Towards Automatically Assessing Children's Picture Description Tasks

Hariram Veeramani[1], *Natarajan Balaji Shankar*[1], *Alexander Johnson*[1], *Abeer Alwan*[1]

[1]University of California, Los Angeles, Department of Electrical and Computer Engineering
{hariram, balaji1312, ajohnson49}@g.ucla.edu, alwan@ee.ucla.edu

Abstract

This paper presents preliminary findings in automatically scoring children's oral assessments while they perform a picture description task. Approximately 200 children aged 9-13 participated in this task in which they tell a story about an image presented to them. We use a BERT-based system to predict assessment scores from input ASR transcripts of the student responses. Finally, we propose next design steps to make the system more applicable to an educational setting.

Index Terms: Children's Speech Recognition, Automatic Assessment, Natural Language Processing

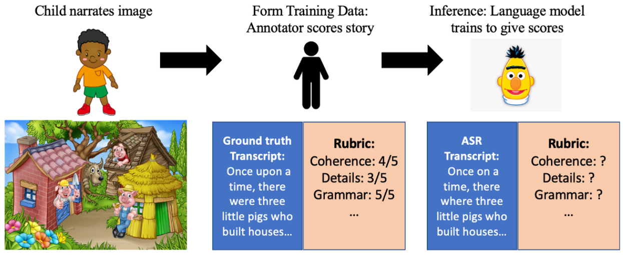
1. Introduction

A large challenge in educational speech technology is creating fair and robust systems that can automatically assess children's oral language proficiency from their spontaneous speech. These systems commonly operate by first using automatic speech recognition (ASR) to transcribe children's spoken answers to a prompt. These systems then apply natural language processing (NLP) to the ASR transcripts in order to calculate metrics relating to correctness of pronunciation and word usage, complexity of grammar patterns used, level of detail expressed, and other factors in order to evaluate the student's language abilities. In recent years, several studies have made great strides in ASR of children's speech as well as NLP for scoring language assessments (e.g., [1]). However, many educational applications of speech systems focus on analyzing read or scripted speech for pronunciation errors, disfluencies, or prosodic inconsistencies [2, 3]. These systems do not capture higher-level aspects of language proficiency like vocabulary size and narrative abilities that are better measured by analyzing spontaneous speech. To accurately transcribe children's spontaneous speech for use in downstream NLP schemes, however, is a difficult task due to the disfluencies, presence of non-speech sounds (gasping, laughter, etc.), and high variability of pronunciation and prosody found in children's speech [4, 5]. The work in [6] proposes a system to transcribe spontaneous speech from English learners and extract features from the transcripts in order to automatically score and evaluate language abilities. This paper, which builds on our previous work in [1], similarly proposes a design for a novel system which extracts transformer-based representations from ASR transcripts of children's spontaneous speech and automatically grades student responses.

2. Picture Description Task

Picture description tasks are often used to elicit spontaneous speech from children. Students are shown a picture with multiple characters or elements relating to a story plot. Images are generally chosen by experts in education to be straightforward to describe and contain enough content for the child to give a lengthy answer. The students are then asked to tell a story about the picture. Students are graded based on completeness of the description, coherence of the story, proper use of grammar, and other aspects relating to narrative language ability. In the preliminary work reported in this paper, we used a picture description task from the Test of Narrative Language [7] in the GSU Kids Speech Corpus [1]. 191 children, aged 9-13, were shown an image containing a character and several elements to describe. The students were then asked to tell a story about the image, making their story as complete as possible. Each child's response to the prompt was recorded, and each child, on average, took about 3 minutes to complete their story. Then, specialists in children's language education graded the assessment as described in [8].

Figure 1: *The proposed framework in which a language model (eg. BERT) is trained to predict the corresponding score for an ASR transcript of a recording of a child's response to the picture description task.*



3. Preliminary Experiment

In our work, thus far, we have focused on training a language model to achieve good performance in giving cumulative scores to children's oral picture description tasks. As shown in Figure 1, we train a language model to take ASR transcripts and human-labeled assessment scores of the 191 utterances and then learn to automatically score the picture description task. In these experiments, the GSU Kids Speech recordings were downsampled to 16kHz and automatically transcribed. We evaluate the large versions of both Whisper [9] and HuBERT [10] for this. We then used a transformer encoder-based language model appended with a fully-connected layer to perform classification of the student's score from the ASR transcript. To measure performance degradation due to ASR error, we also perform the classification from the ground truth transcription. As in [1], we discretized each student's raw assessment score

Model (Size)		BERT (110M)			ALBERT (11M)			DistilBERT (66M)			XLNET (110M)		
Metric	%WER	C. Acc	F1	RMSE	C. Acc	F1	RMSE	C. Acc	F1	RMSE	C. Acc	F1	RMSE
Groundtruth	-	98.0	97.5	0.06	95.5	93.0	0.09	84.0	83.0	0.3	92.0	90.0	0.072
Whisper-Large	22.4	96.5	95.0	0.067	95.5	93.0	0.1	84.0	82.5	0.44	91.3	91.0	0.12
HuBERT-Large	33.5	96.0	96.0	0.12	87.5	85.0	0.22	83.0	83.0	0.27	91.0	90.0	0.16

