CNN Encoding of Acoustic Parameters for Prominence Detection

Kamini Sabu, Mithilesh Vaidya, Preeti Rao

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

Abstract

Expressive reading, considered the defining attribute of oral reading fluency, comprises the prosodic realization of phrasing and prominence. In the context of evaluating oral reading, it helps to establish the speaker's comprehension of the text. We consider a labeled dataset of children's reading recordings for the speaker-independent detection of prominent words using acoustic-prosodic and lexico-syntactic features. A previous well-tuned random forest ensemble predictor is replaced by an RNN sequence classifier to exploit potential context dependency across the longer utterance. Further, deep learning is applied to obtain word-level features from low-level acoustic contours of fundamental frequency, intensity and spectral shape in an end-to-end fashion. Performance comparisons are presented across the different feature types and across different feature learning architectures for prominent word prediction to draw insights wherever possible.

Index Terms: word prominence, children's speech, oral reading evaluation

1. Introduction

The prosodic structure of speech carries important information in terms of the syntax and the meaning, both of which are critical to a listener's ease of comprehension of the spoken message [1, 2, 3]. Phrase boundaries embed sentence syntax through word grouping while prominence or emphasis on specific words signals new information or highlights a contrast. In this paper, we investigate methods for the automatic detection of prominent words in the oral reading of middle-school children in the context of second language learning. The appropriate use of prosody is indicative of the reader's comprehension of the text and hence constitutes a critical component of oral reading evaluation systems [4, 5, 6]. In beginning readers, suprasegmental skills take longer to develop compared to word decoding ability, with phrasing coming earlier than the effective use of prominence.

Prominence is perceived by a listener when a word stands out of its local context in one or more of the suprasegmental attributes such as duration, F0, intensity and spectral shape [7]. The local context itself refers to the phones and syllables within the word as well as a neighborhood of up to several words. Prosody perception is influenced not only by the low-level acoustic cues but also top-down expectations from lexico-syntactic information [8, 9, 10]. The precise combination and relative importance of the cues depends on the speaker, language and speaking style as also on the listener. Various aggregates of the sampled acoustic parameters across the word segment including mean and variance, contour shape descriptors, and differences in these quantities across neighboring words comprise word-level prosodic features. These features are then used to train a conventional supervised classifier for the automatic detection, possibly in combination with lexico-syntactic information [11, 12, 13, 14]. In our own recent work, we applied systematic feature selection within and across the distinct suprasegmental attributes in a random forest ensemble predictor to derive a compact set of interpretable features for speaker-independent boundary and prominence detection on a children's oral reading dataset [15]. With the search space for such 'handcrafted' features being very large, however, the process can miss potentially important features. Further, the pre-selected context windows used in such analyses make it difficult to exploit the long and variable time scale of prosodic relationships across an utterance in any comprehensive manner. The potential for deep learning solutions has therefore been recognized for some time but incorporated successfully in the prominence detection task only more recently, as briefly reviewed next.

Rosenberg et al. [16] used a large number of acoustic-prosodic features and aggregates at word level derived from their previous work [11, 17] (which gave rise to the AuToBI tool) in a BiRNN classifier where the word sequence context was learned over that explicitly provided in the feature vector. They observed a small improvement ($< 1\%$ absolute) in boundary and pitch accent detection over a baseline conditional random forest classifier. Wu et al. [18] also used similar aggregated acoustic features in an LSTM to find an improvement over the use of an SVM classifier. Lin et al. [19] used a hierarchical BLSTM network to aggregate features across phone, syllable and word to model contextual information at multiple granularities in the joint detection of boundaries and prominence.

In a departure from pre-computed word-level features, a recent work [20, 21] uses CNN on frame-based acoustic parameters (energy, F0, loudness, voicing probability, zero crossing rate and harmonic-to-noise ratio) together with a context window of two neighboring words to optimally learn the high-level aggregation features. The max-pooled CNN feature maps are directly classified with a softmax layer. With word position indicators provided in the input segment, they report an improvement of 1-3% points absolute over Rosenberg [11] on lexical stress and phrase boundary detection on the BURNC corpus, with speaker-independent scenarios being more challenging. Zhang et al. [22] also use acoustic contours and MFCC features over 10 s segments as inputs to a CNN with fixed narrow kernel width of 3 frames, with syllable and word position indicators marked at the frame level. The CNN outputs go to a BLSTM classifier to obtain emphasis at frame level.

