Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 25 (2011) 282–306
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

A three-stage approach to the automated scoring of spontaneous spoken responses

Derrick Higgins *, Xiaoming Xi, Klaus Zechner, David Williamson

Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Received 29 June 2009; received in revised form 30 April 2010; accepted 4 June 2010
Available online 15 June 2010

Abstract

This paper presents a description and evaluation of *SpeechRater*^{SM}, a system for automated scoring of non-native speakers' spoken English proficiency, based on tasks which elicit spontaneous monologues on particular topics. This system builds on much previous work in the automated scoring of test responses, but differs from previous work in that the highly unpredictable nature of the responses to this task type makes the challenge of accurate scoring much more difficult.

SpeechRater uses a three-stage architecture. Responses are first processed by a *filtering* model to ensure that no exceptional conditions exist which might prevent them from being scored by SpeechRater. Responses not filtered out at this stage are then processed by the *scoring* model to estimate the proficiency rating which a human might assign to them, on the basis of features related to fluency, pronunciation, vocabulary diversity, and grammar. Finally, an *aggregation* model combines an examinee's scores for multiple items to calculate a total score, as well as an interval in which the examinee's score is predicted to reside with high confidence.

SpeechRater's current level of accuracy and construct representation have been deemed sufficient for low-stakes practice exercises, and it has been used in a practice exam for the TOEFL since late 2006. In such a practice environment, it offers a number of advantages compared to human raters, including system load management, and the facilitation of immediate feedback to students. However, it must be acknowledged that *SpeechRater* presently fails to measure many important aspects of speaking proficiency (such as intonation and appropriateness of topic development), and its agreement with human ratings of proficiency does not yet approach the level of agreement between two human raters.

Keywords: Language testing; English speaking proficiency; Automated scoring; Constructed response scoring; Speech recognition

1. Introduction

The last decade has seen a proliferation of new applications of speech technology in the educational domain, in particular in support of English language learner (ELL) populations. Computer-based learning tools for spoken English allow learners who may have limited access to native or high-proficiency English speakers an opportunity to practice their English and receive feedback on their performance. Automated systems also have the potential to make educational

* Corresponding author at: Educational Testing Service, Constructed Response Scoring Group, Mail Stop 12-R, Rosedale Road, Princeton, NJ 08541, USA. Tel.: +1 609 734 1126; fax: +1 609 734 1090.
E-mail address: dhiggins@ets.org (D. Higgins).

0885-2308/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.06.001

materials for English speaking accessible to a wider range of learners, by reducing their cost (compared to a human tutor), and making them available on a more flexible schedule.

Speech-enabled dialogue systems allow learners to practice their speaking and listening in an exchange with a virtual interlocutor—e.g., *SpeakESL* and Auralog's *Tell Me More* system. Automated tutoring systems for speaking practice provide feedback to learners on their pronunciation, vocabulary, and grammar—e.g., Carnegie Speech *NativeAccent* (Eskenazi et al., 2007), *Saybot* (Chevalier, 2007) and SRI's *EduSpeak* (Franco et al., 2000). The contribution of the work described in this paper is to the field of automated scoring of spoken responses: assessing the level of English speaking proficiency demonstrated in some speaking task used in a testing context.

In particular, this paper describes *SpeechRater*, an automated system for scoring spontaneous spoken responses. *SpeechRater* uses automatically derived measures of fluency, pronunciation, vocabulary diversity, and grammar to score spoken responses to tasks similar to those found on the Test of English as a Foreign Language (TOEFL®), a test of English proficiency for academic settings, used internationally for college admissions. Only the internet-delivered version of the TOEFL test (TOEFL iBT) includes tasks to assess speaking proficiency; the paper-based version of the test assesses only reading, listening, and writing. *SpeechRater* is currently used operationally to score spoken responses to the low-stakes TOEFL Practice Online (TPO) test, but not for the high-stakes TOEFL iBT test.

Where other spoken response scoring systems had previously relied on restricted speaking tasks such as reading a passage aloud, or answering questions to which the range of responses is narrowly circumscribed (Bernstein, 1999; Bernstein et al., 2000; Balogh et al., 2007), *SpeechRater* addresses the more challenging task of scoring responses which are relatively unstructured, unrestricted, and spontaneous (for a set of speakers varying considerably in English proficiency and first language). While much work remains to be done to improve *SpeechRater*'s agreement with human raters and the coverage of important aspects of speaking proficiency in its feature set, the current capability has advanced sufficiently to allow it to be used for the scoring of practice tests used in preparing for the operational TOEFL test.

Section 2 surveys previous work in automated scoring of test items, especially speaking tasks. Section 3 then provides a brief overview of the *SpeechRater* system architecture, and the speech recognition system used as the basis for feature computation. After an overview of the spoken response data is provided in Section 4, Section 5 describes the first component of *SpeechRater*, which filters out non-scoreable responses. Section 6 introduces the scoring model itself, and Section 7 describes the way in which scores are aggregated across tasks, and bounds on prediction error are calculated. Finally, a brief summary of the major findings of the paper is presented in Section 8.

2. Automated scoring of constructed response items

Constructed response test items are those which require the examinee to generate an answer productively (in contrast to *selected response* or *multiple-choice* items). This section will discuss some of the motivation for using constructed-response tasks and scoring them by automated means, and review the previous work in this area.

2.1. Motivation

A number of pedagogical and educational measurement considerations argue for the use of constructed-response (CR) tasks in testing, and have resulted in a dramatic recent expansion of the use of such tasks in high-stakes tests. The TOEFL, the SAT, and the GRE have added constructed-response (speaking and/or writing) sections to the core test in the last decade, and the ACT and TOEIC have added optional constructed-response sections.

Among the arguments for CR tasks is that they are more "naturalistic", in that they require examinees to actually perform a task similar to ones they will be presented with in real-life (Stiggins, 1982; Charney, 1984; Moran, 1987). (For example, they might have to write an essay, rather than answering multiple-choice questions about English grammar.) More naturalistic tasks, presumably, should provide a more accurate and valid measurement of the test *construct*, where the construct is understood as the attribute or quality to be measured: examinees' mastery of the domain of interest.

A related issue is that of tests' "washback" on school curricula (Messick, 1996; Wiggins, 1989; Bailey, 1999). Schools might have some incentive to teach test preparatory material directly, rather than focusing on appropriate coverage of course content, and it has been suggested that rigid multiple-choice format tests with specific content focus are particularly susceptible to pernicious washback effects. Constructed-response items, by contrast, providing a less tightly constrained format for responses, may bring the skills needed for success on the test into better alignment with appropriate curriculum activities. Finally, constructed-response items have the advantage that they can pose more

complex, multifaceted challenges (Wiggins, 1990). For instance, in writing an essay, examinees must demonstrate their facility with grammatical sentence construction, appropriate use of punctuation, use of appropriate organizational scaffolding, and many other skills. Multiple-choice items must generally be more narrowly focused on particular aspects of writing competence.

While constructed-response test items promise many benefits, they also introduce challenges, notably the challenge of scoring the items quickly, reliably, cost-effectively, and in a way which adequately represents the construct to be measured. Human scoring of CR items is by its nature more costly than scoring of multiple-choice items. Where multiple-choice responses can be scored simply by mechanically checking an answer key, human scoring of CR responses (at least for complex item types such as essay questions or spontaneous spoken responses) requires higher-level cognition. Raters must be trained to adhere to an explicit scoring rubric in evaluating CR responses. The process of rater training, scoring of test items, and monitoring of rater performance can be a costly and time-consuming endeavor for a high-volume testing program. Furthermore, even trained and qualified raters may not be free of subjectivity in their ratings, and can be affected by factors such as fatigue, presentation order of responses, and superficial aspects of performance such as essay length (Hoyt, 2000; Murphy and Anhalt, 1992). These influences can threaten both the validity and the reliability of ratings.

These challenges posed by the use of human raters to score constructed-response items have spurred research into the feasibility of scoring certain CR tasks by automated means. If automated scoring can be done in a way which addresses the appropriate test construct, and agrees well with human ratings of the same items, it can offer substantial benefits. In particular, it has the potential to reduce the cost of scoring, and it dramatically reduces scoring time (often sufficiently so that near-instantaneous feedback can be provided). Of course, the context of use is critical in considering where automated scoring is appropriate. For instance, the use of machine scores unsupported by human review may drive students to attempt to game the system by addressing the task in construct-inappropriate ways, if the stakes of the assessment are high enough. For some applications, though, including low-stakes practice tests, the reduced cost and immediate feedback provided by an automated scoring system give it a clear advantage over human scoring.

2.2. Previous work

Much previous work in automated scoring of constructed-response items focuses on written responses. Automated essay scoring (AES) has a long history, beginning with the work of Page (1966, 1968). Modern AES systems include Pearson's *Intelligent Essay Assessor* (Landauer et al., 2003), Vantage *Intellimetric* (Elliott, 2003), and ETS' *e-rater* system (Burstein, 2003; Attali and Burstein, 2006). The Intelligent Essay Assessor uses Latent Semantic Analysis to situate essays in a vector space relative to essays already seen, with distance in this space corresponding to the relatedness between the terms found in each essay. The latter two applications (as well as others) work by using natural language processing to identify essay features which are indicative of writing quality, and using these to predict the appropriate score by statistical means.

Where AES systems are intended to assess the quality of the writing in a response, another class of automated scoring systems, exemplified by *c-rater* (Leacock and Chodorow, 2003), *AutoMark* (Mitchell et al., 2002), the Oxford-UCLES system (Sukkarieh and Pulman, 2005), and applications developed at the University of Portsmouth (Callear et al., 2001) and the University of Manchester (Sargeant et al., 2004), is designed to score responses from a variety of content areas according to whether they express a correct answer to a question, or contain a particular fact to be elicited. These systems fundamentally have to address the problem of identifying when a student answer can be considered a paraphrase of some gold-standard answer or answers.