Table 1: *Percent Classification Accuracy (C. Acc), Percent F1 Score, and Root Mean Square Error of each language model in predicting student scores from the input transcripts (ground truth, Whisper ASR transcript, or HuBERT asr transcripts) along with the word error rate (WER) for each.*

into one of five labels corresponding to a continuous grade of 0%-20%, 20%-40%, 40%-60%, 60%-80%, or 80%-100% (with the data distribution for the individual labels to be 7%, 19%, 38%, 32%, and 4% respectively) before training the language model to predict the student's score range from the transcription text. As coherence and grammar are main factors in the human scoring, we experimented with the language models, BERT [11] and ALBERT [12], which have shown good performance in the next sentence prediction and sentence reordering pre-training tasks respectively, as well as DistilBERT [13] which is trained by distilling the BERT model. We also consider XLNET [14], as its autoregressive structure has proven advantageous over BERT in several text classification tasks. 75% of the GSU Kids Speech was used for training and the other 25% for testing. We performed a 4-fold split to ensure that all data was used in testing and report the average performance over all folds. In addition to the GSU Kids dataset, we also jointly train the system with text from the VHED dataset [15] to augment the size of training corpus. This dataset contains image captions corresponding to sequences of images which form short stories. Each set of captions is also labeled by human annotators with an average quality ranking of the overall story on a scale of 1 to 5. The VHED dataset is a composite of multiple audio-visual story-telling datasets with a wide representation of story tasks. Our intention in using it here is to implicitly teach the model to score story quality in addition to being trained on in-domain data from the GSU Kids training set. We use 80% of the VHED dataset (roughly 10,000 text samples) in training the language model to simultaneously predict cumulative scores of the GSU Kids story-telling assessment and average quality rankings of the VHED text samples. We show the classification accuracy to demonstrate overall system performance, F1-score to demonstrate fair performance across immbalanced classes, and root mean square error to show the magnitude of the machine's errors in Table 1.

4. Conclusions and Next Steps

Our results show that the proposed multitask training scheme used with BERT achieves high accuracy in predicting the overall scores in the GSU Kids story-telling samples. The next sentence prediction training objective and larger parameter size of BERT may contribute to its improved performance over the other models. We note that ALBERT, the model with the fewest parameters, is least robust to an increase in WER in the input transcripts. While transcriptions for chinldren's speech generated from ASR systems still contain several errors, we demonstrate that the usage of language models helps extract high level linguistic features inspite of the high WER of these systems. In the near future, we will train the system to predict individual score components in order to return a detailed score report for each system. For example, the annotators have marked whether or not the child included character names in their story, if they have described key parts of the scene in the picture, and if they keep the same verb tense throughout their story. The promising results in Table 1 imply that, given the annotator labels, the system can be trained to predict these characteristics of the student response individually. Teachers can use use such results to understand which areas a student needs to improve in.

5. References

[1] A. Johnson, H. Veeramani, B. Natarajan, and A. Alwan, "An equitable framework for automatically assessing children's oral narrative language abilities," *Proc. Interspeech*, 2023.

[2] L. Venkatasubramaniam, V. Sunder, and E. Fosler-Lussier, "End-to-end word-level disfluency detection and classification in children's reading assessment," in *IEEE ICASSP*, 2023, pp. 1–5.

[3] B.-C. Yan, H.-W. Wang, Y.-C. Wang, and B. Chen, "Effective graph-based modeling of articulation traits for mispronunciation detection and diagnosis," in *ICASSP*, 2023, pp. 1–5.

[4] S. Dutta *et al.*, "Challenges remain in building asr for spontaneous preschool children speech in naturalistic educational environments," *Proc. Interspeech 2022*, pp. 4322–4326, 2022.

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

[6] L. Chen *et al.*, "Automated scoring of nonnative speech using the speechratersm v. 5.0 engine," *ETS Research Report Series*, vol. 2018, no. 1, pp. 1–31.

[7] R. B. Gillam and N. A. Pearson, *Test of narrative language*. PRO-ED. Inc., 2017.

[8] E. L. Fisher *et al.*, "Executive functioning and narrative language in children with dyslexia," *American journal of speech-language pathology*, vol. 28, no. 3, pp. 1127–1138, 2019.

[9] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[10] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM TASLP*, vol. 29, p. 3451–3460, oct 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, pp. 4171–4186.

[12] Z. Lan *et al.*, "ALBERT: A lite BERT for self-supervised learning of language representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019. [Online]. Available: http://arxiv.org/abs/1910.01108

[14] Z. Yang *et al.*, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] C.-Y. Hsu *et al.*, "Learning to rank visual stories from human ranking data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 6365–6378.