Given the significance of both local acoustic features and longer, more global, contexts spanning several words and possibly different sentences across the utterance in the perception of prominence, it is reasonable to consider architectures combining low-level feature aggregation with sequence models. The CNN learned feature representation of Stehwien et al. [20] was extended recently by Nielsen et al. [23] using full utterances as input and adding an LSTM layer to incorporate greater context.

An improvement in accuracy of 1% was noted for pitch accent detection on the BURNC corpus, with a further similar increase when they added text features by concatenating various dimension Glove word embeddings with the CNN embedding at the LSTM input.

Inspired by the above reviewed works, we investigate specific enhancements to our previously proposed random forest based prominence prediction for children's oral reading evaluation, with its highly optimized acoustic-prosodic word level features [15, 24]. The same dataset of children's read stories in English is used with its 42,138 words across 800 utterances by 35 speakers, recorded at 16 kHz sampling and manually transcribed at word level. The selected speakers have reasonable word decoding ability in English (as second language) but widely varying levels of prosodic skill. The individual utterances comprise between 50-70 words, each word labeled for the presence/absence of prominence by 7 naive listeners using RPT [25]. With a binary prominence decision based on 3 or more votes, we obtain a reasonable figure of 24% of the total words labeled prominent. We begin with replacing the random forest predictor with an RNN sequence model that can, in principle, capture implicit context dependence from across the utterance. Next, lexico-syntactic features based on the linguistic correlates of prominence are included, with new features related to the canonical structure of the text. Finally, motivated by the demonstrated potential of convolutional neural networks to learn discriminative patterns and thus replace any feature engineering, we investigate CNN architectures for our task in the end-to-end learning of prosodic word-level features from low-level acoustic parameters.

2. RNN-based prominence scoring

Using utterances segmented via forced alignment with the transcript, Sabu and Rao [15] obtained a compact and highly optimized set of word-level features by applying random forest model based feature selection on the children's speech dataset considered here. A very large set of acoustic-prosodic features defined across the different suprasegmental attributes of pitch, duration, intensity and spectral balance were considered, with multiple ways of defining temporal context in a ± 2 word neighborhood, to obtain a reduced set of 34 features. A Pearson correlation of 0.69 was obtained between the random forest regression output and the degree of prominence based on proportion of rater votes, and detection F-score of 0.63, in the speaker-independent prominent word detection task. In this section, we report work on using the same set of acoustic-prosodic features with an RNN model with its input given by the variable length sequence of words across an input utterance. We consider further the inclusion of lexico-syntactic information in the input sequence.

2.1. Architecture

We tried different RNN architectures: GRU [26] and LSTM [27] in both unidirectional and bidirectional configurations. The number of layers were varied in the range $\{1, 2, 3\}$ while the number of hidden units were picked from the range $\{48, 96, 128, 256, 512\}$. At each time-step, a feature vector corresponding to a single word is fed as input to the model. A final feed-forward layer linearly transforms the RNN output at each time step to a scalar, which is passed through a sigmoid layer to get the degree of prominence prediction.

2.2. Adding lexical features

The high predictive power of lexico-syntactic information, such as part-of-speech tags, has been noted for several corpora before [28, 29]. Baumann [10] investigated a wide range of non-prosodic factors for prominence prediction in German speech with random forest based feature selection. Lexical information included 12 PoS tags: NN (noun), NP (proper noun), JJ (adjective), RB (adverb), VB (verb), AU (auxiliary verb), MD (modal verb), PR (pronoun), IN (preposition), CC (conjunction), RP (particle) and DT (article). Assuming that this fine break-up may not be suited to our dataset of learners who do not necessarily have high levels of text comprehension, we also consider more coarse groupings of different dimensions. A one-hot encoding is used for the tags per word with the PoS itself determined from English Grammar rules [30] since automatic parsers showed poor performance. We also include the number of phones and syllables per word which indirectly relates to word frequency [10].

