9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Detection of Prominent Words in Oral Reading by Children

Kamini Sabu, Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology Bombay, Mumbai, India
{kaminisabu,prao}@ee.iitb.ac.in

Abstract

The evaluation of oral reading skills is considered an important component of language education in school. Compared with word decoding skill, prosodic fluency typically takes children much longer to achieve. Prosodic fluency, however, is linked to comprehension making its evaluation very useful in an automatic reading assessment system. We consider the detection of prominent words in recordings of oral reading by children, who display good word recognition but varying degrees of prosodic fluency. The manual annotation of prominent words proves to be relatively challenging, likely due to inconsistencies by our speakers with respect to top-down lexical cues. Acoustic-prosodic features drawn from prominence detection research on adult speech are tested on the annotated data using a random forest classifier. Normalized maximum syllable duration and F0 (fundamental frequency) derived features turn out to be important predictors of word prominence with their relative importances being highly speaker dependent.
Index Terms: prominence, prosody evaluation, oral reading

1. Introduction

Oral reading assessment for children in early and middle school grades has typically involved measurements of word decoding accuracy and speaking rate. However, it is recognized that prosodic fluency is an important aspect of good reading, and further, that it is closely linked to comprehension ability [1],[2],[3]. A child who reads with proper phrasing (grouping of words) and prominence (using emphasis for new or important information) is engaged in constructing meaning from the text and reveals comprehension. Such suprasegmental aspects take longer to develop implying that there are children in middle and high school who display automaticity in word recognition but have still to attain prosodic fluency [3].

Phrase boundaries and prominence are signaled by the variation of acoustic attributes across an utterance [4]. The prosodic attributes can be categorized as based on one of F0 (i.e. fundamental frequency), duration, energy and spectral shape. Perceived phrase boundaries and prominence are also influenced by top-down structural cues, such as word identity and part-of-speech, apart from bottom-up acoustic features [5],[6].

The present work is a step towards implementing the rating of prosodic fluency for the automatic assessment of children's oral reading, a topic on which there has been much research over the years but focused chiefly on the detection of lexical miscues [7],[8],[9]. A very recent work [3] reported a study on the measured acoustic attributes of phrasing in an imitation task across older children with similar word decoding-skill but differing comprehension abilities. The better comprehenders were found to provide stronger durational cues to the syntactic structure. In our work, we target the evaluation of prominence in read text by children with adequate word recognition skills. Since the children have not necessarily achieved adult-like prosody, they may not employ acoustic-prosodic features of prominence reliably, apart from a possible lack of consistency with structural cues present in the text. We evaluate various acoustic-prosodic features for prominent word detection in read aloud text by children of different prosodic proficiencies.

We are not aware of any specific previous work on prominence detection in children's speech. In studies of the acoustic correlates of prominence in adult speech corpora, one or more of the following attributes, viz. F0, energy, duration and spectral balance form the basis of the contributing acoustic features [6],[10],[11],[12]. The language specific realization of prominence may vary, but the acoustic features usually come from the same set of attributes [13],[14],[15], as also in infant directed speech [16]. Breen et al. [17] reported that focused words by American English speakers are produced with longer duration, higher F0 and intensity. German speakers are found to use both intensity and pitch accents [18]. In our previous work on Marathi speech, we found that segment duration is the dominant local cue to focus in Indo-Aryan languages accompanied by post focal compression of F0 span [19].

The automatic detection of prominent words has been widely researched with focus on robust acoustic features computed from the measured prosodic attributes for specific test corpora. Features are calculated for individual syllables [11],[20], words [21],[22] or sub-part of words [23]. Temporal context plays a key role in prominence detection and has been achieved by considering features computed across adjacent word or syllable segments [5]. Similarly, features normalized across the intonational phrase are found to be more useful than unnormalized features [5]. Some other approaches like conditional random fields [11] and convolutional neural networks [22] embed temporal context implicitly.

