INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Development of Robust Automated Scoring Models Using Adversarial Input for Oral Proficiency Assessment

Su-Youn Yoon, Chong Min Lee, Klaus Zechner and Keelan Evanini

Educational Testing Service
660 Rosedale Road, Princeton, NJ 08541, USA
{syoon, clee001, kzechner, kevanini}@ets.org

Abstract

In this study, we developed an automated scoring model for an oral proficiency test eliciting spontaneous speech from non-native speakers of English. In a large-scale oral proficiency test, a small number of responses may have atypical characteristics that make it difficult even for state-of-the-art automated scoring models to assign fair scores. The oral proficiency test in this study consisted of questions asking about content in materials provided to the test takers, and the atypical responses frequently had serious content abnormalities. In order to develop an automated scoring system that is robust to these atypical responses, we first developed a set of content features to capture content abnormalities. Next, we trained scoring models using the augmented training dataset, including synthetic atypical responses. Compared to the baseline scoring model, the new model showed comparable performance in scoring normal responses, while it assigned fairer scores for authentic atypical responses extracted from operational test administrations.

Index Terms: automated speech scoring, content scoring, speech recognition, non-native speech, adversarial input

1. Introduction

An automated oral proficiency scoring system can assess spoken responses faster than human raters, and often at lower cost, with the resulting scores being consistent over time. These advantages have prompted strong demand for high-performing automated scoring systems for large-scale language proficiency assessments.

In oral proficiency assessments, some responses may have sub-optimal characteristics. For instance, some test takers may try to game the system by citing memorized responses for unrelated topics, instead of generating his/her own response. Even state-of-the art automated scoring models face challenges in scoring these atypical responses [1, 2, 3], and researchers in the automated scoring field have tried to solve this issue using a two-step approach where an automated filtering model, as a sub-module of an automated scoring system, filters out atypical responses, and only the remaining responses are scored by the scoring model. A spoken canned response detection system in [4], an off-topic response detection system in [5, 6] and a coherence model in [7] are examples of this approach.

However, the percentage of atypical responses is typically very low, and developing high-performing filtering models using authentic atypical responses is an extremely difficult task. Recently, researchers have explored augmenting training data by constructing synthetic samples (e.g., [8]). In the educational field, [9, 10] augmented training data by generating synthetic errors using neural models trained on a small corpus of human-annotated data. Finally, they used the augmented training data for training of a grammatical error detection system.

Similarly, synthesized atypical responses have been used to augment the training data of filtering models. [5, 6] generated off-topic responses by mismatching questions and responses, and [7] generated incoherent responses by randomly shuffling the order of sentences in the original responses. They showed that the filtering models trained on the augmented data, including both normal responses and synthetic atypical responses, achieved promising performance in identifying synthetic atypical responses, but none of these studies reported the performance of the system on the authentic atypical responses. Furthermore, these filtering models were typically specialized for one specific atypical response type, and performance on different types was unknown. Despite the low percentage of atypical responses, the characteristics can vary widely and it is an extremely challenging task to develop and maintain high quality filtering models for each atypical response type.

In this study we developed an automated scoring system to assess holistic speaking proficiency levels from non-native speakers' spontaneous speech, elicited using questions asking about the content of provided stimulus reading or listening materials. Different from the previous studies that develop a series of filtering models to detect each atypical response type, we developed an automated scoring model robust to various atypical responses. This approach has an advantage that the system can be developed more rapidly and maintained more efficiently.

For this purpose, we (a) augment training data to include atypical responses and (b) develop new content features based on the word embeddings. Our contribution is twofold. Instead of augmenting the training data by generating atypical responses based on hypotheses about test takers' behaviors, we started from analyzing the small set of the authentic atypical responses and generated synthesized responses based on the observed characteristics of these responses. Finally, different from the previous studies, we used authentic atypical responses for the system evaluation. Next, we developed a set of content features that are easy to train and maintain. The word-embedding features used in this study do not require any sample training responses for each question.

2. Data

In this study, we used a large collection of spoken responses from an oral proficiency test for non-native English speakers. For each speaker, we used 4 questions for which speakers were prompted to provide around one minute of spontaneous speech per question, resulting in approximately 4 minutes of speech per speaker.