We further propose the canonical information structure (i.e. expected prominence and phrase boundaries) as additional useful features linked to top-down cues. We determine this by applying syntax and givenness rules [3]. Motivated by the observation that expected prosodic events depend on reading speed, the events are labeled mandatory, optional and forbidden. With no known NLP methods for the automatic extraction of this information from arbitrary text, we use the model reading of the story to validate our labels. With reading miscues, albeit few, being a part of our dataset, the PoS and information structure tags are based on the target word after achieving the automatic alignment of the transcript with the text. Other lexical features such as the number of phones and number of syllables in the word are estimated for the uttered word rather and inserted words are marked with a uniform PoS tag.

3. Learning word-level features with CNN

The word-level features of our previous work were computed from word (and sub-word) aligned contours corresponding to the time-varying acoustic parameters of F0, intensity and spectral shape, computed at 10 ms intervals [15]. Utterance based z-score normalization is applied on F0 and intensity parameters. We wish to investigate CNN-based automatic learning of word-level features from the same low-level acoustic contours. Given the previously observed speaker-dependence of the relative importances of the different prosodic attributes, we investigate a 3-channel CNN architecture where attribute-wise embeddings are computed with their own best filters and concatenated for the final representation [22]. The contour groups are F0 (4 contours), intensity (4 contours) and spectral shape including HNR and spectral band energies (7 contours) and each feature group is input to separate CNN filter bank as shown in Figure 1. The output encodings from the filter banks are finally concatenated and fed to the sequence classifier (GRU/LSTM) network together with other word-level features as considered in our experiments.

Perceived prominence depends on the underlying temporal dynamics while CNN computations are time-invariant. To match the information available to the hand-crafted feature extraction, the CNN input is a segment with the selected word context, possibly with word and subword positions within the segment provided explicitly. We explore a range of context choices as also different types of position information, starting with the context of $\pm$ 1 word given the importance of imme-

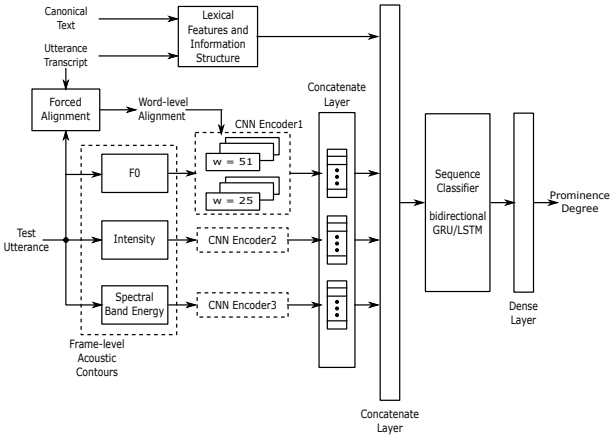

Figure 1: *Prominence degree prediction architecture.*

diately neighboring words in the realization of emphasis on a word [15, 21]. We also consider CNN filter banks with a range of kernel width choices motivated by sub-word units [31]. The 1D convolution output of each of the CNN filters is max-pooled across time to get a scalar value per filter per filter bank. Each filter bank has N filters, each with k kernel sizes, resulting in a kN-dimensional feature encoding, for each channel, corresponding to a word.

4. Experiments and results

4.1. Train-test splits and baseline

The random forest predictor based system reported in [15] serves as a baseline and we adopt the identical training and testing methodology here. The complete dataset of 42,138 words is split into three equal folds with no speaker overlap for 3-fold cross-validation based testing. The hyperparameters are tuned with 4-fold CV on the train split. The model is then trained on the entire train set and results are reported on the corresponding unseen test set. We report the mean and standard deviation across the three test folds. The results for the prominence degree prediction are reported in terms of Pearson correlation between the predictor output and the degree of prominence from the RPT rater votes. Also, the prominence detection F-score values are reported considering prominence present when a word receives 3 or more votes as discussed in Section 1.

