater perception. For example, if we consider prominence labeling based on 3 or more votes, we obtain a more comparable 24% for words considered prominent.

The rater-wise observations indicate 10-30% of the words marked phrase-final by each rater, while for prominence, some raters marked only 6% of the words, while some marked almost 50% of the words as prominent. More consistency was observed in the case of speakers with good (i.e. adult-like) prosody. The apparent inconsistency across raters makes the decision of assigning binary labels less obvious. In any case, prominence is usually considered as a gradual phenomenon, and continuous (Evin et al., 2018) and multi-level prominence approach has been analyzed in literature (Tamburini and Wagner, 2007; Arnold et al., 2012). Researchers have used different approaches like number of votes (Streefkerk et al., 1999) and relative number of votes (prominence score ∈ [0,1]) (Mahrt et al.,

Table 2: Inter-rater reliability with 7 raters per word

Prosodic Event	Fleiss' Kappa
Phrase Boundary	0.59
Prominent Word	0.22

2012) for getting multi-level labels using binary labeling from multiple raters. For automatic detection tasks, however, binary labels are required and continuous or multi-level or multi-rater prominence decisions are converted to binary prominence labels using a threshold which may be one vote (Kakouros and Räsänen, 2016), majority (Evin et al., 2018) or some fraction (Streefkerk et al., 1999).

Our raters found the boundary marking task easier and were seen to be more consistent with each other as compared to the prominence marking task. The consistency values across ratings by 7 raters are measured using Fleiss' kappa and reported in Table 2. Fleiss' kappa measure is used since we have a variable set of multiple raters for every recording chunk (Stemler, 2004; Gisev et al., 2013; Roy et al., 2017). Our values of 0.59 and 0.22 for phrase boundary and prominence marking respectively can be considered as moderate and slight agreement respectively (Landis and Koch, 1977). Kappa value for a phrase boundary marking task is in the reported range of 0.54 to 0.62 (Cole et al., 2010), but the kappa value for prominence marking is lower than 0.4 as reported in (Cole et al., 2010; Kakouros and Räsänen, 2014) or 0.28 for annotation by naive raters (Roy et al., 2017). Our inter-rater agreement for prominence is comparable to the 0.23 reported for Indian listeners marking prominence using RPT (Cole et al., 2017). We must also keep in mind that, perhaps more than raters' perception deviations, our speakers include poor readers who are not adept at the use of prosody and may actually assign conflicting acoustic cues leading to ambiguous perception. This is especially true of prominence marking given that beginning readers typically attain phrasing skills earlier and that the prosodic marking of focus is an aspect that our non-native speakers are less familiar with.

Since the sentence prominence represents new or important information (Brenier et al., 2005), content words (nouns, adjectives, verbs and adverbs) are more likely to be prominent than the function words (articles, pronouns, conjunctions and determiners) (Ananthakrishnan and Narayanan, 2008). Phrase boundaries too are more likely to follow content words in English than function words. Phrase boundaries are related to the syntactic structure and we can expect the phrase boundaries appear with punctuation marks, and at the end of clauses. Based on these observations for typical speakers, prior work on prominence detection has depended on part of speech for ground truth labeling (Wang and Narayanan, 2007). However, in our data we also have, function words like articles, prepositions and conjunctions are marked as phrase boundaries with strong agreements. In our case, we have 8.7% of the strongly agreed phrase breaks coming from function words. Sometimes, phrase boundaries appear just before self-corrections. Strongly agreed upon prominent words include the adverbs and adjectives, similar to the observations by Cole et al. (2019). If an adjective-noun sequence appears, the adjective alone is marked prominent 82% of times, while noun alone is marked prominent only 11% of the times. Similarly, if a verb-adverb sequence occurs, the adverb alone is marked prominent 67% of times, while the verb alone is marked prominent only 25% of the times. While function words, prepositions and conjunctions are not expected to receive prominence, only 70% of the function words in our data are strongly agreed to be non-prominent. These observations indicate that lexico-syntactic features are not useful for prosodic event detection in our task as opposed to the more typical applications to the fluent speech of native speakers and we choose to rely on acoustic features alone.

4. Prosodic Event Detection System

The automatic prosodic event detection system takes as input the audio recording of an utterance together with the word and subword boundaries provided by the ASR. A trained regressor model is applied to the computed word-level acoustic feature vector to predict the degree of boundary and degree of prominence for each word in the utterance as depicted in Figure 2. In the current work, the required segmentation is obtained through forced alignment with the manually corrected word level transcript as presented in Section 3.2.1. This allows us to investigate prosodic event detection performance unobscured by ASR errors. Considering that the independently controlled characteristics of the speech signal such as F0, intensity, segmental durations and spectral balance are all influenced by prosodic factors, as reviewed in Section 2, we see that a large number of possible acoustic features suggest themselves for prosodic event detection. This is particularly true when we further consider ways of incorporating a variety of surrounding context in the feature definitions. Our approach in this work will therefore be towards constructing a large set of well-motivated, but possibly redundant features and applying state-of-the-art feature selection methods to obtain a relatively low-dimensional feature vector with high predictive power. We choose the random forest (ensemble of decision trees) model in our work due to the explicit access it provides to feature importances and its competitiveness with the best available models given modest sized data sets (Fernández-Delgado et al., 2014).

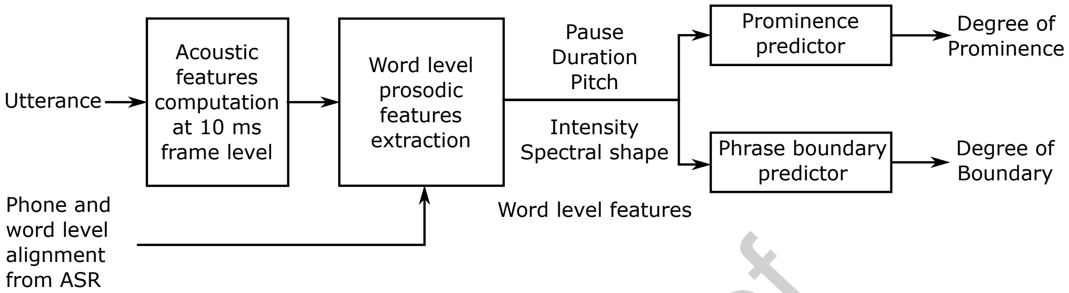

Utterance
Acoustic features computation at 10 ms frame level
Word level prosodic features extraction
Phone and word level alignment from ASR
Pause
Duration
Pitch
Intensity
Spectral shape
Word level features
Prominence predictor
Phrase boundary predictor
Degree of Prominence
Degree of Boundary

Figure 2: Block diagram showing the processing of an utterance for the word-level prediction of boundary and prominence

In this section, we describe the signal processing to obtain the acoustic parameter contours that are then aligned with word boundaries to compute word-level features. We present the feature extraction for both the baseline features (AuToBI of Rosenberg (2010) as implemented by us) and the proposed feature sets. We discuss the training procedure for the random forest predictor including the feature selection process.

4.1. Signal Analyses for Acoustic Contours

The reliable extraction from the signal of the time varying fundamental frequency, energy and spectral shape is needed for effective word level feature computation. We estimate all the parameters at 10 ms intervals to obtain 'acoustic' contours. Segmental durations of the words, subwords and pauses, on the other hand, are obtained from the ASR based forced alignment. Given the high speaker dependence of ranges for pitch, intensity and speaking rate, it is necessary to consider also features computed from speaker-normalized acoustic contours. In our application context of reading assessment across large and diverse groups of speakers, the system may encounter a given speaker in the context of a single story reading only. We therefore use recording level normalization (i.e. across the single speaker-paragraph utterance) to obtain the speaker-normalized (SN) contours.

4.1.1. Duration

All segmental durations are obtained in absolute time based on the alignment boundaries for words and phones. Phone-to-syllable mappings help to get syllable durations. Normalization for speaking rate can be achieved by linearly scaling all durations by a factor related to speech rate, which itself can be estimated in different ways. Phone rate and syllable rate are the most commonly used where, further, the averaging region could be global (i.e. the entire recording) or local (across +/-2 words or pause separated windows). Another method is to use z-score normalization by the local tempo (speech rate across a 3-syllable window surrounding the current syllable (Pfitzinger, 1998)).

4.1.2. F0

Fundamental frequency (F0) is the acoustic correlate of the percept of pitch. In this work, as in similar others, we use F0 and pitch interchangeably. The autocorrelation based pitch and voicing estimator of Boersma (1993) is applied with 20 ms windows and 10 ms hop. F0 values are restricted in the range 100 Hz to 520 Hz, taking into the target population of children. A voicing threshold of 0.45 and silence threshold of -30 dB are applied to determine voiced/unvoiced frames. F0 is linearly interpolated across unvoiced regions. Post-processing and smoothing is performed to correct octave errors and restrict the change in F0 across adjacent frames to within 30 Hz (Chen, 1974; Xu and Sun, 2002; Zhao et al., 2007).

