7th ISCA Workshop on Speech and Language Technology in Education
25-26 August 2017, Stockholm, Sweden

Automatic Assessment of Children's L2 Reading for Accuracy and Fluency

Kamini Sabu, Prakhar Swarup, Hitesh Tulsiani, Preeti Rao

Indian Institute of Technology, Bombay

{kaminisabu,prkhr,hitesh26,prao}@ee.iitb.ac.in

Abstract

This project targets using state-of-the-art in automatic speech recognition technology, coupled with new work in predicting the relevant prosody ratings, to build an oral reading assessment tool. A reliable automatic system can prove invaluable in helping children acquire basic reading skills apart from facilitating the monitoring of literacy programs at large scale. In the present work, we target middle-school learners of English as a second language in a rural Indian setting. We present the design and observed characteristics of our field-collected oral reading dataset to outline the research challenges faced. Recently proposed solutions to the training of robust acoustic models in the face of limited task specific data are evaluated for the prediction of the child's word decoding accuracy and for achieved word-level alignments for prosody scoring. A language model is designed to exploit the known text and observed reading errors while being flexible enough to adapt to new reading material without further training. Based on a scoring rubric proposed by a national mission on literacy assessment in India, we present an automatic system that detects reading miscues and computes fluency indicators at the sentence level which are then correlated with fine-grained subjective ratings by an expert.

Index Terms: speech recognition, prosody, children reading, human-computer interaction

1. Introduction

Recent exercises on cross-country assessment of basic literacy and numeracy skills across primary and middle-school students in India have revealed many disquieting facts such as the low proportion of students who can even meet the expectations of their grade-level [1]. Consistent with the finding that nearly 75% of fifth grade students cannot read second grade level texts in rural India, several parts of the country continue to be vulnerable to high school-dropout rates. Annual surveys undertaken by ASER [1] to measure the literacy level of school children follow a prescribed protocol executed by project volunteers visiting schools across rural India. Using prepared texts of letter and word lists, paragraphs and stories in the selected language, they categorize each student in one of the 5 competency levels , viz. 'story', 'paragraph', 'word', 'letter' and 'nothing' based on the observed word decoding accuracy and reading fluency. For example, word-level mistakes corresponding to common mispronunciations are ignored, and only 3 or more mistakes in a paragraph reading would disqualify a child at the corresponding competency level. The assessment of fluency is carried out on the basis of whether the student reads the words in sentence-style rather than like a "string of words". Testing is carried out in the regional language (medium of instruction in the school) as well as in English due to the high demand for English as a second language. We can easily see that any automation of the reading assessment can contribute greatly to the efficiency and scale of the literacy testing procedure just described. Apart from this, a system which can automatically evaluate the reading skill of a child in terms of word decoding accuracy, fluency and comprehension, also providing feedback, can alleviate the root cause of the low literacy problem - the dismal teacher to student ratio in rural schools.

There has been a fair amount of research on developing automated systems for reading evaluation and feedback using Automatic Speech Recognition(ASR) technology mainly by specific research groups contributing over the years [2], [3], [4], [5]. Black et al. [2] focused on the reading evaluation of isolated word lists for pre-school children. Other groups [3], [6], [7], [4] have addressed the task of evaluating read sentences in story context, sometimes including tracking the child in real-time. The ASR technology used has traditionally been based on GMM-HMM acoustic models. More recently, Deep Neural Network(DNN) based acoustic models have demonstrated superior performances, especially with the availability of large training datasets [8], [9], [10]. In the context of reading assessment, the language modeling (LM) required in the ASR framework typically makes use of the fact that the text to be read is known a priori [3], [7]. Either N-gram LMs trained on the story text [4], or, alternately, task-specific sentence-level LMs [3] with appropriate parallel paths modeling specific miscues are used.