4.2. RNN training and performance

We present the entire input utterance in the form of a sequence of word-level features to an RNN classifier. Various feature combinations are investigated to determine the individual and combined contributions of prosodic and lexical features. We test various RNN architectures as presented in Section 2. For training, we used the AdamW [32] optimizer, an improved version of popular Adam [33] optimizer with a weight decay mechanism that helps with faster training and more generalized models. We used a learning rate of 0.003 and a batch size of 500. Dropout with probability 0.2 is added to each RNN layer except the last layer. The Mean Squared Error between the scaled target score (between 0 and 1) and the predicted score is used as the loss function to be minimized during training.

Table 1 shows the performance with the 34 acoustic-prosodic features ('A34') used in the RNN architectures compared with the same features in the baseline random forest predictor. We note that there is an improvement from the utterance-long context available to the sequence models, especially in the case of the bidirectional models. We employ the BGRU going ahead due to it ease of training and known suitability for lower dataset sizes.

Table 1: *Performance of various models with set of 34 acoustic features. (* indicates sd $<$ 0.01)*

Model	# layers	# units	Correlation	F-score
RFC	-	-	0.69*	0.63*
GRU	2	96	0.68	0.63
LSTM	2	256	0.69	0.63
BGRU	2	96	0.70	0.64
BLSTM	2	256	0.71*	0.64*

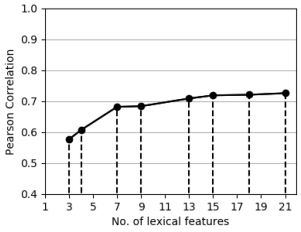

Figure 2: *Correlation performance variation with different sets of lexical features.*

Next, considering separately the lexical features, Figure 2 shows performance for various reduced sets starting from the full set of 21 that includes PoS tags and phone/syllable counts (together termed 'L'), and information structure labels (termed 'I'). From a maximum correlation of 0.72, we note a drop in performance as the 6 information structure and 2 word length features are removed to get to 13 features corresponding to the PoS alone. A further drop is recorded when the PoS tags are grouped in different ways to get to the final feature set of only content/function word distinctions. We note that lexical features alone show a predictive power similar to prosodic features alone (and not much higher as in some previous work [18, 29]), explained by the lower proficiency speakers of our dataset. All the same, the reduction of PoS tags clearly hurts performance. Table 2 summarizes the achieved performance gains as we augment the acoustic-prosodic feature set with lexical (all but information structure) and information structure features with each of the latter two clearly adding value.

Table 2: *Performance with addition of lexical and information structure features. (* indicates sd $<$ 0.01)*

Features	Correlation	F-score
A34	0.70	0.64
A34 + L	0.75*	0.67*
A34 + L + I	0.79*	0.69*

4.3. CNN training and performance

While our multi-channel CNN framework is similar to that of Zhang et al. [22], we expand the search for architecture choices by considering the use of multiple kernel widths in each channel to capture the distinct time scales of acoustic variation. We start

from the 4 kernels with widths [5, 11, 25, 51] similar to that of the sentence parsing CNN architecture of Trang et al. [31], which roughly cover sub-phone, phone, syllable and word, and possibly some context. Given the fixed narrow kernel width of 3 frames used in [20, 22], we add this to our candidates for testing. From the different combinations presented in Table 3, we observe that the syllable and word width kernel sizes (25, 51) helps the performance while including other widths does not change it. For these instances, we fixed the number of filters of each kernel size to 8. The input context to CNN is also fixed to $\pm$ 1 word with position encoding used to indicate current word based on [21]. We find that phone and sub-phone width kernels do not help, and even degrade the performance in some cases. We finalize the use of two kernel widths 25 frames and 51 frames, corresponding to syllable and word widths, for each CNN filter bank. Next, to find the optimal number of kernels for each width, we varied the number of CNN filters in the range $\{4, 8, 16, 20\}$. We found that 8 filters give the best performance and adopted this for all the further experiments.