Outside the realm of natural language processing, automated scoring technologies have also addressed tasks involving simulations or other complex response types (Williamson et al., 2006).

The automated scoring of spoken responses has been addressed in the prior literature as well. However, in much previous work, the responses scored were either very constrained in terms of the content of the response (involving read-aloud or similar tasks), or they were scored only for pronunciation, without any consideration of other aspects of speaking proficiency.

Ordinate, a subsidiary of Pearson, has been developing language tests since the 1990s in which basic language abilities such as reading or repeating are tested (Bernstein, 1999). This is one way of avoiding the high error rate in open-ended speech recognition for spontaneous speech. Bernstein et al. (2000) demonstrated correlations around 0.80 between their tests and other widely used language tests such as the TOEFL.

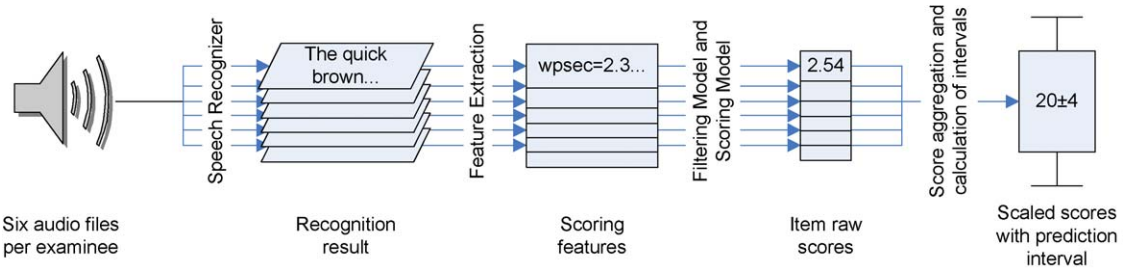

Fig. 1. *SpeechRater* system architecture.

Cucchiarini et al. (1997a,b) developed a speech recognition-based automatic pronunciation scoring system for Dutch by using features such as log likelihood Hidden Markov Model scores, various duration scores, and information on pauses, word stress, syllable structure, and intonation. They also found good agreement between human scores and machine scores.

Stanford Research Institute (SRI) International, similarly, has been developing an automatic pronunciation scoring system, EduSpeak, which measures phone accuracy, speech rate, and duration distributions for non-native speakers who read English texts (Franco et al., 2000). Unlike in Ordinate's test, the texts being read need not be known to the system for prior training. Other pronunciation training programs are described by Eskenazi et al. (2007) and Chevalier (2007).

In our own previous work (Zechner et al., 2009), the *SpeechRater* system for automated scoring of spontaneous spoken responses has been described. While this earlier paper reports on the accuracy of the *SpeechRater* scoring model used for TPO, the current paper extends the analysis in a number of directions. First, the current paper includes a description and analysis of two additional components of the *SpeechRater* system: the model for filtering non-scoreable responses and the model for determining the uncertainty associated with a score estimate. Second, the current paper evaluates *SpeechRater*'s performance on operational TOEFL data in addition to responses from TPO. Finally, the current paper evaluates multiple feature weighting schemes by which the scoring model may be constructed, where the work reported in 2009 investigates only a single regression-based model.

3. System architecture

The *SpeechRater* automated speech scoring system consists of the components outlined in Fig. 1. (The output values depicted in Fig. 1 for each stage of the process are meant solely as illustrative examples.) Responses to each of the six speaking tasks administered in a TPO test are first analyzed by a speech recognizer, the output of which is then processed by a set of feature extraction routines, which derive a set of measures meant to be indicative of speaking proficiency. Based on these scoring features, first the *filtering model* determines whether the response should receive a score, or whether it is in some way anomalous (cf. Section 5). If it is determined to be scoreable, the *scoring model* (cf. Section 6) is then applied to derive the raw score estimate for each response. Finally, the scores for all six items are summed and converted to the 30-point TPO reporting scale, and a prediction interval is calculated to represent the uncertainty inherent in the reported score (cf. Section 7).

As all of the *SpeechRater* components to be described in this paper depend on the speech recognizer used to pre-process the responses, it is described here briefly for reference.

3.1. Audio encoding and processing

The audio for each TPO response is sampled with a sampling rate of 22,050 Hz, mono, 16-bit resolution. The signal then is compressed into the Windows Media format, downsampled to 11,025 Hz and converted to standard PCM (.wav files). Although WMA compression may have a slight adverse effect on speech recognition, the effect of the downsampling should be less noticeable, since the major speech events happen below 5 kHz, and this is consistent with the frequency range for a sampling rate of 11 kHz (0–5500 Hz).

Table 1
Summary statistics of the data sets used for filtering model development and evaluation. **TD** scores are treated as 0 in calculation of mean and standard deviation.

Data set	Num. responses	Num. speakers	Num. topics	Average score	SD of score	Score distribution					
						TD	0	1	2	3	4
TPO train	1595	364	15	2.16	1.31	220	118	58	405	603	191
TPO eval	660	162	9	2.15	1.27	99	41	18	159	289	54
TOEFL train	10,578	1786	36	2.58	0.82	10	41	841	3903	4517	1266
TOEFL eval	10,557	1782	36	2.60	0.83	22	38	814	3816	4466	1401

3.2. Speech recognition system

SpeechRater uses a speech recognizer specifically trained on TPO responses. This recognizer was bootstrapped by using an existing 11 kHz recognizer trained on a large set of transcribed speech of native speakers of American English. The acoustic model was then adapted to a transcribed set of approximately 2000 TPO responses and retrained. The language model was constructed using responses to both prototype and official TOEFL speaking tasks, as well as data from the Linguistic Data Consortium (Garofolo et al., 1997). The out-of-vocabulary rate on a token basis was measured to be 0.8% on a transcribed subset of the data used for the training of the scoring model.

Given the parameterization of the speech recognition system used in conducting the evaluations described below, a word accuracy of 48.9% was observed on the same subset of the scoring model training data.

4. Spoken response data

In developing the candidate models to identify non-scoreable responses (Section 5) and to assign scores to all other responses (Section 6), two sources of spoken response data were used. Responses from the TPO practice test were the main focus of development, as the model was intended for actual use on that test. Responses from the operational TOEFL test were also used for comparison, to determine whether the same features had predictive value in both the practice test condition and the operational test. Previously reported work on *SpeechRater* (Zechner et al., 2009) reported results on TPO responses and on data from a field test conducted prior to launching the newest version of the TOEFL, but no results were provided to indicate *SpeechRater*'s performance on operational data from the TOEFL test.

Both the TPO and the TOEFL contain six speaking items per test form. Two of these items are *independent* speaking items, eliciting a 45-second response to a short, textual stimulus. The other four items are *integrated* speaking items, which elicit 60-second responses to a more complex stimulus including both text and audiovisual lecture material. While there are important differences between these task types, they are not differentiated in the models described below, because the relatively shallow automated features used for scoring are fairly insensitive to these differences.

The set of responses from the TPO and TOEFL data sources was subdivided into a training sample used to explore the features most useful in prediction and to parameterize the models, and a test sample for final evaluation of the models. Each data set was scored by trained content experts. Each train/test split is defined in such a way that there is no speaker overlap between the partitions. That way, the evaluation cannot be influenced by anything that might be tailored to the speaking characteristics of particular individuals.

While TOEFL speaking responses are, for the most part, single-scored only (with a small percentage double-scored for quality control), the TPO responses were scored by two raters each. The first human score for each TPO response was provided within weeks of its recording, so that it could be reported to the examinee along with the rest of his/her scores on the TPO practice assessment. The second human score was assigned to each response in a special scoring session to support this research.

Descriptive statistics regarding the composition of each of these data sets are provided in Table 1. Note that this table indicates two special score classes, **TD** and **0**, which indicate different types of anomalous responses. The code **TD** is applied to responses which are non-scoreable due to a *technical difficulty* (such as equipment or transmission errors, noise levels which are too high, or recording levels which are too low), and a score of **0** is assigned to recordings which represent a failure to respond on the part of the examinee.

Table 2
Confusion matrix showing single-item agreement between two sets of human raters on full set of TPO data. (Three responses are excluded due to lack of a second score.)

Rater 1	Rater 2						Total
	TD	0	1	2	3	4	
TD	75	20	2	3	0	0	100
0	223	137	3	0	0	0	363
1	3	1	25	6	0	0	35
2	6	0	40	285	155	7	492
3	9	1	5	240	541	92	888
4	1	0	1	29	196	147	374
Total	317	159	76	563	892	245	2252

The TPO data sets are smaller than those from the operational TOEFL, but they contain a great many more non-scoreable responses (**TD** s and **0** s) than the TOEFL data sets, for reasons related to the differences between a high-stakes test and a low-stakes practice test environment (see Section 5). The mean score on the test is also lower, but this is also attributable to the larger number of zero scores in the TPO data.

Table 2 shows the confusion matrix for human scoring of the full set of TPO data (training and evaluation combined). Of course, the corresponding agreement information cannot be provided for the single-scored TOEFL data. The overall exact agreement level between raters is 53.7%, and when **TD**s are treated as equivalent to **0** scores, the Pearson correlation between human raters is 0.873 and the quadratic-weighted kappa (Cohen, 1960) is 0.867.

In the training and evaluation of models for scoring spoken responses (Section 6), all responses labeled as either **TD** or **0** by human raters were excluded from the data sets. Descriptive statistics for these subsets of the data sets introduced in Table 1 are provided in Table 3. The prefix **sm-** in the labeling of each data set indicates that they are used in connection with the *scoring model* component of *SpeechRater*.

