CSIT
https://doi.org/10.1007/s40012-018-0202-3

Automatic assessment of children's oral reading using speech recognition and prosody modeling

Kamini Sabu[1] · **Preeti Rao**[1]

Received: 16 May 2018 / Accepted: 7 June 2018

AQ1 **Abstract** This work targets building an oral reading "tutor" that provides automatic and reliable feedback to children learning to read. The work uses state-of-the-art in speech recognition technology coupled with prosody modeling. The system is tested on available datasets of children's readings in English as L2. The expected challenges relate to dealing with children's speech with a variety of skill levels. Both word decoding accuracies and prosody attributes like phrasing and prominence are considered for assessment. The relation between different acoustic features computed from the speech signal and the perceived quality will be investigated. The goal is to have a system that can provide feedback and evaluation that is highly correlated with that of human judges such as language teachers.

Keywords Prosody · Phasing · Prominence · Reading assessment · Child speech

1 Introduction

As per the Annual Survey Education Reports [1] by Pratham over last 4–5 years, basic literacy skills of children in rural India are below par. Almost 70% of the students in 5th standard can't read even a simple English sentence in 2nd standard curriculum book. In spite of 'education for all' movement by government, the reading skills of children have not improved—two major reasons being the shortage of qualified teachers and dismal student–teacher ratio. In this scenario, we aim to build an automatic reading evaluation system to assess children's oral reading in terms of reading miscues, speech rate and fluency [2, 3]. We further aim to emulate the teacher or expert in providing motivating but sound feedback which will help students improve their reading skill.

Reading skills, as per the U.S. National Reading Panel [4], comprise of word decoding accuracy, reading rate and fluency. Word decoding accuracy is indicated through a count of detected miscues in terms of word substitutions, omissions or insertions. Speech rate is measured in terms of number of words read per minute or number of words correctly read per minute. Fluency has two components—lexical fluency and prosodic fluency. Lexical fluency indicates false starts and hesitations, while prosodic fluency refers to expressiveness which can further be considered in terms of phrasing and prominence. The ability of chunking the complete sentence into meaningful word sequences is called phrasing. The emphasis of words indicating new information in order to direct a listener's attention to this is called prominence. Prosody is known to be assistive and indicative of comprehension of the read text [5, 6]

TBall [7], Listen [8] and FLORA [9] are some major research groups working in this area. TBall dealt with kindergarten students reading isolated word lists. They used specially designed language model in ASR to detect lexical disfluency. However, lexical disfluency is not enough to rate the oral reading fluency. The Listen group correlates the prosodic contours (pitch, energy, intensity and latency) of child speech with corresponding adult

MeitY Visvesvaraya PhD scheme.

✉ Kamini Sabu
kaminisabu@iitb.ac.in

Preeti Rao
prao@ee.iitb.ac.in

[1] Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India

speech to rate fluency of 7–10 years old children reading short stories. However, the adult speech may not always be available or usable, especially when student makes substitution or omission. The FLORA group grades overall literacy for 1 min paragraph reading by children of grades 1–4. They use lexical as well as prosodic features for giving overall literacy score. Component-wise grading may, however, prove to be more beneficial in giving feedback to students.

This work is targeted towards children's oral reading assessment. Children are known to be highly inconsistent in articulation. This causes difficulties in modeling their pronunciations and prosody. Further, these are native speakers of Marathi with English being the second language and the language of instruction in school. The native language is seen to have notable influence on their English accent. Since they are new learners of the language, hesitations and filled-pauses are often observed in recordings. This makes the problem scenario unique. Large training speech data in the same context may improve the performance. This is, however, not possible beyond a certain extent. The target use case is within the school environment. This further adds highly variable background noise that must be handled to ensure system reliability.

As observed in the U.S. National Reading Panel report [4], prosody is neglected during reading assessment in schools. Only the word reading accuracy and sometimes reading speed are given importance. The unawareness of prosody results in variety of intonations and accents. Prosody is also highly affected by the regional accent and prosody style.

