Adapting a Duration Synthesis Model to Rate Children's Oral Reading Prosody

Minh Duong, Jack Mostow

Project LISTEN, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA, USA
mnduong@cs.cmu.edu, mostow@cs.cmu.edu

Abstract

We describe an automated method to assess children's oral reading using a prosodic synthesis model trained on multiple adults' speech. We evaluate it against a previous method that correlated the prosodic contours of children's oral reading against adult narrations of the same sentences. We compare how well the two methods predict fluency and comprehension test scores and gains of 55 children ages 7-10 who used Project LISTEN's Reading Tutor. The new method does better on both tasks without requiring an adult narration of every sentence.

Index Terms: education, oral reading, children, prosody, speech synthesis, intelligent tutoring system

1. Introduction

Assessment of children's oral reading fluency is important in education for multiple reasons [1]. Oral reading fluency is the ability to "read text with speed, accuracy, and proper expression" [2]. Educators measure oral reading fluency in two ways. Oral reading rate is the number of words read correctly per minute. This measure is quick and easy to administer, and correlates strongly with children's comprehension test scores [3]. However, it ignores expressiveness. Fluency rubrics [4] rate reading more subjectively and qualitatively against specified criteria.

Previous work on automated assessment of oral reading has focused on oral reading rate [5] or closely related variants such as average inter-word latency [6, 7] or word reading time [8]. In contrast, automatic measurement of oral reading expressiveness would make it possible to assess reading more richly and informatively than oral reading rate, yet more precisely and consistently than human-graded rubrics.

Newer work [1] evaluated oral reading expressiveness by measuring how well the prosodic contours of children's reading correlate in pitch, intensity, pauses, and word reading times with adult narrations of the same sentences. This approach was based on the insight that the more expressive a child's reading of a text, the more the prosody tends to resemble fluent adult reading of the same text [9-11].

In this paper we tackle the same problem but eliminate the need for an adult narration of each sentence. Instead, we adapt prior work [12] that trains models of duration, F0 and intensity in order to map text to prosody. We train similar models, but instead of using them to prescribe a specific prosodic contour, we use them to evaluate children's prosody. We train our model on multiple adult voices so it is not specific to the idiosyncrasies of one speaker. This model also lets us rate readings of new text unnarrated by adults.

What is better for rating children's oral reading prosody – comparison to fluent adult narrations of the same sentences, or a generalized normative model trained on those narrations? To address this question, this paper evaluates both methods.

Our data consist of children's oral reading assisted and recorded by Project LISTEN's Reading Tutor, which listens to children read aloud, and helps them learn to read [13]. The Reading Tutor and the child take turns choosing what to read from a collection of several hundred stories with recorded adult narrations. The Reading Tutor displays text incrementally, adding a sentence at a time. It uses the Sphinx automatic speech recognizer (ASR) [14] to track the child's position in the text [15]. It responds with spoken and graphical feedback when the ASR detects hesitations or miscues, or when the child clicks for help on hard words or sentences. The spoken feedback uses a forced-aligned recording of each sentence by an adult narrator.

The rest of the paper is organized as follows. Section 2 describes the new approach. Section 3 evaluates it against the old approach. Section 4 concludes by summarizing contributions and relating them to prior and future work.

2. Approach

Prosody can be quantified by duration, pitch, and intensity. This paper focuses on duration, because we found duration-based features strongest in predicting paper tests of fluency and comprehension [1], and in detecting prosody improvement [16].

2.1. Duration model for synthesis

Several duration models, either rule-based or statistical, have been shown to work well in speech synthesis. Most well-known among the rule-based methods is Klatt's method [17], which uses rules to model the average duration of a phone given its surrounding context. Examples of good machine learning methods for prosody models are decision trees [18, 19] and the sum-of-products model [20-22]. We had tens of thousands of adult utterances as training data, so we decided to train a decision tree model of phone duration, using tools in the Festival Speech Synthesis System [23]. Given recorded, transcribed utterances, the trainer computes a set of features for each phone and builds a decision tree using these features. Rather than simply using all features, the trainer uses a greedy stepwise approach to select which set of features to use. Each step tests the features to find the best feature to add next. Given a new text to synthesize, the model generates each phone's duration as follows. First it computes the selected features of the phone and its surrounding context, up to the utterance level, to place the phone in the appropriate leaf node. It then uses the mean duration of all training data instances placed into that leaf node as the synthesized duration for the phone. We now describe the features we compute.

