AUTOMATED PROSODY CLASSIFICATION FOR ORAL READING FLUENCY WITH QUADRATIC KAPPA LOSS AND ATTENTIVE X-VECTORS

George Sammit, Zhongjie Wu, Yihao Wang, Zhongdi Wu, Akihito Kamata,[†] *Joseph Nese,*[‡] *Eric C. Larson*

Department of Computer Science, Southern Methodist University, Dallas, TX, USA,
[†]Simmons School of Education, Southern Methodist University, Dallas, TX, USA,
[‡]College of Education, University of Oregon, Eugene, OR, USA

ABSTRACT

Automated prosody classification in the context of oral reading fluency is a critical area for the objective evaluation of students' reading proficiency. In this work, we present the largest dataset to date in this domain. It includes spoken phrases from over 1,300 students assessed by multiple trained raters. Moreover, we investigate the usage of X-Vectors and two variations thereof that incorporate weighted attention in classifying prosody correctness. We also evaluate the usage of quadratic weighted kappa loss to better accommodate the inter-rater differences in the dataset. Results indicate improved performance over baseline convolutional and current state-of-the-art models, with prosodic correctness accuracy of 86.4%.

***Index Terms*—** Automatic Prosody Classification, X-Vectors, Deep Learning

1. INTRODUCTION

Prosody is defined as the rhythm, metre, and intonation by which words are spoken. It may be categorized as symbolic or lexical prosody [1, 2, 3], phrasal breaks [4, 5], and conversational prosody [6, 7, 8]. Most applications of automatic prosody detection require supervised learning, where trained linguists label the interesting segments of audio phrases [9]. Prosody can vary greatly depending upon the context and genre from which it is taken [10, 11]. In this work, we focus on prosody classification of grade-school students in the context of reading specifically crafted passages aloud. The goal is not to classify lexical prosody per se, rather to classify the degree to which the student uses appropriate prosody.

For correctness, we use a a four-point scale developed by a subset of these authors that is based on the National Assessment of Educational Progress (NAEP) rubric [12, 13] and parts of the Multi-Dimensional Fluency Scoring Guide (MFSG) [14]. The goal of classifying correctness of prosody is part of a holistic measure of reading fluency [12, 15]. Such classification has proven to be beneficial in assessing overall oral reading fluency, complimenting the more commonly collected/assessed measures of words per minute (WPM) and words correct per minute (WCPM) [16, 17].

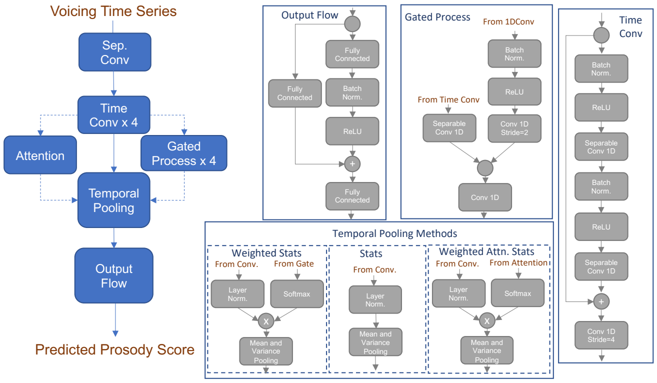

Fig. 1. Overview of the models investigated in the proposed method. Models follow one or more of the dotted paths, encompassing all combinations of temporal pooling.

A number of works (see Section 2) have investigated prosody correctness classification in this domain. One such study [18, 19], using a similar rating scale, achieved lexical accuracy of 73.24% and prosodic accuracy of 69.73% when compared with human ratings. Aside from surpassing this benchmark, we introduce a number of novel concepts: (1) We use the concept of inter-rater reliability in our loss function [20] and leverage a number of concepts from X-vectors [21] and attentive X-vectors [22]. (2) We employ weighted temporal pooling in our convolutional networks. We summarize our contributions as follows: (1) We collected and validated a dataset of prosody correctness classification of 5,841 phrases collected from 1,335 students in 2nd - 4th grade. (2) We evaluate the use of weighted temporal pooling (inspired by X-Vectors[22]) and weighted Kappa loss [20] in prosody correctness classification. (3) We conduct an ablation study to investigate the overall importance of each processing procedure. To the best of our knowledge, this work sets a new state of the art in prosody correctness classification using the largest dataset[1] to date, with accuracy of 86.4%.