Much of the research on technology for reading assessment for children has focused on detecting reading miscues, i.e. word decoding errors. Reading fluency, on the other hand, is indicated by the prosody of speech rather than by misread words [11]. Prosody refers to the supra-segmental aspects of speech. It is linked to the smooth delivery of sentences with appropriate chunking into phrases and the proper marking of word prominence. Good prosody has been associated with successful comprehension [12]. Duong et al. [13] compute pitch, intensity, duration and latency contours for each sentence as sequence of average values per word. These contours of children's speech are correlated with corresponding contours of adult speech to assess reading. To automate the scoring for spontaneous reading by non-native children of age greater than 8 years, [14] tries to combine aspects such as fluency, pronunciation, vocabulary and grammar using various features based on - 1)silence and reading speed, 2)acoustic model score, 3)part-of-speech tags, and 4)number of idioms and/or meaningful phrases. In FLORA [15], for assessment of expressive children reading at paragraph (1-minute reading) level, an SVM classifier is trained on an annotated corpus of children's read speech using lexical and filled-pause based features with prosodic features like pitch and duration.

In the present work, we consider the automation of assessment, at the sentence level, for a story reading task based roughly on the scoring rubric provided by ASER [1]. We would like to detect reading miscues such as target words that are missed or incorrectly uttered, as well as the fluency in delivery.

[1] Authors acknowledge support by MeitY, Govt. of India.

10.21437/SLaTE.2017-21

The reading miscues would ideally be based on the outcome of ASR based decoding which provides a sequence of hypothesized words best matching the acoustic utterance. A comparison of the target sentence text with the hypothesized words provides an immediate means to detect "mistakes" depending on whatever definition of mistake we might choose. On the other hand, prosodic cues extracted from the decoded word segments can help us identify whether an utterance was read "like a sentence". Proper phrasing, including sentence ending, and correct word prominences contribute to the perception of a meaningful sentence form.

In the next section, we present the specific scenario of interest, the ongoing field-data collection and manual labeling methodologies. We discuss the characteristics of the data relevant to the reading assessment task. Following this is a review of the ASR framework with a discussion of the acoustic and language modeling aspects. Next, we discuss the acoustic correlates of the relevant prosodic cues and their implementation. Finally, we evaluate our system in terms of its performance in predicting reading miscues and sentence level fluency indicators as obtained by manual ratings.

2. Database Design and Characterization

Our evaluation data is obtained from a rural middle-school setting similar to our final target scenario. We restrict ourselves to English as a second language, selecting suitable texts from readily available animated video stories with text subtitling [16]. In this section, we present the methodology for data collection and labeling required for the development, tuning and testing of system performance. We also discuss the characteristics of our data to highlight our system design challenges and choices.

2.1. Data collection and ground-truth labeling

Encouraged by the potential of the project, a school in the tribal belt of Western India, where a dialect of Marathi is the native tongue, permitted us to organize oral reading sessions for students of grades 5-8 (aged between 10 to 14 years) as a scheduled activity in school hours [17]. An Android application presents the story in video mode and the child can read and record on a tablet with a headset microphone. The stories are displayed in animated karaoke style, one sentence on each screen with the corresponding subtitles highlighted word-wise at a normal reading pace The tablet application also provides a "listen" mode where the child can hear a narrator read along with the subtitles in a standard Indian English accent [18], [19]. The children are encouraged to listen to the narrator audio before their own recording session.

The sentence level segmentation of the recorded audio is implemented using the karaoke video time-stamps. Manual word-level transcription of each sentence-level audio recording is carried out using a web-based GUI that facilitates marking each displayed target word as one of correct/missed/incorrect. 'Incorrect' words correspond to mispronunciations or substitutions, which are transcribed in terms of phoneme sequences. If the utterance/word is gibberish or otherwise undecipherable, it is not transcribed. Hesitations or sound-outs before a word are separately marked by a 'Disfluency before word' label. Apart from word-level realizations, we also label the noise, if any, that is audible into broad categories such as environmental sounds, breath and microphone noise. The second stage of sentence-level annotation characterizes the speech delivery aspect in terms of the indicative prosodic events related to phrasing and prominence. Phrasing is rated on a 3-level scale as follows: the absence of chunking (1), some attempts at chunking (2), and the correct grouping of words into phrases (3). Sentence ending is marked separately as not realized (1) or realized correctly (2). Finally, prominence is rated per sentence using 3 levels: absence of prominence (1), some perceivable word-level prominence (2), and prominence realized on the correct target words (3). This relatively fine-grained marking of prosodic events was found to reduce the subjectivity in labeling to a great extent. We therefore relied on the ratings of a single English teacher, checked for consistency in randomly chosen samples by one of the authors.