We therefore propose a system to assess lexical as well as prosodic miscues. The lexical miscues are evaluated in terms of the number of detected insertions, deletions and substitutions. The prosodic miscues are indicated in terms of whether the expected words are made prominent and whether phrase breaks are realized where desired. The dataset collections and annotation procedure is described in Sect. 2. The proposed system architecture is explained in Sect. 3. Finally, discussion on results is included in Sect. 4.

2 Data acquisition and annotation

The data is specifically collected from an urban school where students are used to English as language of instruction at school. Students of age group 10–14 years are considered as target class. These students were asked to read 10 sentence stories printed on paper. The story texts were normal English text writings, not annotated in any way for indicating phrasal breaks or important words. The readings were recorded through an Android voice recorder application using a headset microphone. All the recordings were made at 16 kHz sampling frequency. The recordings are then stored and used for further analysis and assessment. As the surroundings were relatively quiet, the SNR (Signal to Noise Ratio) is around 20–30 dB. Twenty students with good English reading ability are selected. Ten stories read by each student make a total of 200 recordings. Table 1 gives detailed statistics of the data used for this study.

The authors manually removed the noisy segments in the data, if any, and aligned data at sentence level. Gibberish and unintelligible words were removed from the audio and the correct transcriptions were written corresponding to the intelligible words. Prosody labeling is done by three experts in the form of phrasing and prominence. For this, they were provided audios with corresponding correct text transcriptions. These transcriptions did not include any punctuation marks or capital letters. The raters marked phrasal breaks (inter-sentence and intra-sentence) and highlighted those words which were perceived as being emphasized .

2.1 Inter-rater agreement

Prosody annotation is a subjective task for which inter-rater agreement is used to estimate the reliability of the ratings. Inter-rater agreement/reliability is usually indicated in terms of Fleiss' Kappa.

For prominence prediction task, Fleiss' Kappa in vicinity of 0.4 has been reported in [10, 11], while for phrase boundary estimation, Fleiss' Kappa in the range 0.54 to 0.62 has been reported in [10]. A similar trend is observed in our task, where agreement ($\kappa = 0.3$) for prominence is lower than that for phrasing ($\kappa = 0.5$).

Ground truth markings from different annotators differ AQ2 at many places. Taking the opinions of different annotators into account wisely is also important. Different approaches for determining the ground truth reference like majority voting or some specified number or all the raters agree condition [12]. We mark the ground truth for prominence for any given word as in [13, 14] . If any of the raters marks a word as prominent, the word will be labeled prominent.

Table 1 Dataset statistics

Parameter	Rating
Number of students	20
Number of stories	10
Number of recordings	200
Total duration of recordings	2.2 h
Average duration of recordings	40 s

Journal : Large 40012 | Dispatch : 16-6-2018 | Pages : 5
Article No. : 202 | LE | TYPESET
MS Code : CSIT-D-18-00031 | CP | DISK

The ground truth for phrasal break detection is also determined in the same way (Table 2).

There is large variability in reading abilities among students in the same school grade. While the word decoding accuracy is good for most of the students considered at the given text difficulty level, the prosody is highly variable. As per the feedback from raters, only a few students appear to comprehend the text while reading as revealed by their expressiveness via the prosody. Many students neglect punctuation marks completely while reading. Most of the students don't know the importance of stressing particular words as can be judged from the cadence or rhythm in their reading. Some students have tendency to read each and every word accurately, consciously and separately, while some other students read flatly in monotonous voice giving no stress at all. Accordingly, we made two groups of students and calculated Fleiss' kappa values for prominence separately for 11 students with good prosody and 9 students with poor prosody. The agreement values were 0.28 and 0.25 respectively. This indicates that raters agree slightly better for students with good prosody. This may be due to the reduced inconsistency between the acoustic cues and expected top-down syntax cues in good prosody.