For each question, speakers read and/or listened to a passage and then provided answers based on the information in the passage. We used a total of $49,490$ responses prompted by a set of 80 questions (hereafter, Original set).

http://dx.doi.org/10.21437/Interspeech.2019-1711

All responses were scored by trained raters using a 4-point scoring scale from 1 to 4 with 4 indicating the highest proficiency. We used the TOEFL iBT Speaking Test Rubrics[11]. The rubrics consisted of three major performance categories: delivery (pronunciation, intonation, rhythm, and fluency), language use (vocabulary and grammar), and topic development (content and coherence). In addition, raters provided a score of 0 when test takers did not exhibit any intention to directly respond to the question. Inter-rater agreement was calculated from 10% double scored data, and both Pearson correlation and quadratic weighted kappa were 0.61.

The average of the human scores was 2.58, and the most frequent score was 3 (48%) and followed by 2 (39%), 4 (8%), 1 (4%), and 0 (1%). The percentage of responses with score of 0 was low, while their characteristics were varied. The most frequent categories of 0-score responses included (a) no-responses including cases with no speech other than fillers or simple sentences (e.g., "I don't know"); (b) responses in non-target language; (c) off-topic; (d) canned responses[1]; and (e) repetition of the question. These responses tended to have serious content issues.

Our main goal was to develop a reliable model for scoring these atypical responses. However, due to the skewed distribution, it was difficult to train and evaluate the model on the atypical responses in the original dataset alone. In order to address this issue, we used two additional data sets: (a) the simulated off-topic response set (hereafter, OFF) and (b) the atypical response set (hereafter, ZERO).

For the OFF set, we artificially generated question-response-mismatches. We first selected 677 questions that did not overlap with the previous 80 questions used for the Original set. For each question in the Original set, we randomly selected approximately 60 responses from the responses to the 677 non-overlap questions. This resulted in a total of 4,939 new, simulated responses (approximately 10% of Original set). Each question asked about substantially different content from other questions, and, therefore, mismatched responses had substantial content issues to the extent to be considered off-topic. We did not re-score these responses as answers for the new question we randomly assigned. Because of the inappropriate content for the new questions, we assigned a score of 0 for these responses during model training. Next, for the atypical response set, we extracted a large numbers of responses with score of 0 from the same English proficiency assessment, but from a larger number of test administrations (ZERO set). The size of the data set is presented in Table 1.

Table 1: *Number of responses for each partition*

Purpose	Partition	# responses
Train	Original	24,278
	OFF	2,376
Eval	Original	25,212
	OFF	2,563
	ZERO	986

During the question generation, expert assessment developers generated a list of key points to guide the creation of the reading and listening passages. These key points were provided to and used by human raters to evaluate content of the spoken responses. We also used them for the content feature development. They created three key-points per question: key-point1 was about the mentioning of the concepts introduced in the source materials or the general opinions voiced (i.e., agree or disagree with a situation/change/proposal). Depending on the nature of the task questions, key-point2 and key-point3 involved brief definitions of the concepts, reasons provided for the opinions voiced, or detailed examples that illustrated the topics or concepts discussed. Thus, key-point1 was shorter and contained simpler content, while key-point2 and key-point3 contained more complicated content.

We normalized key-points by removing stop words and disfluencies. Next, we created a word list containing all words (ALL) for each key-point. While some words (e.g., the topic or the concept name) appeared in multiple key-points, some words were unique to a particular key-point. Under the assumption that these unique words may be more important for detecting the absence of the specific key-points, we created two additional word lists: a unique word list (Unique) and a shared word list (Shared) that contained the words that appeared in the particular key-point and also other key-point(s). This resulted in a total of three word lists (ALL, Unique, Shared) for each key-point.

3. Method

3.1. Features

We used two sets of features: speech-driven features and content features. For a given spoken response, a state-of-the-art automated proficiency scoring system[12] generated 35 speech-driven features assessing fluency, pronunciation, and prosody. They were: (a) speech rate features (3 features), (b) pronunciation quality features[2] (6 features), (c) pause pattern features (9 features), (d) prosody features[3] (11 features), and (e) duration features[4] (6 features). The detailed descriptions of the selected features were provided in [13].