The prominence detection task has been implemented in supervised, semi-supervised and unsupervised fashion, supervised often yielding the best results as expected [12]. Top-down lexical and syntactic cues have also been profitably incorporated in prominence detection tasks in corpora like the BU Radio corpus [24]. Different machine learning algorithms and classifiers are used on the combination of lexical and prosodic features such as Support Vector Machine (SVM) [6], Gaussian Mixture Model-Neural Network (GMM-NN) [23], Bidirectional Neural Network (BiRNN) [25], Probabilistic Graphical Model (PGM) like Conditional Random Field (CRF), Conditional Neural Field (CNF) [20]. Different combinations of features in different normalization conditions with different classifiers have been investigated [10],[11],[26].

In the next section, we present our dataset of children's oral reading and the manual annotation method and accompanying observations. This is followed by a review of the acoustic-prosodic features implemented and tested in this work in section 3. Prominence detection experiments using the more distinctive

10.21437/SpeechProsody.2018-64

features in a random forest classifier are presented in section 4. We conclude with a discussion of the automatic detection performance and its implications for automatic assessment of oral reading.

2. Dataset and Annotation

Recordings of oral reading of short stories by 20 children in Grades 5-7 were made using a headset mic in a quiet room at 16 kHz sampling frequency. Each story, presented as a printed text, was recorded in a separate session. The children were native speakers of Marathi, but had studied in schools with English as the language of instruction since Grade 1. The children, selected for their satisfactory word recognition ability based on a preliminary reading test, displayed various levels of prosodic fluency in terms of phrasing and prominence. The different kinds of prosody observed ranged from adult-like variation to the hyper-articulation of every word on one extreme and a rhythmic sing-song style at the other extreme. Although we did not carry out comprehension testing, it seemed obvious on listening that the latter two types of readers had relatively poor comprehension of the text.

Due to the required manual annotation, we picked a limited set of 6 speakers for the current study with data characteristics as shown in Table 1. The total speech duration for each speaker ranged from 160 sec to 570 sec with a single story reading session averaging 37 sec. We picked the top 3 speakers in Table 1 based on their larger data duration for the speaker-specific observations reported later. Of the 3 speakers, two had good phrasing in terms of grouping of words and sentence ending cues, while the third had poor phrasing and a rhythmic style of reading with alternating stress unrelated to the text syntax.

Table 1: *Details of speakers in our dataset*

Sr. No.	**Initials**	**No. of Stories (Words)**	**Speaking Rate**	**Phrasing**
1	AB	15 (1511)	Normal	Good
2	HS	10 (1047)	Fast	Good
3	VR	9 (893)	Slow	Poor
4	PP	5 (550)	Fast	Poor
5	PR	5 (521)	Fast	Good
6	SH	5 (508)	Normal	Good

Manual labeling of prominence is known to be difficult, so much so that some previous works have relied on part-of-speech measures to obtain the reference markings [6]. This is not an available option in our work since we certainly cannot rely on lexical or syntactic cues with our inexpert speakers. In fact, the possible conflicting cues from the top-down expectancies and bottom-up signal cues make the manual labeling task particularly challenging. We had 3 raters, all of whom rated all the data in Table 1. All three raters were fluent speakers of Indian English (only one of them had formally taught the language), but had not been exposed to prosody labeling before.

The raters were provided each story recital with the actual text transcript, but with all punctuation marks and capitalization removed to facilitate unbiased listening. They were asked to mark words which are perceived to be intentionally emphasized by the speaker. The raters also marked phrase boundaries but this is not used in the current work. Repeated listening during annotation was permitted. The raters found the boundary marking task easier and were seen to be more consistent with each other as compared to the prominence marking task. We observed that each individual rater marked 20% to 30% of the total words (5030), uttered across the 6 speakers, as prominent. Table 2 shows the percentage of words that received a given number of votes for prominence (out of a maximum possible of 3 votes). We also assign a gradient level of prominence to each word depending on the number of votes of prominence based on the consideration that prominence is not strictly binary but rather offers a few more linguistic distinctions [27],[28],[29]. Further, to obtain categorical ground-truth for the classification experiments, we consider a word as prominent if it receives one or more votes (similar to a strategy used in [12]). From Table 2, we see that this corresponds to about 45% of the total words.