F0 is not defined for detected silence frames. Further, we convert the F0 values in Hz to the logarithmic scale of semitones with reference to the mean F0 across all the voiced frames of the recording. Both Hz and logarithmic

F0 have been found useful in the past work (Arnold et al., 2011; Mahrt et al., 2012). Besides, we also compute the instantaneous pitch slope contour (Rosenberg, 2010) in every frame as the first-order derivative of F0 considering two frames on either side of the current frame. Finally, speaker normalization for pitch range is implemented by mean-variance normalization across the recording for each of the three contours (F0 in Hz and in semitone, and slope).

4.1.3. Intensity

To compute energy related features, the signal is first amplitude normalized by division with the absolute maximum value across the utterance. A contour of energy is computed by taking squared sum of amplitude values in 10 ms windows. Intensity is then calculated as the logarithmic energy with respect to the average energy content in the detected silence regions (obtained from the ASR alignment) after lower limiting the energy to that typical for silence regions (-30 dB). Smoothing with 100 ms windows is performed prior to the z-score normalization across the recording (SN) for both energy and intensity contours in order to overcome the recording specific biases due to microphone gain or the natural loudness of a speaker.

4.1.4. Spectral Shape

The acoustic correlates of perceived loudness are intensity as well as the balance of energy in the higher frequency bands relative to lower regions (Campbell and Beckman, 1997). Rosenberg (2010) computes energy in the Bark bands 2 to 20 (200 Hz - 6500 Hz) as a representation of the spectral shape. However several alternate representations for spectral shape have been proposed and we consider those listed below in this work. The spectral energy in a selected frequency region is computed from the short-time magnitude spectrum relative to the total frame energy. Smoothing of the resulting contour is carried out over 100 ms windows.

1. 0 to 500 Hz, 500 to 1000 Hz, 1000 to 2000 Hz, 2000 to 4000 Hz (Sluijter and van Heuven, 1996; Miao et al., 2006; Christodoulides and Avanzi, 2014)
2. 60 to 400 Hz, 400 to 2000 Hz, 2000 to 5000 Hz, 5000 to 8000 Hz (Patil, 2014; Patil and Rao, 2016)
3. Sonorant band (300 to 2300 Hz) (Wang and Narayanan, 2007)
4. Total energy in Bark bands 2 to 20 (200 to 6500 Hz) (Rosenberg, 2010)

Further, we also consider the first MFCC coefficient c_1 in a cepstral analysis of speech, capturing spectral tilt, and found to be a robust indicator of spectral balance for prominence detection (Kakouros et al., 2017). Speaker normalization is not applied to the spectral shape contours.

In Figure 3, we show the analyses of two utterances from our data set by speakers of distinct speaking styles illustrating the different acoustic realizations of the perceived prosodic events.

4.2. Word-level Feature Extraction

The acoustic contours computed at 10 ms intervals across the utterance and further aligned with the word boundaries form the basis for the computation of word-level features for prosodic event detection. The contour samples across the word segment are aggregated in different ways to produce a single feature value per word. The distinction between a given word's features and the corresponding features of the other words in its immediate surroundings provides the cues to its status as a phrase-final word or a sentence prominence. This relative behavior can be captured by either (i) the explicit computation of the differences or (ii) by the normalization of the feature with reference to values in the specified neighborhood of the word (Brenier et al., 2005; Rosenberg, 2010; Evin et al., 2018). For instance, the pitch and intensity resets associated with phrase boundary can be captured via the differences in the corresponding features between the phrase final word and its succeeding neighbor. A systematic way to consider all reasonable temporal contexts around a word is to carry out z-score normalization within the 8 distinct context windows mentioned by Rosenberg (2010). The windows are such that 0, 1 or 2 words are considered before and after the current word. That is, if n specifies the index of the current word, the different context subsets are drawn from the set of words ($n-2$, $n-1$, n, $n+1$, $n+2$). Another mechanism to bring in context is to compute the feature itself over the contour across the entire window spanning the words. An example is the slope feature obtained by the linear fit across the context window segment. Finally, we deal with the missing features for words at the beginning and end of the utterance by replacing these with utterance averages (Brownlee, 2019). We review next the baseline set of features followed by a description of the expanded feature set investigated by us.

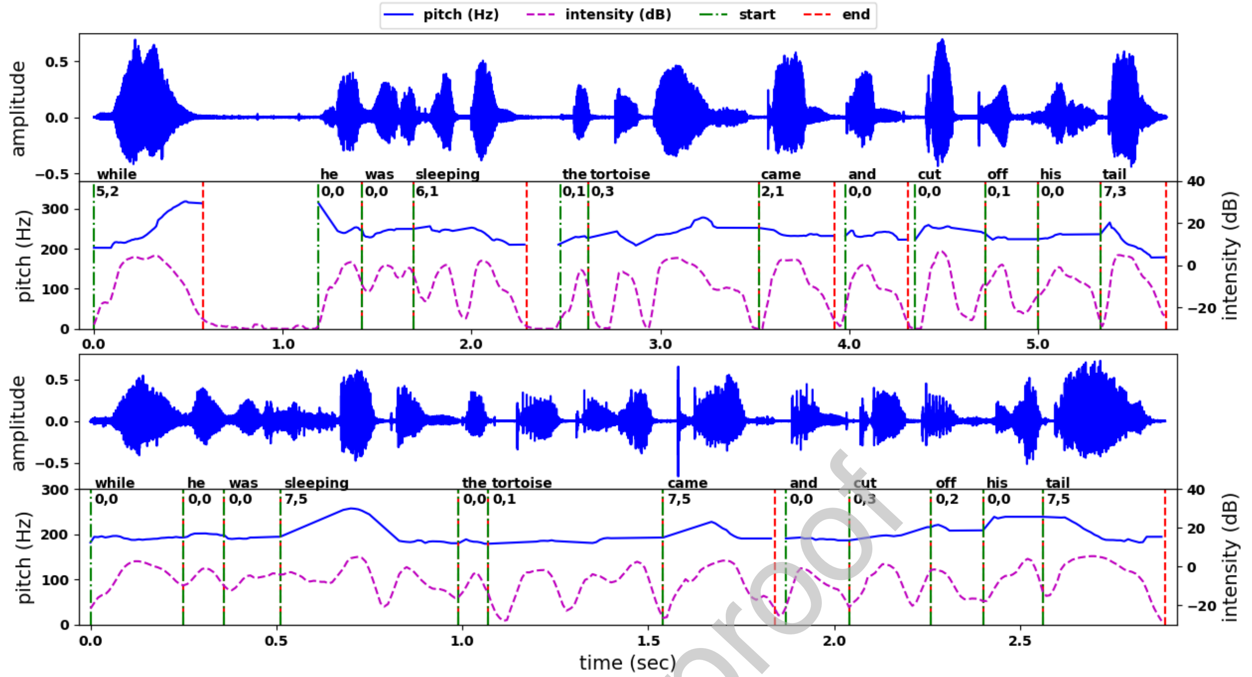


Figure 3: Audio examples for a sentence uttered by two different speakers. Below the waveform, the aligned words appear with boundary and prominence votes by the 7 raters indicated for each word. F0 in Hz (blue) and intensity in dB (pink) contours are shown along with word boundaries indicated by green and red vertical lines. We note from the time axes, the large difference in speaking rates between the utterances. Description: The first utterance has the words ‘while’, ‘sleeping’ and ‘tail’ strongly perceived as phrase-final. We note pauses marking phrase boundaries, but also the pitch slope at the end of the word, rising for ‘while’ and falling for ‘sleeping’ and ‘tail’. Mild prominence is perceived on words ‘tortoise’ and ‘tail’ apparently realized by the pitch span and higher intensity of the words. The second utterance is fast paced and devoid of pauses. The words ‘sleeping’, ‘came’ and ‘tail’ are perceived as phrase final from the both pitch and intensity falling towards word ending. Prominences are also perceived on the same words with lower prominence on ‘cut’. All the strong prominences appear to be linked to within-word pitch variation in an otherwise monotonous utterance. The prominent words also show duration elongation.

4.2.1. Baseline Feature Set - AuToBI System

To benchmark the performance of our system, we treat AuToBI (Automatic ToBI label prediction) system (Rosenberg, 2010) as a baseline. While they used logistic regression to predict binary ToBI labels for presence/absence of each of the prosodic events, we use the same feature set (based on the detailed implementation available in Rosenberg’s doctoral thesis (Rosenberg, 2009)) but with the random forest model adopted in this work. Different sets of features are specified for boundary and prominence although there is a large overlap.

For prominence detection, we have the functionals of word segment contour values in the form of the minimum (min), maximum (max), mean, standard deviation (s.d.) and word-level z-score normalized maximum (zmax). This is used on the pitch contour in Hz, intensity contour in dB as well as their SN versions. The max and mean values are also computed with z-score normalization across each of the 8 temporal context windows. An instantaneous slope contour is computed for each of the pitch and intensity contours. It is represented by its min, max, mean, s.d. and zmax across the word segment together with the mean and max z-score normalized across the different temporal context windows. The similar features are computed for a specific spectral balance contour (intensity in bark band 2 to 20) and relative balance contour (fraction of intensity in bark band 2 to 20) with no SN. The total number of features for prominent word detection is 210. Phrase break detection shares some features with prominence detection. These include min, max, mean, s.d. and zmax for the pitch contour in Hz, intensity contour in dB and their SN versions. They also compute similar features for the instantaneous slope of pitch and intensity contours. Distinctive features for phrase breaks are the word duration and the silence duration after the word. The presence/absence of silence after the word is further indicated by a separate flag. Boundary detection uses 83 features in all. We keep in mind that while the original AuToBI system targeted adult native speech, our baseline system uses the same feature definitions but now trained and tested on non-native children’s read speech using a random forest regressor. Further, our targets are perception based labels from multiple naive raters instead of the expert annotations of Rosenberg (2009).