2.2. Features in duration model

Our duration model uses features of the phone itself, as well as contextual features about the syllable structure and the word that the phone belongs to. Phone level features include the

phone name, its position in the syllable, whether it occurs before or after the syllable's nucleus, and whether it's a consonant or a vowel. If it's a vowel, we compute its length (short, long, diphthong or schwa), its height (high, mid, or low), its frontness (front, mid, or back), and its roundedness (rounded or unrounded). If it is a consonant, we include its type (stop, fricative, affricative, nasal, lateral, or approximant), its place of articulation (labial, alveolar, palatal, labio-dental, dental, velar, or glottal) and its voicing (voiced or unvoiced). To account for coarticulation effects, we also compute some of these features for the previous two phones and the next two phones. Syllable level features include number of phones in onset and coda, position in word, distance to end of the phrase, number of syllables from previous and next phrase breaks, number of stressed syllables from previous and next phrase breaks, and the lexical stress of the syllable. At the word level, we use the context-sensitive part of speech of the word and the number of syllables in the word.

2.3. Adapting synthesis model into normative model

So far we have simply described common approaches in prosodic synthesis. The novelty of our method is in adapting the synthesis model into a normative model to rate children's oral reading. To this end, we train the model similarly to what is done for synthesis, but use it differently. Instead of using the mean duration at each leaf node of the decision tree to prescribe a duration for the phone being synthesized, we use the mean and the standard deviation of all the instances at the leaf node to compute the likelihood of the phone's actual duration.

Figure 1 depicts a fragment of a simple (made-up) tree covering the two instances of the phone /IH/ in speaking the sentence *This is an example.* At the "Syllable initial?" node, the /IH/ in *This* follows the left branch because it occurs in mid-syllable. Conversely, the /IH/ in *is* follows the right branch. Each phone contributes to the training instances used to estimate the mean and standard deviation at its leaf node.

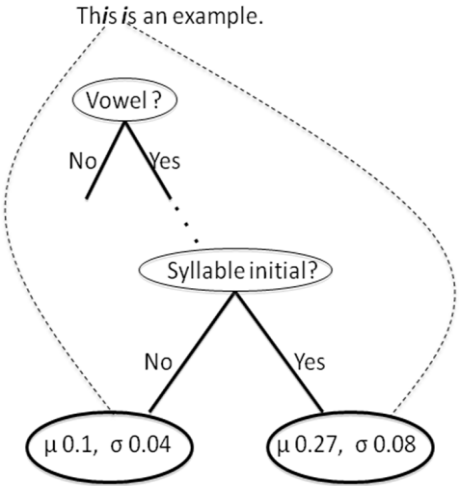

Figure 1: *Decision tree fragment with two leaf nodes*

Given a child's utterance and a trained decision tree, we evaluated different formulas for how to aggregate phone-level ratings into an overall rating of the utterance. In the following equations, u is the utterance to rate, containing n phones; the p_i's are phones in that utterance; d_i is the actual duration produced by the child for phone p_i; and μ_i and σ_i are the mean and standard deviation of the training data for that phone.

1. Average log likelihood

$$avg\_LL = \frac{1}{n}\sum_{p_i \in u} \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(d_i-\mu_i)^2}{2\sigma_i^2}\right)\right)$$

2. Average z-score

$$avg\_zscore = \frac{1}{n}\sum_{p_i \in u} \frac{d_i-\mu_i}{\sigma_i}$$

3. Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{p_i \in u}(d_i-\mu_i)^2}$$

4. Mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{p_i \in u}|d_i-\mu_i|$$

5. Pearson correlation of phone durations

$$phone\_dur\_correl = \frac{1}{n-1}\sum_{p_i \in u}\left(\frac{d_i-\overline{d}}{s_d}\right)\left(\frac{\mu_i-\overline{\mu}}{s_\mu}\right)$$

where s_d and s_μ are the standard deviations of the actual durations and mean durations, respectively, of the phones in this utterance.