In addition to the **sm-eval** evaluation sets, one additional set of data was used to evaluate the scoring models for the TPO data. Because only a limited number of TPO responses were available, only 520 could be reserved exclusively for evaluation. Within this **TPO sm-eval** set, there were very few speakers who had complete test forms of six scoreable responses. Because aggregation across responses substantially improves the correlation between independent scores, it was desirable to have a data set for which we could aggregate reliably across a full complement of six items. The additional data partition is the **TPO rectrain+eval** set, which consists of the TPO evaluation data combined with a set of responses which had been reserved for training of the speech recognizer. This **TPO rectrain+eval** set contains substantially more candidates with a full set of six completed items (308) than does the **TPO sm-eval** set (58).

Because some responses were used in recognizer training, the **rectrain+eval** set is not technically a completely independent test set. Nevertheless, any scoring differences observed in this data set could only be a result of a difference in the recognizer's word error rate, and this is unlikely to be a strong predictor of scoring accuracy, because our features are not dependent on a highly faithful recovery of the words in the response.

For the data in Table 2, the quadratic weighted kappa for human agreement, with **TD** and **0** scores excluded, is 0.537, and the Pearson correlation is 0.547. Further statistics regarding human agreement in scoring these data sets

Table 3
Summary statistics of the data sets used for scoring model development and evaluation.

Data set	Num. responses	Num. speakers	Num. topics	Average score	SD of score	Score distribution			
						1	2	3	4
TPO sm-train	1257	281	15	2.74	0.77	58	405	603	191
TPO sm-eval	520	123	9	2.73	0.69	18	159	289	54
TPO rectrain+eval	2427	507	24	2.79	0.72	70	709	1300	348
TOEFL sm-train	10,527	1786	36	2.59	0.80	841	3903	4517	1266
TOEFL sm-eval	10,497	1782	36	2.61	0.82	814	3816	4466	1401

are provided (together with human–*SpeechRater* agreement statistics) in Table 13. As mentioned above, humans' agreement in scoring sets of multiple responses is much greater than their agreement on single responses. For the **TPO rectrain+eval** set, the correlation between human ratings is 0.742 on sets of six responses. For the **TOEFL sm-eval** set, the correlation between human ratings is 0.843 on sets of three responses.[1]

Note that while we have measures of inter-rater agreement for our human ratings, we do not have measures of intra-rater agreement, as our data collection was not designed to support this metric. Previous work on rating of oral proficiency interviews (Clark and Swinton, 1979; Shohamy, 1983) indicates that intra-rater and inter-rater reliabilities in speaking proficiency assessment tend not to diverge greatly, with the intra-rater reliability serving as an upper bound on inter-rater reliability. The following subsections outline the scoring procedures used in order to ensure the reliability of human scores.

4.1. Training and certification of human raters

Human raters selected to score the TOEFL and TPO tests are typically experienced teachers or specialists in English as a Second Language, but they may also have otherwise demonstrated experience with scoring of non-native English responses. Before being certified as TOEFL raters, they must complete an online tutorial and certification test.

The tutorial is designed to familiarize raters with the different types of items on the test (two independent and four integrated tasks), and the guidelines for scoring them. The scoring of each task type is governed by an explicit rubric, which indicates criteria typical of responses at each score level. As mentioned above, the score scale for each item ranges from **1** (the lowest) to **4** (the highest), with the codes **0** and **TD** used for special cases. In addition to the rubric, raters become familiar with the use of *Topic Notes* and *Key Points*, documents which supplement the rubrics by discussing how specific response types should be handled, and indicating what relevant material from the item stimulus should be noted in responses to integrated task items, respectively. *Benchmark* responses, prototypical examples of items at each score point, are also provided to raters to facilitate their understanding of how the rubrics are to be applied in practice.

During the tutorial, raters have an opportunity to practice scoring responses, and to check their scores against those provided by certified raters. Finally, once the training session has been completed, participants can take a certification test consisting of additional scoring exercises to become a TOEFL rater.

4.2. Human scoring procedures

In the process of operational TOEFL scoring, raters will typically score a single task type for a session of about two hours in length, with no more than two task types scored in a single day. Each scoring session dedicated to a task is preceded by a short calibration test to ensure that raters' scoring standards are in alignment, and followed by a short break. Raters who do not pass the calibration test on the first or second attempt will not be permitted to score that item type on that day.

Each group of 8–10 raters also has access to a scoring leader, a former rater with demonstrated consistency in scoring and knowledge of scoring procedures, who they may consult for guidance on the handling of particular responses or recurring issues.

The scoring of the TPO test was conducted largely according to these same procedures, with two exceptions. First, the initial set of scores that were assigned for reporting to examinees were not assigned in topic-specific scoring sessions. Rather, raters were allowed to score responses to multiple topics in a single session. Second, item type specific calibration testing was not done for either set of scores prior to a scoring session.

5. Filtering model

Capturing spoken responses to test items can be complicated by technical problems of many sorts: equipment malfunctions, excessive levels of ambient noise, and data transmission errors, to name a few. In addition, there is the

[1] Because only a random sample of TOEFL speaking responses is double-scored, it was not possible to determine human agreement on full sets of six responses.

possibility that certain examinees may fail to respond to particular test items (by remaining silent). Both of these issues are aggravated in a practice test environment, where the equipment and testing environment are less tightly controlled, and examinees' motivation to score well is not as high as it would be for a high-stakes test.

For these reasons, it is necessary to use a *filtering model* before actually scoring responses, in order to identify ones which are anomalous in some way that should preclude their scoring. In the TPO test, examinees are permitted to re-record responses which are flagged by this filtering model. This section describes the development of a statistical model to identify responses which should be assigned a **TD** or **0** designation, and therefore should not be scored by *SpeechRater*.

Section 4 described the data used to construct and evaluate the *SpeechRater* filtering model, including human raters' agreement in assigning scores to these spoken responses. For the purposes of the filtering model, though, the most informative statistics relate to the agreement in drawing distinctions between scoreable and non-scoreable responses.

If we conflate the non-scoreable classes **TD** and **0** on the one hand, and the score classes 1–4 on the other, human agreement on this distinction is quite good. For the TPO data presented in Table 2, the two sets of raters agreed with one another in classifying responses as scoreable or non-scoreable 98.7% of the time ($\kappa = .961$). Agreement in classifying non-scoreable responses as either **0** or **TD** (the upper left four cells of the table) was much lower; the exact agreement was only 46.6% ($\kappa = .095$).

This poor agreement in distinguishing **TD** s from **0** s indicates some inconsistency in the rater training processes used in the first and second scoring sessions. (The two scores were assigned in distinct rating sessions, as described in Section 4.) The second set of raters showed a clear preference for the label **TD**, while the first set of raters chose to assign **0** more often. Nevertheless, it is unlikely that humans could achieve acceptable agreement levels in making this distinction, given the nature of the criteria specified in the scoring rubric to distinguish the two classes.

The rubric used for the TOEFL Speaking test defines the scores of **0** and **TD** in such a way that they are quite difficult to distinguish reliably, even for trained human scorers. Briefly stated, a score of **0** is assigned if "the speaker was unwilling or unable to respond or made no attempt to answer the question," while a **TD** is assigned under a number of special conditions, such as if the response is too loud, too quiet, contains noise or feedback, or contains complete silence. The difficulty in distinguishing between the two score classes arises because by far the largest class of anomalous responses is those which consist almost completely of silence. In such cases, the distinction between the two classes hinges on whether the scorer hears evidence of the candidate's presence, such as breathing. If the candidate is thought to be at the microphone, the response is scored as a **0**. Otherwise they receive a **TD**.

Because of the conceptual difficulty and lack of human agreement in distinguishing between **TD** s and **0** s, the filtering model for the TPO assessment was designed to discriminate only between scoreable responses (receiving a score of **1** –**4**) on the one hand and non-scoreable responses (**TD** s or **0** s) on the other. This level of granularity is appropriate for the TPO test, as the distinction between **TD** s and **0** s does not have consequences for examinees on this test. In operational TOEFL testing, while **0** s are taken into account in computing the total scores, candidates are offered another chance to take the test if two or more **TD** s occur out of the six responses. However, on the TPO test, examinees are allowed to re-record both **TD** and **0** responses.

5.1. Features

A total of 44 predictive features were available for consideration in the development of the filtering model (see Table 4). These features included a number of measures developed primarily for their relevance to the scoring of responses, such as those encoding the length of the response, its fluency, the diversity of vocabulary used, the speaker's pronunciation and grammar, and the confidence scores of the speech recognizer. In addition, a set of low-level signal processing features encoding statistical moments of the response audio's pitch and power were considered, as they were thought to be promising for the detection of responses which had audio levels that were too high, too low, or inconsistent.

5.2. Analysis

In selecting features for inclusion in the models to predict whether responses are scoreable, the primary criterion used was the association of each feature with the scoreable/non-scoreable distinction. In addition to this empirical

Table 4
Candidate features for the development of the filtering model.