[1] `https://s2.smu.edu/~eclarson/prosody.html`

2. RELATED WORK

Project LISTEN, a reading tutor, ([23],) began in the 1990s and is regarded as seminal research on automated analysis of children's spoken reading. Much research in this field builds upon it. Ananthakrishnan and Narayanan [24] propose augmenting automatic speech recognizers (ASR) by adding symbolic alphabet annotations of prosodic events. In doing so, they identify relevant prosodic features, particularly: 1) 6 F0-related; 2) 3 RMS energy-related; and 3) vowel duration. Mostow and Duong [25] and then Duong, *et al.* [26] compared a child's oral reading to that of an adult (of the same text) by analyzing the contours in pitch, intensity, pauses and word reading times. Their research is grounded in the observation that a child's expressive reading tends to mirror that of an adult's as the child progresses [17] and culminates with a trained model. In terms of scale, the closest to ours is the work of Sitaram and Mostow [27] (which builds on [28]) who mined Project LISTEN's database to evaluate oral prosody in an effort to predict fluency and comprehension. In all 85,209 sentences were evaluated and compared against a corpus of 4,558 sentences. In total, 158 pitch-related, 115 intensity-related, and 166-duration related features were assessed. Bolaños, *et al.* [18, 19] combine lexical and prosodic features to analyze children's oral reading based on the NAEP rating scale [12], a more standard and recognized scale than earlier studies. These feature sets were necessarily force-aligned, but this was accomplished in an automated fashion. They were able to obtain 73.24% lexical and 69.73% prosodic (76.05% overall) classification accuracy when compared with human ratings. The most recent work in this area is from Sabu and Rao [29]. They build a reading tutor for identifying lexical and prosodic miscues during oral reading similar to those identified by a trained professional. They laud a cross-validated precision-recall scores of 73.2%/73.0% for prominence and 59.2%/80% for phrasal break.

This work builds upon the aforementioned, but differs in several respects. First, our goal is solely to assess prosody according to a novel 4-scale rubric using conventional neural networks. Similar to many of these studies, our automated assessment will be compared against expert judgement. However, we incorporate the disagreement of those judges into our model. Our sample size is considerably larger than previous studies both in terms of recordings and participants. While previous studies have informed our feature selection, we concentrate on known prosodic low-level descriptors (LDDs) in the audio signal.

3. DATASET

In order to design and evaluate our prosody classification algorithm, we collected audio samples of oral readings from a variety of schools in the Pacific Northwest region of the USA. Passages were written by an expert who also co-wrote the original easyCBM oral reading fluency and reading comprehension passages [30]. Each passage is an (1) original work of fiction, (2) has a beginning, middle, and end, (3) follows either a "problem/resolution" or "sequence of events" format, and (4) contains minimal use of dialogue and symbols. We used 150, 50 at each of grades 2-4 consisting of 20 long (80-90 words) and 30 medium (45-55 words), passages in the experiment. Passages were distributed evenly (50% ± 1.6%) across the grade, passage length, and overall audio sample size dimensions.

Although NAEP only applied the scoring rubric to Grade 4, our research team made the decision to use the study-generated rubric and grade-calibrated passages which focus on phrasing, adherence to the author's syntax, and expressiveness to assess prosody across Grades 2 through 4. In total, 49 audio samples were identified and scored by the research team as exemplars and used for training annotators and for certification of raters. A total of 63 human prosody raters were recruited and completed two training sessions, meeting the prosody certification. Raters score prosody on a four-point prosody scale, with the option to score between if they were not certain (1, 1.5, 2, ...) thus creating a seven-point scale. Independent scores were averaged before taking the floor to provide the final score. Recording were reviewed by 138 groups of raters paired randomly in batches of approximately 50. Each rater scored between 38 and 747 recordings (mean 217.7, SD 177.3). Initial analysis of the ratings showed inter-rater agreement of 95.8% within 1.0 point of disagreement (42.5% 0.0; 31.2% 0.5; 22.1% 1.0). Disagreement of larger than 1.0 was deemed abnormal by the expert author and held-out for re-review (255 samples). In this manner, 5,841 audio recordings, each scored by two raters, were made available for model training. Given the inherent ordinal nature of these classifications, intra-class correlation (ICC) [31, 32] was used to validate acceptability of rater agreement. An acceptable mean of 0.74 (SD 0.12) on the 7-point scale and 0.71 (SD 0.12) on the final 4-point scale is observed.

Table 1 provides summary statistics of the data collected from a total of 1,335 students, and these follow national averages [33, 34] with the exception of race which is highly skewed toward white children (1.6 times higher). A stratified random sampling was applied to select audio recordings equally across grade, gender, ethnicity, race, and special needs (not shown). All data collection was completed with IRB approval.