6. Pearson correlation of word durations

$$word\_dur\_correl = \frac{1}{\#words-1}\sum_{word_j}\left(\frac{D_j-\overline{D}}{s_D}\right)\left(\frac{\mathrm{M}_j-\overline{\mathrm{M}}}{s_\mathrm{M}}\right)$$

where D_j and M_j represent the actual and prescribed durations of $word_j$, respectively, and s_D and s_M are their standard deviations

7. Weighted average of word-level correlations

$$avg\_word\_correl = \frac{1}{n}\sum_{word_j} \#phones_j \times word\_correl_j$$

Measures 1-5 combine the ratings of individual phones (including silences) directly into an utterance level rating. Measure 1 (average log likelihood) indicates how close the children's durations are to the model, and assumes that all the phones in the utterance are independent. It takes into account both the means and the standard deviations that the model computes from the training data. Measure 2 (z-score) also uses the standard deviations as well as the means, specifically to normalize the distance of an instance from the mean.

Measures 3-5 consider only the mean durations, measuring how far the children's durations are from these means. Measures 3 (root mean squared error) and 4 (mean absolute error) consider only the difference between the model's mean durations and the children's actual durations. Measure 5 (correlation of phone durations) looks in addition at whether the two sequences follow similar contours.

Measures 6 and 7 first compute ratings at the word level, and then combine these word level ratings into utterance level ratings. For Measure 6, we first compute each word's actual and prescribed durations by adding up the durations of its phones (including the preceding silence) – respectively, the actual durations of the phones spoken by the child, and the durations prescribed by the synthesis model (namely, the mean durations of the training examples for the corresponding leaf nodes). We then compute the correlation between the actual and prescribed durations of the words in the utterance. For Measure 7, we first correlate the actual and prescribed durations of the phones in each word, obtaining one

correlation for each word. We then average these correlations, weighted by the number of phones in each word.

When we compute these ratings, we must handle cases where a child misread, skipped, or repeated words. We consider only the child's first attempt to read a word, and exclude words that the child skipped, or that the tutor assisted before the child could attempt them. We count the number of remaining words in each utterance, and use the percentage of included words to weight the rating of each utterance when we combine the ratings of all the utterances. This weighting scheme gives higher weights to sentences with fuller observations of the child's performance.

3. Evaluation

Ideally we would evaluate our rating method directly against a gold standard measure for the prosody of each read sentence. The obvious candidate for such a measure is a human-graded rubric to evaluate oral reading fluency. One such rubric [24] rates expression, phrasing, smoothness, and pace on separate 4-point scales. However, these labor-intensive ratings would cost too much to obtain for large amounts of data. Moreover, when two members of Project LISTEN used the rubric to rate a sample of 200 read sentences, their inter-rater reliability was low, especially at the level of individual sentences [1]. (Perhaps reading professionals would agree more.)

In the absence of a reliable gold standard, we evaluated our rating method indirectly by how well the trained model predicts students' performance outside the Reading Tutor on highly reliable, psychometrically validated tests of fluency and reading comprehension, individually administered at the beginning (pretest) and end (posttest) of the semester.

We combined the utterance-level ratings for each student as described in Section 2.3, to get 7 different ratings for each student. We then used linear regression to predict students' test scores, with these ratings as predictors.

We compared our normative model against a strong baseline – our previous rating method [1] that, when combined with pretest scores, predicted fluency posttest scores with adjusted $R^2 > 0.9$ It rated the child's prosodic contour for a sentence by correlating it against the adult narration of the same sentence. To compare methods more fairly, we limited the baseline to duration-based features, including child-adult correlations for word production, latency, and duration, both raw and after normalizing for word length. Here production is the time to pronounce the word; latency is the time between successive text words, including "false starts, sounding out, repetitions, and other insertions, whether spoken or silent" [6, 7]; and word duration is the sum of production and latency.