Feature category	Feature name	Feature description
Low-level signal processing	*powmean*	Global mean of power
	powmeandev	Mean absolute deviation of power
	powvar	Variance of power
	powstddev	Standard deviation of power
	powmin	Global minimum of power
	powmax	Global maximum of power
	powdelta	*powmax - powmin*
	pitmeandevnorm	Mean absolute deviation of pitch normalized by mean pitch
	pitminnorm	Minimum pitch normalized by mean pitch
	pitmaxnorm	Maximum pitch normalized by mean pitch
	pitdeltanorm	Difference of maximum pitch and minimum pitch normalized by mean pitch
Response length	*numwds*	Number of words
	numtok	Number of tokens (including disfluencies)
	globsegdur	Duration of entire response
	segdur	Duration of words in response
	uttsegdur	Duration of response without inter-utterance pauses
	types	Number of unique words
Fluency	*wdpchk*	Average length of speech chunks
	secpchk	Average duration of speech chunks
	wpsec	Speech articulation rate
	wpsecutt	Speaking rate
	secpchkmeandev	Mean absolute deviation of chunk durations
	wdpchkmeandev	Mean absolute deviation of chunk lengths
	numsil	Number of silences in the response
	silpwd	Duration of silences normalized by response length in words
	silpsec	Duration of silences normalized by total word duration
	silmean	Mean duration of silences
	silmeandev	Mean absolute deviation of silence durations
	silstddev	Standard deviation of silence durations
	longpwd	Number of long pauses normalized by response length in words
	longpmn	Mean duration of long pauses
	longpmeandev	Mean absolute deviation of long pause durations
	longpstddev	Standard deviation of long pause durations
	numdff	Number of disfluencies
	dpsec	Number of disfluencies normalized by total word duration
	numrep	Number of repeated words
	repfreq	Number of repeated words normalized by response length
Fluency and vocabulary diversity	*tpsec*	Unique words normalized by total word duration
	tpsecutt	Unique words normalized by speech duration
	ttratio	Type-token ratio
Pronunciation	*amscore*	Acoustic Model score; compares the pronunciation of non-native speech to a reference pronunciation model
Grammatical accuracy	*lmscore*	Language Model score; compares the language of non-native speech to a reference language model
ASR confidence	*confavg*	Average over all confidence scores
	conftimeavg	Time-weighted average of confidence scores

criterion used to assess features, it was desirable that the model ultimately deployed for the TPO test be simple in structure, both so that the basis on which responses are flagged is transparent, and to simplify the process of maintaining and updating the model.

To identify a small set of features for inclusion in each model, we used a correlation-based feature selection criterion which selects features highly associated with the target variable of scoreability, but which are minimally redundant

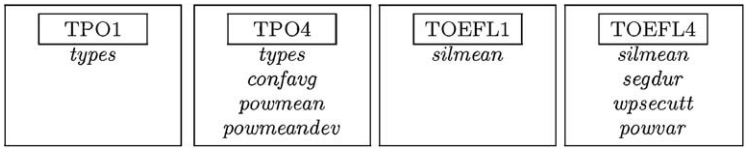

Fig. 2. Features in feature sets labeled TPO1, TPO4, TOEFL1 and TOEFL4.

(Hall, 1998). This resulted in the feature sets TPO4 and TOEFL4 in Fig. 2. For comparison, we also defined feature sets for modeling which consist only of the top-ranked feature for each data set (using chi-square as a measure of association with the target variable).[2] These are the feature sets TPO1 and TOEFL1 in Fig. 2. The single-feature models will serve as baselines, to demonstrate the performance level which can be achieved in filtering non-scoreable responses by very simple means.

Two different types of models were also compared in developing the filtering models described below. Because they are widely used in machine learning, have demonstrated superior performance across a range of classification tasks, and are well-suited to the structure of the response-filtering problem (with continuous feature values and potential interdependencies between features), we used support vector machines (Vapnik, 1995) as one model class. Radial basis function kernels were used in these SVM models, with a cost parameter optimized to give good performance on the training data. In accordance with the aim of balancing model accuracy with model simplicity, we also used a simple *Classification by Regression* (CbR) procedure (Frank et al., 1998), in which the classification decision is made based on a cutoff value specified on a linear weighting of the model features. Feature values were standardized before model training for maximal interpretability of the weights. Both model types were implemented using the Weka machine learning toolkit (Witten and Frank, 2005).

5.2.1. Models for TOEFL practice online

In the construction of models to identify non-scoreable responses in the TPO data set, three sets of features were considered. The TPO1 feature set consists only of the *types* feature, while the TPO4 feature set consists of the four most predictive features in the TPO training data (as indicated in Fig. 2). Intuitively, the TPO4 features indicate that the likelihood of a response being labeled anomalous will increase as

- it contains fewer distinct recognizable words (*types*),
- the recognizer is less sure of its hypothesis (*confavg*),
- it contains less sound (*powmean*), and
- the variability of the sound level is lower (*powmeandev*).

The complete set of predictive features (ALL) was also made available to the SVM classifier in order to investigate whether predictive information was being lost in using a restricted feature set.

The results of model building on the TPO data are shown in Table 5, both for cross-validation within the **TPO train** set, and for evaluation of the final models on the **TPO filter-eval** set. In these evaluation tables, *overall accuracy* is the proportion of all responses which each model classifies correctly. *Precision* is the proportion of the responses classified as anomalous which are indeed anomalous. *Recall* is the proportion of anomalous responses which the model correctly finds. *False positive rate* is the proportion of scoreable responses which are misclassified as being anomalous. (This number is of greatest importance for our application; while it is undesirable for anomalous responses to receive a score, the cost of incorrectly filtering out a legitimate response is even greater.) The final evaluation statistic is the area under the ROC curve (AUC), a measure of discrimination across all possible cut points between scoreable and non-scoreable responses.

Table 5 shows that the test set is somewhat easier than the training set (in that the models' accuracy is higher than on the training data), and that all of the models generalize very well. Each model finds over 95% of the anomalous responses in the test set, while keeping the false positive rate well under 1%. On the test set, all models agree with

[2] Feature values were first discretized by the default MDL method used in the Weka toolkit (Fayyad and Irani, 1993).

Table 5
Filtering model results on TPO data.

Eval. set	Model type	Feature set	Overall accuracy	Precision	Recall	False positive rate	AUC
XVal	CbR	TPO1	97.9%	96.9%	93.2%	0.8%	0.980
	CbR	TPO4	98.3%	99.1%	92.9%	0.2%	0.985
	SVM	TPO4	98.2%	98.1%	93.5%	0.5%	0.984
	SVM	ALL	98.2%	97.5%	93.8%	0.6%	0.984
Test	CbR	TPO1	98.8%	98.5%	95.7%	0.4%	0.988
	CbR	TPO4	99.2%	100%	96.4%	0.0%	0.993
	SVM	TPO4	98.8%	97.8%	96.4%	0.6%	0.994
	SVM	ALL	98.9%	98.5%	96.4%	0.4%	0.995

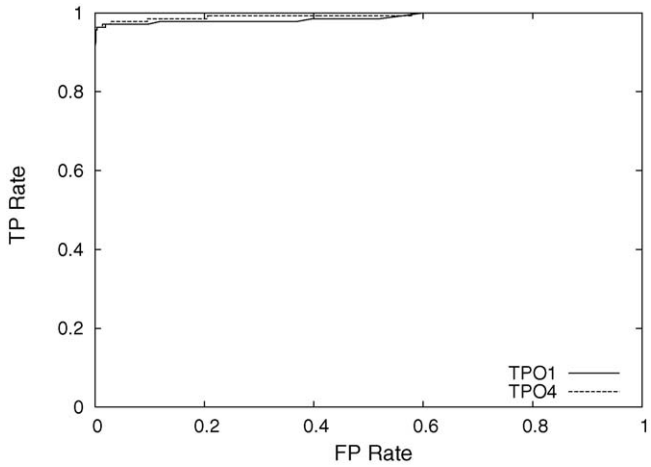

Fig. 3. ROC curves for filtering models evaluated on TPO test set.

human ratings of scoreability approximately as well as humans agree with one another. (Human–human agreement was 98.9% in making this binary distinction on the **TPO eval** set.)

Even the model based only on the *types* feature (TPO1) demonstrates very good performance; by the AUC measure, no other model is significantly better than this simple one.[3] The strong performance of the simple classification by regression models is illustrated in Fig. 3, which shows the ROC curves for these models on the **TPO eval** set. For the CbR model based on the TPO4 feature set, the standardized feature weights were approximately .56 for *types*, .09 for *confavg*, .14 for *powmean*, and .22 for *powmeandev*, indicating that the *types* feature remains the most important factor in the model. It is clear from these results that a model with a small feature set is sufficient to classify responses as scoreable or non-scoreable, and that there is no benefit from using a more sophisticated model such as an SVM on this task and data set.

In the production application deployed to score the TPO test, the four-feature CbR model is used. The decision to use this model was motivated by its slightly higher agreement with human raters (compared to the TPO1 model), and by the model's relative simplicity (compared with the SVM-based models).

5.2.2. Models for TOEFL data

The building of models for the TOEFL data was conducted in a similar fashion to the model building on the TPO data set. In addition to feature sets consisting of the most predictive single feature (TOEFL1) and set of features (TOEFL4), we also retained the feature sets chosen based on their association with the target variable in the TPO data (TPO1 and

[3] Significance calculations for AUC are performed with a one-tailed Z-test, using the Hanley-McNeil estimate of the standard error (Hanley and McNeil, 1982).

Table 6
Filtering model cross-validation results, for models built on TOEFL training data.

Model type	Feature set	Overall accuracy	Precision	Recall	False positive rate	AUC
CbR	TPO1	99.5%	0.0%	0.0%	0.0%	0.949
CbR	TPO4	99.5%	0.0%	0.0%	0.0%	0.950
CbR	TOEFL1	99.5%	0.0%	0.0%	0.0%	0.938
CbR	TOEFL4	99.5%	0.0%	0.0%	0.0%	0.906
SVM	TPO4	99.6%	63.0%	33.3%	0.1%	0.927
SVM	TOEFL4	99.7%	77.4%	47.1%	0.1%	0.920
SVM	ALL	99.7%	68.4%	51.0%	0.1%	0.910

TPO4) in order to assess the sensitivity of the task to the operational testing conditions. Ideally, a single model could be used for the identification of non-scoreable responses in any test with similar response types.

Table 6 shows the performance of a selected set of models on the classification task, using cross-validation within the **TOEFL train** set. Because the frequency of anomalous responses is so much lower in the operational TOEFL data (in which examinees are more motivated because of the higher stakes of the test, and the audio equipment is more standardized), measures of classification accuracy tend to be uninformative on this data set. None of the classification by regression models succeed in correctly classifying a single non-scoreable response, largely because the sheer number of scoreable responses increases the risk for the model to flag a response. The most meaningful measure on the TOEFL data is the AUC, which assesses the discriminative power of the model at all possible decision points.

