Atypical Inputs in Educational Applications

Su-Youn Yoon, Aoife Cahill, Anastassia Loukina,
Klaus Zechner, Brian Riordan, Nitin Madnani
Educational Testing Service
660 Rosedale Road, Princeton, NJ
syoon, acahill, aloukina, kzechner, briodan, nmadnani@ets.org

Abstract

In large-scale educational assessments, the use of automated scoring has recently become quite common. While the majority of student responses can be processed and scored without difficulty, there are a small number of responses that have atypical characteristics that make it difficult for an automated scoring system to assign a correct score. We describe a pipeline that detects and processes these kinds of responses at run-time. We present the most frequent kinds of what are called non-scorable responses along with effective filtering models based on various NLP and speech processing technologies. We give an overview of two operational automated scoring systems —one for essay scoring and one for speech scoring— and describe the filtering models they use. Finally, we present an evaluation and analysis of filtering models used for spoken responses in an assessment of language proficiency.

1 Introduction

An automated scoring system can assess constructed responses such as essays to open-ended questions faster than human raters, 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 various educational applications. However, even state-of-the-art automated scoring systems face numerous challenges when used in a large-scale operational setting. For instance, some responses have atypical characteristics that make it difficult for an automated scoring system to provide a valid score. A spoken response, for example, with a lot of background noise may suffer from frequent errors in automated speech recognition (ASR), and the linguistic features generated from the erroneous ASR word hypotheses may be inaccurate. As a result, the automated score based on the inaccurate features may differ greatly from the score a human expert would assign. Furthermore, it may substantially weaken the *validity* of the automated scoring system.[1] More recently, some studies have systematically evaluated the impact of atypical inputs, particularly gaming responses, on the validity of automated scoring of essays (Lochbaum et al., 2013) and short-answers (Higgins and Heilman, 2014). They showed that automated scoring systems tend to be more vulnerable than human raters to students trying to game the system. Consistent with these findings, Zhang (2013) argued that the ability to detect abnormal performance is one of the most important requirements of a high-stakes automated scoring system. However, despite its importance, and compared to the large body of work describing the empirical performance of automated scoring systems, there has been little discussion of how NLP tools and techniques can contribute to improving the detection of atypical inputs.

In this paper we present a typical processing pipeline for automated scoring of essay and spoken responses and describe the points in the process where handling of atypical inputs and system failures can occur. In particular, we describe some of the NLP technologies used at each of these points and the role of NLP components as *filters* in an automated scoring pipeline. We present a case study on building automated filters to detect problematic responses in an assessment of spoken English.

2 Detecting Atypical Inputs

In this section, we give an overview of an automated scoring pipeline and describe how atypical

[1] Test validity is the extent to which a test accurately measures what it is supposed to measure.

Proceedings of NAACL-HLT 2018, pages 60–67
New Orleans, Louisiana, June 1 - 6, 2018.

inputs can be detected and processed in applications that automatically score the language quality of spoken responses and essays.

A typical automated scoring pipeline has three major components: (1) the student response is captured by the input capture module; (2) the system computes a wide range of linguistic features that measure various aspects of language proficiency using NLP and/or speech processing technologies; and (3) a pre-trained automated scoring model predicts a proficiency score using the linguistic features. The components above the dotted line in Figure 1 illustrate the typical processing sequence.

However, this scoring pipeline is a simplified version of what happens in a real-life operational system. In a large-scale assessment, there are usually a small number of atypical responses, where the automated scoring system would have difficulty in predicting *valid* scores.[2] We call these problematic responses *non-scorable*. In order to handle these problematic inputs in real-time, we can add *filtering models* (FMs) as sub-modules of the automated scoring systems. The filtering models detect and process different kinds of problematic inputs at different points in the typical pipeline. Figure 1 indicates three points at which filtering models could be employed in an operationally-deployed automated scoring system (below the dotted line). The points at which FMs could be introduced are after (1) input capture, (2) feature generation and (3) score generation. Responses that are flagged by the FMs are handled in different ways depending on the application context. The responses can receive an automated score (with a warning that it is unreliable), they can be rejected and receive no score, or they can be sent to a human rater for manual scoring.