Table 2: *Percentage distribution of words in terms of number of prominence votes obtained from three raters*

Prominence votes	0	1	2	3
Percentage of words	54.6	22.8	13.8	8.8

We compute the pair-wise inter-rater agreement in terms of Cohen's kappa [30]. The average agreement is found to be 0.36 as shown in Table 3, where we also report average inter-rater agreement separately for each of the 3 larger data speakers in Table 1. The lowest kappa is obtained on the rhythmic-style speaker while the other two obtain close to fair agreement ($\kappa >= 0.4$).

The manual comparison of the labels across raters for the same recital revealed that very often the marked words were shifted one place indicating that while the region of prominence in an utterance was clear, the precise word was ambiguous leading to different choices across raters. Compensating for this ambiguity improved the average inter-rater agreement from kappa of 0.36 to a substantially better 0.64. This could also possibly indicate the frequent use of "prosodic words" by the speakers where a function word is prosodically attached to an adjacent content word [31].

Table 3: *Average of pairwise inter-rater Cohen's kappa agreement for three raters. Overall value is across all six speakers*

Speakers	All	AB	HS	VR
Kappa	0.36	0.39	0.45	0.35

3. Acoustic-Prosodic Feature Extraction

A perception of prominence arises from surprise or unusualness in the local prosodic attributes in a region of the utterance. Thus word-level acoustic measures with suitable temporal context can potentially serve as features in a word prominence classifier. Further, normalizations for more global influences such as speaking rate, recording volume and F0 range are required.

Features derived from each of the different prosodic attribute classes, viz. duration, F0, energy and spectral shape, are computed for each word in the transcript-aligned audio recording. Since the speakers have good word decoding ability, the transcript corresponds closely with the presented text. The ocassional missed word is manually deleted. The forced alignment of the audio at the phone level is achieved with a state-of-the-art Automatic Speech Recognition (ASR) system. The ASR system uses an Indian English pronunciation dictionary and is trained on bilingual children reading English and Hindi texts, amounting to 5 hours of speech, to obtain 47 acoustic models

of the phones and silence fillers expected in the English speech of native Hindi or Marathi speakers. The system uses hybrid DNN-HMM acoustic models configured as described in [32].

3.1. Duration

The relative syllable duration is considered a relevant cue to word prominence [6],[11],[33]. A word is perceived as prominent when one of its syllables is elongated over the rest. Given the phone level alignment of the text, we obtain the duration of every syllable and number of syllables contained in a word using a manually constructed pronunciation dictionary. The average syllable duration in the word and duration of the longest and the shortest syllables in the word are calculated. The measured durations are normalized by average speech rate (number of syllables per second after deleting the pauses between words) across the story audio recording. The silence duration (in seconds) just preceding and following the word are also computed [11].

3.2. F0

F0 (in Hz) is computed at every 10 ms hop using the autocorrelation method in Praat [34] with the required post-processing to reduce the number of octave jumps and voiced-unvoiced transitions. The raw F0 values are linearly interpolated across the detected unvoiced frames, and further converted to the logarithmic scale of semitones or cents with a reference frequency of 50 Hz. Both raw and logarithmic F0 have been found useful in past work [28],[29]. We apply min-max normalization to [0,1] range on each of the F0 contours [16] across the story audio recording.

For a given word, the six statistical values, viz. mean, median, minimum, maximum, standard deviation and span (maximum - minimum) [11],[16],[35] are calculated for each of the two F0 contours over the word duration. The absolute differences in each of the statistical values between the current word and it's two neighbours are also computed in order to incorporate the immediately local temporal context.