Unlike the CbR models, the SVM-based models are able to draw a sharp enough distinction between scoreable and non-scoreable responses to correctly classify some non-scoreables as well. This superiority seems to be limited to the condition in which false positives and false negatives are weighted equally, though, as the SVM classifiers have AUC statistics among the worst of all models compared in Table 6. (Since SVMs by their nature maximize classification accuracy by maximizing the classification margin around a specific cutoff, they will not necessarily produce models which perform with high accuracy when a different threshold is chosen.) The AUC values in Table 6 also seem to indicate that the predictive features chosen based on the TPO data (**TPO1** and **TPO4**) are more effective than those selected based on the TOEFL training data, which would suggest that the composition of the TOEFL data, in which non-scoreable responses are scarce, may be suboptimal for the purposes of feature selection. However, none of the AUC differences in Table 6 are significant.

The performance of these models on the **TOEFL eval** data set is very similar to their performance in cross-validation. As Table 7 shows, the AUC differences between models on the held-out data are very small, and again none of these differences are statistically significant. As was observed in the TPO data sets, it would appear that neither using an expanded feature set, nor using a more complex (SVM) classifier confers any advantage on the filtering model for the TOEFL data. The full ROC curves, shown in Fig. 4 show that the simpler CbR models using restricted feature sets demonstrate very comparable performance to the SVM models.

Table 7
Filtering model results on TOEFL test data, for models built on TOEFL training data.

Model type	Feature set	Overall accuracy	Precision	Recall	False positive rate	AUC
CbR	TPO1	99.4%	0.0%	0.0%	0.0%	0.899
CbR	TPO4	99.4%	0.0%	0.0%	0.0%	0.917
CbR	TOEFL1	99.4%	0.0%	0.0%	0.0%	0.911
CbR	TOEFL4	99.4%	0.0%	0.0%	0.0%	0.899
SVM	TPO4	99.6%	95.5%	35.0%	0.0%	0.882
SVM	TOEFL4	99.5%	67.7%	35.0%	0.1%	0.916
SVM	ALL	99.6%	78.1%	41.7%	0.1%	0.891

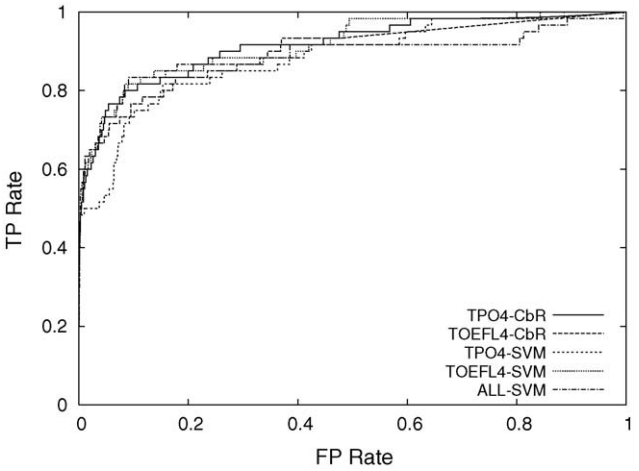

Fig. 4. ROC curves for filtering models evaluated on TOEFL test set.

In addition to the question of which model type and feature set are best suited to identifying non-scoreable responses, the question of model sensitivity to the source of training data is an important one for implementation. Ideally, a single model could be used across multiple testing programs or testing situations (perhaps with a different cutoff value for each, depending on the frequency of non-scoreable responses). If the optimal model parameterization differs greatly from one test to another, though, it may not be possible to apply a single model in all cases.

Table 8 shows the performance of two classification by regression models with restricted feature sets on the **TOEFL eval** data set, under the condition in which they were trained on TPO data, and under the condition in which they were trained on TOEFL data. These two models were chosen for comparison because they are conceptually simple, and did not differ significantly in performance from the other models shown in Table 7. (Note that the model listed in the first row of Table 8 is the one operationally deployed in the use of *SpeechRater* for TPO.)

While the models trained on TOEFL responses have higher AUC values in Table 8, there is no significant difference in this critical statistic between the TPO training and TOEFL training conditions, although naturally the accuracy, precision, and recall statistics differ as a result of the models' differential expectation of the frequency of non-scoreable responses. This result suggests that the models' parameterization can be reasonably applied across the two different tests, and that only the appropriate flagging threshold may need to be adjusted. As in Table 6, the models using features selected on the basis of TPO data perform somewhat better, especially among models which are also trained on TPO responses. Again, though, these differences slightly exceed the 0.05 probability level required for significance.

To summarize the findings of the investigation into filtering non-scoreable responses, it seems that a very simple model, using four predictive features, performs very well in identifying non-scoreable responses on the TPO test. This model furthermore generalizes fairly well to responses drawn from the operational TOEFL test, and in fact no competing model investigated (SVM models, models with features selected by predictive value on TOEFL data, or models with an expanded feature set) performs better by the AUC measure, although the model's utility is somewhat diminished by the low frequency of problematic responses on the operational test. Setting a threshold value for the flagging of non-scoreable responses can be done in a test-specific fashion, but model estimation seems otherwise insensitive to the distinction between testing conditions.

Table 8
Filtering model results on TOEFL test set, for models built on TPO and TOEFL data.

Training set	Model type	Feature set	Overall accuracy	Precision	Recall	False positive rate	AUC
TPO	CbR	TPO4	99.6%	100.0%	38.3%	0.0%	0.910
	CbR	TOEFL4	99.4%	42.9%	35.0%	0.3%	0.825
TOEFL	CbR	TPO4	99.4%	0.0%	0.0%	0.0%	0.917
	CbR	TOEFL4	99.4%	0.0%	0.0%	0.0%	0.899

6. Scoring model

The model for estimating the score to be assigned to a spoken response is the core of the *SpeechRater* application. The usefulness of the system's feedback depends primarily on this model's success in providing scores which are similar to those which human raters would provide, and which accurately reflect the important aspects of speaking proficiency reflected in the rubrics for the speaking tasks.

These criteria for success imply two different sets of evaluation criteria for *SpeechRater*. The desirability that *SpeechRater*'s scores agree with human ratings suggests an empirical evaluation, in which automated scores and human scores are obtained for an evaluation set of responses, and the correlation between the two is measured (as well as other agreement statistics). The importance of the relationship between *SpeechRater*'s scores and the aspects of speaking proficiency to be measured suggests a qualitative evaluation by subject-matter experts, who can assess the degree to which the model structure reflects the generalizations appropriate to the domain.[4]

The former, empirical evaluation of the model will be the main focus of this section, while the latter evaluation of the construct-appropriateness of the model structure will be touched on only in passing. Nevertheless, this criterion is important to mention, as it influenced the choice of modeling techniques applied to the scoring task, the features selected for inclusion in the model, and the feature weighting schemes discussed below.

6.1. Modeling approach

The scoring models compared in this section are all based on a simple multiple regression, in which the score assigned to a response is estimated as a weighted linear combination of a selected set of features. The features are in some cases subjected to transformations to improve their empirical correlation with human scores, or to improve their conformance to a normal distribution (since normality of the independent variables is an assumption of the multiple regression model).

Other modeling approaches, such as decision trees (Breiman et al., 1984; Quinlan, 1986) or support vector machines (Vapnik, 1995) have been considered for automated speech scoring applications as well (cf. Zechner and Bejar, 2006; Xi et al., 2008), and seem to offer certain advantages. For instance, decision trees may provide a more direct representation of the considerations human raters use in scoring responses, and SVMs have demonstrated very strong empirical accuracy in a range of classification tasks. However, for the scoring of the TPO Speaking application, it was determined that a multiple regression model would be the most flexible and transparent option. Because regression models provide a score estimate on a continuous scale, it is possible to re-scale or shift the score distribution if necessary to account for changing population characteristics. Also, regression methods naturally produce a model which is not *biased*, in the sense that its score estimates are neither too low nor too high, on average, for the training population. Enforcing such a requirement is not easy to do in a principled manner for other model types. Finally, the regression model is attractive because it is simple. With a single weight associated with each scoring feature, the contribution of each feature to a particular score can be determined easily, and this reduces the danger of feature use which is inappropriate or not easily defensible. (For instance, if the model assigns negative weights to features which conceptually ought to contribute to a higher score, this will be immediately apparent.)

The modeling is conducted in a generic way, without regard to the topic which is being addressed by the speaker. It may be that there is some benefit to tailoring scoring models to specific topics, but the features used in the current work address aspects of speaking proficiency which are unlikely to show much topic-dependence, such as pronunciation.

The data used to train and evaluate the spoken response scoring model was described in Section 4. So for the TOEFL data, the **TOEFL sm-train** data set was used to select features and parameterize the scoring models, and the **TOEFL sm-eval** as a held-out test set to assess the models' performance. Similarly, a train and test partition were defined for the TPO data (and the **TPO rectrain+eval** set was added to support better evaluation of scores aggregated across multiple items.) Refer to Table 3 for details of these data partitions.

[4] This second set of evaluative criteria corresponds to the requirement of *validity* in educational measurement (American, 1999; Bennett and Bejar, 1998; Yang et al., 2002). Scores which are reliable and strongly predictive of human ratings may nevertheless fail to be assigned in a way which respects the nature of the skill to be measured. To take a commonly cited example, essay length is strongly predictive of human ratings of written essays (Page, 1966), but as a measure of writing skill, essay length alone has very limited construct validity.

Table 9
Construct and empirical evaluations of features considered for inclusion in multiple-regression scoring model.

Feature	Feature class	Correlation with human ratings	Aggregate CAC rating
wpsec	Fluency	0.449	13.9
tpsecutt	Fluency and Vocab	0.408	15.6
tpsec	Fluency and Vocab	0.296	14.7
wdpchk	Fluency	0.106	15.6
wdpchkmeandev	Fluency	0.097	11.6
longpmn	Fluency	−0.204	12.1
silmean	Fluency	−0.282	13.5
silpwd	Fluency	−0.294	13.4
lmscore	Grammar	−0.295	12.4
longpwd	Fluency	−0.327	11.4
amscore	Pronunciation	−0.445	13.1

Table 10
Final feature set used in scoring models, with CAC feature weighting scheme.