In our experiments, we trained the decision tree on a corpus of adult narrated speech data used in the Reading Tutor. The corpus consisted of 24,816 sentences, with 811,418 phones, that were read by 20 narrators. Each story was narrated by a single narrator, so in the baseline method, each correlation is computed using a single adult voice; despite individual variation, adults' prosodic contours for a given text correlate very strongly, even across geographical regions (Paula Schwanenflugel, personal communication, 10/18/2008). Our children's data came from 235 students, who spoke 399,285 utterances comprising 8,320,114 phones. We used SPSS's linear regression function, with either the "stepwise" or "enter" option for selecting features. The enter option simply includes all the features in the regression, whereas the bidirectional stepwise option inserts or removes one feature at each step based on an F-test. This greedy technique sometimes does worse, so we tried both ways and reported the higher adjusted R^2 of the two.

Table 1 uses adjusted R^2 to measure how accurately the two types of ratings predict students' test scores. The correlational method uses features computed by correlating the child's and adult narrator's prosodic contour for each sentence. The normative method uses ratings output by the synthesis model trained on the same set of sentences.

Table 1. *Adjusted R^2 for competing methods*

Dependent variable	Normative	Correlational
Posttest fluency	0.572	0.565
Posttest comprehension	0.369	0.362

As Table 1 shows, the normative method slightly surpassed the baseline on both tasks. Although the difference is small, the new approach is qualitatively superior in that it eliminates the requirement for an adult to narrate each sentence in order for the computer to rate it. The opposite result would have suggested that the phone features employed by the synthesis model failed to capture enough information about the sentence text to rate its prosody as well as comparing it to the adult narration. Evidently the smoothness added by generalizing over multiple narrators and sentences more than compensates for the information lost by ignoring sentence details unrepresented by the phone features in the synthesis model.

Pretest scores are typically strong predictors of posttest scores, so we also tested whether normative ratings plus pretest score predicted posttest score better than pretest score alone. Table 2 shows that pretest scores achieved high adjusted R^2, but adding normative ratings accounted for about 0.01 additional variance.

Table 2. *Pretest with vs. without normative ratings*

Dependent variable	Pretest	Normative + pretest
Posttest fluency	0.852	0.866
Posttest comprehension	0.792	0.802

We noticed that the first (and sometimes only) feature selected by stepwise regression was always average log likelihood for the utterance (Measure 1) – one of the only two measures to incorporate the standard deviation of phone durations. This finding demonstrates the value of exploiting this information.

4. Conclusions

This paper introduces and evaluates a method to take a prosodic synthesis model trained on fluent adult narrations and adapt it to rate children's oral reading prosody. We first show how to extend the trained synthesis model to rate each phone. We then investigate seven different formulas to combine phone ratings to rate utterances. Two of these formulas exploit standard deviation statistics readily computed from the data collected for each leaf of the decision tree as a byproduct of training it, but unused by the synthesis model. The method trains normative models of oral reading prosody that generalize to sentences without adult narrations.

We evaluate this approach against a previous approach [1] that required an adult narration of each sentence in order to rate how well the child read it. We compare the two approaches by their ability to predict students' scores on fluency and comprehension tests. The new approach beats the old one on both tasks. Although the difference is small, the

fact that the normative method out-predicted the correlational method means it gained more by generalizing across sentences than it lost by ignoring the sentence details it did not capture.

Our method could be used to rate prosody in other contexts, such as language learning or public speaking. Given a corpus of transcribed speech with exemplary prosody, one can train a decision tree for each prosodic attribute (duration, pitch, and intensity) and then adapt the trees just as we did to assess the corresponding prosodic attributes of other voices.

Our method leaves room for further improvement. For example, we base our decision tree estimates (mean and standard deviation) solely on the statistics at the leaf level. Although we had a rule to stop splitting the tree whenever there are fewer than 20 training instances, our estimates might still suffer from data sparseness. A principled approach to remedy this problem is deleted interpolation [25], which smoothes sparse estimates of leaf-level probabilities, conditioned on many features of the phone, by combining them with better-estimated but less specific probabilities at higher levels in the tree. Besides implementing deleted interpolation, future work includes training similar models for pitch and intensity, in order to rate more aspects of children's oral reading prosody.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080628. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. We also thank the educators, students, and LISTENers who helped generate, collect, transcribe, annotate, and analyze our data.

References

[1] Mostow, J. and M. Duong. Automated Assessment of Oral Reading Prosody. *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED2009)*, 189-196. 2009. Brighton, UK.