3.3. F0 Contour Shape

Five Gaussian likelihood features correspond to correlation between interpolated F0 contour of current word and Gaussian shaped signal of same length with standard deviation 0.2, 0.5, 1.0, 2.0 and 5.0 [6]. Four likelihood features correspond to correlation between the raw F0 contour of current word and a signal of same length with rising, falling, peak, and valley shape respectively. The peak and valley points in these reference contours are decided by the peak and valley points in the word F0 contour. Similar features are also computed for the F0 contour over a 3-word (previous, current and successive word) interval.

3.4. Energy and Spectral Shape

Short-time spectra are obtained using a 20 ms Hamming window at 10 ms hop across the audio recording. The short-time full-band energy and energy in four distinct frequency bands, 0-0.5K (band 1), 0.5-1K (band 2), 1-2K (band 3), 2-4K (band 4) are computed and min-max normalised across the recording [11]. Further spectral tilt, implemented as the Mel Frequency Cepstral Coefficient of index 1 [16], is computed for each frame and min-max normalized. It represents the dominance of high frequencies over low frequencies in the signal.

Relative spectral energy in each band (energy in band divided by total energy) are used as features. Six statistical values, viz. mean, median, minimum, maximum, standard deviation and span [16], are computed across the word segment on the frame-level energy, spectral tilt and each of the four spectral band energy contours. The absolute difference in these statistical values of current word with previous and successive words are also computed to incorporate local temporal context.

We eventually have 111 features for every word; this set is subjected to feature selection as described in the next section.

Table 4: *Reduced feature set based on importance scores*

Attribute class	**Description**
Duration	Average, minimum and maximum syllable duration in word normalized by average speech rate (syllables per second) across story audio; Silence duration after the word (in sec)
F0	Span and standard deviation of F0 in semitone, maximum and median of F0; Difference between maximum, standard deviation and span of F0 values of current word and adjacent words; Difference between mean and median of current word and successive word, difference between min of current and previous word
F0 contour shape	Likelihood in the form of correlation of F0 contour shape with peak and valley contours and Gaussian contours of same length and variance 0.2, 1.0, 5.0; Correlation of 3-word (+/-1 word) F0 contour shape with peak and valley contours and Gaussian contour with variance 5.0
Energy and spectral shape	Minimum and span of energy, band3 energy, standard deviation, span and median of band1 energy, span and median of band3 and 4, minimum of band4; Maximum, standard deviation, span and median of spectral tilt, difference between mean and standard deviation values of spectral tilt of current and successive word

4. Classification Experiments

Given the large number of potential acoustic-prosodic features, as reviewed in the previous section, we choose a random forest classifier due to the associated Gini feature importance [36] of tree-based methods that can help determine the most rele-

Table 5: *Precision-recall values (%) for prominent word detection obtained with the reduced feature sets in a random forest classifier, and Pearson correlation (with $p < 0.0001$) of linear regression output with gradient level prominence*

Features	**Precision**	**Recall**	**Correlation (r)**
Duration	62.4	66.6	0.42
F0	64.5	68.5	0.42
F0 contour shape	56.2	60.6	0.19
Energy and spectral shape	64.6	68.6	0.43
All features	68.4	72.1	0.53

Table 6: *Top six features with feature importance score (in %) for overall data and for 3 individual speakers*

