Interspeech 2018
2-6 September 2018, Hyderabad

Automatic detection of expressiveness in oral reading

Kamini Sabu, Kanhaiya Kumar, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India
kaminisabu@ee.iitb.ac.in, kanhaiyanathani8@gmail.com, prao@ee.iitb.ac.in

Abstract

We present a Computer-Aided Language Learning (CALL) system that assesses a child's oral reading skill including the prosodic aspects. With children who have otherwise achieved word decoding automaticity, prosodic fluency is a reliable predictor of comprehension. Prosody includes attributes such as pace, phrasing and expression. Based on the acoustic correlates of prosodic events, we propose and test features that discriminate expressive speech from monotonous speech, and further detect whether the expression is meaningful or simply a rhythmic cadence with no relation to the underlying syntax or semantics of the text. Finally the system based on processing short samples of recorded oral reading and providing feedback on the goodness of both lexical and prosodic aspects is described.
Index Terms: CALL, oral reading assessment, prosody, expression

1. Introduction

Oral reading is an important component in language learning by children across early- and middle-school grades. Both, word recognition skill and the effective use of prosody in speech, can be taught by providing suitable feedback on the child's oral reading of text. While word decoding is a cognitively demanding task in the case of early readers, practice leads to better fluency accompanied by increasing attention to prosodic aspects such as phrasing and prominence [1], [2]. Phrasing refers to the syntactically correct grouping of words and is a necessary component of any intelligible speech communication. Another important aspect of the information structure of speech is focus or prominence. Stressed or prominent words indicate new information in an utterance in contrast with given information. The acoustic correlates of prominence are the variations in word duration, intensity and F0 movement [3]. Such prosodic variations constitute expressiveness in reading or speaking. Several studies have shown that prosody is a strong indicator of comprehension and teachers are advised to pay attention to the implementation of phrase boundaries and expression in assessing oral reading skills [1], [2].

Most research in CALL technology has focused on the detection of lexical miscues as related to word decoding abilities. The automatic assessment of prosodic fluency has achieved much less attention. In this work, we address detection of expressiveness in short samples of read speech. While the presence of significant prosodic variation would signal expressiveness, good comprehension is expected to be indicated by variation that is matched to the actual underlying syntax and semantics. Towards the detection of such "meaningful" expression, we investigated acoustic cues to perceived prominence in a previous work on children's reading [3], [4]. In the current work, we present a system that discriminates different categories of expressiveness based on the same cues but now applied to capture the nature of global prosodic variation in an utterance.

Supported by Visvesvaraya PhD scheme by Ministry of Electronics and Information Technology of India

2. Dataset and Annotation

From an available database [3] of oral reading recordings in English by middle-school children in Mumbai, we selected a set of 14 speakers who had good lexical fluency but displayed a variety of prosodic skills. The recordings had been carried out in quiet room with a headset microphone on a tablet at 16 kHz sampling frequency. All the selected utterances showed some degree of phrasing, they differed in reading pace from slow to fast and also in expressiveness. While the children who read all words uniformly without any prosodic variation sounded unexpressive, an interesting finding was that children who read enthusiastically did not necessarily show the correct expression. There were some children who displayed adult-like prosody in terms of phrasing and prominence that was consistent with the syntax and semantics of the text. There were others who had a fixed cadence (rather rhythmic or sing-song) in that they did not appear to be interpreting the text meaningfully but simply connecting words.

We selected recordings of one story read by all the children for this study. Each story audio recording was divided into smaller tokens of 1-3 sentences with 15-35 words in all. Three listeners classified the tokens into 3 categories - monotonous reading, expressive reading with sing-song/rhythmic style and expressive reading with meaning. We thus had 5 readers with rhythmic/sing-song style of reading. Of the remaining 9 speakers, 5 had good expressive reading, while 4 read in monotonous way. Two of the last category stressed every word while two read with flat prosody. We had in total 20 tokens of no-expression type, and 25 each of poor and good expression types for our further experiments. The students also had different reading speeds. Of the 4 non-expressive readers, 3 were slow readers and 1 normal. Of the good readers, 2 were fast readers, while 3 read with normal pace. All of the rhythmic style readers were relatively fast.

3. Method

The overall system block diagram is shown in Figure 1. Audio recording by the student is input to the system. Syllable and phone level alignments are obtained from an ASR (Automatic Speech Recognition) engine. The Language Model (LM) in the ASR is the combination of a trigram LM trained on the target story text and a garbage model (unigram LM trained on a large set of stories). Acoustic models are obtained from a previous system trained on a large dataset of children's speech [4].