In addition, the system generated 28 features assessing the quality of the content. The first feature group within this set was designed to assess lexical similarity with high-scoring responses. These features used a tf-idf weighted cosine similarity score between a test response vector and the question-specific tf vector. The question-specific tf vector was a vector whose elements were the frequency of each word in the entire sample responses with score of 4 that answered the same question. The second feature group was designed to measure semantic similarity with key-points. First, a response was segmented into a sequence of word n-grams [5] with 5 words overlap between two consecutive n-grams. For each word n-gram, the similarity with a particular key-point was calculated using three word-embedding based metrics: Word Mover's Distance ([14]), the cosine similarity between two averaged word embedding vectors ([13]), and query-document similarity metrics ([15]). For each embedding metric, we selected the minimum value among the all n-grams in a response. We repeated this process for 3 key-point word lists (Shared, Unique, All) and 3 key-points, resulting in a total of 27 features (3 embedding-based simi-

[1]Responses that only include memorized segments from external sources. The sources were irrelevant to the question, and the responses were likely to be off-topic.

[2]This group of features measures how much the test takers' pronunciation deviates from the native norms.

[3]This group of features measures patterns of variation in time intervals between syllables or phonemes.

[4]This group of features captures variation in the duration of vowels and consonants.

[5]n = the number of words in a key-point after the normalization. The number of words in key-points was presented in Table 2.

larity metrics * 3 word lists per key-point * 3 key-points). We used the publicly available word embedding vectors trained on the Google News corpus by [16] for all word-embedding based features and WM-distance implementation in the gensim package[17] for WM-distance calculation.

For content feature generation, both spoken responses and key-points was normalized by removing stop words and disfluencies. Table 2 provides the average length of the spoken responses in the Original partition and key-points before and after the normalization process.

Table 2: *Number of words in the spoken responses and the key-points*

	Text characteristics	mean	STD
Original	original	129.5	31.1
	normalized	53.1	13.7
key-point1	original	24.6	10.0
	normalized	13.9	5.0
key-point2	original	32.3	12.0
	normalized	17.6	6.9
key-point3	original	36.2	10.9
	normalized	19.3	6.4

The distribution of the response length in the OFF set was comparable to that of the Original partition. The distribution of the response length in ZERO set was substantially different from both sets, and it was presented in Table 5 in Section 5.

After the normalization process, the length of the key points and responses were reduced to 55% and 40% of the original texts on average.

3.2. Model

We trained models in two different conditions:

- Unaugmented: the model was trained on the Original Train partition.
- Augmented: the model was trained on the combination of the Original and OFF Train partition containing approximately 10% of synthetic atypical responses.

We trained regression models using both speech-driven and content features as the independent variables and the human score as the dependent variable.

4. Experiment

For each spoken response, the system performed speech processing including speech recognition, forced-alignment, and pitch/power analysis. It first generated word hypotheses for each response using an automated speech recognition (ASR) system. A gender independent acoustic model (AM) was trained on 800 hours of spoken responses extracted from the same English proficiency test using the Kaldi toolkit [18]. The AM training dataset consisted of 52,200 spoken responses from 8,700 speakers, and this dataset did not overlap with any datasets described in section 2. It was based on a 5-layer DNN with p-norm nonlinearity using layer-wise supervised back-propagation training. The language model (LM) was a trigram model trained using the same dataset used for AM training. This ASR system achieved a Word Error Rate of 23% on 600 held-out responses. Detailed information about the ASR system is provided in [19].

Next, the system generated both speech-driven features and content features. In order to compare the performance of the content features with the speech-driven features, we trained three models: Speech (model based on 35 speech-driven features), Content (model based on 28 content features), and ALL (model based on both speech-driven and content features, 63 features in total). Finally, a total of 6 models (3 feature groups * 2 training data sets) were trained using the RandomForestRegressor algorithm[6] in scikit-learn[20].

5. Results

First, we evaluated the performance of the automated scoring models for scoring the normal responses using the Original Evaluation set. Table 3 presents the agreement between the human scores and the automated scores for each model.

Table 3: *Correlations, quadratic weighted kappas, and root mean squared error (RMSE) between the automated scores and human scores for normal responses*

Train set	Features	corr	κ	RMSE
Unaugmented	Content	0.534	0.419	0.592
	Speech	0.644	0.528	0.536
	ALL	0.660	0.542	0.526
Augmented	Content	0.510	0.440	0.609
	Speech	0.600	0.439	0.607
	ALL	0.642	0.546	0.538

Among the models trained on the unaugmented train set, the Speech-based model substantially outperformed the Content-based models. The combination of the two feature groups achieved a slight further improvement and this improvement was statistically significant at 0.01 level ($p < 0.01$), based on the Steigers Z-test for dependent correlations. We observed very similar trends from the models trained on the augmented Train set: ALL> Speech> Content. However, the combination of the feature groups resulted in much larger improvement over the individual-feature based models, and the increases in the correlation and weighted kappa over the best performing individual model were 0.042 and 0.107, respectively.