All speakers	AB	HS	VR
maxsyldurnorm (14.9)	F0semitonespan (15.4)	F0max (13.2)	maxsyldurnorm (30)
F0semitonespan (9.7)	maxsyldurnorm (5.7)	maxsyldurnorm (9.8)	stdF0diffafter (4.3)
F0max (4.6)	F0semitonestd (3.4)	F0semitonespan (8.5)	wordgaussian5.0 (4.2)
energyspan (3.4)	F0max (3.4)	spectralband4span (4.6)	F0semitonespan (3.4)
spectralband4span (4.6)	spectralband4span (3.4)	energyspan (3.2)	energymin (2.6)
stdF0diffAfter (3.1)	avgTiltdiffafter (3.1)	peaklikelihood (3.1)	spectralband4median(2.5)

vant features [11],[33],[37],[38]. The classifier is implemented using scikit-learn toolkit [39] in Python. Within each prosodic attribute category (i.e. duration, F0, F0 contour shape, and energy and spectrum shape) classification is performed on the 6-speaker data corpus of Table 1 using 5-fold cross-validation. Feature importance scores are obtained and only features with scores greater than 4% are retained to represent the specific prosodic attribute. The reduced set of 42 features, as shown in Table 4, is used in the binary classification of words which, as described in Sec. 2, are ground-truth labeled as prominent if any one of the raters assigned prominence. Further, the same set of features is used in the linear regresssion based prediction of prominence level of a word as determined by the number of raters labeling it as prominent.

Table 5 shows the classification results in terms of precision and recall of prominent words. It also shows the Pearson correlation coefficient between the predicted and ground-truth levels of prominence. We see the different performances obtained with features restricted to each single attribute category, as well as the overall superior performance obtained with the full set of features across all categories. We observe a positive correlation between the predictor output and the assigned word prominence levels.

Table 6 shows the top-ranked 6 features in terms of importance scores as obtained from the classifier using 5-fold cross validation on the full dataset, and then again separately on the data of each of the 3 larger data speakers of Table 1. On the all speakers dataset, we see that duration and F0 cues appear to have comparable importance in cuing prominence. Energy and spectral shape are also represented although with lower importance. However, the variability across the 3 speakers in terms of feature importance is notable. While speakers AB and HS share most of the top ranked features, the case is quite different for speaker VR. We recall that AB and HS have good phrasing skills and hence are more likely to show adult-like prominence. These two speakers use maximum syllable duration and a small set of F0 features (span and maximum chiefly) with the latter contributing more than the former. In the case of VR, there is a single highly dominant feature, namely maximum syllable duration. F0 cues play a relatively weak role. As noted in Sec. 2, this speaker has a reading style that sounds rhythmic; we can now relate this perception to a regular stressing pattern achieved by varying syllable duration and not much more.

5. Discussion and Conclusion

In this work, we considered the manual annotation, followed by the automatic detection, of prominent words in children's oral reading of short stories. With prosodic fluency serving as an important indicator of comprehension, a study of prominence can benefit methods used for reading assessment in school literacy programs. On a dataset of 5000 words by 6 middle school students of different prosodic proficiencies, but good word recognition skill, we studied the consistency in manual annotation for prominence by 3 naive raters who are good Indian English speakers. The raters marked on average 25 percent of the words prominent in the binary labeling task. While pair-wise inter-rater agreement (Cohen's kappa) for prominence marking task on adult speech corpora is usually reported to be in the range 0.57 to 0.88 [11],[40],[41], the average agreement in our case, however, is comparatively low, $\kappa = 0.36$ with slighly better agreement on the utterances by children with better phrasing. From the annotations, we observe that although the raters don't agree on the precise words, they still agree on the regions of prominence. By extending prominence labeling to adjacent words, a larger agreement, $\kappa = 0.65$ was obtained. This could be attributed to the formation of prosodic words or, possibly, to ambiguous phonetic realizations of prominence by our poor readers and inconsistencies with structural cues in our data. We eventually considered a word prominent if any rater marked it so, and also assigned a gradient level based on fraction of prominence ratings obtained across the 3 raters.

