Automatic assessment of children's oral reading skills

Kamini Sabu

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

Abstract

Fluent reading is a critical component of literacy skills and necessary for overall personality development. Proper assessment and feedback to students are therefore essential. This project is an effort to automate reading skill evaluation. We collected data of English stories read by Indian children who are L2 speakers of English, and asked teachers to rate it on different lexical and prosodic attributes. The ratings were then predicted using a machine learning model trained with different acoustic-prosodic features. The work aims at determining the optimal set of predictive features for rating the paragraphs for the scoring attributes. The trained machine learning models are expected to accurately mimic the teachers' ratings on unseen test data.

Index Terms: oral reading, children's speech, acoustic-prosodic features

1. Motivation

Reading is the first step in acquiring literacy skills. The literacy level all over the world, especially in India, is a grave concern. An annual education survey made by the NGO Pratham shows that 80% of the fifth-grade students can't read simple English sentences [1]. The situation is more worrisome, considering the skewed teacher-student ratio and the scarcity of good teachers. In this scenario, having an automated tool to evaluate students' reading skills will be helpful to a great extent. Presently, the evaluation is done manually, where the teachers ask students to read the text in one minute and count the number of words read correctly. The teachers further use their own judgment to determine the students' reading fluency based on prior experience. Maintaining unbiased scoring with reliability and scalability is therefore difficult. The use of automated tools can tackle this issue with added benefits.

The word pronunciation as well as speaking style matter in spoken communication [2]. Therefore, reading skills, too, need to be practiced to refine both lexical and prosodic aspects. The reading assessment strategies are being developed to focus on both these traits [3, 4, 5]. The automation of reading assessment has been considered before in many research works. However, most of these concentrate on WCPM (word correct per minute) computation, which is a combined measure of word decoding accuracy and the reading speed [6, 7]. It is based on the assumption that students reading fast are good readers. However, the good word decoding is not always indicative of good prosodic reading. Instead, there is a wide range of prosodic skills, and so are the comprehension skills [8] among the good word decoders. Three major research projects, viz. TBALL [9], Listen [10] and FLORA [11] have worked on the reading prosody evaluation for children. TBALL deals with isolated word lists reading. Listen and FLORA score the continuous speech. All these are for L1 of the speakers. They use both lexical and acoustic features for the evaluations.

2. Thesis Proposal

We propose to design an automatic reading skills evaluation system that will mimic the ratings by human experts. The system is expected to rate the students on different lexical and prosodic attributes [12]. The scoring can also be used to give feedback to students on their performance and the aspects they can improve on.

This work involves L2 reading evaluation in the same way as teachers usually do for their class. The assessment is on continuous paragraph reading. Based on the literature and discussion with experts, we consider the following attributes for scoring: pace, phrasing, prominence, confidence, cadence (reading style), and comprehensibility. Reading research shows that the reader's comprehension can also be determined by the reader's ability to determine where the phrase breaks or prominent words should appear in the text [13]. We use this approach to rate the comprehension level. The target population is beginner level children, who are not expected to follow the syntactic and semantic rules of the language. Therefore, the lexical features such as 'part of speech' may prove to be misleading while assessing the prosodic skills. Hence, we use the prosodic features alone.

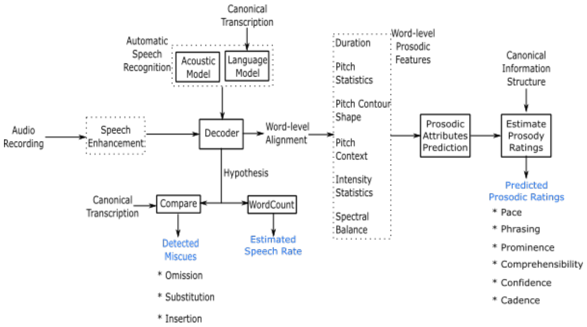

Figure 1: *Proposed System Block Diagram*

The proposed system is shown in Figure 1. It takes the audio recoding as input and preprocesses the same for noise removal or suppression. The automatic speech recognition [14] is used to get the decoded text hypothesis, along with the word-level and phone-level alignments. Acoustic contours for duration, pitch, intensity, and spectral balance are extracted from the audio recording at 10 ms frame level. Word level alignments are used to compute the word-level features, which are further used to predict prosodic event positions (phrase breaks and prominent words) [15, 16]. The detected prosodic event positions are compared to the expected locations from the canonical text of the story. These, along with other utterance level prosodic features, form the input to the high-level prosodic rating predictors.