4.2.2. Proposed Feature Set

In this work, we consider a large number of features for both prosodic events in common and use feature selection (as described later in Section 4.3) to customize the set for each of the boundary and prominence detection tasks. We augment the baseline feature set considerably based on other literature and our own data observations. A brief description of the features grouped by the prosodic attribute is provided here, especially highlighting the differences from the AuToBI baseline feature set.

- Pause Features - Pause after a word is known to be an important cue in realizing phrase boundary. We obtain the silence duration from the ASR based forced alignment. AuToBI (Rosenberg, 2010) system uses the pause after the word as a feature for phrase boundary detection. We redefine pause to include only those silence regions which are greater than 200 ms, based on the maximum duration observed for stop closures. Further, a silence region larger than 500 ms, termed 'long pause' has often been shown to be associated with phrasing (Liscombe, 2007), and therefore included here as a separate feature.

  Pause and long pause durations *before* the word are also included in our feature list as Indian language speakers are known to employ pre-focal pauses (Féry et al., 2016; Rao et al., 2017). This effect could therefore appear as an L1 transfer to L2. Another important enhancement to pause representation is normalization by the mean pause duration across the utterance thus accounting for speaker dependency including the influence of reading speed.

- Duration Features - Phrase-final words are usually elongated or uttered at lowered tempo. The duration aspect considered relevant for cueing prominence and phrasing is the syllable duration (Wang and Narayanan, 2007; Christodoulides and Avanzi, 2014; Arnold et al., 2012). A word is perceived as prominent when one or more of its syllables are elongated over the syllables in the neighborhood (Lunden, 2017). On the other hand, AuToBI (Rosenberg, 2010) system uses only one feature in this group: word duration. The inherent duration of a word, however, depends on the number of sub-word units within, such as the syllables. Therefore, we consider different subword entities (syllable and phone) as well. The features include the functionals: mean, minimum and maximum of the duration of the entity across the word. Since focus on a word leads to non-uniform duration changes in multi-syllable words (Rao et al., 2017), we incorporate the ratio as well as the difference between the longest and the shortest syllable durations in the word. If the two durations are equal, the word is unlikely to be prominent.

  While syllable duration is common in prominence detection, research shows that the syllable variation is primarily due to vowel elongation (Lunden, 2017). In our data set of L2 reading, however, we also see the individual phones sometimes stretched by the speaker in cases of word decoding difficulty, boredom or as filled pauses. This motivates the inclusion of the duration of the surrounding phones to help discriminate such non-linguistic causes of syllable elongation.

  The above features are computed for the raw and SN duration sequences. The speech rate can vary across a recording possibly due to the occasional presence of difficult words. The speech rate inconsistency is usually accompanied by pauses with pauses acting as the memory refreshing positions. Each segment between two pauses is expected to have its own reference for speech rate. Therefore, in addition to the +/-2 word windows, we normalize the duration features in the pause separated context windows.

- Pitch and Intensity Features - Phrase boundaries typically involve resets of both pitch and intensity. Prominence is also indicated through increased loudness in addition to the increase in pitch relative to the surrounding words. We compute the statistical functionals, viz. min, max, mean, median, s.d. and span for F0 and intensity contours across the word, similar to AuToBI. Further, the instantaneous slope of intensity contour is also computed across word. Context normalization is applied across each of the 8 temporal context windows.

- Pitch Contour Shape Features - The pitch is expected to decline across a declarative sentence, and rise for yes-no questions. The pitch contour shape at the phrase-final word of an intermediate phrase has a rising trend. Prominent words too have pitch accent which is usually visible through a peaky shape on the word. Sometimes, when short-duration words are emphasized, it is difficult to realize the prominence cue aligned within the word boundaries, which gives rise to prominence regions spreading across words. In such cases, the pitch contour for a prominent word may take just a rising or falling shape while the peaky shape appears across a contextual window. We exploit the relationship of prosodic events with the pitch contour shape by correlating the F0 contour shape with simple ideal contour shapes.

1. F0 contour shape is represented by the likelihood of the ideal contour shapes: rising, falling, peak and valley (Tamburini, 2003). We also consider Gaussian shaped peaky contours with different variances (0.2, 0.5, 1.0, 2.0 and 5.0) (Wang and Narayanan, 2007).
2. F0 contour shape likelihoods are also computed for the contours across the longer temporal context of 3-word windows (+/-1 words).
3. Phrase breaks usually produce a steep slope in F0 contour on the phrase-final word. Therefore, an 'ending slope' feature is computed as the slope of F0 contour across last 100 ms voiced interval of the word.
4. The high ending slope can be due to a phrase break and/or as a part of prominence peak. This is especially strong when a phrase final word is also prominent. To discriminate the two events, an additional 'pitch reset' feature is computed as the difference between average F0 value across 100 ms long voiced region at the end of the word and that across 100 ms long voiced region at the start of the next word.

Besides, F0 shape features also include F0 slope across word as well as across each of the 8 temporal context windows. The word-level statistical functionals (min, max, mean, median, s.d. and span) are also computed for the instantaneous F0 slope contours and their SN versions along with the temporal context normalized features similar to AuToBI.

- Spectral Shape Features - The percept of loudness is linked to both increased intensity as well as a relative increase in the higher frequency harmonics from the more rapid glottal closure. The spectral shape change is captured in the contours of the various spectral bands. Statistical functionals (min, max, mean, median, s.d. and span) are computed for each of the spectral bands as well as for the spectral tilt contours along with the z-score normalization across temporal context similar to AuToBI. Several distinct divisions of the spectrum are investigated as listed in Section 4.1.4. The slope of line fit of the corresponding contour across the word as well as across the different temporal context windows is also incorporated.

Those features computed purely from word segments without considering further temporal context are augmented by the corresponding difference features with respect to the adjacent word. The proposed feature set, common for both prosodic event detection tasks, eventually comprises *2524* features across the different attributes and types of normalization and temporal context. We next present the feature selection process applied to eliminate the redundancies in this large set of features with minimal loss of predictive power.

4.3. Feature Selection and Classification

Our approach to feature selection is a combination of knowledge- and data-driven methods. The acoustic correlates of each of the main suprasegmental attributes, viz. duration, pitch and loudness can be viewed in terms of speech production as the independently controlled characteristics of the speech. Acoustic features related to a specific attribute are more likely to be highly correlated with each other rather than with features representing a different attribute. This is especially true in our case of feature construction where several statistical functionals of the same acoustic contour are derived for a given segment, again with normalization in several overlapping temporal contexts. We therefore group features together based on the underlying acoustic contour and the associated perceptual property to define the five feature groups: pauses, duration, pitch, intensity and spectral shape. We process each 'feature group' individually to derive a compact and non-redundant set of features representing the particular suprasegmental attribute. This is achieved by removing highly correlated features based on cross-validation prediction performance on the training set. 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. This two-stage feature selection was observed to provide better generalization of the trained model over purely data-driven feature selection from the large combined set of features.

Table 3: Description of the data set with three splits specified for the speaker-independent train-test cross-validation experiments

Data set	Phrase Boundary			Prominence		
Split	#Speakers	#Utterances	#Words	#Speakers	#Utterances	#Words
Full	35	791	41326	35	808	42138
Fold 1	11	270	14118	11	285	14837
Fold 2	11	268	14012	11	270	14105
Fold 3	13	253	13196	13	253	13196

As reviewed in Section 3.2.2, the data labeled for phrase boundary prediction has 41,326 words from 791 recordings by 35 speakers and that for prominent word prediction has 42,138 words from 808 recordings by 35 speakers. Every word was labeled by each of 7 raters for the presence/absence of a boundary (following the word) and presence/absence of prominence on the word. We thus have 7 votes per word for each of the two detection tasks. Rather than using binary labels as targets for model training, we prefer to work with estimates of 'degree of prominence' (or 'degree of boundary') obtained directly from the counting of the votes per word. This allows us the flexibility to investigate various choices for binary labeled targets at a later stage. We thus set up our system for the prediction of the degree per word with the target labels ranging from 0 to 7, depending on the number of votes received for the presence of the corresponding event. The trained model can be applied to the detection of binary labels by selecting a suitable threshold for the predicted value.

The available data is split into 3 folds with no speaker overlap as shown in Table 3 ensuring the similar distribution of the target labels in the different folds that further matches the distribution across complete available data set as provided in Table 1. The training data set is created out of two of the three folds with the third fold serving as the held-out test data. The training data set is further partitioned into 4 non-overlapping speaker subsets, again maintaining target label distribution, for the 4-fold cross-validation exercise during feature selection and model tuning. As mentioned previously, we adopt the random forest ensemble model for the prediction of the target labels.