In both conditions, the best performing model was the model based on both feature groups (ALL). The correlation and RMSE of the unaugmented-ALL model was better than the augmented-ALL model, while the weighted kappas of the two models were comparable.

Next, we analyzed automated scores for atypical responses on the OFF ad ZERO sets. All responses in both sets contained serious content issues, and thus, the scores were expected to be 0. However, automated scoring models susceptible to atypical responses tend to assign inflated scores. In order to examine whether the automated models in this study have the score inflation tendency for atypical responses, we compared the average automated scores of the models. Table 4 provides the mean and standard deviation of the automated scores in both OFF and ZERO set.

In general, the automated scores for the unaugmented-models were high. Despite the serious content issues of the simulated off-topic responses, the average score for the Speech model was comparable to the average human scores of

[6]The RandomForestRegressor was selected because of its superior performance over different machine learning algorithms in pilot experiments.

Table 4: *Mean and standard deviation of the scores predicted by the models for synthetic (OFF) and authentic (ZERO) atypical responses*

Train set	Features	OFF		ZERO	
		mean	STD	mean	STD
Unaugmented	Content	1.594	0.219	1.359	0.226
	Speech	2.683	0.448	1.514	0.432
	ALL	2.191	0.438	1.409	0.336
Augmented	Content	0.288	0.466	0.898	0.471
	Speech	2.378	0.389	1.356	0.399
	ALL	0.207	0.464	1.016	0.375

the Original dataset (2.58). This was the expected result since the Speech model did not include any features to model content abnormality. The average score of the Content model was substantially lower than the Speech model, but it was still 1.594. The ALL model scored slightly lower than the Speech model demonstrating the impact of the content features on modeling the content issue. On the contrary, the automated scores for the augmented-models were generally low; the average scores for both the Content model and the ALL model were lower than 0.5. However, the average score of the Speech model was still high (2.378) and only slightly lower than that of the unaugmented-model.

The comparison of ALL models with Speech models demonstrate the impact of both content features and an augmented training dataset. In both augmented- and unaugmented-conditions, the automated scores of ALL models were consistently lower than those of Speech models, and this showed that the content features can identify content abnormality and assign more accurate scores. However, the impact of content features in the unaugmented condition was substantially weaker than in the augmented condition suggesting the importance of encountering more atypical responses during training.

Finally, we evaluated whether the automated scoring models assign fair scores for authentic atypical responses using the ZERO set. In general, the average scores of the automated models were low in all conditions, but those of the augmented models were even lower.

We further analyzed model performance on scoring of atypical responses with content abnormality. Because these responses were longer than no-speech responses, we investigated the relationships between the automated scores and response length. Table 5 presents the distribution of the zero response length. Figure 1 and 2 present the average automated scores of the authentic atypical responses by response length.

Table 5: *Distribution of the number of words in zero responses*

Number of words	# of responses	Percent (%)
0-20	343	35
20-40	271	27
40-60	158	16
60-80	88	9
80-100	66	7
100-120	34	3
120-172	26	3

In the unaugmented condition, the automated scores of models were relatively low for the short responses, but they increased substantially as the response length increased. However, in the augmented condition, the automated scores except Speech model were consistently low, and even the average scores for long responses (e.g., number of words > 80) was lower than 1.0.