2.2. Dataset characteristics

While we try to ensure that there is no immediate source of noise in the vicinity of the child who is recording, it is difficult to control the more distant noises such as children playing and falling rain. For the present study, we consider the labeled subset (about 20% of our overall field-collected data) that is relatively free of background noise, disfluencies and untranscribed gibberish; this comprises of 7 stories read by 68 distinct speakers. We separate the data into groups based on story; then a subset of 3 stories, considered the "test dataset", is set aside for all the system evaluation reported in this paper. The distribution of ground-truth word-level labels in this test dataset (1371 utterances across 52 speakers spanning a duration of 64 minutes) is shown in Table 1 where the substituted words are further partitioned into (i) predictable (i.e. acceptable) substitutions such as common mispronunciations and word inflections, and (ii) out-of-vocabulary (OOV) substitutions. We note that close to 30% of the target words are incorrectly read (i.e. either missed or substituted). Among the predictable substitutions, widely observed substitutions were grapheme to phoneme errors and replacement of English (L2) phones by native language, i.e. Marathi (L1), phones. Observing the mispronunciations at word-level, we found that the children rarely substituted the common function words. It was also observed that unfamiliar content words were more often replaced with OOV words whereas the common content words were more likely to be replaced by inflected forms when misread.

Category	**Word distribution**
Correct	5817 (69%)
Substituted	1573 (19%) (OOV: 13.3%, Predictable: 5.7%)
Missed	1039 (12%)

Table 1: *Data characterization in terms of observed miscues for the test set of 1371 utterances comprising 8429 words by 52 speakers across 3 stories*

The remaining subset, comprising 978 utterances across 4 stories by 38 speakers (spanning a total duration of 58 minutes), is used as task-specific adaptation data for the ASR acoustic models. This dataset serves to tune the ASR acoustic models trained on more general data to the target population speech with its specific L1 influence. This data is partitioned in a suitable manner to ensure that the reported results always correspond to both speakers and stories being non-overlapping in the adaptation and test sets.

Prosody-based evaluation is reported on a subset of the test dataset described in Table 1 derived as follows. We discard story title-author utterances and further consider those utterances of the remaining sentences that are devoid of omissions (substituted words allowed). Apart from this, we reject text sentences with less than 10 utterances in the dataset. This gives us a total of 688 utterances across 40 unique sentences. Of these, 19 sentences comprise 2 or more phrases such that phrasal breaks can be uniquely specified; further there are two question forms, one a Wh-question, and the other yes/no. The 688 utterances (total duration of 30 min) come from 52 speakers giving an average of 12 utterances per speaker. Table 2 displays the distribution of subjective ratings for each of the 3 prosodic events across the set of 688 utterances. Not all the attributes are rateable for all utterances, e.g. prominence for list-form reading. Sentence endings are not rated for wrongly segmented utterances. We note that a reasonable representation of the different rating levels is available in our data. We further observe that while most sentence endings are realized correctly, improper phrasing (i.e. rating levels 1 or 2) is observed in 32% of the cases. Word prominence is usually not realized at all (rating 1) or placed on the wrong words (rating 2). We expect students to give prominence on same words as narrator. It was observed that sentences that ended with a prominent word were the most prosodically challenging for the children.