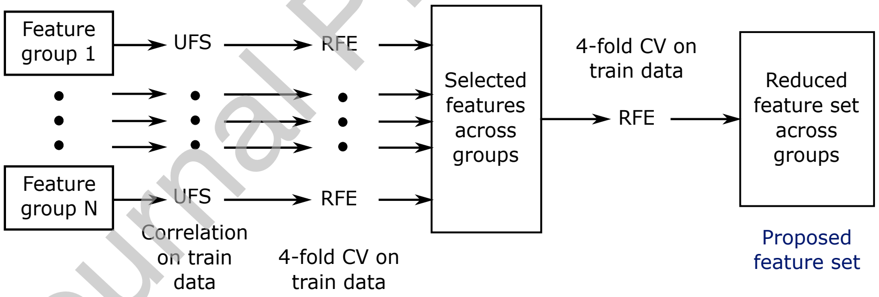


Figure 4: Feature selection procedure with univariate selection (UFS) followed by two stages of recursive feature elimination (RFE)

The feature selection algorithm as implemented on the train data set is illustrated in Figure 4. First, a stage of feature pruning serves to retain only those features that show a strong enough relationship with the output variable by means of a statistical test applied to the individual features on the training data. For this step of univariate feature selection (UFS), we compute the Spearman's rank-order correlation that measures the strength of the monotonic relation between the feature and the target variable (number of votes for event presence) without any assumptions on the nature of the relationship. Features showing absolute correlation value (r) less than 0.15, or showing correlation with insufficient significance ($p > 0.005$) are discarded. Next, the feature-feature similarity within each feature group is determined using Pearson's correlation which captures the extent of a linear relationship between the two variables. If two features show a Pearson's coefficient higher than 0.95, the feature exhibiting the lower Spearman's correlation with the target is discarded.

The retained features from the UFS are next processed to identify the optimal subset of features in terms of maximizing the model's performance as determined in cross-validation testing within the train set. Recursive Feature Elimination (RFE) (scikit-learn, 2019) is a procedure that embeds the desired model as well as the actual task performance criterion. One feature is eliminated at a time recursively, where the eliminated feature is the one ranked lowest in importance at that point in the recursion. We use the random forest regressor with each feature group as input to derive the performance measure versus number of features in 4-fold CV on the training set. The mean squared error (MSE) is used as the loss function and R^2 score (coefficient of determination) as the performance criterion for RFE. The final number of features selected is that beyond which the R^2 score does not improve by more than 0.001 with every added feature. The random forest regressor is trained with a default set of hyperparameters: 1000 trees, each trained on 15000 samples using all the features.

As shown in Figure 4, the selected features from each of the feature groups are next combined and subjected to another pass of the RFE in 4-fold CV on the train set. This stage is expected to reduce the extent of strongly covarying features across groups that are therefore not necessarily informative together. Starting with the 2524 features defined in Section 4.2.2, we obtain an overall reduction in the number of features by a factor of about 3 with the UFS stage and then again by another factor of about 30 in the two succeeding stages of RFE. The final set of features, termed the proposed feature set in Figure 4 is obtained for each of the two tasks, boundary and prominence prediction. Table 4 and Table 6 provide the number and description of features eventually retained in each feature group together with its 4-fold CV performance in terms of the Pearson's correlation between predicted and target variables.

5. Evaluation of Prosodic Event Detection

Our goal is to compare the performance of the proposed feature set with that of the baseline feature set on each of boundary and prominence detection on our data set. This is achieved by training the random forest regressor model with each of the two feature sets separately for each task. In each case, first, a stage of model hyperparameter tuning is carried out on the training set using 4-fold cross-validation. Next the tuned model is fitted to the entire train split and its performance is evaluated on the corresponding test split. The process is repeated for the remaining two train-test splits (all drawn from the 3 fold splits specified in Table 3) to obtain mean and standard deviation across the 3 trials for the Pearson correlation between the predicted and target labels. We recall that the target label is given by the actual number of votes for event presence obtained for the word. We also report the best F-score in the similar manner for each of the distinct binarizations of the target labels. The different binarizations correspond to assigning the event present/absent label based on different thresholds applied to the actual number of votes. Given the imbalanced nature of the labeled data set inherent to our task, the F-score is a more relevant indicator of detection performance compared to classification accuracy, capturing instead the balance between the recall and precision of the system. We start by reviewing the features selected within each feature group. We also estimate predictability for the individual feature groups for insights on the relative contributions of the different suprasegmental attributes in the realization of perceived boundary and prominence in our data set.

5.1. *Phrase Boundary Detection*

The feature selection within and across feature groups, as presented in Section 4.3 results in the reduced set of 28 features listed in Table 4 with a few overlaps with the AuToBI baseline set. We note that all the suprasegmental attributes are represented in the selected feature set. We observe that the individual feature groups contain sufficient information to produce highly correlated predictions with reference to the target label on their own. With intensity and spectral shape forming the acoustic correlates of loudness, we see the corresponding features produce the highest correlations. The same is true about F0 (including F0 contour shape) features, the acoustic correlate of pitch. Relative changes in the features between the current word and the next word directly capture the acoustic reset that marks a phrase boundary and are among the selected features. The F0 contour across the word in terms of its rising or falling shape as well as its actual slope also play a role. The spectral shape features that matter most come from the high band in terms of its variation across the word as well as mid band energy in terms of difference with the succeeding word given that an increase in loudness is expected across a boundary. Pause features are also highly predictive of phrase boundary with speaker normalization turning out to play an important role in their effectiveness. This is clearly different from the baseline feature set representation of pause by merely its absolute duration. Further local variation in pause duration by the differences in pre- and post-word occurrences is important. The normalization of segmental durations by local speech rate estimates appears critical as seen from the selected features. Although an increase in word duration is known to be a characteristic of phrase finality, we see a weaker dependence of boundary on duration cues in our data set.

Table 4: The proposed set of features for phrase boundary detection with performance evaluated feature group wise in terms of prediction correlation with target labels. The features common with AuToBI baseline system are indicated in plain text while new features appear in bold

Sr.No.	Feature Group (No. of Features)	Correlation	Selected Features
1	Pause (3)	0.60	**Difference between SN post-word pauses of current and next word; Difference between pre-word and post-word pauses (both with/without SN)**
2	Duration (3)	0.42	**Max syllable duration SN by articulation rate; Mean phone duration context normalized by phone rate; Difference in longest vowel duration within a word with that of the previous word, both SN by articulation rate**
3	F0 (4) F0 Contour Shape(6)	0.68	Word-level F0 functionals (mean, min) context normalized and as differences with next word's corresponding functionals **Differences between current and next word in instantaneous slope means, rising and falling likelihoods of word F0 contour; Slope of line fitted to F0 contour of word and difference from its next word; Word-level max of SN instantaneous F0 slope contour**
4	Intensity (6)	0.69	Word-level intensity functionals (min, span, slope) in context and the same as differences with next word
5	Spectral Shape (5)	0.59	**Word-level span for energy in band 5kHz to 8kHz; Word-level min of spectral tilt; Difference of band energy for 1kHz to 2kHz between current and next word; Difference in across-word fitted slope of the same between current and next word**

Table 5: Average performance (correlation and F-score) for phrase boundary detection across three folds. Train data performance is in 4-fold cross validation mode. F-score is computed for different methods for binary labeling of the target. All reported values exhibit s.d.< 0.01 across the folds.

Sr. No.	Feature Group (No. of Features)	Train Data				Test Data			
		Correlation	F-score			Correlation	F-score		
			$\geqslant$2 Votes	$\geqslant$3 Votes	$\geqslant$4 Votes		$\geqslant$2 Votes	$\geqslant$3 Votes	$\geqslant$4 Votes
1	AuToBI (83)	0.82	0.77	0.76	0.76	0.83	0.78	0.76	0.75
2	Proposed (28)	0.85	0.79	0.78	0.77	0.85	0.80	0.79	0.78

The performance of the overall proposed feature set is compared with that of the baseline features for boundary prediction in Table 5. We report performance on, both, the train data set in 4-fold cross-validation mode and performance on the held-out test data. We see that the proposed feature set provides a performance that either meets or exceeds that of the baseline features at a significantly lower input feature vector dimensionality. In terms of F-scores, the average across 3 test splits obtained with the best definition (in terms of performance) for recall and precision are 0.77, 0.79 for the baseline set and 0.79, 0.81 in the case of the proposed features. We see that both the precision and recall are improved over the baseline features.

Comparing the Tables 4 and 5, we also note that combining features across attributes boosts the correlation coefficient significantly (to 0.82) over that achieved with any single feature group in Table 4. This underlines the complementarity of the features computed from distinct suprasegmental attributes achieved by our process of feature selection.

5.2. *Prominent Word detection*

The final set of features selected from each feature group are enlisted in Table 6. We eventually have 34 features per word for prominence detection, several of which are not represented in the baseline feature set. The regression classifier is trained on the selected features from each feature group. The corresponding 4-fold cross-validation performance is reported in Table 6. Performance values are reported in terms of the Pearson correlation between the ground truth number of event presence votes and predicted output. We see that the independent contributions of the suprasegmental attributes is nearly uniform except for the pause features which do not surface in the final set. The pause features were motivated by the observations in previous literature related to the focus realization in Marathi, the native tongue of our speakers (Rao et al., 2017). However this does not appear to play a role in the current context of children's spoken English.