Analogous to the FMs, non-scorable responses can also be classified into three main groups. Non-scorable responses in the feature generation and score generation groups are mostly system-initiated non-scorable responses, where system components had critical errors at the feature generation or score generation stages. Non-scoreable responses in the input capture group can be further classified into (a) system-initiated and (b) user-initiated non-scorable responses. User-initiated non-scorable responses can occur for a number of reasons, both for good faith (where the students try their best to answer the question appropriately) and bad faith (where the students do not make a reasonable attempt to answer the question) attempts. Students making good-faith attempts to answer questions may still present an automated scoring system with atypical input. For example, in a speaking assessment, inappropriate distance to the microphone may result in a distorted recording; or in a written assessment a student may misread the question, and unintentionally write an off-topic response. Bad faith responses can come from unmotivated students, or students trying to game the system. These responses represent a wide range of (often very creative) inputs that can be troublesome for an automated scoring system that is ill-prepared to handle them. If system components fail or the system is not confident in its prediction, the automated processing pipeline also needs to be able to handle this correctly and fairly.

Next we describe some of the techniques that have been proposed for detecting non-scorable responses at each of the stages.

2.1 Input Capture Filtering Models

Some system-initiated errors that should be flagged at the input capture stage are: severe audio problems typically caused by equipment malfunction, background noise, or problems with test takers' recording level (speaking proficiency assessment) or text capture failures (writing proficiency assessment).

There is also a wide range of potential user-initiated non-scorable responses. Some of the most frequent categories include (a) response in non-target language; (b) off-topic; (c) generic responses;[3] (d) repetition of the question; (e) canned responses;[4] (f) banging on the keyboard; and (g) no-response.

Five of the categories mentioned relate to topicality. Off-topic responses and generic responses are unrelated to the prompt, while the prompt-repetition responses and canned responses can be considered repetition or plagiarism. For automated essay scoring, off-topic detection systems

[2]Many scoring systems could produce *some* score for these problematic responses, however it is unlikely to be correct. It is therefore important for the overall validity of the test and automated scoring system, to be able to identify such responses and treat them correctly.

[3]Responses that only include simple unrelated sentences such as "I don't know," "this is too difficult" "why do I have to answer," etc.

[4]Responses that only include memorized segments from external sources (often websites)

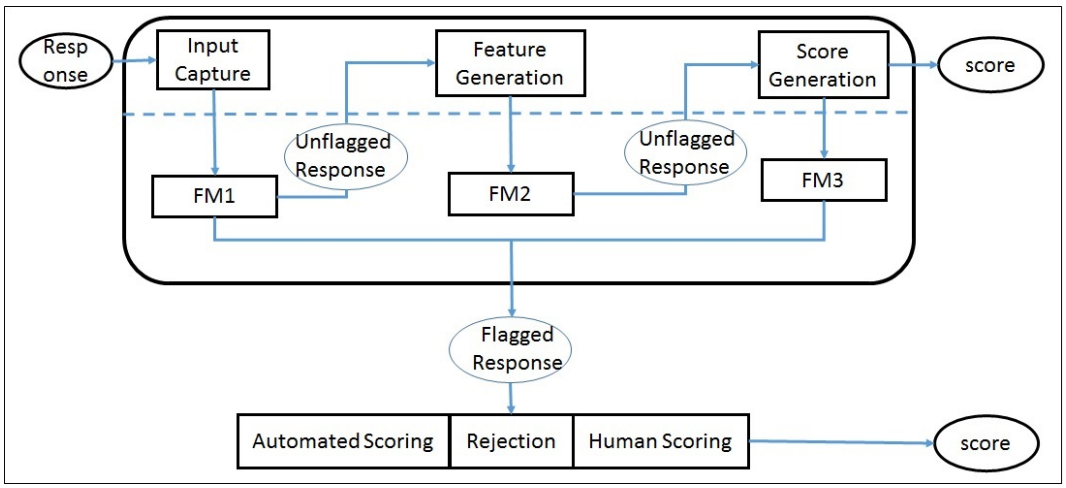

Figure 1: A diagram of the overall architecture of a generic automated scoring pipeline. Above the dotted line are the key stages in automated scoring. Below the dotted line are the possible additions to the pipeline to handle atypical inputs using filtering models (FMs).

have been developed based on question-specific content models, such as a standard vector space model (VSM) built for each question (Bernstein et al., 2000; Higgins et al., 2006; Louis and Higgins, 2010).