[2] NRP. Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. 2000, National Institute of Child Health & Human Development. At www.nichd.nih.gov/publications/nrppubskey.cfm: Washington, DC.

[3] Deno, S.L. Curriculum-Based Measurement: The emerging alternative. *Exceptional Children*, 1985. *52*(3): p. 219-232.

[4] Pinnell, G.S., J.J. Pikulski, K.K. Wixson, J.R. Campbell, P.B. Gough, and A.S. Beatty. Listening to Children Read Aloud: Oral Reading Fluency. 1995, National Center for Educational Statistics: Washington, DC.

[5] Balogh, J., J. Bernstein, J. Cheng, and B. Townshend. Automatic Evaluation of Reading Accuracy: Assessing Machine Scores. *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE)* 2007. Farmington, PA.

[6] Mostow, J. and G. Aist. The sounds of silence: Towards automated evaluation of student learning in a Reading Tutor that listens. *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, 355-361. 1997. Providence, RI.

[7] Beck, J.E., P. Jia, and J. Mostow. Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning*, 2004. *2*(1-2): p. 61-81.

[8] Beck, J.E. and J. Mostow. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. *9th International Conference on Intelligent Tutoring Systems*, 353-362. Best Paper Nominee. 2008. Montreal.

[9] Schwanenflugel, P.J., E.B. Meisinger, J.M. Wisenbaker, M.R. Kuhn, G.P. Strauss, and R.D. Morris. Becoming a fluent and automatic reader in the early elementary school years. *Reading Research Quarterly*, 2006. *41*(4): p. 496-522.

[10] Schwanenflugel, P.J., A.M. Hamilton, M.R. Kuhn, J.M. Wisenbaker, and S.A. Stahl. Becoming a Fluent Reader: Reading Skill and Prosodic Features in the Oral Reading of Young Readers. *Journal of Educational Psychology*, 2004. *96*(1): p. 119-129.

[11] Miller, J. and P.J. 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*, 2008. *43*(4): p. 336–354.

[12] Jurafsky, D. and J.H. Martin. *Speech and Language Processing*. 2nd ed. 2000, Upper Saddle River, NJ: Pearson Prentice Hall. Chapter 8.

[13] Mostow, J., G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, M.B. Sklar, and B. Tobin. Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 2003. *29*(1): p. 61-117.

[14] CMU. The CMU Sphinx Group Open Source Speech Recognition Engines [software at http://cmusphinx.sourceforge.net]. 2008.

[15] Mostow, J., S.F. Roth, A.G. Hauptmann, and M. Kane. A prototype reading coach that listens [AAAI-94 Outstanding Paper]. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 785-792. 1994. Seattle, WA.

[16] Duong, M. and J. Mostow. Detecting Prosody Improvement in Oral Rereading. *Second ISCA Workshop on Speech and Language Technology in Education (SLaTE)* 2009. Wroxall Abbey Estate, Warwickshire, England.

[17] Klatt, D.H. Synthesis by rule of segmental durations in English sentences. In B.E.F. Lindblom and S. Ohman, Editors, *Frontiers of Speech Communication Research*. Academic: 287-299, 1979.

[18] Quinlan, J.R. Induction of Decision Trees. *Machine Learning*, 1986. *1*: p. 81-106.

[19] Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. 1984, Pacific Grove, CA: Wadsworth & Brooks.

[20] van Santen, J.P.H. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 1994. *8*: p. 95-128.

[21] van Santen, J.P.H. Segmental duration and speech timing. In Y. Sagisaka, N. Cambell, and N. Higuchi, Editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer 1997.

[22] van Santen, J.P.H. Timing. In R. Sproat, Editor, *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*, 115-140. Kluwer 1998.

[23] Hunt, A. and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. *Proceedings of ICASSP 96*, 373-376. 1996. Atlanta, GA.

[24] Zutell, J. and T.V. Rasinski. Training Teachers to Attend to Their Students' Oral Reading Fluency. *Theory into Practice*, 1991. *30*(3): p. 211-17.

[25] Magerman, D.M. Statistical decision-tree models for parsing. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 276-283. 1995. MIT, Cambridge, MA.