Rating	**Phrasing**	**Prominence**	**Sentence Ending**
1	86 (13%)	261 (40%)	98 (16%)
2	124 (19%)	295 (46%)	511 (84%)
3	445 (68%)	90 (14%)	-
Total	655	646	609

Table 2: *Distribution of prosody ratings for the rateable utterances (out of 688) of 40 unique sentences by 52 speakers*

3. The ASR Framework

3.1. Acoustic Model Training Data

An important predictor of ASR performance is the quality of training data in terms of how well it represents the expected test data. Given the paucity of usable field data in the present project, we present the considerations that have gone into creating suitable training data by more easily available means. Since the test data is L1-influenced Indian English, we need acoustic models representing both L1 and L2 phones. This motivated the use of a phonetic inventory of 47 Hindi and English phones. Hindi is chosen due to the availability of a Hindi dictionary and its phonetic overlap with several Indo-Aryan languages including Marathi. Considering that our target population is children, we recorded a variety of English and Hindi text (200 distinct phonetically rich sentences drawn from middle-school level material) read by 30 fluent English and 11 fluent Hindi speakers in the age group of 10-14 years who are students in an urban school. A total duration of 5.27 hours of speech was transcribed with minimal effort by rejecting utterances inconsistent with the text in any way (as in [20]). The training speakers are not a completely homogeneous set due to the relatively large age range apart from other speaker dependencies [21]. Most importantly, the test data from the rural school children differs from the training data in both fluency and accent.

3.2. Acoustic Modeling

Due to their capacity for highly nonlinear classification, Deep Neural Network (DNN) based acoustic models have been shown to outperform conventional Gaussian Mixture Model (GMM) based acoustic models on many speech recognition tasks [22]. Two distinct modes in which DNNs are used for acoustic modeling are the Hybrid and the Tandem mode. In the latter, a deep network serves as a nonlinear feature extractor feeding into a conventional GMM-HMM back-end. When speaker-normalized MFCC features are concatenated with the extracted "bottleneck" features, and all features are speaker-normalized by SAT, the resulting Tandem SAT matches or exceeds the performance of a Hybrid SAT system [23], [24], [25]. The former further allows an additional beneficial stage of model adaptation with task-specific data due to its GMM-HMM back-end. The training procedure for the DNN Tandem SAT models is provided below [25].

1. A baseline SAT GMM-HMM system was trained by estimating fMLLR transforms [26] for each speaker in the training set using raw MFCC features.
2. Speaker-normalized features were generated by transforming each speaker's data through the estimated fMLLR matrix.
3. A DNN with a 40 dimensional bottleneck(BN) layer was trained using these speaker-normalized features. This was used to obtain lower-dimensional discriminative features from the speaker-normalized features.
4. These 40 dimensional BN features were appended with the speaker-normalized features obtained in Step 2 to get "DNN Tandem SAT" features.
5. These features were used to train a SAT GMM-HMM system using the same procedure as in Step 1.

Finally, we use MAP adaptation of the GMM means in our DNN Tandem SAT model [27] with the task-specific adaptation described in Section 2.2. The GMM-HMM model had 1000 context-dependent tied HMM states with 8000 Gaussians shared across them. A single global speaker-specific fMLLR transform is estimated for each speaker in the training set. The decoding process used is the standard two-pass unsupervised fMLLR decoding process which uses the first pass decoding hypothesis of the SAT GMM-HMM system as the transcription labels for speaker-specific transform estimation. During decoding we estimate fMLLR transform at the speaker-story level. The DNN architecture used for extracting BN features consists of 6 hidden layers with 1024 neurons each with the penultimate hidden layer replaced by a 40-dimensional layer to lower dimensionality. It is trained using standard cross entropy loss function on 40-dimensional speaker-normalized LDA-MFCC features with the context of +/-5 frames. All hidden neurons use the sigmoid activation function.

3.3. Language Modeling

To identify word-level miscues, the ASR decoder hypothesis for an utterance must serve to indicate whether each of the text words is omitted or uttered incorrectly. Since we consider only omission and unpredictable substitutions (referred to as OOV substitutions) of a word as mistakes, we would like to discriminate between OOV substitutions and predictable substitutions. A good language model would capture all the expected variations in an utterance from the known text with the appropriate probabilities. To achieve our aim of detecting miscues, we use a sentence specific LM with paths corresponding to the options for each word (correct, omitted, substituted with a predicted form or substituted with an OOV) with appropriate probabilities. We choose fragment based modeling of OOVs [28] where the fragments are phone sequences determined in a data-driven way from an English and a Hindi dictionary [29], [30].

