Speech Communication 93 (2017) 43–52

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Combining human and automated scores for the improved assessment of non-native speech

Su-Youn Yoon*, Klaus Zechner

Educational Testing Service, 660 Rosedale Road, Princeton NJ 08541 United States

ARTICLE INFO

Article history:
Received 11 August 2016
Revised 5 June 2017
Accepted 1 August 2017
Available online 2 August 2017

Keywords:
Automated oral language proficiency scoring
Automated speech recognition
Non-scorable spoken responses
Filtering system
Hybrid scoring system
Non-native speech

ABSTRACT

In this study, we propose an efficient way to combine human and automated scoring to increase the reliability and validity of a system used to assess spoken responses in the context of an international English language assessment. A set of filtering systems are used to automatically identify various classes of spoken responses that are difficult to score with an automated scoring system, for example, due to a high level of noise or imperfections in components of the overall system. Finally, these flagged responses are then routed to and scored by human raters. The vast majority of responses are not flagged by the filtering system and receive scores by the automated scoring system, resulting in a hybrid scoring approach. The overall hybrid speech scoring system presented here is comprised of multiple subprocesses, including the recording of spoken responses, transcription generation based on an automated speech recognizer, linguistic feature generation, filtering of problematic responses, automated score generation, human rater scoring, and final score combination. We evaluate this scoring approach with pilot data from a novel international English proficiency assessment. It achieves a substantial improvement in scoring performance and score validity with a limited amount of human scoring and most responses scored automatically; the correlation between the baseline system (baseline filtering with imputation) with human raters' scores is 0.72, and using an extended filtering model, the performance improves to 0.82. The improvement can be attributed in part to the extended filtering model itself that identified more classes of non-scorable responses, and in part to the combination of machine and human scores in our hybrid system.

1. Introduction

An automated scoring system can assess constructed responses, such as spoken or written language produced by a test taker, faster than human raters and at a lower cost. In addition, in contrast to human raters, automated systems are not subject to fatigue and emotion, and the resulting scores are always consistent over time (Engelhard, 2002). These advantages have created a high demand for high-performing automated scoring systems for various applications. However, even state-of-the-art automated scoring systems face numerous challenges to their use as a sole rater in operational test programs. Zhang (2013) pointed out the lack of background knowledge and difficulty in scoring creativity, logic, and quality of ideas were weaknesses of automated scoring systems. Lochbaum et al. (2013) also reported that automated scoring systems tended to be more vulnerable to students' gaming strategies. Based on a careful review of current state-of-the art automated essay scoring systems, Zhang (2013) proposed that an automated scoring system needs to meet three criteria in order to be used as the sole rater of a high-stakes assessment: (a) a transparent internal mechanism, (b) a broad base of validity evidence,[1] and (c) a quality control system that detects possible aberrant performances. In order to meet the third criterion, we developed an automated filtering system able to detect responses that are likely to cause aberrant performances in an automated scoring system for a low-stakes spoken language test for non-native speakers of English.

Aberrant performances may occur for various reasons. The automated speech scoring system in this study is comprised of multiple sub-processes, including the recording of a test taker's response, processing a spoken response with an automatic speech recognition (ASR) system, and generating a large set of linguistic features (e.g., related to fluency, pronunciation, prosody, vocabulary, grammar). Inaccuracies or errors may be introduced in any of

* Corresponding author.
E-mail address: syoon@ets.org (S.-Y. Yoon).

[1] For example, evidence may point to a strong correlation between automated scores and certain test-takers' verbal behaviors in some settings of communication that are deemed to be indicative of his/her language proficiency.

http://dx.doi.org/10.1016/j.specom.2017.08.001
0167-6393/

these processes and may cause problems for the subsequent automated scoring system to assign a proper score to a given response. In addition, responses from uncooperative test takers (e.g., responses from test takers who try to game the automated scoring system) are another main cause of aberrant performances. This study explores both types of responses.

First, we analyze a large collection of spoken responses from a pilot administration and classified problematic responses associated with aberrant performances (hereafter, non-scorable responses) into four groups: (a) responses with a high level of noise; (b) empty responses (silence); (c) responses for which the automated speech scoring system cannot produce a score; and (d) responses with certain characteristic speech recognition output that signals that scoring inaccuracies are more likely to occur. The automated scores for these problematic responses are likely to be inaccurate and reduce the validity of the automated scores.