We observe that changes in acoustics with reference to the previous word are important in word prominence prediction, as we can see in the case of F0, intensity as well as segmental duration. This can be contrasted to boundary detection where the word is compared with its succeeding neighbour. The z-score normalizations computed across temporal context, however, involve the succeeding words. This is similar to the finding by Evin et al. (2018) that right context is more representative of prominence than left context although both combined gives even better results. We find that the F0 contour serves in both Hz and semitone units. The maximum and the span of F0 over the word or context window are the most effective F0 functionals. Similarly the variation of intensity over the course of the word as well as word intensity functionals relative to the surrounding context turn out to be important. It is interesting to note that nearly all spectral regions surface in one or other feature representing the spectral shape; this is in contrast to the baseline representation that is restricted to the relative energy of the specific band above 200 Hz. Both the span across the word and the slope are selected features. The change in spectral energy distribution going to the next word is strongly represented in the feature set. This can possibly be related to a drop in loudness post the prominence. The rate of variation of energy in the formant band (400 Hz to 2 kHz) comes up as well. The durations of subword units (syllable, phones) clearly help cue word prominence. The duration features benefit from both speaker normalization as well as that with reference to the local tempo estimated in pause-separated segments.

Table 6: The proposed set of features for prominence detection with performance evaluated feature group wise in terms of prediction correlation with target labels. The features common with AuToBI baseline system are indicated in plain text while new features appear in bold

Sr.No.	Feature Group (No. of Features)	Correlation	Selected Features
1	Pause	-	-
2	Duration (5)	0.55	**Max-min of SN vowel durations in the word normalized by local tempo in pause-separated window; Max of vowel duration, Max of syllable duration and its difference with that of previous word, Max-min difference for phone duration with that of previous word, all SN by phone rate**
3	F0 (8) F0 Contour Shape(3)	0.61	Word-level F0 functionals (min, max, span) in context; **max and span as the differences with previous word** **Slope of line fitted to semitone F0 contour of word; Similarity of word contour to valley**
4	Intensity (6)	0.52	Word-level intensity functionals (mean, span) in context and **the same as differences with previous word; Slope of line fitted to energy contour across context window**
5	Spectral Shape(12)	0.59	**Mean for bands: 0 to 500Hz and 400Hz to 2kHz; Span across temporal context for intensity in specific bands: 2kHz to 5kHz, 5kHz to 8kHz; Slope of band 400Hz to 2kHz intensity contour across word; Difference of specific band energy between current and next word (bands 0 to 500 Hz, 60 Hz to 400 Hz, 1kHz to 2kHz, 2kHz to 4kHz, 5kHz to 8kHz)**

The performance of the overall proposed feature set is compared with that of the baseline features in Table 7. The table presents performance on, both, the train data set in 4-fold cross-validation and on the held-out test data. We see that the proposed feature set provides a performance that either meets or exceeds that of the baseline features at a significantly lower input feature vector dimensionality, as in the case of boundary detection. We also note that combining features across attributes boosts the correlation coefficient over that achieved with any single feature group in Table 6, although this effect is less strong compared to that observed for boundary detection. We also see a large dependence of detection F-score on changing ground-truth definitions. With event presence defined as prominence voted by at least 2 raters (out of the 7), we obtain the highest detection performance while the definition based on majority votes obtains the least. The average across 3 test splits obtained with the best definition (in terms of performance) for recall and precision are 0.86, 0.65 for the baseline set and 0.86, 0.69 in the case of the proposed features. We see that the precision is improved over that with the baseline features.

Table 7: Average performance (correlation and F-score) for prominence detection across three folds. Train data performance is in 4-fold cross validation mode. F-score is computed for different methods for binary labeling of the target. All reported values exhibit s.d.< 0.01 across the folds.

Sr. No.	Feature Group (No. of Features)	Train Data				Test Data			
		Correlation	F-score			Correlation	F-score		
			⩾2 Votes	⩾3 Votes	⩾4 Votes		⩾2 Votes	⩾3 Votes	⩾4 Votes
1	AuToBI (210)	0.66	0.75	0.62	0.48	0.66	0.75	0.61	0.48
2	Proposed (34)	0.69	0.76	0.63	0.49	0.69	0.76	0.63	0.49

6. Investigating Speaker Variability

As mentioned in Section 2, speech prosody is characterized by significant speaker and language dependence. We expect this to be particularly true in our task given the diversity of oral skills arising from the L2 language learner context and the target population of children. While speaker adaptation is commonly used in automatic speech recognition, this has not been extended to prosody detection to the best of our knowledge. In this section, we attempt to use the rich information available to us via the regression model trained on our data set to obtain insights on the relevant speaker characteristics that could influence an individual's prosody realization. The random forest ensemble (of decision trees) allows us to assess the relative importance of individual features with respect to the predictability of the target variable through the relative depth of the feature used at a decision node in a tree. The relative importances of the different suprasegmental attributes are examined for speaker dependence and the possible grouping of speakers.

6.1. Speaker Specific Observations

To observe the importance or ranking of the distinct suprasegmental attributes in the acoustic cueing of prosodic events by a given speaker, we must train our regression model on the individual speaker's data. For this exercise, we consider only those speakers who have a comparatively large number of recordings. We have, in our data set, 23 speakers with at least 10 recordings (i.e. 34921 words in total for phrase boundary and 35643 words for prominence) each. We trained the random forest regressor model with the proposed feature set on each speaker's entire set of utterances for each of boundary and prominence prediction. The feature importance values, read off the trained model, sum to 1.0 across the input features. We combine features belonging to each of the groups, as specified in Table 4 and Table 6 for phrase boundary and prominence respectively, to obtain the relative importances of the different feature groups. The feature group importances for boundary and prominence prediction are thus derived separately for each of the 23 speakers. The 5-dimensional vector of importances (values ranging between 0 and 1) tells us about the relative strengths of the different suprasegmental attributes in the acoustic realization of the associated prosodic event for the given speaker. We compare the importance vectors across speakers with the gray scale images shown in Figure 5 and Figure 6 for phrase boundary and prominent word detection respectively.

An immediate observation is the strong variation across speakers in terms of the most dominant feature group. For phrase boundary in Figure 5, nearly all feature groups play some role, with pause features dominating several speakers' importance profiles. A few speakers show a complete absence of usage for pause features with importances distributed instead over the remaining groups, most notably the pitch group. Pause and pitch features appear to trade-off with speakers clearly preferring one over the other as seen in the complementary behavior of the corresponding rows of Figure 5. Intensity and spectral shape, the acoustic correlates of loudness, have a relatively uniform presence across speakers which is consistent with our previous observations on their contribution to the speaker-independent prediction of phrase boundary in Table 4.

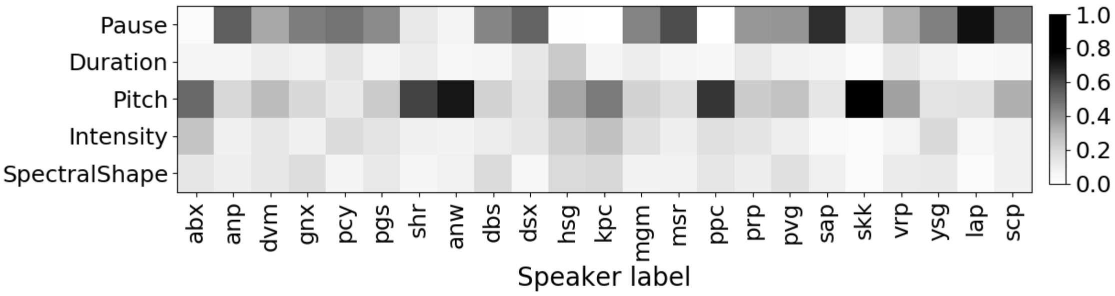


Figure 5: Profile of attribute-wise feature importance for boundary, for each of the 23 speakers, in gray scale with black corresponding to 1.0

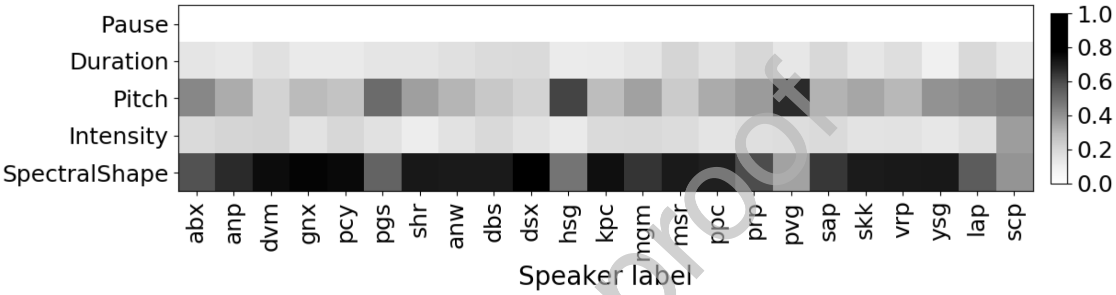


Figure 6: Profile of attribute-wise feature importance for prominence, for each of the 23 speakers, in gray scale with black corresponding to 1.0

In the case of prominent word prediction in Figure 6, we see that spectral shape and pitch are the two most dominant feature groups across speakers, again consistent with the performances reported in Table 6 for the speaker-independent prediction of prominence. While spectral shape features are exploited by most speakers, the importance of pitch features is more variable with a few speakers trading off spectral shape for pitch. Duration features are used by about half the speakers while pause features do not play a role in cueing prominence across speakers.