3 System implementation

The overall proposed system is as shown in Fig. 1. The audio recording is passed through a voice activity detector to obtain sentence-level utterances. The test utterance is then enhanced for noise suppression using GAN (Generative Adversarial Network) [15]. The enhanced utterance is passed through the ASR decoder which converts the speech into a text hypothesis using previously trained acoustic and language models. The hypothesis is compared with canonical text to estimate the speech rate and yield detected miscues in the form of omission and substitution. The ASR decoder also provides word-level alignments which are used to get word-level prosodic features. Prosodic features related to pitch, intensity and spectral balance contours are used. These features are fed to the prosodic attribute prediction classifier to get prosody rating estimates in the form of phrasing and prosody.

Table 2 Statistics of expert raters' agreement for phrasing and prominence marking

No. of raters	0	1	2	3
No. of prominent words	7982	4197	2982	1328
No. of phrase breaks	7102	3802	3012	2573

3.1 Automatic speech recognition

For this task, we use state-of the art ASR system discussed below.

3.1.1 Acoustic model

The acoustic model is trained on 5 h of children reading data recorded by fluent English and Hindi speakers. Models for 47 Hindi and English phones are trained. Hindi phones are included to take care of native language phones which appear in our dataset, but are not part of English language. First speaker normalized features are obtained from MFCC features using SAT GMM-HMM model. These are passed through DNN Tandem model to get bottleneck features. The bottleneck features and speaker normalized features are appended to get the final feature vectors. A SAT GMM-HMM model is trained with these feature vectors. Finally MAP adaptation is performed with some other story reading data to tune the acoustic model parameters for current task [16]. The ground truth transcriptions based word segmentation are obtained through forced-alignments using these acoustic models.

3.1.2 Language model

The lexical dictionary used for language model has all the words observed in the transcribed recordings. Besides, some other possible variations expected in reading are also added manually. A language model is built separately for each story. First a universal garbage model is trained as a general unigram model. The training is performed on canonical text of all the available stories. This LM will have only those words which occurred at least once in the training text. Other words which did not occur in this text are also added in the LM as parallel paths. The newly added unseen words are given the probability same as the words which occurred only once in the training text. The probability values for other words are then adjusted accordingly using Good Turing method for smoothing and discounting [17]. Now a trigram model is trained on the canonical text of the target story. This trigram model, with the garbage model in parallel, forms the final language model.

3.2 Prosody evaluation

For prosody evaluation, we need to compute certain acoustic-prosodic features. First the pitch, intensity and spectral tilt are calculated every 10 ms across the utterance to get prosodic contours. Statistical measures are calculated—mean, minimum, maximum, median, span and standard deviation—for these contours across every word

Journal : Large 40012 Dispatch : 16-6-2018 Pages : 5
Article No. : 202 LE TYPESET
MS Code : CSIT-D-18-00031 CP DISK