Next, we develop the automated filtering models to identify these problematic responses. We propose two different filtering systems: a baseline filter and an extended filter. The baseline filter is a statistical model trained on a large dataset using various features derived from intensity, pitch and automated speech recognition (ASR) output focusing on identifying response groups (a), (b), and (c). The extended filter is comprised of both the baseline filter and three rules additionally addressing responses in groups (d). The extended filter filters out a higher percentage of responses than the baseline filter. Finally, in order to maintain score quality, we propose a hybrid scoring approach for spoken responses in a language assessment that was a combination of both human and automated scoring based on this automated filtering system. Responses flagged as non-scorable by the filtering system are routed to human raters for scoring, whereas the remainder of responses are scored by the automated scoring system. We compare this hybrid scoring approach to a baseline where all scores for responses flagged as non-scorable by the filtering system are imputed with the average of the non-flagged automated scores for a given test taker. We will show that the extended filtering model in conjunction with our hybrid scoring method can produce a substantial improvement in scoring performance and score validity with a limited amount of human scoring.

2. Related work

As automated scoring is increasingly used in educational assessments, quality control processes that identify the aberrant performances for the automated scoring systems become a key element in maintaining the quality of automated scores. Bejar (2011) provided deep insight into diverse defects that can cause aberrant performances in the operational use of automated scoring systems. He first made a distinction between design defects and quality defects. Design defects result from problems in the design of the systems, which can only be resolved by system changes. On the contrary, quality defects can be prevented by proper quality control systems that monitor automated scoring systems. He further classified the scoring process into two steps (evidence generation and evidence synthesis) and provided examples of defects and the quality control solutions for each step. In automated speech scoring systems, the generation of diverse linguistic features, such as pronunciation quality and syntactic complexity, is an evidence-generation step, while generation of an automated score using a model based on the combination of multiple linguistic features is an evidence-synthesis step.

Several studies explored the quality control of the evidence-generation step. The main issues in this step are audio quality problems that occur while recording spoken responses. Audio quality problems, such as very loud background noise, may cause serious challenges in both human scoring and automated scoring, rendering the resulting scores less reliable. Jeon and Yoon (2012) and Cheng and Shen (2011) developed automated systems that identified responses with bad audio quality using signal processing and automated speech recognition technology.

Responses from uncooperative test takers are another main cause of aberrant performances. Yoon and Xie (2014) analyzed a large collection of spoken responses from a high-stakes, international English proficiency assessment and found four different types of problematic responses: responses that included no valid speech, off-topic responses, responses in test takers' native languages, and plagiarized responses that used the memorized speech of others. Furthermore, there are responses that may belong to more than one type. When automated scoring is used in the context of a high-stakes language proficiency assessment for which the scores are used to make consequential decisions about the examinees, some test takers may have an incentive to try to game the system in order to artificially inflate their scores. Except for responses without valid speech, all other types may be used by test takers who try to game the automated scoring system. By using these strategies, test takers can generate fluent speech more easily and the automated proficiency scoring system, which utilizes fluency as one of the important factors, may assign a score that is higher than warranted based on the actual proficiency of the test taker. These responses from uncooperative test takers, in particular responses that aim to game automated scoring systems, are serious problems that have a strong impact on the validity of the automated scores.

In order to address this issue, van Doremalen et al. (2009), Lo et al. (2010), and Cheng and Shen (2011) developed automated systems that detect off-topic responses. They used speech recognition based features such as normalized confidence scores, acoustic model scores, and language model scores. These studies addressed diverse gaming strategies and achieved promising performance when applied to constrained items (i.e., where the response is predictable, such as read-aloud items). Cheng and Shen (2011) extended these studies and combined an acoustic model score, a language model score, and a garbage model score with confidence scores. They applied this new filter to less constrained items (e.g., picture descriptions) and identified off-topic responses with high accuracy.

Although normalized confidence scores achieved promising performance in restricted speech, they were not appropriate for unconstrained speech. For items that elicit spontaneous speech, Yoon and Xie (2014) developed an automated filtering system that used text similarity measures that have shown promising results in automated essay scoring (Foltz et al., 1999; Attali and Burstein, 2006), automated speech scoring (Xie et al., 2012), and off-topic essay detection (Higgins et al., 2006). Inclusion of text similarity measures substantially improved the accuracy of non-scorable response detection. In addition, for the detection of responses in a language other than the target language, Yoon and Higgins (2011) developed a method based on language identification technology.

However, problematic audio quality and test takers' uncooperative behaviors represent only some of the problems that can cause aberrant performance of automated scoring systems. The coverage of previous studies was rather narrow and could address only a limited range of scoring issues. Recently, Yoon et al. (2014) and Metallinou and Cheng (2014) developed quality control systems that targeted a wider range of issues. Both studies targeted any response for which an automated score is likely to be substantially different from the criterion score, which included but was not limited to the subset of non-scorable responses containing audio quality problems or gaming strategies.

A few studies tried to identify defects in the evidence synthesis stage based on various statistical methods. Zu et al. (2014)