To assign probabilities to various parallel paths around each target word, we use heuristics based on our more general observations of the story reading by children about dependence on word category i.e. whether it is a function or content word (parts of speech) and the complexity of the word, typically the length in phones. For example, our observations indicate that function words are more likely to be missed rather than substituted, whereas the reverse holds for content words. Compared with function words and short content words, long content words have a higher probability of miscues relative to correct utterances. We use a set of heuristic probabilities for each of the above word classes, and employ the probabilities assigned to miscue (omission and OOV combined) relative to correct (including predicted substitutions) to tune the achieved miscue detection versus false alarm rate of the system.

3.4. Reading Miscue Detection

To evaluate the performance of the ASR system, we compute the traditional measure, the Word Error Rate(WER). Here we look for the precise word while considering a correct detection of OOV by the fragment bigram model as a correct recognition. For the task of miscue detection however, we report the results in terms of detection rate/recall (DR) and false alarm rate (FAR) of miscues [2], [3], [7] in a 2-fold cross-validation mode using the total adaptation and test data divided in a manner such that there is no overlap of speakers or stories. A small amount of data was kept aside as a validation set in each of the 2 folds. This validation set was only used to tune the LM weight and Word Insertion Penalty(WIP) in the ASR. The definition of miscue is motivated by [1] as including omission and substitution by OOV. Also, since we are interested in obtaining accurate alignments for the subsequent extraction of word-level prosodic cues, we report our results on a location accuracy based metric (as done in [28]). This is the fraction of words in reference ground-truth (GT) alignments that have some hypothesized word in decoder output whose both start and end boundaries fall within +/- 50ms of the GT word. This metric captures the information about alignments, useful for prosodic evaluation, irrespective of underlying word. For the location accuracy, GT boundaries were obtained by forced alignment with the GT transcripts using MAP adapted DNN Tandem SAT acoustic models.

The reported figures in Table 3 correspond to the test set described in Table 1. We obtain an overall miscue detection rate of 68% with false alarm rate close to 10% which is comparable to the reported performance of reading assessment systems built with significantly higher amounts of training data [3], [31]. The analysis of errors of our system shows that our OOV model (i.e. fragment bigram) sometimes eats up words adjacent to actual OOVs indicating that further topological constraints on phone/fragment bigrams may be warranted.

4. Prosodic Event Detection and Scoring

Duong et al. [13] uses correlation with measured adult speech prosody features for the same text to assess the child's reading prosody. In the interest of a more general system, and also given that our data is characterised by word omissions and substitutions, we prefer to base our automatic ratings on the known generic acoustic correlates of prosodic attributes. Based on the previous discussion of subjective rating of prosodic attributes (phrasing, sentence ending and prominence) in Sec. 2.1, we investigate acoustic features at the sentence level for the prediction of subjective ratings. This is followed by a discussion of the implementation and evaluation of the automatic scores on the prosody test dataset.

4.1. Acoustic Cues

Correct phrasing refers to the grouping of words so as to indicate phrase breaks between the correct groups as per the text. A phrasal break can be expressed by a pause and/or a pitch reset [32]. The latter refers to a large pitch difference (>15Hz) between end of a phrase and the start of the next [33]. Syllable lengthening at the end of the phrase is also an important cue to phrasal break perception [32]. It has been observed that number of pauses with respect to the expected number as per the text, mean and standard deviation of pauses and word-level average syllable duration are important acoustic cues for the detection of phrasing [11]. We observe in our data that students at beginner level tend to read in list form. Besides monotonous pitch, the list form reading may be perceived through relatively large pauses between consecutive words and/or unusual lengthening of every syllable in the sentence. Our observations show that if the average syllable duration for each word is more than 300ms, the utterance is perceived as list form (i.e. the lowest subjective rating for phrasing). Sentence endings are typically cued by the pitch contour slope and trend (rise, fall or flat) over the segment corresponding to the final word[32]. Next, perceived prominence depends on acoustic features at the word level such as the RMS energy, average intensity, syllable duration, pitch span, maximum pitch, average pitch, and pitch difference across adjacent words [34], [35].