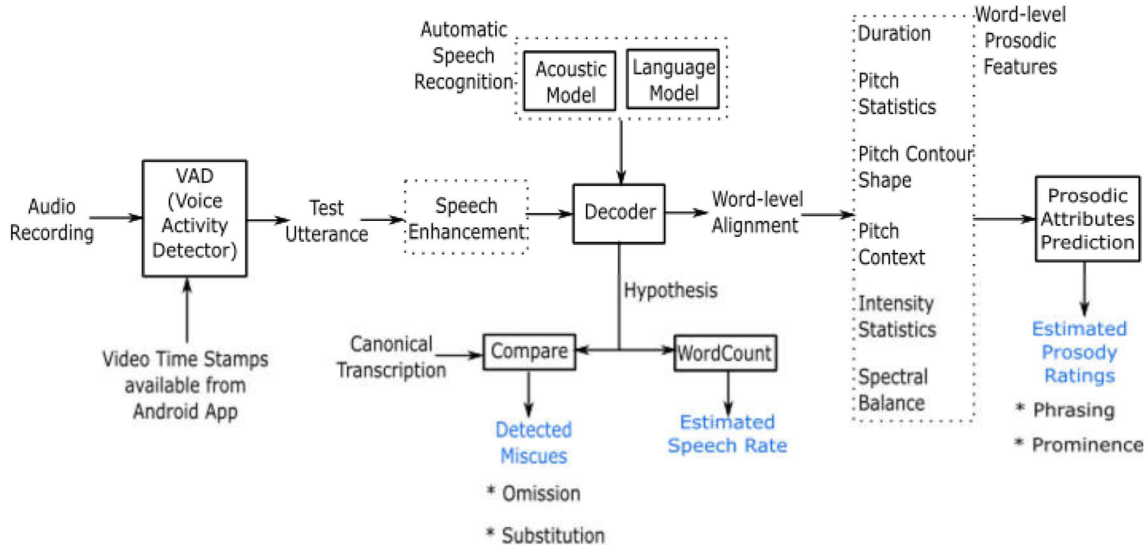

Fig. 1 System block diagram

interval. Short time magnitude spectrum is also computed for every word with 10 ms window. Then fixed band energy contours are computed for four frequency bands—0–500 Hz, 500–1000 Hz, 1000–2000 Hz and 2000–4000 Hz. Total and relative energy are calculated for these bands and above referred six statistical measures are also computed. Context based features are calculated corresponding to the above features as difference with values for previous and successive words. Features indicative of pitch contour shape involve correlation with some ideal contours—rising, falling, peak, valley and Gaussian with five different variance values—0.2, 0.5 1.0, 2.0 and 5.0. These ideal contours are of same length as the word duration. First four contours are taken such that their maximum and minimum also coincide with the word pitch contour. Further, duration related features are also computed which include average, maximum and minimum syllable duration for the word and pause/silence durations before and after the word.

These features are then used for training random forest classifiers [14, 18, 19] for phrasal break detection and for prominent word detection. The detection decisions for the given utterance can be used further to predict ratings for high level prosodic attributes and hence to give feedback whether an utterance was read “like a sentence” and whether the intended important words were stressed and intended phrasal breaks were realized.

4 Results and discussion

Performance of ASR is judged in terms of Word Error Rate (WER). With leave one story out based cross validation, we obtained WER of 3.44% for the system. In terms of children’s reading assessment, number of word miscues is considered as the measure of evaluating ASR performance. Word miscues include substitution, insertion and deletion. The precision-recall values for word miscue detection are noted in Table 3.

For prosody assessment, 5-fold cross-validation is performed. We noted individual feature importance in classification using Gini criteria [18]. The top important features with best Gini values were used for final classification and the corresponding results are shown in Table 3 in terms of precision-recall for phrase boundary detection and prominence detection.

From manual expert ratings and acoustic analysis of audio chunks, we observed and further verified from experiments that pitch trends are important in determining intonations and signify proper realization of sentence endings. Phrasal breaks can be estimated using features like pitch contour shape across the word, maximum syllable duration in the word and pause/silence duration after the

Table 3 Results of reading assessment in terms of precision and recall

	Precision (%)	Recall (%)
Word miscue detection	70.4	68.9
Prominent word detection	73.2	73
Phrasal break detection	59.2	80

Journal : Large 40012 | Dispatch : 16-6-2018 | Pages : 5
Article No. : 202 | ☐ LE | ☐ TYPESET
MS Code : CSIT-D-18-00031 | ☑ CP | ☑ DISK

word. The context based features, especially the difference values with the next word, are found to be important for phrasal breaks. Pitch span, maximum intensity and maximum syllable duration are important features in the prominence detection task.

As can be seen from Table 3, precision for phrasal break estimation is still very low. We also need to work for further improvement in prominence detection task too. While analysing the results, we observe that the inter-dependence of prosody attributes is an important aspect that affects the phrasal break and prominent word detection performance. e.g. syllable lengthening (indicated by maximum syllable duration) and pitch decline on phrase-final word (indicated by pitch span) are characteristics of proper phrasing. The same characteristics (duration elongation and high pitch span) are cues to prominent words as well. This causes false detection of phrase-final words as prominent and vice versa. Further, lexical disfluencies like hesitation and over-extended phones also have longer durations and lead to prominence marking.