For speaking tests eliciting highly or moderately restricted speech, filtering models based on features derived from ASR systems such as normalized confidence scores and language model (LM) scores can achieve good performance in identifying topic-related non-scorable responses (van Doremalen et al., 2009; Lo et al., 2010; Cheng and Shen, 2011). However, this approach is not appropriate for a speaking test that elicits unconstrained spontaneous speech. More recently, similar to techniques that have been applied in essay scoring, systems based on document similarity measures and topic detection were developed to detect spoken non-scorable responses. In addition, neural networks and word embeddings, which have the advantage of capturing topically relevant words that are not identical, have been used in Malinin et al. (2017) and Yoon et al. (2017), and this has resulted in further improvements over systems using only traditional lexical similarity features.

Unlike off-topic responses, canned responses include pre-existing material. These can often be identified by matching responses to test preparation websites or other student responses. Potthast et al. (2014) give an overview of approaches to detecting plagiarism in written texts based on the systems that competed in the PAN-2014 shared task on plagiarism detection. Wang et al. (2016) developed a spoken canned response detection system using similar techniques applied in essay plagiarism detection.

In addition, various speech processing and NLP techniques have also been used to detect other types of non-scorable responses: language identification technology for non-English detection (Yoon and Higgins, 2011) and speaker recognition technology for automated impostor detection (Qian et al., 2016). "Banging on the keyboard" can be identified by analyzing part-of-speech sequences and looking for ill-formed sequences (Higgins et al., 2006).

2.2 Feature Generation Filtering Models

The most typical way for a response to be flagged at the Feature Generation stage is for an internal component to fail. For example, in an automated speech scoring system the ASR system, or the speech-signal processing component may fail. In addition, parsers and taggers also sometimes fail to produce analyses, particularly on ill-formed language-learner responses. In order to detect sub-optimal ASR performance, filtering models have been developed using signal processing technology and features derived from ASR systems, e.g., confidence scores and normalized LM scores (Jeon and Yoon, 2012; Cheng and Shen, 2011).

It should be noted that while some of the user-initiated non-scorable responses undetected at the input capture stage would likely also cause fea-

ture generation failures (e.g., no-speech may cause empty ASR hypothesis which result in feature generation failure), other types (e.g., gaming responses) would simply cause subtle differences in feature values leading to potential inflation of the automated scores without causing clear subprocess failures.

2.3 Score Generation Filtering Models

It is also possible to identify responses that may not have received a correct score from the automated scoring system by looking at the output of the scoring model directly. van Dalen et al. (2015) developed an automated essay scoring system that uses a Gaussian process to not only generate proficiency scores, but also give a measure of the uncertainty of the generated score. They proposed a process that uses the uncertainty measure to filter responses with high uncertainty and send them to human raters for scoring.

2.4 Adjusting Scoring for non-scorable responses

We consider two main scoring scenarios:

- Human raters in the loop: there are several ways that human scoring can be combined with automated scoring. The two most common situations are co-grading (a majority of responses are scored by both human raters and the automated scoring system) and hybrid scoring (a majority of the responses are scored by the automated scoring system, while only subset of responses are scored by human raters for quality control purposes).

- Sole automated scoring: all responses are scored by only the automated scoring system; there are no humans involved in scoring. Such situations could include practice tests or classroom tools.

If a response is flagged as non-scorable in a scoring situation that has a human in the loop, the most typical behavior is for the response to be sent to a human rater for additional scoring. The score from the automated system may or may not be combined with human scores, depending on the use case and the kind of flag.

If a response is flagged as non-scorable in a sole scoring situation, there are two main ways to process the response. Either no score is given and a message is returned to the user that their response could not be successfully processed. Or alternatively, a score is given with a warning that it is unreliable.

3 Practical Implementation of Filtering Models

In this section we describe two systems for automated scoring of CRs: (1) e-rater – an automated scoring system for essays and (2) $SpeechRater^{SM}$ – an automated scoring system for spoken responses. We describe the kinds of filters used by both systems.