Feature	CAC-assigned weight
amscore	4
wpsec	2
tpsecutt	2
wdpchk	1
lmscore	1

6.2. Features

The features to be included in the multiple-regression scoring models were selected from those in Table 4, with those from the categories "Low-level Signal Processing", "Response Length", and "ASR Confidence" excluded due to their limited validity and construct-relevance. In some cases, when candidate features were strongly intercorrelated, features were excluded to avoid redundancy. The feature set was then further refined based on the correlation which each feature demonstrated with human scores, and also in response to feedback from the Content Advisory Committee convened to consult in the development of *SpeechRater*.

The Content Advisory Committee (CAC) is a body consisting of experts in the scoring of spoken responses such as those found on the TOEFL and TPO tests. They provided advice during the development of *SpeechRater* regarding the rubrics used for human scoring of spoken responses, and how closely the proposed scoring features and models match the considerations weighed by human judges in rating a spoken response. In considering the scoring features available for incorporation into a model, CAC members were asked to complete a questionnaire, in which features were rated according to how well they represent the TOEFL scoring rubric, a particular class of features (e.g., Fluency features), and a particular dimension of the construct (e.g., Delivery) on a six-point Likert scale. The aggregate rating for each feature is indicated in Table 9 (with higher ratings corresponding to better representation). This table also shows the correlations between features and human scores within the **TPO sm-train** data set.

The CAC ratings and empirical correlations in Table 9 were taken into consideration in selecting a set of five final features to be used in the scoring model (cf. Table 10). In addition to the empirical and construct considerations used to assess the merits of individual features, the final feature set was also influenced by the CAC's recommendation to cover the construct of speaking proficiency as broadly as possible within the constraints of the available features.

Before building the final regression models, features were statistically examined to assess their conformance to a normal distribution. Multiple regression models assume normality of the independent variables, and to the extent that this assumption is violated it can degrade model performance. Features which diverged from a normal distribution (as revealed by Q–Q plots) were transformed to improve their normality and their correlation with human scores. In particular, a logarithmic transformation was used for *wdpchk*, while an inverse transformation was used for *amscore* and *lmscore*.

6.3. Model variants

In order to evaluate the sensitivity of our speech scoring models to the exact weighting scheme used, three different ways of assigning weights to each of the scoring features in a multiple-regression model were investigated. If the accuracy of the predicted scores depends strongly on the optimization of the feature weights, it may be necessary to fully re-train the model when there are changes to test items or the population of examinees. If the model's accuracy does not depend greatly on the exact weights, though, it may be sufficient to reassess the scaling of the model (the mean and variance of the assigned score distribution). Attali and Burstein (2006) found that for automated scoring of essays, the exact feature weights chosen were not particularly critical, and models using fixed feature weights demonstrated similar performance to ones using optimal weights.

The three types of feature weightings considered here are a standard *empirical* weighting, an *equal* weighting of the features, and a weighting scheme devised by the *CAC* (shown in Table 10). For each weighting considered, features are first standardized so that they have a mean of zero and unit variance.

The empirical weighting is simply the optimal least-squares weighting, in which each feature f_i is assigned a distinct weight a_i, in the regression formula (1) used to assign scores.

$$\text{Score} = \sum_i a_i f_i + b \tag{1}$$

In the equal-weights model, all of the a_i are constrained to be equal, so that the only free parameters in the model are a slope parameter m and intercept b:

$$\text{Score} = m \sum_i f_i + b \tag{2}$$

Finally, in the model using expert weights determined by the CAC, the weights a_i' assigned to each feature are fixed, so that again the only free parameters are the slope and intercept:

$$\text{Score} = m \sum_i a_i' f_i + b \tag{3}$$

In assigning the weights in Table 10, the CAC used their judgement regarding the relative importance of the construct area covered by each feature to the overall assessment of speaking proficiency, and their evaluation of how well each feature represented that construct area. They were also aware of empirical evaluations of the features, including their correlation with scores assigned by human raters.

6.4. Results

Multiple-regression models using these three weighting schemes were trained on responses from TPO (the **TPO sm-train** set), and the TOEFL (the **TOEFL sm-train** set). The TPO models were then used to estimate scores on responses from the **TPO sm-eval** set and the **TPO rectrain+eval** set; the TOEFL models were used to score responses to the **TOEFL sm-eval** set. Evaluation statistics were then calculated to estimate the performance of each model on each associated test set.

In addition to assessing how well *SpeechRater* scores correspond to human ratings for single items, it is worthwhile to examine the effect of machine scoring in estimating the scores for sets of items, and especially for a complete speaking test consisting of six responses by the same examinee. Depending on precisely how scores are reported to examinees and used, error in the scoring of individual items may be tolerable, so long as it does not result in an unacceptable level of inaccuracy in the estimation of a candidate's score for a full test. In order to assess *SpeechRater*'s performance on larger sets of items, the results of this section include statistics relating to the aggregate score for pairs, triples, and sets of six items. In each case, the score for the aggregate is simply the sum of the scores assigned to each item.

Table 12 presents some descriptive statistics for the *SpeechRater* models evaluated on each data set, as well as for human scores assigned to the same responses. (The sample sizes on which these statistics are based for each data set and level of score aggregation are provided in Table 11.) Looking first at human scores, Table 12 shows that the **TOEFL sm-eval** set has a somewhat lower mean score than the data sets from TPO, and also a higher standard deviation. This

Table 11
Number of response sets available for each level of aggregation considered in scoring model evaluation.

	TPO sm-eval	TPO rectrain+eval	TOEFL sm-eval
Single scores	520	2427	10,497
Sets of 2 items	232	1129	5176
Sets of 3 items	163	757	3396
Sets of 6 items	58	308	1615

difference makes the TOEFL data more statistically appropriate for evaluation of *SpeechRater*, as the TPO data sets are affected somewhat by a restriction of range. (Higher-proficiency candidates are more likely to seek out opportunities for practice, such as the TPO, than are lower-proficiency candidates.) Of course, results on the TOEFL data will also be more stable due to the higher sample size.

Turning to the descriptive statistics for *SpeechRater*'s automated scoring, it seems that all three model variants perform similarly in Table 12. The statistics for sets of one, two, three, and six aggregated responses are presented in rows between double lines, while the statistics for the three different test data sets are presented in columns between double lines. The three columns for a given data set correspond to the results of *SpeechRater* models with equal weights, CAC-assigned expert weights, and least-squares optimal weights, respectively. (For human scores, the weighting distinction does not apply, and only one set of statistics is therefore provided.)

Regardless of the feature weighting scheme used, the models match the mean score assigned by humans very closely (which is the expected result for a multiple-regression model), but display substantially lower score variance. Because of the imperfect relationship between *SpeechRater*'s scores and those assigned by humans, the models tend to be more conservative, and assign scores closer to the mean of the observed distribution than humans do.[5]

Table 13 presents the agreement between *SpeechRater* scores and human scores, as well as that between two independent human raters. Two statistics are provided here as measures of the agreement between raters (human or machine). The first of these is the Pearson coefficient of correlation between ratings, which is based on unrounded scores. The second measure is the quadratic-weighted kappa between scores (Cohen, 1960), which is based on scores rounded to the nearest integer. Because information is lost in the rounding process, this is a less accurate measure, but since reported scores are likely to be rounded as well, it may provide an additional indicator of the usefulness of the score feedback provided to examinees.

The first generalization which is clear in Table 13 is that the agreement between human raters is substantially higher for the TOEFL data sets than for the data from the TPO test. In part this stems from the range restriction in the TPO data alluded to above, but it also reflects the different training and scoring procedures used in the high-stakes version of the test. (As described in Section 4, in scoring the TOEFL responses, human raters perform item type specific calibration, and rate only one item type during a session; the process for rating the TPO is less tightly controlled.)

SpeechRater's agreement with human raters is also higher on the **TOEFL sm-eval** set than on the TPO data sets, due to the greater human score variance on that set, and the higher reliability of human scores on the TOEFL data relative to the TPO data.

Both for human agreement and for the agreement between human raters and *SpeechRater*, the correlation and weighted kappa improve as more responses are aggregated together, although the level of improvement is greater for human raters than for *SpeechRater*. For instance, on the **TOEFL sm-eval** set, while the correlation between human ratings and the CAC-weighted model rises from 0.592 for single ratings to 0.707 for full sets of six ratings, the agreement between human raters rises from 0.667 to 0.843 for sets of three ratings (and would likely be even better for sets of six ratings if the data set were sufficient to calculate this).

Overall, the scoring accuracy of the models examined here is moderate for sets of six items combined, with correlations of about 0.7 with human ratings on the **TOEFL sm-eval** set, and correlations between 0.5 and 0.6 on the TPO-based test sets.

[5] While it is possible to correct for this tendency by re-scaling the model (increasing the score variance to match that of human raters), this would increase the model error, and magnify the effect of misclassifications.

Table 12
Descriptive statistics of the human and *SpeechRater* scores for single responses and aggregations of multiple responses across candidates.