4.2. Implementation

The acoustic features required for automatic prosody ratings are estimated at the word level using the segmentation provided by the ASR decoder.

4.2.1. Feature Estimation

The pitch contour is estimated at 10 ms intervals across the utterance using an autocorrelation based pitch detector over 20ms Hamming windowed speech segments. Unvoiced regions are detected based on low pitch salience and energy. The resulting pitch contour is smoothened further based on [36] where the complete pitch contour is divided into distinct continuous parts

WER (%)	**OOV substitution**		**Omission**		**Miscue (OOV + Omission)**		**Location Accuracy (%)**
	DR(%)	**FAR(%)**	**DR(%)**	**FAR(%)**	**DR(%)**	**FAR(%)**	
20.9	35.0	4.9	83.7	4.0	67.9	9.9	84.7

Table 3: *Miscue detection results on test set (Table 1) in terms of WER, DR/FAR and localization metric using MAP adapted DNN Tandem SAT acoustic models*

subject to the conditions: 1) the adjacent pitch values should not deviate more than 12%, and 2) such continuity should exist for at least 50ms. Octave errors are corrected by doubling or halving small pitch regions appropriately wherever abrupt pitch changes are observed. We also compute a short-time intensity contour at 10 ms intervals across the utterance.

The ASR decoder hypothesis provides word onset and offset and intervening silence boundaries. For our work, silence regions greater than 150 ms are deemed as 'pause' as in [11]. The number of pauses, mean and standard deviation of pauses, and position of the pauses are calculated using the silence intervals. In order to obtain the average syllable duration for a given word, we divide word duration by number of syllables in the hypothesised word. From the computed pitch and intensity contours, we obtain the maximum value of pitch, mean pitch, pitch span, standard deviation of pitch, RMS energy and average intensity for each word segment.

4.2.2. Prosodic Event Scoring

For phrasing estimation, we first check for "list form" reading. For this, we appropriately threshold each of the following features: at the sentence level, we consider number of pauses, standard deviation in pause durations and average syllable duration; at word level, for words other than phrase-final words, we consider the average syllable duration and pitch span. For the remaining utterances (i.e. those not falling in the "list form" category), we search for the position where the pitch reset or pause is observed. If all the candidates are found at the expected positions as per the target story text, we assign rating 3. If number of phrasal breaks are more or less than the number of expected pauses or if the break is at the wrong position, rating 2 is given.

In order to score sentence ending, pitch shape and span over the last word are examined. The pitch shape is said to be 'flat', if the pitch span across the word is less than 5 Hz. In all other cases, the pitch shape is considered either rising or falling. For the rating level 2 (correct ending), a rising shape is expected for the yes-no question and falling otherwise [33]. In case the pitch slope on last word is close to zero, the pitch declination over the complete utterance is considered. If the latter exceeds 10Hz, sentence ending is rated correct.

For prominence prediction, we train a decision tree using 90 utterances with prominence on the correct words. We obtain 192 prominent and 254 non-prominent words in the training set. The classifier uses the word-level features of mean pitch, standard deviation of pitch and pitch span, all normalized by average pitch at the utterance level. RMS energy and average intensity are also normalized with sentence level RMS energy and average intensity respectively. Average pitch difference among neighboring words and average syllable duration per word are other important features in the prominence classifier. We then test words in the remaining utterances of the prosody test dataset (utterances with subjective rating 1 and 2) using this classifier. If any one word in an utterance is found to be prominent, we assign rating 2 to the utterance; if no word is found prominent in the whole utterance, it is marked as rating 1.