3.1 Automated Essay Scoring

The e-rater system (Attali and Burstein, 2006) automatically evaluates the writing quality of essays by taking key aspects of writing into account. It aligns the writing construct, via scoring rubrics, to NLP methods that identify various linguistic features of writing. The feature classes include the following: (a) grammatical errors (e.g., subject-verb agreement errors), (b) word usage errors (e.g., their versus there), (c) errors in writing mechanics (e.g., spelling), (d) presence of essay-based discourse elements (e.g., thesis statement, main points, supporting details, and conclusions), (e) development of essay-based discourse elements, (f) a feature that considers correct usage of prepositions and collocations (e.g., powerful computer vs. strong computer), and (g) sentence variety. The features are combined in a simple linear model learned from an appropriate data sample for the target populations.

In a high-stakes testing scenario with e-rater, there is a human in the loop. The first step in the pipeline is that a human rater reviews the essay (*FM1*). If they deem the essay to be non-scorable (e.g., because it is off-topic, or gibberish), the essay is immediately sent to another human for adjudication and there is no automated score produced.

If the first human rater assigns a valid score to the essay, it is then passed to the e-rater engine. The e-rater engine applies a filter that does automated garbage detection as its first step and filters out responses that it detects as being non-English (*FM1*). This filter uses unusual POS tag sequences to identify non-English responses. The number of such responses should be very low with a human as the first filter. Non-garbage responses are then passed to the next stage of processing – feature extraction. At this point there are two differ-

ent kinds of filters. The first kind of filter compares the response to other responses seen during the training of the model. If the current response is too dissimilar (in terms of length, vocabulary, etc.), it is flagged (*FM2*). These filters rely on typical textual-similarity techniques including content-vector analysis (Salton et al., 1975) and distributional similarity. The second kind of filter flags responses for which the engine cannot reliably determine a score because an internal component (e.g., a parser) has indicated that its output may be unreliable (*FM2*). In both cases flagged responses are sent to a human for adjudication.

Table 1 gives a summary of the frequency of each of the types of filters applied in a high-stakes standardized test of English proficiency. The values were computed using a large sample of over 2 million candidate essays. The final score for the responses can range from 0–5. It can be seen that the average final score assigned to most of the essays flagged is very low. The average final score assigned to essays automatically flagged as being too dissimilar to the training data is higher. Typically this category of flags are designed to be conservative and sometimes flag perfectly reasonable essays, simply to err on the side of caution.

	Freq. (%)	Avg. Score
FM1 (human)	0.47	0.04
FM1 (automated)	0.02	0.96
FM2 (dissimilar)	0.95	2.65
FM2 (engine uncertainty)	0.06	1.09

Table 1: Frequency of different kinds of filters in high-stakes e-rater deployment with the average final score assigned to the responses flagged by each filter.

3.2 Automated Speech Scoring

The $SpeechRater^{SM}$ system is an automated oral proficiency scoring system for non-native speakers' of English (Zechner et al., 2009). It has been operationally deployed to score low-stakes speaking practice tests. In order to score a spoken response, the input capture module in $SpeechRater^{SM}$ records the audio. Next, the ASR system generates word hypotheses and time stamps. The feature generation modules create a wide range of linguistic features measuring fluency, pronunciation and prosody, vocabulary and grammar usage based on the ASR outputs and NLP and speech processing technologies (e.g., a POS tagger, a dependency parser, pitch and energy analysis software). In addition, it generates a set of features to monitor the quality of ASR and the audio quality of input responses.

Because of the low-stakes nature of the tests, only limited types of non-scorable responses have so far been observed in the data. There were some system-initiated non-scorable responses. Of the user-initiated non-scorable responses, the majority are no-response and the proportion of gaming responses is close to none. As a result, the filtering models in $SpeechRater^{SM}$ system are much simpler than the e-rater system; it consists of just one FM at the location of FM2, comprised of a set of rules along with a statistical model based on a subset of $SpeechRater^{SM}$ features. A detailed description and evaluation of its performance is summarized in Section 4. Finally, non-flagged responses are scored by the automated scoring moduleand flagged responses are not scored.

4 Case Study: Developing filtering models for an Automated Speech Scoring System

In this section, we will introduce a filtering model for $SpeechRater^{SM}$ developed for a low-stakes English proficiency practice test comprised of multiple questions which elicit unconstrained and spontaneous speech with duration of 45 to 60 seconds. All responses were scored by the automated scoring system (sole automated scoring scenario).