			TPO sm-eval			TPO rectrain+eval			TOEFL sm-eval		
		Weighting scheme →	Equal	CAC	Least-squares	Equal	CAC	Least-squares	Equal	CAC	Least-squares
Single scores	Human scoring	Mean		2.73			2.79			2.61	
		SD		0.69			0.71			0.82	
	SpeechRater scoring	Mean	2.78	2.78	2.78	2.78	2.78	2.73	2.61	2.61	2.61
		SD	0.32	0.33	0.31	0.32	0.33	0.31	0.45	0.48	0.50
Totalscore on 2 items	Human Scoring	Mean		5.44			5.58			5.23	
		SD		1.13			1.23			1.44	
	SpeechRater scoring	Mean	5.56	5.55	5.47	5.56	5.55	5.47	5.22	5.22	5.22
		SD	0.57	0.58	0.53	0.57	0.58	0.53	0.86	0.92	0.94
Totalscore on 3 items	Human Scoring	Mean		8.28			8.41			7.86	
		SD		1.56			1.69			2.05	
	SpeechRater scoring	Mean	8.39	8.38	8.25	8.39	8.38	8.25	7.84	7.85	7.85
		SD	0.79	0.83	0.77	0.79	0.83	0.77	1.25	1.34	1.39
Totalscore on 6 items	Human Scoring	Mean		16.66			16.94			15.79	
		SD		2.71			3.02			3.83	
	SpeechRater scoring	Mean	16.82	16.82	16.60	16.82	16.82	16.60	15.74	15.76	15.77
		SD	1.50	1.56	1.41	1.50	1.56	1.41	2.41	2.59	2.68

Table 13
Agreement statistics for human and *SpeechRater* scores for single responses and aggregations of multiple responses across candidates.

			TPO sm-eval			TPO rectrain+eval			TOEFL sm-eval		
		Weighting scheme →	Equal	CAC	Least-squares	Equal	CAC	Least-squares	Equal	CAC	Least-squares
Single scores	Human–human agreement	Pearson r		0.506			0.547			0.667	
		Weighted κ		0.486			0.537			0.667	
	Human-SpeechRater agreement	Pearson r	0.461	0.491	0.469	0.446	0.468	0.452	0.557	0.592	0.613
		Weighted κ	0.301	0.322	0.354	0.311	0.325	0.329	0.443	0.480	0.503
Totalscore on 2 items	Human–human agreement	Pearson r		0.586			0.632			0.787	
		Weighted κ		0.551			0.614			0.787	
	Human-SpeechRater agreement	Pearson r	0.492	0.520	0.498	0.508	0.531	0.517	0.627	0.661	0.682
		Weighted κ	0.350	0.399	0.405	0.416	0.445	0.418	0.538	0.582	0.609
Totalscore on 3 items	Human–human agreement	Pearson r		0.648			0.679			0.843	
		Weighted κ		0.605			0.656			0.842	
	Human-SpeechRater Agreement	Pearson r	0.534	0.557	0.520	0.536	0.562	0.556	0.653	0.686	0.708
		Weighted κ	0.431	0.435	0.410	0.448	0.478	0.472	0.573	0.621	0.650
Totalscore on 6 items	Human–human agreement	Pearson r		0.625			0.742			–	
		Weighted κ		0.560			0.710			–	
	Human-SpeechRater agreement	Pearson r	0.555	0.567	0.509	0.546	0.574	0.565	0.674	0.707	0.729
		Weighted κ	0.469	0.473	0.459	0.488	0.507	0.467	0.603	0.653	0.678

The effects of different feature weighting schemes on the scoring accuracy of *SpeechRater* are not immediately apparent: while the optimally weighted model has the highest association with human ratings on the TOEFL data set, it also has the lowest agreement figures of the three model types on the **TPO sm-eval** set, and lags behind the CAC-weighted model on the **TPO rectrain+eval** set.

However, a closer examination reveals that the optimally weighted model does seem to have a slight advantage over the CAC-weighted model (in terms of agreement with human raters), and that the CAC-weighted model is in turn superior to the model using equally weighted scoring features. Concentrating on the agreement statistics for a full set of six responses (the most meaningful ones, as this is the only speaking score ultimately reported to TPO test-takers), none of the differences between the correlation coefficients for alternative model weightings are significant for the **TPO sm-eval** set, and the only significant difference on the **TPO rectrain+eval** set is that between the CAC-weighted model and an equal-weights model ($p < 0.001$ using a dependent-sample t-test). On the **TOEFL sm-eval** set, all differences in correlation with human ratings between the three models are significant at the $p < 0.001$ level.

While these significant differences on the TOEFL data set show a preference for the optimally weighted model over the expert weights provided by the CAC, and for the CAC weights over equal weights, the magnitude of differences in correlation is rather small, which is consistent with previous findings that multiple regression models tend to be insensitive to small variations in their weights (Wainer, 1976).

The fact that the least-squares weighted model shows an improvement over alternative weightings only on the TOEFL data may be related to the greater reliability of human scores on TOEFL responses. Weights set on the basis of less reliable TPO scores are likely to reproduce human scoring behavior less effectively.

6.5. Discussion

Overall, the best estimate for the agreement of *SpeechRater* scores with human ratings on sets of six completed items comes from the **TOEFL sm-eval** set in Table 13. While *SpeechRater* is ultimately used to score the TPO, and not the TOEFL, the data drawn from the TPO program itself are less than optimal for evaluation, as they are characterized by less reliable human ratings, a smaller sample size, and a distribution of human ratings displaying a significant restriction of range. Intuitively, it also seems reasonable that responses from the high-stakes TOEFL test will present a more direct reflection of examinees' speaking ability than a practice test such as the TPO, during which examinees' attention and motivation may be inconsistent.

The correlations on sets of six items for the three differently weighted models are all close to 0.70 on the **TOEFL sm-eval** set, with the optimal-weights model showing the best agreement with human raters ($r = 0.729$), and the CAC-weighted model a close second ($r = 0.707$). This level of agreement is moderate, as a correlation around 0.7 indicates that any of the weighted automated scoring models captures approximately half of the score variance in human ratings ($r^2 \approx 0.7 \times 0.7 = 0.49$).

Nevertheless, given the goals of the TPO assessment, to allow students to familiarize themselves with the TOEFL test format and to receive instantaneous feedback about their performance, it was determined that this level of agreement was sufficient to make *SpeechRater*'s automated feedback preferable to human scoring for the TPO. Using *SpeechRater* has the advantage of allowing students to get their scores almost immediately, whereas human rating of the TPO had previously required them to wait for up to two weeks. Automated scoring also allows for more reliable scaling of the scoring system during times of high system load. Furthermore, a number of conditions were placed on the use of *SpeechRater*, in order to ensure that examinees are fully informed about the nature of the scores they receive, and that the scores are interpreted appropriately.

First, the only speaking scores reported on the TPO test are for complete sets of six items, due to the lower agreement of *SpeechRater* with human scores on single items or smaller item sets. Second, a Frequently Asked Questions document about *SpeechRater* was drafted and made easily accessible on the TPO web site[6], so that examinees could read about the basis of the machine scoring of their spoken responses and the statistical performance of the *SpeechRater* models. Finally, in addition to the total Speaking score reported by *SpeechRater*, a prediction interval is reported to examinees, which expresses the uncertainty surrounding the relationship between *SpeechRater* scores and human scores. The construction of this interval is the subject of Section 7.

[6] http://toefl.startpractice.com/programs/toefl/toefl_faq.htm.

The expert weighting scheme devised by the CAC was selected for use in the deployed *SpeechRater* scoring model for TPO, because the difference in statistical performance between the CAC model and the optimally weighted model was small, and the faithfulness of construct representation was judged to be superior in the CAC model.

7. Construction of prediction intervals

Once the raw scores for each of an examinee's six responses have been calculated by the scoring model, they are summed together to produce the complete TPO Speaking section score. Note that these item scores and the complete section raw score are unrounded values; rounding occurs only when the section raw score is converted to a scaled score (on the TPO 0–30 scale), in order to make the most efficient use possible of the score estimates provided by *SpeechRater*.

In addition to the score prediction provided by the expert-weighted scoring model, the TPO score report provides an indication of the expected amount by which this predicted score might differ from the score a human would assign to a response, or a set of responses. In short, it provides a prediction interval. Given the limited nature of the speaking construct currently covered by *SpeechRater*, and the imperfect alignment of human and machine scores, it is important to convey to examinees the uncertainty in the score estimates provided. This prediction interval is calculated on the basis of the Speaking section raw score, and the endpoints of the interval are then converted to the 0-30 scale using the same procedure which is applied to the score estimate itself.

One way to estimate the prediction interval would have been to provide a symmetric band, centered around the predicted score, based on the standard error of measurement. However, to account for the possibility that the expected difference between human and machine scores might not be constant across the entire scoring scale, we chose instead to provide an interval based on a cumulative logit model, which estimates the likelihood of a predicted score corresponding to a human score at each point in the scale.[7] Given these probabilities associated with each point in the score scale, we then identify a central region with a probability mass above a certain value, and report it to the examinee as the interval within which a human's rating would be expected to fall with a certain probability. A 90% prediction interval was chosen, because it contains enough of the probability mass that the examinee could be fairly certain that their score would fall within that range, while allowing for relatively simple explanation (e.g., "nine times out of ten, a trained human rater would score your response within this range").

The general form of a cumulative logit model is shown in (4), in which the logit (or log-odds) of a variable Y taking on a value less than some cutoff value k, given some features $x^0, \ldots, x^m$ is modeled as a linear function of these features.

$$\text{logit}(P(Y \leq k|x^0, \ldots, x^m)) = \zeta_k - \sum_{i=0}^{m} \eta_k^i \times x_k^i \tag{4}$$

In defining the prediction interval for *SpeechRater*, the variable Y corresponds to the human score assigned to a response, and there is only one predictive feature x, namely the estimated score for the response. Thus, the cumulative logit model in (4) simplifies to (5) in this case, which can be transformed to (8) to express the probability that the human score is less than k, given *SpeechRater*'s score estimate x.

$$\text{logit}(P(Y \leq k|x)) = \zeta_k - \eta_k \times x_k \tag{5}$$

$$\log\left(\frac{P(Y \leq k|x)}{1 - P(Y \leq k|x)}\right) = \zeta_k - \eta_k \times x_k \tag{6}$$

$$\frac{P(Y \leq k|x)}{1 - P(Y \leq k|x)} = \mathrm{e}^{\zeta_k - \eta_k \times x_k} \tag{7}$$

$$P(Y \leq k|x) = \frac{\mathrm{e}^{\zeta_k - \eta_k \times x_k}}{1 + \mathrm{e}^{\zeta_k - \eta_k \times x_k}} \tag{8}$$

[7] An alternative would be to calculate intervals based on the conditional standard error of measurement(Feldt and Brennan, 1989; Qualls-Payne, 1992) at each point on the scale, but this would require estimation the difficulty parameters associated with each test item, for which the available data may not suffice.