6.2. Speaker Grouping

To understand better the possible structure of the prosody features' space as suggested by the shared feature importance profiles within groups of speakers in Figures 5 and 6, we apply statistical clustering techniques. We choose the hierarchical clustering method of agglomerative clustering (Day and Edelsbrunner, 1984) which starts out with each data point (one speaker, in our case) as a cluster, successively merging clusters based on the Euclidean distance between the importance profile vectors until a specified number of clusters remains. We study the cases obtaining 4, 3 and 2 clusters out of the 23 speaker vectors. We evaluate cluster quality with the silhouette score, which quantifies intra-cluster distance (similarity of a sample with the samples within its cluster) with reference to inter-cluster distance (similarity of a sample with the samples of other clusters) (Kaufman and Rousseeuw, 1990). With values in the range -1 to +1, a higher positive value indicates better separated clusters. We see that the silhouette score is highest for the 2-cluster configuration as shown in Table 8 with a skewed distribution of speakers. Increasing the number of clusters leads to the further splitting of each of the two clusters with the smaller cluster splitting first, and with lowered separation quality.

We examine the case of two clusters with its higher silhouette scores for both the boundary and prominence detection tasks by plotting the obtained cluster centroids in Figure 7. We now see very clearly the distinct characteristics of the clusters. In the boundary detection task, we have speakers who predominantly use pitch as opposed to the larger majority who use pauses mainly, as seen in Figure 7a. An informal listening exercise comparing speakers from the two clusters revealed that the Cluster 1 speakers, who made very little use of pause to signal boundaries, were relatively fast-paced in terms of their speaking rate. Such speakers naturally prefer pitch reset and F0 contour shape to indicate phrase boundary instead of inserting pauses between the words in their fast-paced delivery. The Cluster 2

Table 8: Silhoutte scores for different numbers of speaker clusters with cluster sizes and observable speaker characteristics for each prosodic event

Prosodic Event	No. of Clusters	Silhoutte Score	Cluster Sizes	Speaker Characteristics
Boundary	2	0.634	7,16	fast vs normal or slow pace
	3	0.582	3,4,16	
	4	0.375	3,4,7,9	
Prominence	2	0.549	6,17	periodic pitch modulation vs other
	3	0.483	2,4,17	
	4	0.299	2,4,7,10	

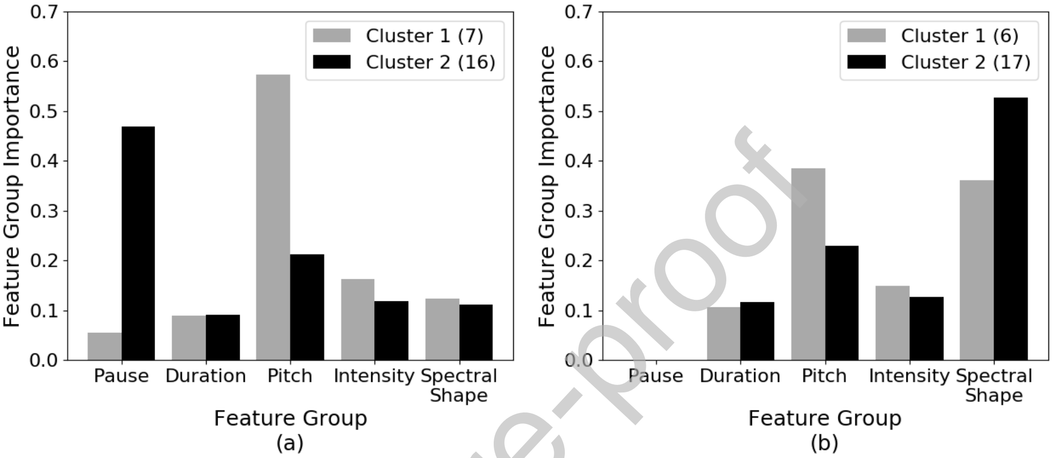


Figure 7: Feature group importance profiles for each of the two clusters for (a) phrase boundary and (b) prominence detection models

speakers were observed to have normal or slow speaking pace and while they relied more on pauses, they also used pitch features in the expected manner across the boundary.

The prominence detection clusters, as represented by the centroid profiles in Figure 7b, separate speakers who mainly employ spectral balance from the minority who use both pitch and spectral balance more or less equally. Listening to the different speakers revealed a variety of speaking styles that possibly underlie the variation in the acoustic cues to prominence. We have good speakers who use loudness, pitch and duration cues to clearly convey the meaning in the text. There are also speakers who sound monotonous in terms of rather restricted pitch and intensity variation apart from that used to signal prominent words. These speakers also manage to convey the meaning while reading but otherwise sound unenthusiastic or self-conscious (lacking confidence). We have, at another extreme, speakers with a sing-song style of regular and substantial variation of pitch (unrelated to the information structure of the text). This is common to children accustomed to learning their lessons by rote and reciting these without attention to the meaning. Such speakers are seen to belong to Cluster 1 for prominence realization. Their large and regular pitch swings are perceptually salient and lead to strongly perceived emphasis on the associated words by the raters. The accurate detection of the prominent words can potentially identify such learners for the necessary corrective feedback in line with our longer term goals for this work in overall reading assessment.

6.3. Speaker Group-based Modeling

Finally, we conduct an experiment to explore whether training the model separately for each speaker cluster helps performance for within-cluster speakers (in leave-one-speaker-out cross-validation mode). We compare the average performance of the cluster-based training with that obtained with a model trained purely on outside-cluster speakers. Given our data limitations, we restrict the experiment to a number of speakers per cluster matching that of the smaller cluster. The speaker groups and corresponding performance are presented in Table 9 and Table 10 respectively for phrase boundary detection and prominent word detection. The speakers were chosen from the larger cluster, in each case, by random sampling. We report the F-score for the majority vote for target label binarization.

We see some improvements in the performance with speaker-dependent training in the case of the smaller clusters. This encouraging result needs more thorough investigation with a much larger data set of speakers with different speaking styles. Further the results can be exploited by developing a speaker (or speaking style) representation based on characteristics such as, for example, speech rate and pitch modulation regularity that can serve in the speaker adaptation of the prosody predictor.

Table 9: Performance of phrase boundary detection with speaker grouping in terms of correlation (F-score in parantheses)

Cluster	#Speakers	#Utterances	#Words	Within Cluster	Outside Cluster
1	7	202	10626	0.82 (0.78)	0.80 (0.76)
2	7	211	11009	0.85 (0.76)	0.83 (0.76)

Table 10: Performance of prominence detection with speaker grouping in terms of correlation (F-score in parantheses)

Cluster	#Speakers	#Utterances	#Words	Within Cluster	Outside Cluster
1	6	145	7566	0.70 (0.54)	0.68 (0.53)
2	6	162	8383	0.69 (0.49)	0.69 (0.49)

7. Conclusions

We addressed the problem of automatic detection of phrase boundary and prominence in oral reading by children. The data set differs in many ways from those used in the similar task in previous research, viz. adult native speakers engaged in either reading or spontaneous speech, with prosodic labels supplied by trained annotators. Our speakers are middle school students in L2 English classes with reasonable word recognition skills but wide variability in their use of prosody. We employ the method of 'rapid prosody transcription' with several untrained annotators per word in our data set of 921 paragraph-sized recordings across 35 speakers. Our goal was to predict prosodic events based on the signal acoustics alone given the inconsistent adherence to the text syntax and semantics observed in our data set. We implemented and tested AuToBI features (Rosenberg, 2010) on our data set in a speaker-independent manner using a random forest ensemble predictor. We also propose a more compact set of word-level features obtained by applying model based feature selection on a very large set of acoustic features across the prosodic attributes of pitch, duration, intensity and spectral balance considering multiple ways of defining temporal context to reliably detect the local variations that cue phrase finality or prominence. While obtaining a speaker-independent performance that exceeds that of the baseline, we used speaker-dependent models to derive valuable insights on speaker specific characteristics by way of the model importance scores for the different suprasegmental attributes, potentially useful in speaker adaptation apart from their value to linguistic studies.

In terms of next steps, the currently semi-automatic pre-processing blocks of noise removal and word segmentation need to be automated using robust ASR trained on a large amount of representative data. The automatically detected phrase boundaries and prominent words can be compared with the expected prosodic events based on the syntax and semantics of the text to assess the speaker's comprehension (or equivalently, the comprehensibility of the speech), a very important attribute of good oral skills. The prediction of overall 'goodness' at the utterance level would depend on modeling the ratings of the same by an expert, such as a language teacher, obtained in a manner similar to our recent work in reading confidence prediction (Sabu and Rao, 2020).