We collected $6,000$ responses from $1,000$ test takers, and expert human raters assigned a score on a scale of 1 to 4, where 1 indicates a low speaking proficiency and 4 indicates a high speaking proficiency. In addition, the raters also annotated whether each response fell into the input capture non-scorable group.[5] A total of 605 responses (10.1%) were annotated as being non-scorable responses. The majority of them were due to recording failures (7.0%), followed by no-response (3.0%) and non-English (0.1%). The data was randomly partitioned into Model Training ($4,002$ responses) and Model Evaluation ($1,998$ responses).

The ASR system was based on a gender-independent acoustic model and a trigram lan-

[5]The human raters did not annotate any errors caused by system component failures (system errors at the feature generation and scoring model stage).

guage model trained on 800 hours of spoken responses extracted from the same English proficiency test (but not overlapping with the Model Building set) using the Kaldi toolkit (Povey et al., 2011). In a separate study, the ASR system achieved a Word Error Rate (WER) of 37% on the held-out dataset (Tao et al., 2016). Although the performance of our ASR on non-native speakers' spontaneous speech was a state-of-the art, it showed substantially high WER for a small number of responses. Our analysis showed that the feature values and the scores for such responses were unreliable.

Initially, we developed two FMs. The Baseline FM was comprised of a statistical model trained on the Model Training partition and a set of rules to detect no-responses and responses with poor audio quality (recording errors). The extended FM was comprised of the baseline FM and an additional rule, ASRErrorFilter to detect responses for which the ASR result is likely to be highly erroneous.

The performance of two FMs on the evaluation dataset is given in Table 2.

	% flagged	acc.	pre.	recall	fscore
Baseline	9%	0.96	0.83	0.80	0.81
Extended	13%	0.95	0.66	0.90	0.76

Table 2: Performance of FMs in detecting non-scorable.

The accuracy and fscore of the baseline FM was 0.96 and 0.81, respectively. The extended model achieved a higher recall with slightly lower accuracy and fscore than the baseline. This was an expected result, since the extended model was designed to detect responses with high ASR errors that are likely to cause high human-machine score difference (HMSD) and we choose to err on the side of caution and flag more responses at the risk of flagging some good ones.

In order to investigate to what extent the extended FM can identify responses with a large HMSD, we also calculated an absolute HMSD for each response. After excluding responses annotated as non-scorable by human raters, the remaining 1,775 responses in the evaluation set were used in this analysis. Table 4 shows the distribution of absolute system-human score differences for flagged and unflagged responses by the extended FM.

	# response	HMSD		
		mean	SD	max
Flagged	77	0.83	0.58	2.38
Unflagged	1698	0.48	0.34	1.96

Table 3: Average human-machine score differences.

The average HMSD for the flagged responses was quite large (0.83), given that the proficiency score scale was 1 to 4. Furthermore, it was 1.73 times higher than that of unflagged responses (0.48). The extended FM indeed correctly identified responses for which the automated scores were substantially different from the human scores. In contrast, the number of responses flagged by the extended FM was substantially higher than the baseline FM. 4% of responses scorable by human raters were flagged and they could not receive scores.

5 Discussion and Conclusions

We discussed the issue of atypical inputs in the context of automated scoring. Non-scorable responses can cause critical issues in various subprocesses of automated scoring systems and the automated scores for these responses may differ greatly from the correct scores. In order to address this issue, we augmented a typical automated scoring pipeline and included a set of filtering models to detect non-scorable responses based on various NLP and speech processing technologies. Finally, we described two automated scoring systems deployed to score essay and spoken responses from large scale standardized assessments.

An alternative to the approach presented here (individual filtering models for different kinds of inputs), is to simply train one single classifier to detect non-scorable responses (perhaps as an additional score point). Depending on the context of the automated scoring system, this may be sufficient, however for our purposes, it was important to have more fine grained control over the different kinds of filters (setting thresholds, etc.). This gives us the freedom to treat and route each of the types of flags differently. This is particularly important if a system is being used in both high-stakes and low-stakes testing scenarios since the number and type of non-scorable responses varies by scenario.