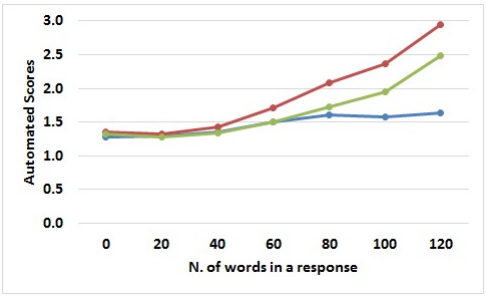

Figure 1: *Average score of the responses in ZERO set predicted by the unaugmented models: Content (blue), Speech (red), and ALL (green)*

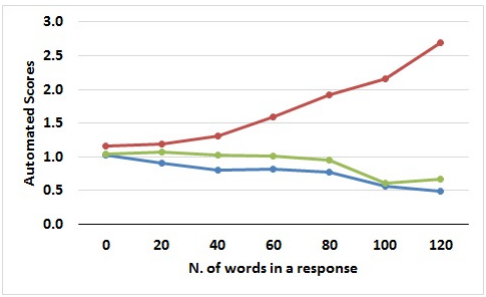

Figure 2: *Average score of the responses in ZERO set predicted by the augmented models: Content (blue), Speech (red), and ALL (green)*

The consistently low scores of the augmented ALL model indicated that the combination of the new content features and an augmented training dataset can indeed improve the robustness of automated models in modeling authentic atypical responses. This approach therefore results in more accurate and reliable automated scores.

6. Conclusions

We developed an automated scoring model that is robust to atypical responses for oral proficiency test eliciting spontaneous speech. The atypical responses in this test were typically associated with content abnormality, and even a state-of-the art scoring model had serious scoring errors if it was not explicitly designed to handle these responses. In order to address this issue, we first developed a set of content features based on word-embeddings. Next, we augmented the training dataset by adding synthesized off-topic responses. The inclusion of new content features in the unaugmented condition resulted in a slight but statistically significant further improvement over the baseline scoring model. Furthermore, the combination of the content features and the augmented training data, including synthetic atypical responses, substantially reduced scoring errors for the authentic atypical responses while maintaining comparable performance on scoring of the normal responses.

7. References

[1] D. Higgins and M. Heilman, "Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior," *Educational Measurement: Issues and Practice*, vol. 33, no. 3, pp. 36–46, 2014.

[2] H. Yannakoudakis and T. Briscoe, "Modeling coherence in ESOL learner texts," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, 2012, pp. 33–43.

[3] S.-Y. Yoon, A. Cahill, A. Loukina, K. Zechner, B. Riordan, and N. Madnani, "Atypical inputs in educational applications," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, vol. 3, 2018, pp. 60–67.

[4] X. Wang, K. Evanini, J. Bruno, and M. Mulholland, "Automatic plagiarism detection for spoken responses in an assessment of english language proficiency," in *Proceedings of the Workshop on Spoken Language Technology Workshop (SLT)*, 2016, pp. 121–128.

[5] A. Malinin, K. Knill, and M. J. Gales, "A hierarchical attention based model for off-topic spontaneous spoken response detection," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 397–403.

[6] C. M. Lee, S.-Y. Yoon, X. Wang, M. Mulholland, I. Choi, and K. Evanini, "Off-topic spoken response detection using siamese convolutional neural networks." in *INTERSPEECH*, 2017, pp. 1427–1431.

[7] Y. Farag, H. Yannakoudakis, and T. Briscoe, "Neural automated essay scoring and coherence modeling for adversarially crafted input," *arXiv preprint arXiv:1804.06898*, 2018.

[8] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.

[9] S. Kasewa, P. Stenetorp, and S. Riedel, "Wronging a right: Generating better errors to improve grammatical error detection," *arXiv preprint arXiv:1810.00668*, 2018.

[10] M. Rei, M. Felice, Z. Yuan, and T. Briscoe, "Artificial error generation with machine translation and syntactic patterns," *arXiv preprint arXiv:1707.05236*, 2017.

[11] Educational Testing Service. (n.d.) Integrated speaking rubrics. [Online]. Available: https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

[12] L. Chen, K. Zechner, S.-Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma *et al.*, "Automated scoring of nonnative speech using the speechrater sm v. 5.0 engine," *ETS Research Report Series*, vol. 2018, no. 1, pp. 1–31, 2018.

[13] S.-Y. Yoon, A. Loukina, C. M. Lee, M. Mulholland, X. Wang, and I. Choi, "Word-embedding based content features for automated oral proficiency scoring," in *Proceedings of the Third Workshop on Semantic Deep Learning*, 2018, pp. 12–22.

[14] M. Kusner, Y. Sun, N. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 957–966.

[15] S. Kim, W. J. Wilbur, and Z. Lu, "Bridging the gap: a semantic similarity measure between queries and documents," *arXiv preprint arXiv:1608.01972*, 2016.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[17] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[19] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 6140–6144.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.