References

Ananthakrishnan, S., Narayanan, S.S., 2008. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. IEEE Transactions on Audio, Speech, and Language Processing 16, 216–228. doi:10.1109/TASL.2007.907570.

Andreeva, B., Barry, W., Koreman, J., 2014. A cross-language corpus for studying the phonetics and phonology of prominence, in: Proceedings of International Conference on Language Resources and Evaluation, Reykjavik, Iceland. pp. 326–330.

Arnold, D., Möbius, B., Wagner, P., 2011. Comparing word and syllable prominence rated by naïve listeners, in: Proceedings of INTERSPEECH, Florence, Italy. pp. 1877–1880.

Arnold, D., Wagner, P., Möbius, B., 2012. Obtaining prominence judgments from naïve listeners - Influence of rating scales, linguistic levels and normalisation, in: Proceedings of INTERSPEECH, Portland, OR, USA. pp. 2394–2397.

Beckman, M.E., Elam, G.A., 1997. Guidelines for ToBI Labelling version 3.0. Technical Report. The Ohio State University Research Foundation.

Black, M.P., Bone, D., Skordilis, Z.I., Gupta, R., Xia, W., Papadopoulos, P., Chakravarthula, S.N., Xiao, B., Segbroeck, M.V., Kim, J., Georgiou, P.G., Narayanan, S.S., 2015. Automated evaluation of non-native english pronunciation quality: Combining knowledge- and data-driven features at multiple time scales, in: Proceedings of INTERSPEECH, Dresden, Germany. pp. 493–497.

Bock, J.K., Mazzella, J.R., 1983. Intonational marking of given and new information: Some consequences for comprehension. Memory & Cognition 11, 64–76. doi:10.3758/BF03197663.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in: Proceedings of the Institute of Phonetic Sciences, Amsterdam. pp. 97–110.

Bolaños, D., Cole, R.A., Ward, W.H., Tindal, G.A., Schwanenflugel, P.J., Kuhn, M.R., 2013. Automatic assessment of expressive oral reading. Speech Communication 55, 221–236. doi:10.1016/j.specom.2012.08.002.

Breen, M., Fedorenko, E., Wagner, M., Gibson, E., 2010. Acoustic correlates of information structure. Language and Cognitive Processes 25, 1044–1098. doi:10.1080/01690965.2010.504378.

Breen, M., Kaswer, L., Dyke, J.A.V., Krivokapić, J., Landi, N., 2016. Imitated prosodic fluency predicts reading comprehension ability in good and poor high school readers. Frontiers in Psychology 7. doi:10.3389/fpsyg.2016.01026.

Brenier, J.M., Cer, D.M., Jurafsky, D., 2005. The detection of emphatic words using acoustic and lexical features, in: Proceedings of INTERSPEECH, Lisbon, Portugal. pp. 3297–3300.

Brownlee, J., 2019. Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End.

Campbell, N., Beckman, M., 1997. Stress, prominence and spectral tilt, in: Proceedings of Intonation: Theory, Models, and Applications, Athens, Greece. pp. 67–70.

Chen, G., 1974. The pitch range of English and Chinese speakers. Journal of Chinese Linguistics 2, 159–171. URL: https://www.jstor.org/stable/23752908.

Chen, K., Hasegawa-Johnson, M., Cohen, A., 2004. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada. pp. I–509–I–512. doi:10.1109/ICASSP.2004.1326034.

Christodoulides, G., Avanzi, M., 2014. An evaluation of machine learning methods for prominence detection in french, in: Proceedings of INTERSPEECH, Singapore. pp. 116–119.

Christodoulides, G., Avanzi, M., Simon, A.C., 2017. Automatic labelling of prosodic prominence, phrasing and disfluencies in French speech by simulating the perception of naïve and expert listeners, in: Proceedings of INTERSPEECH, Stockholm, Sweden. doi:10.21437/Interspeech.2017.

Cole, J., Hualde, J.I., Smith, C.L., Eager, C., Mahrt, T., de Souza, R.N., 2019. Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. Journal of Phonetics 75, 113–147. doi:10.1016/j.wocn.2019.05.002.

Cole, J., Mahrt, T., Roy, J., 2017. Crowd-sourcing prosodic annotation. Computer Speech and Language 45, 300–325. doi:10.1016/j.csl.2017.02.008.

Cole, J., Mo, Y., Hasegawa-Johnson, M., 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. Laboratory Phonology 1, 425–452. doi:10.1515/labphon.2010.022.

Danne, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., Oranje, A., Goldstein, A., 2005. The Nation's Report Card: Fourth-Grade Students Reading Aloud: NAEP (The National Assessment of Educational Progress) 2002 Special Study of Oral Reading. Technical Report. National Center for Education Statistics, U.S. Department of Education.

Day, W.H.E., Edelsbrunner, H., 1984. Efficient algorithms for agglomerative hierarchical clustering methods. Journal of Classification 1, 7–24. doi:10.1007/BF01890115.

Dowhower, S.L., 1987. Effects of repeated reading on second-grade transitional readers' fluency and comprehension. Reading Research Quarterly 22, 389–406. doi:10.2307/747699.

Endress, A.D., Hauser, M.D., 2010. Word segmentation with universal proosdic cues. Cognitive Psychology 61, 177–199. doi:10.1016/j.cogpsych.2010.05.001.

Ericsson, C., 2003. Predicting Prosodic Phrase Boundaries for Speech Synthesis. Master's thesis. Kungiga Tekniska Hogskolan. Stockholm.

Escudero-Mancebo, D., González-Ferreras, C., Vivaracho-Pascual, C., Cardeñoso-Payo, V., 2014. A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling. Computer Speech and Language 28, 326–341. doi:10.1016/j.csl.2013.08.001.

Evin, D., Cossio-Mercado, C., Torres, H.M., Gurlekian, J., Mixdorff, H., 2018. Automatic prominence detection in Argentinian Spanish, in: Proceedings of Speech Prosody, Poznan, Poland. pp. 680–684. doi:10.21437/SpeechProsody.2018-138.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15, 3133–3181. doi:10.5555/2627435.2697065.

Féry, C., Pandey, P., Kentner, G., 2016. The prosody of focus and givenness in Hindi and Indian English. Studies in Language 40, 302–339. doi:10.1075/sl.40.2.02fer.

Gisev, N., Bell, J.S., Chen, T.F., 2013. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy 9, 330–338. doi:10.1016/j.sapharm.2012.04.004.

González-Ferreras, C., Escudero-Mancebo, D., Vivaracho-Pascual, C., Cardeñoso-Payo, V., 2012. Improving automatic classification of prosodic events by pairwise coupling. IEEE Transactions on Audio, Speech and Language Processing 20, 2045–2058. doi:10.1109/TASL.2012.2194284.

Groen, M.A., Veenendaal, N.J., Verhoeven, L., 2018. The role of prosody in reading comprehension: evidence from poor comprehenders. Journal of Research in Reading 42, 37–57. doi:10.1111/1467-9817.12133.

Hasegawa-Johnson, M., Chen, K., Cole, J., Borys, S., Kim, S.S., Cohen, A., Zhang, T., Choi, J.Y., Kim, H., Yoon, T., Chavarria, S., 2005. Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus. Speech Communication 46, 418–439. doi:10.1016/j.specom.2005.01.009.

Kakouros, S., Räsänen, O., 2014. Perception of sentence stress in English infant directed speech, in: Proceedings of INTERSPEECH, Singapore. pp. 1821–1825.

Kakouros, S., Räsänen, O., 2016. 3PRO - an unsupervised method for the automatic detection of sentence prominence in speech. Speech Communication 82, 67–84. doi:10.1016/j.specom.2016.06.004.

Kakouros, S., Räsänen, O., Alku, P., 2017. Evaluation of spectral tilt measures for sentence prominence under different noise conditions, in: Proceedings of INTERSPEECH, Stockholm, Sweden. pp. 3211–3215. doi:10.21437/Interspeech.2017-1237.

Kaufman, L., Rousseeuw, P.J., 1990. Finding groups in data: An introduction to cluster analysis. Wiley. Wiley Series in Probability and Statistics, p. 87.

Kochanski, G., Grabe, E., Coleman, J., Rosner, B., 2005. Loudness predicts prominence; fundamental frequency lends little. Journal of the Acoustical Society of America 118, 1038–1054. doi:10.1121/1.1923349.

Koreman, J., Andreeva, B., Barry, W.J., Sikveland, R.O., van Dommelen, W., 2009. Cross-language differences in the production of phrasal prominence in Norwegian and German, in: Proceedings of the Xth Conference of Nordic Prosody 2008, Peter Lang, Helsinki, Finland. pp. 139–150.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174. doi:10.2307/2529310.

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

Levow, G., 2005. Context in multi-lingual tone and pitch accent recognition, in: Proceedings of INTERSPEECH, Lisboa, Portugal. pp. 3301–3304.

Levow, G., 2009. Investigating pitch accent recognition in non-native speech, in: Proceedings of the ACL-IJCNLP, Suntec, Singapore. pp. 269–272.