	Phrasing	**Prominence**	**Sentence Ending**
PR(%)	57	55	47
RC(%)	71	57	72
Acc(%)	64	58	71

Table 4: *Prediction of prosody ratings evaluated in terms of Precision-Recall and Accuracy using decoder alignments of prosody evaluation dataset from Table 2*

4.3. Evaluation With Respect to Subjective Ratings

The automatically computed prosodic event scores are compared with the corresponding subjective ratings at the sentence level. Since the goal is to flag reading errors, we report our results in terms of precision-recall and accuracy(shown in Table4) in the detection of reading errors for each of the three subjectively rated events, viz. phrasing (where both rating levels 1 and 2 constitute reading errors), sentence ending and prominence (where rating level 1 constitutes a reading error in each case).

The sentence ending reading error shows that we can rely on pitch contour of the last word to determine the proper sentence ending. Cases where the sentence ending error goes undetected are typically associated with the occurrence of final word prominence. Another case that needs further investigation is that of flat pitch ending, which sometimes gets perceived as correct sentence ending.

Due to the interdependence of phrasing and prominence, students tend to lengthen the syllables though maintaining proper chunking of utterance. The higher average syllable duration of the sentence then leads to poor phrasing decision (rating 1) by our system. Some other false alarms arise from improper word segmentation by the decoder due to the confusion of recording noise, breath noise, fillers like 'ummm' with phones in adjacent words. This shows that syllable duration and pause related features obtained by our ASR may need refinement for phrasing.

In the prominence estimation, syllable duration lengthening is found to be the most important feature followed by pitch span, RMS energy, and change in average pitch across adjacent words. The pitch span feature is expected to be large for a prominent word. However, for the last word in a sentence, we expect large pitch decline for sentence end realization, and hence large pitch span. Same is true for standard deviation of pitch. The overlapping characteristics of sentence ending and prominence on final word need to be addressed with better features. Further, if even a single word is wrongly marked prominent, the sentence-level decision is affected. Finally, we note that errors in pitch estimation arising from the challenges of signal quality and pitch range can affect the accuracy of the automatic prosodic event ratings.

The whole set of experiments for prosody evaluation is repeated with word-level alignments obtained from known annotated ground truth transcription. The results are quite similar suggesting that the decoder hypothesis are reliable enough for prosody features estimation.

5. Conclusions

We presented ongoing work on the development of a reading assessment system with a discussion of our field data collection and labeling methods. We defined reading errors to match the expectations of a specific literacy monitoring project in terms of word decoding accuracy and fluency. An ASR framework was used to detect word omissions and substitutions, as well as to obtain word-level segmentation for prosodic events of interest. We achieve reasonable accuracy on detection of reading miscues and on the prosody attributes related to phrasing and sentence-ending detection. Word prominence detection needs further work. Future work is targeted towards training and testing with larger datasets on possibly more diverse speakers and environmental conditions.

6. References

[1] "ASER: The annual status of education report (rural)," (http://img.asercentre.org/docs/Publications/ASER%20Reports/ASER%202016/aser_2016.pdf, ASER Centre, 2016.

[2] P. Black, J. Tepperman, and S. Narayanan, "Automatic prediction of children's reading ability for high-level literacy assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1015–1028, 2011.

[3] 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.

[4] L. Xiaolong, J. Yun-Cheng, L. Deng, and A. Acero, "Efficient and robust language modeling in an automatic children's reading tutor system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hawaii, USA, 2007.

[5] J. Mostow, S. Roth, and A. Hauptmann, "Demonstration of a reading coach that listens," in *Proceedings of the 8th annual ACM symposium on User interface and software technology*, New York, USA, 1995.

[6] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Communication*, vol. 49, no. 12, pp. 861–873, 2007.

[7] J. Mostow, S. Roth, A. Hauptmann, and M. Kane, "A prototype reading coach that listens," in *Proceedings of the National Conference on Artificial Intelligence*, Washington, USA, 1994.

[8] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, 2016.

[9] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners," in *Proceedings of INTERSPEECH*, Singapore, 2014.