Table 3: *Performance for different sets of kernel width choices with the corresponding lit. reference indicated.*

Kernel widths	Correlation	F-score
3,5,11,25,51	0.67	0.62
5,11,25,51 [31]	0.67	0.62
25,51	0.67*	0.62*
11,25,51	0.67*	0.62
11 [23]	0.65	0.61*

Next, we consider different choices for position encoding. Stehwien et al. [20] found it useful to indicate the frames corresponding to the word as distinct from the context frames from the neighboring words within the input segment. However, it has been observed that the preceding and following words influence perceived prominence in different ways [15]. Therefore, we decided to change the position encoding to distinguish all the three words. To test whether a special encoding to indicate inter-word pauses can help, we applied 5-bit one-hot encoding across the segment of 3 words and 2 pauses. The results are shown in Table 4. Taking position information to intra-word level, we explore one-hot encoding to indicate syllable number (up to a maximum of 7 syllables) in the word as indicated in last row of the table. As can be seen, the pause encoding improves the performance thereby indicating that it is important indeed to differentiate between next and previous word as well as pauses. Syllable marking helps too, and further experiments are reported using the 5-bit encoding along with the 7-bit syllable position.

Table 4: *Performance with various positional encoding approaches. Input to CNN has $\pm$ 1 word context.*

Positional encoding	Correlation	F-score
1-bit [21]	0.67*	0.62*
3-bit (word order)	0.68	0.62
5-bit (word/pause order)	0.68	0.62
5-bit + syllable position	0.68	0.63

We also experimented with choosing fixed duration input segments, centered at the word, as an alternative to the variable (3 word) segments considered so far to avoid the zero-padding at the CNN input. We find that the duration that works best corresponds to the average 3-word segment and that there is no performance gain with fixing input segment length.

To capture the contribution of multi-channel processing of the attribute-wise contour, we compare it with single-channel processing of the combined contours in Table 5 with all else kept unchanged. We note a drop in performance. Feature pooling of multi-channel outputs also reduced performance compared to simple feature concatenation suggesting the importance of retaining all attribute variations in the input to the RNN model.

Table 5: *Performance with different CNN architectures*

Architecture	Correlation	F-score
Multichannel (concatenate)	0.69	0.63
Multichannel (pooling)	0.66	0.62
Single-channel	0.67*	0.62

4.4. Overall performance

Given our overall goal of investigating the automatic learning of features for a prosodic prominence detection task on a challenging dataset of children's read speech, we explored the cascade of CNN and RNN with word and sentence level inputs respectively. Our baseline was a random forest ensemble predictor with hand-crafted acoustic-prosodic features optimized with to exploit the best of acoustic parameter aggregation, normalization and local context for prominence detection of a word within a long utterance. Using a bidirectional GRU with the sequence of acoustic-prosodic features ('A34') helped improve F-score by 1% over the baseline. As we see in row 2 of Table 6, the CNN-learned features in the same setting fall slightly short.

Adding features assumed to be harder to learn automatically, but important for word prominence, such as the actual and speech-rate normalized word, syllable and pause durations, we concatenate a 12-dimensional duration feature vector ('DP-12') with the CNN-learned features. We further select 10 contour shape features ('A10', a subset of A34) that require fitting the temporal variation of the acoustic parameters. The added features are seen to bring distinct additional value. Finally, the inclusion of text features (L, I) boosts performance for both feature sets as we see in the last 2 rows of Table 6 with higher performance for A34.

Our outcomes are overall consistent with those of the few previous works that have compared automatically learned and hand-crafted features on the same dataset and task. As is well known, automatic feature learning is limited by dataset sizes and future work must examine the use of larger, possibly unlabeled, data and architectures that learn additional useful information such as speaker identity or employ attention for better context learning.

Table 6: *Performance of CNN encoding concatenated with different word-level features as RNN input. (* indicates sd $<$ 0.01)*

Features	Correlation	F-score
A34	0.70	0.64
CNN	0.69	0.63
CNN + D-P12 + A10	0.71	0.64
CNN + D-P12 + A10 + L + I	0.77*	0.68
A34 + L + I	0.79*	0.69*

5. References

[1] J. K. Bock and J. R. Mazzella, "Intonational marking of given and new information: Some consequences for comprehension," Memory & Cognition, vol. 11, no. 1, pp. 64–76, 1983.