Li, K., Mao, S., Li, X., Wu, Z., Meng, H., 2018. Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. Speech Communication 96, 28–36. doi:10.1016/j.specom.2017.11.003.

Li, K., Wu, X., Meng, H., 2017. Intonation classification for L2 English speech using multi-distribution deep neural networks. Computer Speech and Language 43, 18–33. doi:10.1016/j.csl.2016.11.006.

Li, X., Ju, Y., Deng, L., Acero, A., 2007. Efficient and robust language modeling in an automatic children's reading tutor system, in: Proceedings of ICASSP, Honolulu, Hawaii.

Liscombe, J., 2007. Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency. Ph.D. thesis. Columbia University.

Lochrin, M., Arciuli, J., Sharma, M., 2015. Assessing the relationship between prosody and reading outcomes in children using the PEPS-C. Scientifc Studies of Reading 19, 72–85. doi:10.1080/10888438.2014.976341.

Lunden, A., 2017. Duration, vowel quality, and the rhythmic pattern of English. Laboratory Phonology 8, 27. doi:10.5334/labphon.37.

van Maastricht, L., Zee, T., Krahmer, E., 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, Stockholm, Sweden. pp. 364–368. doi:10.21437/Interspeech.2017-1150.

Mahrt, T., Cole, J., Fleck, M., Hasegawa-Johnson, M., 2012. F0 and the perception of prominence, in: Proceesings of INTERSPEECH, Portland, OR, USA. pp. 2422–2425.

Maxwell, O., 2014. The Intonational Phonology of Indian English: An Autosegmental-Metrical Analysis Based on Bengali and Kannada English. Ph.D. thesis. School of Languages and Linguistics, The University of Melbourne.

McDowell, B., Settles, B., 2019. Duolingo - CEFR Checker. Available at: https://cefr.duolingo.com/. Accessed: July 28, 2020.

Miao, Q., Niu, X., Klabbers, E., van Santen, J., 2006. Effects of prosodic factors on spectral balance: Analysis and synthesis, in: Proceedings of Speech Prosody, Dresden, Germany. p. 107.

Miller, J., 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, 336–354. doi:10.1598/RRQ.43.4.2.

Mishra, T., Sridhar, V.R., Conkie, A., 2012. Word prominence detection using robust yet simple prosodic features, in: Proceedings of INTERSPEECH, Portland, OR, USA. pp. 1864–1867.

Moniz, H., Mata, A.I., Hirschberg, J., Batista, F., Rosenberg, A., Trancoso, I., 2014. Extending AuToBI to prominence detection in European Portuguese, in: Proceedings of Speech Prosody, Dublin, Ireland. pp. 280–284. doi:10.21437/SpeechProsody.2014.

Ostendorf, M., Veilleux, N., 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. Computation Linguistics 20, 27–54.

Paige, D., Rupley, W., Smith, G., T. Rasinski, W.N., Magpuri-Lavell, T., 2017. Is prosodic reading a strategy for comprehension? Journal for Educational Research 141, 245–275. doi:10.1177/1086296X14535170.

Patil, V.V., 2014. Automatic Classification of Obstruents of Marathi and Hindi. Ph.D. thesis. Department of Electrical Engineering, IIT Bombay, India.

Patil, V.V., Rao, P., 2016. Detection of phonemic aspiration for spoken Hindi pronunciation evaluation. Journal of Phonetics 54, 202–221. doi:10.1016/j.wocn.2015.11.001.

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.

Rao, P., Mixdorff, H., Deshpande, I., Sanghvi, N., Kshirsagar, S., 2016a. A quantitative study of focus shift in Marathi, in: Proceedings of Speech Prosody, Boston, USA. pp. 936–940. doi:10.21437/SpeechProsody.2016.

Rao, P., Sanghavi, N., Mixdorff, H., Sabu, K., 2017. Acoustic correlates of focus in Marathi: production and perception. Journal of Phonetics 65, 110–125. doi:10.1016/j.wocn.2017.07.002.

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

Rosenberg, A., 2009. Automatic Detection and Classification of Prosodic Events. Ph.D. thesis. Columbia University.

Rosenberg, A., 2010. AuToBI - A tool for automatic ToBI annotation, in: Proceedings of INTERSPEECH, Makuhari, Japan. pp. 146–149.

Rosenberg, A., Cooper, E., Levitan, R., Hirschberg, J., 2012. Cross-language prominence detection, in: Proceedings of Speech Prosody, Shanghai, China. pp. 278–281.

Rosenberg, A., Fernandez, R., Ramabhadran, B., 2015. Modeling phrasing and prominence using deep recurrent learning, in: Proceedings of

INTERSPEECH, Dresden, Germany. pp. 3066–3070.

Roy, J., Cole, J., Mahrt, T., 2017. Individual differences and patterns of convergence in prosody perception. Laboratory Phonology 8, 22. doi:10.5334/labphon.108.

Roy, S., 2017. A systematic review of Hindi prosody. arXiv:1705.03247.

Sabu, K., Rao, P., 2018a. Automatic assessment of children's oral reading using speech recognition and prosody modeling, in: CSI Transactions on ICT, pp. 221–225. doi:10.1007/s40012-018-0202-3.

Sabu, K., Rao, P., 2018b. Detection of prominent words in oral reading by children, in: Proceedings of Speech Prosody, Poznan, Poland. pp. 314–318. doi:10.21437/SpeechProsody.2018-64.

Sabu, K., Rao, P., 2020. Automatic prediction of confidence level from children's oral reading recordings, in: Proceedings of INTERSPEECH, Shanghai, China.

Sabu, K., Swarup, P., Tulsiani, H., Rao, P., 2017. Automatic assessment of children's L2 reading for accuracy and fluency, in: Proceedings of SLaTE, Stockholm, Sweden. pp. 121–126. doi:10.21437/SLaTE.2017-21.

Schwanenflugel, P.J., Hamilton, A.M., Wisenbaker, J.M., Kuhn, M.R., 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, 119–129. doi:10.1037/0022-0663.96.1.119.

scikit-learn, 2019. Feature selection – scikit-learn 0.23.1 documentation. Available at: http://scikit-learn.org/stable/modules/feature_selection.html#rfe. Accessed: July 28, 2020.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tür, G., 2000. Prosody-based automatic segmentation of speech into sentences and topics. Speech Communication 32, 127–154. doi:10.1016/S0167-6393(00)00028-5.

Sinambela, S.E., 2017. Prosody as a tool for assessing reading fluency of adult ESL students. Advances in Language and Literary Studies 8, 83–87. doi:10.7575/aiac.alls.v.8n.6p.83.

Sluijter, A.M.C., van Heuven, V.J., 1996. Spectral balance as an acoustic correlate of linguistic stress. Journal of the Acoustical Society of America 100, 2471–2485. doi:10.1121/1.417955.

Sridhar, V.K.R., Bangalore, S., 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, 797–811. doi:10.1109/TASL.2008.917071.

Stehwien, S., Schweitzer, A., Vu, N.T., 2020. Acoustic and temporal representations in convolutional neural network models of prosodic events. Speech Communication 125, 128–141. doi:10.1016/j.specom.2020.10.005.

Stemler, S.E., 2004. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation 9, 1–11. doi:10.7275/96jp-xz07.

Streefkerk, B.M., Pols, L.C.W., ten Bosch, L.F.M., 1999. Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's, in: Proceedings of EUROSPEECH, Budapest, Hungary. pp. 551–554.

Tamburini, F., 2003. Prosodic prominence detection in speech, in: Proceedings of International Symposium on Signal Processing and Its Applications, Paris, France. pp. 385–388. doi:10.1109/ISSPA.2003.1224721.

Tamburini, F., Bertini, C., Bertinetto, P.M., 2014. Prosodic prominence detection in Italian continuous speech using probabilistic graphical models, in: Proceedings of Speech Prosody, Dublin, Ireland. pp. 285–289. doi:10.21437/SpeechProsody.2014.

Tamburini, F., Wagner, P., 2007. On automatic prominence detection for German, in: Proceedings of INTERSPEECH, Antwerp, Belgium. pp. 1809–1812.

Wagner, P., Tamburini, F., Windmann, A., 2012. Robust tracking for automatic reading tutors, in: Proceedings of INTERSPEECH, Portland, OR, USA.

Wang, D., Narayanan, S.S., 2007. An acoustic measure for word prominence in spontaneous speech. IEEE Transactions on Audio, Speech, and Language Processing 15, 690–701. doi:10.1109/TASL.2006.881703.

Whalley, K., Hansen, J., 2006. The role of prosodic sensitivity in children's reading development. Journal of Research in Reading 29, 288–303. doi:10.1111/j.1467-9817.2006.00309.x.

Wightman, C.W., Ostendorf, M., 1994. Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing 2, 469–481. doi:10.1109/89.326607.

Xu, Y., Sun, X., 2002. Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America 111, 1399–1413. doi:10.1121/1.1445789.

Zhao, X., O'Shaughnessy, D., Minh-Quang, N., 2007. A processing method for pitch smoothing based on autocorrelation and cepstral F0 detection approaches, in: Proceedings of International Symposium on Signals, Systems and Electronics, Montreal, Canada. pp. 59–62. doi:10.1109/ISSSE.2007.4294413.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.