The important cases where the proposed system fails, include the confusion of fricatives with noise by ASR leading to faulty word-alignment and pitch contour irregularities. This clearly indicates the importance of accurate extraction of prosodic features (pitch, energy and duration) for predicting prosodic events (phrase boundary and prominent word). Different techniques to extract these features and different functionals obtained from them can improve/degrade the performance of prosodic events prediction system.

This is an ongoing research work towards automating the highly reliable and consistent assessment of student's reading skills on known text. To reach a reliable level of evaluation, we need to work further on improving both phrasing as well as prominence predictions. Attempts towards accurate prosodic feature extraction, use of other lexical features from ASR and training on larger dataset can prove to be helpful. Besides, different symmetric and asymmetric temporal contexts and different normalization techniques can also be tested. Regional accent related features may help in further improvements.

The long term goals include building a text independent system to give assessment based on one of the standard reading fluency scales and evaluating comprehension ability from the prosody. The system can be developed to work on other Indian accents as well by adapting to the desired regional accent. The real world scenario with different noises should also be considered and efforts need to be made towards building a noise robust system.

References

1. ASER Centre (2016) ASER: the annual status of education report (Rural India). http://img.asercentre.org/docs/Publications/ASER. ASER Centre
2. Rao P, Swarup P, Pasad A, Tulsiani H, Das G (2016) Automatic assessment of reading with speech recognition technology. In: Proceedings of international conference on computers in education, Mumbai, India
3. 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
4. Panel NR (2000) Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Tech. rep., The Eunice Kennedy Shriver National Institute of Child Health and Human Development
5. Miller J, Schwanenflugel P (2008) A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. Read Res Q 43(4):336
6. Breen M, Fedorenko E, Wagner M, Gibson E (2010) Acoustic correlates of information structure. Lang Cognit Process 25(7–9):1044
7. Wang D, Narayanan S (2007) An acoustic measure for word prominence in spontaneous speech. IEEE Trans Audio Speech Lang Process 15(2):690
8. Duong M, Mostow J, Sitaram S (2011) Two methods for assessing oral reading prosody. ACM Trans Speech Lang Process 7(4):14.1
9. Bolanos D, Cole R, Ward W, Tindal G, Schwanenflugel P, Kuhn M (2013) Automatic assessment of expressive oral reading. Speech Commun 55(2):221
10. Mo Y (2008) Acoustic correlates of prosodic prominence for nave listeners of American English. In: Proceedings of Berkeley linguistics society
11. Kakouros S, Rasanen O (2014) Perception of sentence stress in English infant directed speech. In: Proceedings of INTERSPEECH, Singapore
12. Tamburini F, Bertinetto CBP (2014) Prosodic prominence detection in Italian continuous speech using probabilistic graphical models. In: Proceedings of spech prosody, Dublin, Ireland
13. Kakouros S, Rasanen O (2016) 3PRO: an unsupervised method for the automatic detection of sentence prominence in speech. Speech Commun 82:67
14. Sabu K, Rao P (2018) Detection of prominent words in oral reading by children. In: Proceedings of speech prosody, Poznan, Poland
15. Sabu K, Nathani K, Rao P (2018) Improving the noise robustness of prominence detection for children's oral reading assessment. In: Proceedings of national conference on communications, Hyderabad, India
16. Swarup P (2017) Acoustic model training and adaptation for children's read speech recognition. Tech. rep, MTP Report, Department of Electrical Engineering, IIT Bombay
17. Church K, Gale W (1991) A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of English bigrams. Comput Speech Lang 5:19
18. Breiman L (2001) Random forests. Mach Learn 45(1):5
19. Christodoulides G, Avanzi M (2014) An evaluation of machine learning methods for prominence detection in French. In: Proceedings of INTERSPEECH, Singapore

Journal : Large 40012 | Dispatch : 16-6-2018 | Pages : 5
Article No. : 202 | LE | TYPESET
MS Code : CSIT-D-18-00031 | ✔ CP | ✔ DISK