In the interest of assessing generalizability of final model, both passages and students were held out of the training data, and used solely for model testing. One medium and one long passage from each grade was selected at random totaling 1,185 samples. The students were held out in an iterative manner that preserved the original distribution in the dataset. In all, 102 students were held out, totaling 656 samples. Both sets were reconciled, yielding 4,128 (70.7%) training/validation samples and 1,713 (29.3%) testing samples.

Table 1. Summary of the collected dataset (in %)*

Trait	Student	Audio	2nd	3rd	4th
N=	1,335	5,841	460	427	448
Female	45.8	49.1	48.0	46.8	52.5
Male	53.6	49.8	50.4	52.5	46.7
White	80.5	80.9	84.1	79.9	78.6
Non-White	19.0	18.1	14.4	19.4	20.5
Latinx	31.7	25.4	26.3	27.6	22.3
Non-Latinx	67.7	73.6	72.2	71.7	76.8

* Failure to sum to 100% due to personal data withheld

Training samples were augmented by: 1) adding Gaussian noise; 2) adjusting gain; 3) applying a high/low/band-pass filter using the Audiomentations Python library. Augmentation was applied randomly over a range of the adjustments taking care not to distort the student's voice. Less frequent classes were over-sampled to ensure balanced classes. In total, 14,092 new samples were added to training data (240% increase). From these samples, voicing-related low-level descriptors (LDDs) were extracted using the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)[35].

4. METHODS

In our design, we borrow concepts from X-Vectors [21] and attentive X-Vectors [22] for use in our architectures. However, we do not employ the time-context layers in our work—we only build variations of temporal pooling methods that weight the convolution output activations before mean (μ) and variance (σ^2) are calculated into an embedding. We hypothesize the use of different temporal weighting schemes can help the model learn to ignore insignificant phrasal breaks in the spoken passage, while emphasizing other more costly prosodic mistakes. Several model architectures are constructed and investigated, described as follows:

(1) We employ a **baseline** network that uses residual connections with identity mappings [36] and 1D time convolution, as shown in Figure 1. Each time convolutional block uses a number of separable and strided convolutions for downsampling. This baseline network does not follow any dotted paths in Figure 1 and does not use any temporal pooling before entering the output flow of the network. That is, the pooling layer is replaced by a flattening operation. (2) We employ an **x-vector**-based network that uses traditional μ and σ^2 pooling of the convolutional output activations over time. Note that there is no silence detection employed as pauses are critical for the classification of prosody. (3) We employ a **weighted x-vector** architecture that uses the center path and right dotted path in Figure 1. This dotted path uses processing gate blocks to multiply the μ and σ^2 before pooling. This weighting is achieved through multiple 1D convolutions followed by a softmax layer to force the network to focus on certain time segments before entering the output flow. (4) We employ a network using convolutional **self-attention** instead of temporal pooling as shown in the left dotted branch of Figure 1. (5) Finally, we employ a **self attention weighted x-vector** architecture that uses portions of each dotted branch in Figure 1. Here self attention is used to calculate the weighted vector to weight the segment before pooling, rather than the gate process blocks in method 2.

To analyze each model, the result must be compared to human scoring results through a loss function. Most works use the Categorical Cross Entropy(CCE) as a loss function, which is a classic loss function for many classification tasks. However, due to the nature of prosody scoring, classes have a quantitative relation between each other that CCE cannot represent. Therefore, we also conducted experiments with a modified version of loss function: Quadratic Weighted Kappa (QWK) [20].

QWK loss is related to the calculation of Inter-Rater Reliability (IRR) that is typically measured between two human raters. QWK quantifies the seriousness of the disagreement between human rating and model output as:

$$\kappa = 1 - \frac{\sum_{i,j} \omega_{i,j} O_{i,j}}{\sum_{i,j} \omega_{i,j} E_{ij}} \tag{1}$$

where O, ω and E are the confusion matrix, penalty weights matrix, and outer product of histogram of raters, respectively. O, the confusion matrix, corresponds to the number of answers that receive a score i by the first rater and a score j by the second rater. A quadratic penalty weight matrix can be expressed as are:

$$\omega_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{2}$$

where N is the total number of possible classes. Matrix O and matrix E are normalized to sum to 1.

When optimizing with QWK, [20] showed that the problem can be reformed as a minimization problem of L by:

$$L = \log(1 - \kappa + \epsilon) \tag{3}$$

where $L \in (-\infty, \log(2)]$ since $\kappa \in [-1, 1]$, the log serves to decouple the numerator and denominator calculations, which in turn eases the computation of the gradient [20]. The ϵ is a small value that avoids calculating $\log(0)$ for the loss function. In our dataset, we use the multiple scores from multiple raters as ground truth. When there is disagreement among raters, the QWK allows the loss to take into account this disagreement. We hypothesize that such behavior is advantageous for prosody classification.