Section 4, in its comparison between two FMs, highlights an important trade-off between the accuracy of automated scores and the percentage of

filtered responses by the filtering model. By proactively filtering out non-scorable responses, the automated scoring system (the extended FM, in this study) can prevent the generation of erroneous scores and improve the quality and validity of the automated scores. However, this may result in a higher percentage of scoring failures which could cause higher costs (e.g., additional human scoring, or providing free retest or refund to the test takers). These two important factors should be carefully considered during system development.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e–rater V.2. *Journal of Technology, Learning, and Assessment* 4(3).

J. Bernstein, J. DeJong, D. Pisoni, and B. Townshend. 2000. Two experiments in automated scoring of spoken language proficiency. In *Proceedings of the Workshop on Integrating Speech Technology in Learning*.

Jian Cheng and Jianqiang Shen. 2011. Off-topic detection in automated speech assessment applications. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. pages 1597–1600.

Derrick Higgins, Jill Burstein, and Yigal Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering* 12(2):145–159.

Derrick Higgins and Michael Heilman. 2014. Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice* 33(3):36–46.

Je Hun Jeon and Su-Youn Yoon. 2012. Acoustic feature-based non-scorable response detection for an automated speaking proficiency assessment. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association*. pages 1275–1278.

Wai-Kit Lo, Alissa M Harrison, and Helen Meng. 2010. Statistical phone duration modeling to filter for intact utterances in a computer-assisted pronunciation training system. In *Proceedings of the Conference on Acoustics, Speech and Signal Processing*. pages 5238–5241.

Karen E Lochbaum, Mark Rosenstein, Peter Foltz, and Marcia A Derr. 2013. Detection of gaming in automated scoring of essays with the IEA. *Presented at 75th Annual meeting of NCME* .

Annie Louis and Derrick Higgins. 2010. Off-topic essay detection using short prompt texts. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*. pages 92–95.

Andrey Malinin, Kate Knill, Anton Ragni, Yu Wang, and Mark JF Gales. 2017. An attention based model for off-topic spontaneous spoken response detection: An initial study. In *Proceedings of the 7th Workshop on Speech and Language Technology in Education*. pages 144–149.

Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, and Benno Stein. 2014. Overview of the 6th international competition on plagiarism detection. In *Proceedings of the CLEF Conference on Multilingual and Multimodal Information Access Evaluation*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *Proceedings of the workshop on Automatic Speech Recognition and Understanding*.

Yao Qian, Jidong Tao, David Suendermann-Oeft, Keelan Evanini, Alexei V Ivanov, and Vikram Ramanarayanan. 2016. Noise and metadata sensitive bottleneck features for improving speaker recognition with non-native speech input. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*. pages 3648–3652.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* pages 613–620.

Jidong Tao, Shabnam Ghaffarzadegan, Lei Chen, and Klaus Zechner. 2016. Exploring deep learning architectures for automatically grading non-native spontaneous speech. In *Proceedings of the Conference on Acoustics, Speech and Signal Processing*. pages 6140–6144.

Rogier C. van Dalen, Kate M. Knill, and Mark J. F. Gales. 2015. Automatically grading learners english using a gaussian process. In *Proceedings of the Workshop on Speech and Language Technology in Education*. pages 7–12.

Joost van Doremalen, Helmet Strik, and Cartia Cucchiarini. 2009. Utterance verification in language learning applications. In *Proceedings of the Workshop on Speech and Language Technology in Education*.

Xinhao Wang, Keelan Evanini, James Bruno, and Matthew Mulholland. 2016. Automatic plagiarism detection for spoken responses in an assessment of english language proficiency. In *Proceedings of the Workshop on Spoken Language Technology Workshop (SLT)*. pages 121–128.

Su-Youn Yoon and Derrick Higgins. 2011. Non-English response detection method for automated proficiency scoring system. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. pages 161–169.

Su-Youn Yoon, Chong Min Lee, Ikkyu Choi, Xinhao Wang, Matthew Mulholland, and Keelan Evanini. 2017. Off-topic spoken response detection with word embeddings. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*. pages 2754–2758.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication* 51:883–895.

Mo Zhang. 2013. Contrasting automated and human scoring of essays. *ETS R & D Connections* 21:1–11.