A large set of acoustic-prosodic features was considered for automatic detection of discrete prominence as well as of the gradient level in our dataset. Syllable duration and F0 cues appear to contribute most but the relative importances show a high speaker dependence. Prominent words were detected with precision=68% and recall=71%; we obtain a correlation of 0.53 between predicted and actual gradient levels. Precision-recall values in the range 64% to 87% have been reported for prominence estimation task in case of adult corpora [11],[12],[22], [26]. The correlation of prominence levels to duration, pitch and intensity are reported to be in the range 0.3 to 0.7 [28],[29]. The dataset used in this work is clearly very limited and meaningful comparisons with the aforementioned studies demand a much larger speaker set and larger set of raters exploiting more efficient mechanisms such as rapid rating for prosody [29].

Some interesting speaker-dependent characteristics were observed. A speaker with a rhythmic cadence was found to rely entirely on duration variation suggesting that specific prosody deficits may be identifiable through specific signal characteristics. Given that phrase boundary cues have been found to replace pitch accent cues in focused words in Indian English dialects [31], the acoustic correlates of phrase boundaries and the interaction of phrasing and prominence are also topics for future work. Finally, it is the prediction of goodness of reading rendition that is important. Such a system would use the proposed work with acoustic-prosodic features selected based on matching subjective ratings of goodness at the sentence level.

6. Acknowledgements

The authors acknowledge support by Ministry of Electronics and Information Technology, Govt. of India through Visvesvaraya scheme.

7. References

[1] R. Hudson, P. Pullen, H. Lane, and J. Torgesen, "The complex nature of reading fluency: A multidimensional view," *Reading & Writing Quarterly*, vol. 25, no. 1, pp. 4–32, 2008.

[2] T. Rasinski, A. Rikli, and S. Johnston, "Reading fluency: More than automaticity? more than a concern for the primary grades," *Literacy Research and Instruction*, vol. 48, pp. 350–361, 2009.

[3] M. Breen, L. Kaswer, J. V. Dyke, J. Krivokapic, and N. Landi, "Imitated prosodic fluency predicts reading comprehension ability in good and poor high school readers," *Frontiers of Psychology*, vol. 7, pp. 1–17, 2016.

[4] S. Shattuck-Hufnagel and A. Turk, "A prosody tutorial for investigators of auditory sentence processing," *Journal of Psycholinguistic Research*, vol. 25, no. 2, pp. 193–247, 1996.

[5] J. Brenier, D. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *Proceedings of INTERSPEECH*, Lisboa, Portugal, 2005.

[6] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 690–701, 2007.

[7] M. Black, J. Tepperman, S. Lee, and S. Narayanan, "Estimation of childrens reading ability by fusion of automatic pronunciation verification and fluency detection," in *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008.

[8] E. Yilmaz and J. Pelemans, "Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model," in *Proceedings of INTERSPEECH*, Singapore, 2014.

[9] J. Proenca, C. Lopes, M. Tjalve, A. Stolcke, S. Candeias, and F. Perdigao, "Automatic evaluation of reading aloud performance in children," *Speech Communication*, vol. 94, pp. 1–14, 2017.

[10] A. Rosenberg, E. Cooper, R. Levitan, and J. Hirschberg, "Cross-language prominence detection," in *Proceedings of Speech Prosody*, Shanghai, China, 2012.

[11] G. Christodoulides and M. Avanzi, "An evaluation of machine learning methods for prominence detection in french," in *Proceedings of INTERSPEECH*, Singapore, 2014.

[12] S. Kakouros and O. Rasanen, "3pro an unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[13] J. Koreman, B. Andreeva, W. Barry, R. Sikveland, and W. Dommelen, "Cross-language differences in the production of phrasal prominence in norwegian and german," in *Proceedings of Nordic Prosody*, Frankfurt, Germany, 2009.

[14] A. Endress and M. Hauser, "Word segmentation with universal proosdic cues," *Journal of Cognitive Psychology*, vol. 61, no. 2, pp. 177–199, 2010.

[15] B. Andreeva, W. Barry, and J. Koreman, "A cross-language corpus for studying the phonetics and phonology of prominence," in *Proceedings of the International Conference on Language resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014.