3. Dataset

We collected the required data from the rural and urban regions surrounding Mumbai, India. Marathi is the mother tongue for most of these children. The target population is children aged 10 to 14 years - the age of L2 English learning onset in Maharashtra. The text to read consisted of short paragraphs ($\sim$50 words) from 50 English stories at B2 level on the CEFR scale [17]. The students were asked to read these from an Android application using a headset microphone. Recordings were made in a perceptually quiet environment in the school. The recordings were then sentence aligned and transcribed in Roman script. Devanagari script was used for transcribing the utterances which were intelligible, but not valid in English. Different tags were used to indicated mumblings and filled pauses. We have collected about 3000 recordings by 300 students from 7 different schools with a total recording duration around 12 hrs.

The recordings from good word decoders (reading more than 80% of the words correctly) were used for prosodic analysis. The audio and the corresponding transcribed text were given to naive but fluent English raters for marking the prosodic events (positions of perceived phrase breaks and emphasized words). The LMEDS web page [18] was used to get ratings from over 25 raters through rapid prosody transcription. Each rater was assigned a random set of recordings. The text to be marked was given without any capitalization or punctuation marks, which would have led to a top-down bias otherwise. Further, two teacher experts were given all the recordings to provide high-level ratings for pace, phrasing, prominence, confidence, cadence (reading style), and comprehensibility. Around 2000 recordings have been assessed by two expert raters for high level ratings, while subset 800 recordings has been annotated for word-level phrase break and sentence prominence by more than 7 raters. The raters' inter-reliability, as measured in Fleiss' kappa, is found to be 0.59 and 0.22 for phrase break detection and prominent word detection, respectively, which is in line with [19]. The Fleiss' kappa for high-level ratings by the two teachers is around 0.2 for all the scoring attributes.

4. Results and Discussion

We computed different word-level acoustic-prosodic features to extract the statistical information about the word-level acoustic contours as well as information about the contour shape. Different normalization techniques were employed to suppress speaker-specific and recording-specific variabilities. The redundant and non-informative features were removed using feature selection techniques. A random forest classifier was used to train the model using the teacher's ratings. For the prosodic event detection, votes from 7 raters were treated as the degree of event occurrence to predict. We obtained a Pearson's correlation of 0.85 in predicting the degree of phrase break, while 0.69 in prominence degree prediction. Considering the majority votes (3 out of 7) based event detection, we could get F-score of 0.77 and 0.49 respectively for phrase break and prominence [16]. Considering the ground truth criteria as more than 2 votes indicate prominence, an F-score of 0.63 was obtained.

We also tried predicting the high-level ratings using the prosodic features computed across sentence long windows as well as across the entire recorded utterance. We found that recording level features helped the prediction more than the aggregates of features computed across the long windows. We also observed that the two raters differed on the acoustic features they perceived important for the given attribute and hence their ratings differed. We were able to discriminate the non-confident readers with 80% accuracy. Further, we decided to mimic each individual rater separately and were able to predict the confidence level ratings with 65% accuracy [20].

5. Future Plans

The immediate task is to quantify the results of the high-level expert ratings. The cases of erroneous performances need to be analyzed in detail, which will help achieve further performance improvement.

The results depend on the accuracy of the word alignments and the acoustic features. Improving their extraction accuracy is therefore critical for the performance. Further, only two raters have given the high-level ratings and the inter-rater reliability is quite low. Ratings from more raters will prove to be valuable in improving the reliability and scalability of the system. The generalization for different L1 and L2 can also be considered in the future with the aim to cater to diverse regions in and outside of India. The designed system can assess the students' reading skills so that teachers can decide to give special attention to poor students. The system can be further developed to provide suggestions for the students' improvement automatically.

The use of system in the field will further need noise enhancement and speech-silence detections. Noise-robust acoustic feature extraction will also be important for this application. Porting the system on mobile applications can also be considered to allow the system's usage anytime, anywhere, as per the user's convenience. The user-friendly interface can also be built for handy usage of the application.