5. RESULTS

A summary of the performance of each architecture is presented in Table 2 where each row represents a separate trained

Table 2. Results summary showing accuracy and IRR using Cohen's Linear Kappa.

Classifier	Loss	In-domain		Cross-domain	
		Acc	IRR	Acc	IRR
Baseline	CCE Loss	80.6%	0.76	45.8%	0.27
	κ-Loss	68.7%	0.69	47.7%	0.36
X-Vec.	CCE Loss	82.8%	0.80	50.1%	0.34
	κ-Loss	81.5%	0.79	48.0%	0.31
W. X-Vec.	CCE Loss	75.9%	0.70	46.6%	0.27
	κ-Loss	80.4%	0.78	56.4%	0.42
X-Vec. +SA	CCE Loss	**86.4%**	**0.84**	52.6%	0.39
	κ-Loss	74.9%	0.73	57.2%	0.44
W. X-Vec. +SA	CCE Loss	77.9%	0.73	48.5%	0.30
	κ-Loss	79.5%	0.77	**60.2%**	**0.46**

model and each "loss" column indicates whether the model was trained using CCE loss or QWK loss (κ-loss). The leftmost "results" column shows performance using in-domain phrases—that is, known phrases are used in the training set. The rightmost column shows results across-domain phases where phrases are used that do not exist in the training set. In both scenarios, the training and testing sets are separated according to students (as previously described) such that a student is never in both the training and testing sets. For performance, the overall accuracy is shown for classifying prosody into four scales, as well as inter-rater reliability (IRR, linear κ) assuming that the model is another prosody rater. We use the average of human raters as a ground truth for accuracy and IRR. The best performing models (boldface in Table 2) per-domain employ self-attention only and weighting with self-attention, respectfully.

We organize the discussion of results by research questions. First we ask: *Can automated prosody classification for oral reading fluency be applied within or across domains reliably?* Within domain, the models perform similarly to (in many cases better than) human raters. Therefore, we conclude their use reasonable in this context. However, when applied across domain, this performance drops considerably. Therefore, automated cross-domain performance is still an open research topic for the community. Second we ask: *Do X-vector architectures provide an advantage over baseline convolutional models for prosody classification?* Based on the performances over baseline, we can conclude that the X-Vector architecture provides a significant advantage in prosody classification. Third, we ask: *Do X-vector weighting methods provide a distinct benefit?* Here the results are not as straightforward. Using attention has an advantage, but weighting does not seem to provide an advantage. Therefore, we conclude that the most significant method for performance is attention. We note that, when applying attention, the number of weights in the model is reduced which might influence performance due to the dataset size. Finally, we ask: *Does using κ loss provide a distinct benefit over traditional cross entropy?* For most models, there is not a clear advantage, but a performance boost is observed in others. In general, we encourage other members of the speech processing community to employ and evaluate the QWK loss when subjective scores are used.

6. CONCLUSION

In conclusion, we presented a new dataset for prosody classification in the context of oral reading fluency. The highest recorded performance for in-domain classification was achieved using X-Vectors and self-attention, resulting in a new state-of-the-art in prosody classification of 86.4%. An interesting future work to increase across domain generalization may be to include the concept of an average difference feature extraction, whereby the "ideal" prosody of a known phrase is estimated using text-to-speech (TTS) models. Thus, the X-Vector model can focus upon the difference of a spoken phrase to a suggested baseline. We leave this exploration to future work.

7. REFERENCES

[1] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. on SAP*, vol. 2, no. 4, pp. 469–481, 1994.

[2] M. Hasegawa-Johnson, K. Chen, J. Cole, and et al., "Simultaneous recognition of words and prosody in the boston university radio speech corpus," *Speech Comm.*, vol. 46, no. 3-4, pp. 418–439, 2005.

[3] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Trans. on ASLP*, vol. 16, no. 1, pp. 216–228, 2007.

[4] S. Pascual and A. Bonafonte, "Prosodic break prediction with rnns," in *Adv. in Speech and Lang. Techn.* Springer, 2016, pp. 64–72.

[5] A. Rosenberg, *Automatic detection and classification of prosodic events*, Columbia University, 2009.

[6] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *ICASSP*. IEEE, 2005, vol. 1, pp. I–1061.

[7] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," in *ICASSP*. IEEE, 2012, pp. 5089–5092.