[16] S. Kakouros and O. Rasanen, "Perception of sentence stress in english infant directed speech," in *Proceedings of INTERSPEECH*, Singapore, 2014.

[17] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 1044–1098, 2010.

[18] F. Tamburini and P. Wagner, "On automatic prominence detection for german," in *Proceedings of INTERSPEECH*, Antwerp, Belgium, 2007.

[19] P. Rao, N. Sanghavi, H. Mixdroff, and K. Sabu, "Acoustic correlates of focus in Marathi: production and perception," *Journal of Phonetics*, pp. 110–125, 2017.

[20] F. Tamburini and C. B. P. Bertinetto, "Prosodic prominence detection in italian continuous speech using probabilistic graphical models," in *Proceedings of Spech Prosody*, Dublin, Ireland, 2014.

[21] K. Chen, M. H.-J. A. Cohen, and J. Cole, "A maximum likelihood prosody recognizer," in *Proceedings of Speech Prosody*, Nara, Japan, 2004.

[22] S. Stehwien and N. Vu, "Prosodic event recognition using convolutional neural networks with context information," *arXiv preprint arXiv:1706.00741*, 2017.

[23] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[24] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task dependent high level information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 1009–1024, 2009.

[25] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning," in *Proceedings of INTERSPEECH*, Dresden, Germany, 2015.

[26] H. Moniz, A. Mata, J. Hirschberg, F. Batista, A. Rosenberg, and I. Trancoso, "Extending autobi to prominence detection in european portuguese," in *Proceedings of Spech Prosody*, Dublin, Ireland, 2014.

[27] B. Streefkerk, L. Pols, and L. Bosch, "Acoustical features as predictors for prominence in read aloud dutch sentences used in ANN's," in *Proceedings of EUROSPEECH*, Budapest, Hungary, 1999.

[28] D. Arnold, B. Möbius, and P. Wagner, "Comparing word and syllable prominence rated by naïve listeners," *Proceedings of INTERSPEECH*, 2011.

[29] T. Mahrt, J. Cole, M. Fleck, and M. Hasegawa-Johnson, "F0 and the perception of prominence," in *Proceesings of INTERSPEECH*, Portland, USA, 2012.

[30] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, no. 1, pp. 159–174, 1977.

[31] C. Fery, P. Pandey, and G. Kentner, *The prosody of focus and givenness in Hindi and Indian English*. John Benjamins Publishing Company, 2016.

[32] K. Sabu, P. Swarup, H. Tulsiani, and P. Rao, "Automatic assessment of children's l2 reading for accuracy and fluency," in *Proceedings of SLaTE*, Stockholm, Sweden, 2017.

[33] D. Arnold, P. Wagner, and R. Baayen, "Using generalized additive models and random forests to model prosodic prominence in german," in *Proceedings of INTERSPEECH*, Lyon, France, 2013.

[34] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, pp. 341–345, 2001.

[35] Y. Ning, Wu, X. Lou, H. Meng, J. Jia, and L. Cai, "Using tilt for automatic emphasis detection with bayesian networks," in *Proceedings of INTERSPEECH*, Dresden, Germany, 2015.

[36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[37] "Feature selection – scikit-learn," http://scikit-learn.org/stable/modules/feature_selection.html, scikit-learn documentation.

[38] T. Mishra, V. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features," in *Proceedings of INTERSPEECH*, Portland, USA, 2012.

[39] "scikit-learn:mmachine learning in python," http://scikit-learn.org/stable/, scikit-learn.

[40] J. Buhmann, J. Caspers, V. Heuven, H. Hoekstra, J. Martens, and M. Swerts, "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the spoken dutch corpus," in *Proceedings of the third International Conference on Language resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, 2002.

[41] M. Avanzi, A. Simon, J.Goldman, and A. Auchlin, "C-prom. an annotated corpus for french prominence study," in *Proceedings of Conference on Speech Prosody*, Chicago, USA, 2010.