[10] J. Cheng, X. Chen, and A. Metallinou, "Deep neural network acoustic models for spoken assessment applications," *Speech Communication*, vol. 73, pp. 14–27, 2015.

[11] J. Liscombe, "Prosody and speaker state: Paralinguistics, pragmatics, and proficiency," Ph.D. dissertation, Columbia University, 2007.

[12] J. Miller and P. Schwanenflugel, "A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children," *Reading Research Quarterly*, vol. 43, no. 4, pp. 336–354, 2008.

[13] M. Duong, J. Mostow, and S. Sitaram, "Two methods for assessing oral reading prosody," *ACM Transactions on Speech Language Processing*, vol. 7, no. 4, pp. 14.1–14.22, 2011.

[14] L. C. K. Hassanali, S. Yoon, "Automatic scoring of non-native children's spoken language proficiency," in *SLaTE*, Leipzig, Germany, 2015.

[15] D. Bolaos, R. Cole, W. Ward, G. Tindal, P. Schwanenflugel, and M. Kuhn, "Automatic assessment of expressive oral reading," *Speech Communication*, vol. 55, no. 2, pp. 221 – 236, 2013.

[16] "Bookbox: A book for every child in her language (2016)," www.bookbox.com.

[17] P. Rao, P. Swarup, A. Pasad, H. Tulsiani, and G. Das, "Automatic assessment of reading with speech recognition technology," in *Proceedings of the 24th International Conference on Computers in Education*, Mumbai, India, 2016.

[18] R. Bansal and J. Harrison, *Spoken English*. Orient BlackSwan, 1972.

[19] P. Pandey, G. Leitner, A. Hashim, and H. Wolf, *Communicating with Asia: The Future of English as Global Language*. Cambridge University Press, Cambridge, 2016, ch. 4. Indian English Prosody.

[20] R. Pascual and R. Guevara, "Developing a children's filipino speech corpus for application in automatic detection of reading miscues and disfluencies," in *Proceedings of TENCON 2012 IEEE Region 10 Conference*, Cebu, Philipines, 2012.

[21] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of Acoustic Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[22] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[23] S. Rath, K. Knill, A. Ragni, and M. Gales, "Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages," in *Proceedings of INTERSPEECH*, Singapore, 2014.

[24] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *IEEE workshop on automatic speech recognition and understanding*, Hawaii, USA, 2011.

[25] T. Yoshioka, A. Ragni, and M. Gales, "Investigation of unsupervised adaptation of DNN acoustic models with filter bank input," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, 2014.

[26] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *Computer Speech & Language*, vol. 20, no. 1, pp. 107–123, 2006.

[27] C. Lee and J. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minnesota, USA, 1993.

[28] I. Bazzi, "Modelling out-of-vocabulary words for robust speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[29] "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict, Carnegie Mellon University.

[30] K. Samudravijaya, P. Rao, and S. Agrawal, "Hindi speech database," in *Proceedings of International Conference on Spoken Language Processing*, Beijing, China, 2000.

[31] Y. Tam, J. Mostow, J. Beck, and S. Banerjee, "Training a confidence measure for a reading tutor that listens," in *Proceedings of INTERSPEECH*, Geneva, Switzerland, 2003.

[32] T. Gibson, "Prosody and intonation," https://ocw.mit.edu/courses/brain-and-cognitive-sciences/9-59j-psycholinguistics-spring-2005/lecture-notes/0414_intonation.pdf.

[33] P. Schwanenflugel, A. Hamilton, J. Wisenbaker, M. Kuhn, and S. Stahl, "Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers," *Journal of Educational Psychology*, vol. 96, no. 1, pp. 119–129, 2004.

[34] 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.

[35] F. Tamburini, "Prosodic prominence detection in speech," in *International Symposium on Signal Processing and its Applications*, Paris, France, 2003.

[36] Y. Xu and X. Sun, "Maximum speed of pitch change and how it may relate to speech," *Journal of acoustic Society of America*, vol. 111, no. 3, pp. 1399–1413, 2002.