6. Research Contribution

This work is novel in terms of reading assessment of L2 learning children's prosody. The work is focused on a challenging group with native accents. The work tries to mimic teacher's ratings in terms of novel attributes like confidence and speaking style. A novel approach of using prosodic events for predicting comprehension and comprehensibility level has been considered. The work is based on the prosodic features alone. The major research contributions are:

1. The task-specific dataset was collected and annotated by several raters. The dataset includes speakers from diverse backgrounds and a wide variety of reading styles ranging from hesitating readers to readers with adult-like prosody, monotonous to sing-song style readers.
2. Different acoustic-prosodic features were computed based on the literature and our observations. Feature selection techniques were applied to get the optimal set of features to use for each task (rating each scoring attribute).
3. A prosodic event detection system was developed for children's speech, which yielded a better performance over a baseline [21].
4. Comprehensibility prediction was tried by comparing the realized positions of prosodic events against the expected positions.
5. We tried using speaker-specific characteristics to group speakers and train one model for each group to get improved results in prosodic event detection.
6. A machine learning based prediction model was trained for high-level expert ratings so as to mimic the individual raters.

7. References

[1] ASER Centre, “ASER: The Annual Status of Education Report (Rural) 2016,” http://img.asercentre.org/docs/Publications/ASER Reports/ASER 2016/aser_2016.pdf, Pratham Education Foundation, Tech. Rep., 2017.

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

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

[4] J. Zutell and T. Rasinski, “Training teachers to attend to their students’ oral reading fluency,” *Theory into Practice*, vol. 30, no. 3, pp. 211–217, 1991.

[5] T. Rasinski, *Assessing Reading Fluency*. Pacific Resources for Education and Learning (PREL), 2004.

[6] “Pearson - Versant spoken language tests, patented speech processing technology, and custom test services,” https://www.pearson.com/english/versant/tests.html, 2016, Pearson Education Inc.

[7] K. Zechner, D. Higgins, X. Xi, and D. Williamson, “Automatic scoring of non-native spontaneous speech in tests of spoken english,” *Speech Communication*, vol. 51, no. 3, pp. 883–895, 2009.

[8] M. Breen, L. Kaswer, J. A. V. Dyke, J. Krivokapić, and N. Landi, “Imitated prosodic fluency predicts reading comprehension ability in good and poor high school readers,” *Frontiers in Psychology*, vol. 7, no. 1026, 2016.

[9] M. Black, “Automatic quantification and prediction of human subjective judgements in behavioral signal processing,” Ph.D. dissertation, University of Southern California, 2012.

[10] J. Mostow, “Why and how our automated reading tutor listens,” in *Proceedings of International Symposium on Automatic Detection of Errors in Pronunciation Training*, Stockholm, Sweden, 2012.

[11] D. Bolaños, R. A. Cole, W. H. Ward, G. A. Tindal, P. J. Schwanenflugel, and M. R. Kuhn, “Automatic assessment of expressive oral reading,” *Speech Communication*, vol. 55, no. 2, pp. 221–236, 2013.

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

[13] 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.

[14] K. Sabu, P. Swarup, H. Tulsiani, and P. Rao, “Automatic assessment of children’s L2 reading for accuracy and fluency,” in *Proceedings of SLaTE*, Stockholm, Sweden, 2017, pp. 121–126.

[15] K. Sabu and P. Rao, “Automatic assessment of children’s oral reading using speech recognition and prosody modeling,” in *CSI Transactions on ICT*, vol. 6, no. 2, 2018, pp. 221–225.

[16] ——, “Prosodic event detection in children’s read speech,” *(submitted for reveiew) Computer Speech and Language*, 2020.

[17] “The cefr levels,” https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions, 2019, common European Framework of Reference for Languages (CEFR), Council of Europe.

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

[19] J. Roy, J. Cole, and T. Mahrt, “Individual differences and patterns of convergence in prosody perception,” *Laboratory Phonology*, vol. 8, no. 1, p. 22, 2017.

[20] K. Sabu and P. Rao, “Automatic prediction of confidence level from children’s oral reading recordings,” in *Proceedings of INTERSPEECH*, Shanghai, China, 2020.

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