[2] L. van Maastricht, T. Zee, E. Krahmer, and M. Swerts, "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, 2017, pp. 364–368.

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

[4] S. E. Sinambela, "Prosody as a tool for assessing reading fluency of adult ESL students," Advances in Language and Literary Studies, vol. 8, no. 6, pp. 83–87, 2017.

[5] D. Paige, W. Rupley, G. Smith, W. N. T. Rasinski, and T. Magpuri-Lavell, "Is prosodic reading a strategy for comprehension?" Journal for Educational Research, vol. 141, no. 4, pp. 245–275, 2017.

[6] H. K. Keskin, G. Ari, and M. Bastug, "Role of prosodic reading in listening comprehension," International Journal of Education & Literacy Studies, vol. 7, no. 1, pp. 59–65, 2019.

[7] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," Language and Cognitive Processes, vol. 25, no. 7/8/9, pp. 1044–1098, 2010.

[8] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," Laboratory Phonology, vol. 1, no. 2, pp. 425–452, 2010.

[9] P. Wagner, F. Tamburini, and A. Windmann, "Robust tracking for automatic reading tutors," in Proceedings of INTERSPEECH, Portland, OR, USA, 2012.

[10] S. Baumann and B. Winter, "What makes a word prominent? Predicting untrained German listeners' perceptual judgments," Journal of Phonetics, vol. 70, pp. 20–38, 2018.

[11] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.

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

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

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

[15] K. Sabu and P. Rao, "Prosodic event detection in children's read speech," Computer Speech and Language, vol. 68, pp. 1–19, 2021.

[16] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning," in Proceedings of INTERSPEECH, Dresden, Germany, 2015, pp. 3066–3070.

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

[18] Y. Wu, H. Li, and S. Li, "Automatic pitch accent detection using long short-term memory neural networks," in Proceedings of International Symposium on Signal Processing Systems, Beijing, China, 2019.

[19] B. Lin, L. Wang, X. Feng, and J. Zhang, "Joint detection of sentence stress and phrase boundary for prosody," in Proceedings of INTERSPEECH, Shanghai, China, 2020, pp. 4392–4396.

[20] S. Stehwien and N. T. Vu, "Prosodic event recognition using convolutional neural networks with context information," in Proceedings of INTERSPEECH, Stockholm, Sweden, 2017, pp. 2326–2330.

[21] S. Stehwien, A. Schweitzer, and N. T. Vu, "Acoustic and temporal representations in convolutional neural network models of prosodic events," Speech Communication, vol. 125, pp. 128–141, 2020.

[22] L. Zhang, F. M. Jia Jia and, S. Zhou, W. Chen, C. Zhang, and R. Li, "Emphasis detection for voice dialogue applications using multi-channel convolutional bidirectional long short-term memory network," in Proceedings of INTERSPEECH, Hyderabad, India, 2018.

[23] E. Nielsen, M. Steedman, and S. Goldwater, "The role of context in neural pitch accent detection in English," in Proceedings of Conference on Empirical Methods in Natural Language Processing, 2020.

[24] K. Sabu and P. Rao, "Detection of prominent words in oral reading by children," in Proceedings of Speech Prosody, Poznan, Poland, 2018, pp. 314–318.

[25] J. Cole, T. Mahrt, and J. Roy, "Crowd-sourcing prosodic annotation," Computer Speech and Language, vol. 45, pp. 300–325, 2017.

[26] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," CoRR, vol. abs/1412.3555, 2014. [Online]. Available: http://arxiv.org/abs/1412.3555

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

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

[29] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, pp. 216–228, 2008.

[30] P. Wren, H. Martin, and N. P. Rao, High School English Grammar and Composition. New Delhi, India: S. Chand & Company Ltd., 1936.

[31] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Parsing speech: A neural approach to integrating lexical and acoustic-prosodic information," in Proceedings of NAACL-HLT, New Orleans, Louisiana, 2018.

[32] S. Gugger and J. Howard, "Adamw and super-convergence is now the fastest way to train neural nets," https://www.fast.ai/2018/07/02/adam-weight-decay/, 2018, accessed: 2021-04-01.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.