The feature extraction module extracts different prosodic features from the segmented syllables obtained from the phone alignments. Both phrasing and expressiveness are cued by supra-segmental attributes such as duration, pitch and intensity. Pitch and intensity are detected at 10 ms intervals throughout

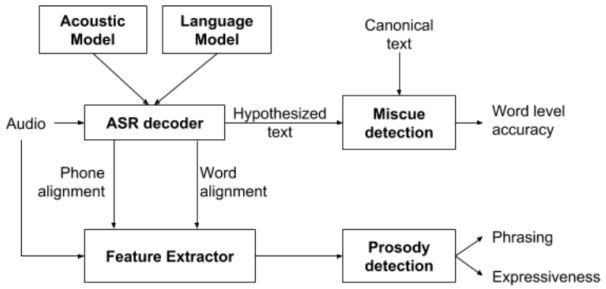

Figure 1: *System block diagram.*

Table 1: *Features that discriminate expressive and non-expressive speech. Number of stars indicates significance level.* * = $p < 0.05$, ** = $p < 0.01$ *and* *** = $p < 0.001$.

Sr.No.	Feature with p value
1	Standard deviation of vowel duration***[5]
2	Mean of absolute differences of adjacent vowel duration*[5]
3	Mean syllable duration***[5]
4	Standard deviation of syllable duration***[5]
5	Standard deviation of syllable pitch***
6	Mean of absolute differences of pitch of adjacent syllables*
7	Degree of periodicity for syllable pitch**
8	Mean syllable intensity*
9	Degree of periodicity for syllable intensity**
10	Autocorrelation lag period for syllable pitch**

the voiced regions of the utterance. These are min-max normalized across the utterance token and the mean pitch and intensity per syllable are computed. We thus obtain a temporal sequence of syllable-level pitch and intensity. Given our task to estimate the extent and type of prosodic variation in an utterance, we calculate different features that summarize the variability and further include traditional speech rhythm features [5]. Thus, the mean, standard deviation and periodicity measures (normalized autocorrelation coefficient at the lag peak) are computed for each of the temporal sequences. Of the total 31 features, the 10 features that show high significance in a Welch's t-test between expressive and non-expressive utterances are listed in Table 1. Further, the features that discriminate the most between the two categories of expressive readers (in terms of meaningful expression) are observed to be features 1, 2, 5, 6 and 9 in Table 1 indicating more repetitive variation in the sing-song style.

Finally, the selected features are used in a decision tree classifier to score the expressiveness (between monotone to expressive) and, if expressive, another decision tree scores the meaningfulness. The scores are based on the confidence levels output by the decision tree.

4. System User Interface

We have created an app for Android devices that displays the text to be read on the screen and records the reading via a headset [6]. On submission of the recording, the lexical and prosodic evaluation is carried out as discussed in the previous section. Audio feedback (comparison with a model recording of the same text) and visual feedback are provided as depicted

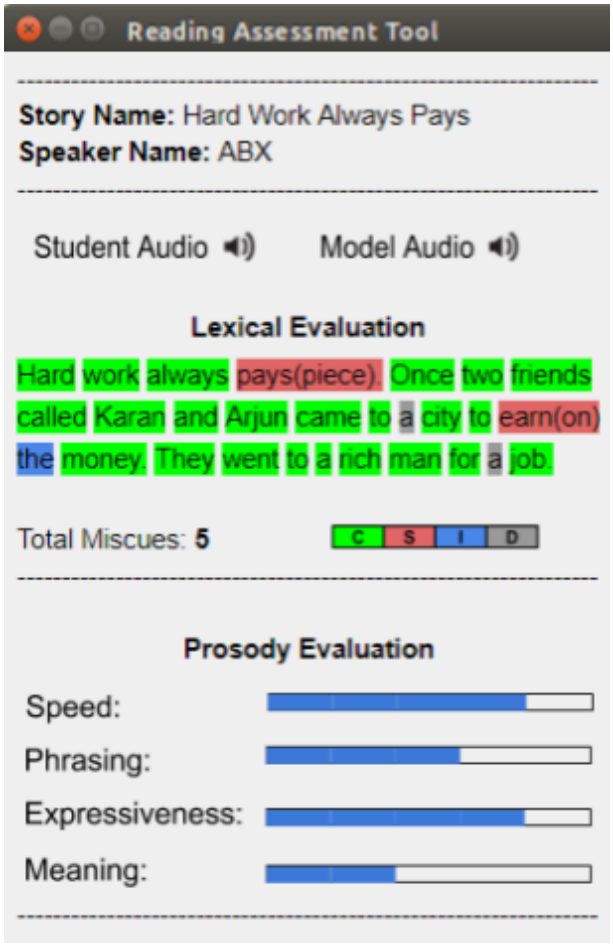

Figure 2: *System audio-visual feedback interface.*

in Figure 2. The display is designed using Tkinter library [7] of python. The lexical miscues are marked on the displayed story text using colour codes. The phrasing (to be implemented) and expressiveness ratings are provided below the text.

5. References

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

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

[3] K. Sabu and P. Rao, "Detection of prominent words in oral reading by children," in *Proceedings of Speech Prosody*, Poznan, Poland, 2018.

[4] 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.

[5] R. Fuchs, *Speech Rhythm in Varieties of English*. Springer, 2016, ch. The Concept and Measurement of Speech Rhythm, pp. 35–86.

[6] 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.

[7] "Tkinter - python interface to tcl/tk," https://docs.python.org/2/library/tkinter.html.