[8] D. Ortega and N. T. Vu, "Lexico-acoustic neural-based models for dialog act classification," in *ICASSP*. IEEE, 2018, pp. 6194–6198.

[9] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, and et al., "Tobi: A standard for labeling english prosody.," in *ICSLP*, 1992, vol. 2, pp. 867–870.

[10] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Comm.*, vol. 32, no. 1-2, pp. 127–154, 2000.

[11] E. Shriberg, B. Favre, J. Fung, and et al., "Prosodic similarities of dialog act boundaries across speaking styles," *Ling. Patt. in Spont. Speech*, , no. A25, pp. 213–239, 2009.

[12] S. White, J. Sabatini, B. J. Park, and et al., "The 2018 naep oral reading fluency study," *NCES*, 2021.

[13] M. C. Danne, J. R. Campbell, W. S. Grigg, and et al., "Fourth-grade students reading aloud: Naep 2002 special study of oral reading. the nation's report card. nces 2006-469.," *NCES*, 2005.

[14] T. Rasinski, A. Rikli, and S. Johnston, "Reading fluency: More than automaticity? more than a concern for the primary grades?," *Lit. Res. and Instr.*, vol. 48, pp. 350–361, 10 2009.

[15] J. Shin, "Completing the triangle of reading fluency assessment: Accuracy, speed, and prosody," *Chall. in Lang. Test.*, p. 307, 2021.

[16] T. G. Morrison and B. Wilcox, "Assessing expressive oral reading fluency," *Edu. Sci.*, vol. 10, no. 3, 2020.

[17] P. Schwanenflugel, A. Hamilton, J. Wisenbaker, and et al., "Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers," *Journal of Edu. Psych.*, vol. 96, pp. 119–129, 03 2004.

[18] D. Bolaños, R. Cole, W. Ward, and et al., "Automatic assessment of expressive oral reading," *Speech Comm.*, vol. 55, pp. 221–236, 02 2013.

[19] D. Bolaños, R. Cole, W. Ward, and et al., "Human and automated assessment of oral reading fluency," *Journal of Edu. Psych.*, vol. 105, pp. 1142, 11 2013.

[20] J. de La Torre, D. Puig, and A. Valls, "Weighted kappa loss function for multi-class classification of ordinal data in deep learning," *Pattern Rec. Letters*, vol. 105, pp. 144–154, 2018.

[21] D. Snyder, D. Garcia-Romero, G. Sell, and et al., "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*. IEEE, 2018, pp. 5329–5333.

[22] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv:1803.10963*, 2018.

[23] A. Hauptmann, J. Mostow, S. F. Roth, and et al., "A prototype reading coach that listens: Summary of project LISTEN," in *HLT Workshop*, 1994.

[24] S. Ananthakrishnan and S. Narayanan, "Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition," *IEEE Trans. on ASLP*, vol. 17, no. 1, pp. 138–149, 2009.

[25] J. Mostow and M. Duong, "Automated assessment of oral reading prosody," in *AI in Education*, NLD, 2009, p. 189–196, IOS Press.

[26] M. Duong, J. Mostow, and S. Sitaram, "Two methods for assessing oral reading prosody," *ACM Trans. on SLP*, vol. 7, no. 4, Aug. 2011.

[27] S. Sitaram and J. Mostow, "Mining data from project listen's reading tutor to analyze development of children's oral reading prosody," *IFAIRS*, 01 2012.

[28] M. Duong and J. Mostow, "Adapting a duration synthesis model to rate children's oral reading prosody," in *INTERSPEECH 2010*, 2010.

[29] K. Sabu and P. Rao, "Automatic assessment of children's oral reading using speech recognition and prosody modeling," *CSI Trans. on ICT*, vol. 6, no. 2, pp. 221–225, 2018.

[30] C.-F. Lai, J. Alonzo, and G. Tindal, "easycbm® reading criterion related validity evidence," *Behav. Research and Teach.*, 2013.

[31] K. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tut. in Quant. Methods for Psych.*, vol. 8, pp. 23–34, 07 2012.

[32] N. Gisev and J. Bell, "Interrater agreement and interrater reliability: Key concepts, approaches, and applications," *RSAP*, vol. 9, 06 2012.

[33] "Nces. cdp05.1 and cdp05.3: 2014-18, geography: United states, population group: Relevant children – enrolled (public)," [Online].

[34] "Nces. english language learners in public schools and students with disabilities: Annual reports, the condition of education," [Online].

[35] F. Eyben, K. R. Scherer, B. Schuller, and et al., "Gemaps for voice research and affective computing," *IEEE Trans. on Aff. Comp.*, vol. 7, pp. 190–202, 2016.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," 2016.