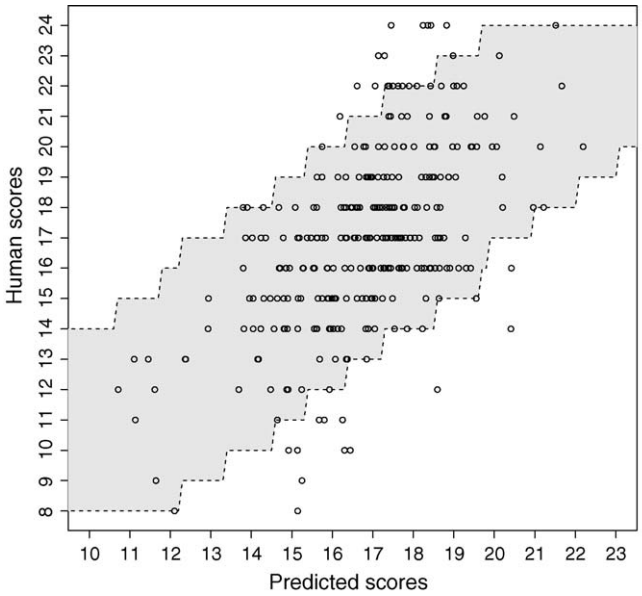

Fig. 5. Prediction intervals for the full range of predicted TPO speaking section scores.

The cumulative logit model was trained on the combination of the scoring model evaluation set and the speech recognizer training set (the **TPO rectrain+eval** set), because the **TPO sm-eval** set alone did not contain enough candidates with complete sets of six task scores. It is less than optimal to use data on which the speech recognizer was trained in further model building, but there is unlikely to be any significant risk of overtraining which would corrupt the calculation of prediction intervals. (On the other hand, using the scoring model training set for the parameterization of the prediction intervals would have been problematic, as we would expect lower prediction error to begin with on these responses.) Fig. 5 shows the data from this combined set, plotting each human score-predicted score correspondence for a set of six tasks as a point on the graph. The sparsely occupied region in the lower left of the graph shows that our TPO data contained very few examinees with aggregate scores of 12 or below.

Fig. 5 also shows the 90% prediction intervals associated with each predicted score in the **TPO rectrain+eval** set, for a full set of six tasks. Because the only speaking score reported on the TPO assessment is the total score for the speaking section, no prediction intervals were created for individual task scores.

In this figure, the prediction interval for a given score predicted by *SpeechRater* is to be read vertically, from the lower dotted line to the upper dotted line. For example, the the interval for a predicted score of 14 can be found by reading across the x-axis to the "14" tick-mark, and then following a vertical line straight up through the dotted lines at 10 and 18. Thus, for a predicted score of 14, the 90% prediction interval is [10,18]. The 90% prediction interval averages about eight score points on the 0-24 raw total score scale. (It is a bit wider when this score is converted to the 0-30 scale on which TOEFL section scores, as well as TPO scores, are reported.)

Fig. 6 shows the same prediction intervals (from the cumulative logit model trained on TPO data) applied to the TOEFL evaluation data introduced in Section 5 above. Because the overall proficiency level in the population of TOEFL examinees differs so much from that of TPO examinees, the *SpeechRater* score estimates are systematically too high, especially at the low end of the score scale. The score region covered by the prediction intervals is correspondingly too high, as shown by the large number of points falling below the grey region within the prediction intervals in Fig. 6. This indicates that the uncertainty expressed by the score intervals is not sufficient to subsume the effect of this population shift on the underlying score estimates.

The prediction interval is an important component of the score feedback provided by *SpeechRater* on the TPO assessment; the limitations of the current automated scoring technology for spontaneous spoken responses make it especially crucial to draw examinees' attention to the uncertainty inherent in the score estimates provided. The accuracy of these intervals is of course strongly dependent on the accuracy of the underlying score estimates, so that to the extent

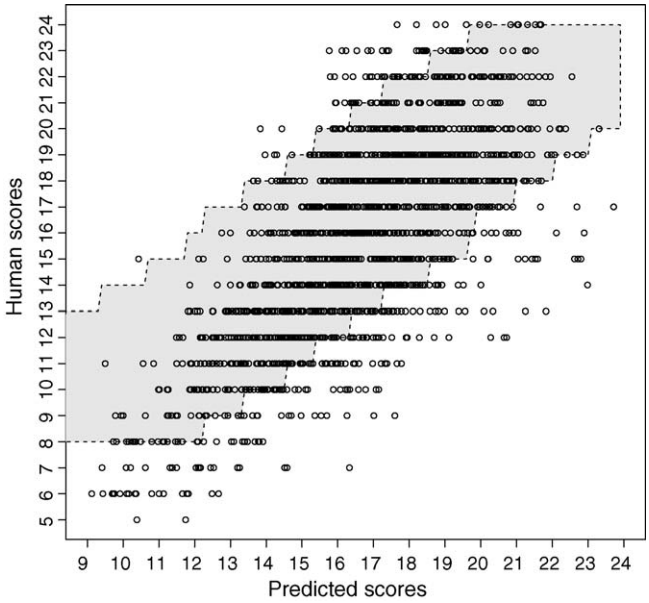

Fig. 6. Prediction intervals applied to spoken responses from the TOEFL.

that the scoring model requires recalibration for a new population, the prediction interval for each score will also need to be re-estimated.

8. Discussion

This paper has described the development and structure of *SpeechRater*, a system for automated scoring of spontaneous spoken responses such as those provided to TOEFL Speaking tasks.

The system processes responses according to a three-stage process. First, a filtering model screens out responses which are not scoreable, because of technical difficulties or the examinee's failure to respond appropriately to the question. Second, the scoring model uses a set of features related to fluency, pronunciation, vocabulary diversity, and grammar to estimate the examinee's score on a 1–4 scale. Finally, the six scores on a test form are summed to produce a total speaking score for the examinee, and a prediction interval is reported which expresses the uncertainty of the relationship between the *SpeechRater* score and the score a human rater might have assigned.

It proved feasible to construct a filtering model with good discrimination between responses which humans scored 1–4 on the one hand, and those which received a zero score or "technical difficulty" rating on the other, although the low frequency of anomalous responses in the TOEFL test limits the range of the model's operational application. *SpeechRater*'s scoring model displays moderate accuracy in terms of agreement with human scores; on sets of six completed items, it exhibits a correlation of about 0.7 with human ratings on the data set with the most reliable set of human scores for comparison (TOEFL).

Based on the strength of this relationship with human scores, the construct basis for *SpeechRater*'s scoring, and a set of preconditions for use (such as reporting of a prediction interval rather than a score estimate alone, and an automated scoring information sheet for examinees), *SpeechRater* was approved for use on a low-stakes practice test known as the TOEFL Practice Online (TPO).

Future work planned for the development of *SpeechRater* will focus on increasing the coverage of the speaking proficiency construct beyond the relatively limited feature set currently used. *SpeechRater*'s current features are primarily related to speech delivery, although there remain important aspects of proficiency to capture even there, such as stress and intonation. In future work, it is hoped that *SpeechRater*'s construct coverage can progress beyond delivery to more fully encompass higher-level aspects of speaking proficiency, such as the complexity of linguistic structures, the appropriate development of the speaking topic, and the coherence of the response. This is a very challenging research

agenda, of course, as it depends on effective speech recognition for a population of non-native speakers of varying L1 backgrounds and proficiency levels.

Acknowledgements

The authors would like to thank an anonymous reviewer, and their ETS colleagues Isaac Bejar, Yeonsuk Cho, and Dan Eignor for their thoughtful reviews which helped to improve the content and readability of this paper. We are also indebted to the Technical and Content Advisory Committees at ETS, whose advice was invaluable in developing the models described here. Mike Wagner and Ramin Hemat deserve special mention for their help in collecting and processing the data for this project, and Pam Mollaun, Arthur Denner and William Lee have our sincere appreciation for their assistance with the human scoring needed for this study. Of course, any errors remaining in the paper are the sole responsibility of the authors.

References

American Educational Research Association, 1999. National Council on Measurement in Education. Standards for Educational and Psychological Testing. American Educational Research Association, Washington, DC.

Attali, Y., Burstein, J., 2006. Automated essay scoring with e-rater v.2. Journal of Technology, Learning and Assessment 4 (3).

Bailey, K.M., 1999. Washback in Language Testing. Tech. Re MS-15, TOEFL Monograph.

Balogh, J., Bernstein, J., Cheng, J., Townshend, B., 2007. Automated evaluation of reading accuracy: assessing machine scores. In: Proceedings of The International Speech Communication Association Special Interest Group on Speech and Language Technology in Education (SLaTE), Farmington, PA.

Bennett, R.E., Bejar, I.I., 1998. Validity and automated scoring: it's not just the scoring. Educational Measurement, Issues and Practice 17 (4), 9–17.

Bernstein, J., 1999. PhonePass testing: Structure and Construct. Tech. Rep., Ordinate Corporation, Menlo Park, CA.

Bernstein, J., DeJong, J., Pisoni, D., Townshend, B., 2000. Two experiments in automated scoring of spoken language proficiency. In: Proceedings of InSTILL (Integrating Speech Technology in Language Learning), Dundee, Scotland.

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA.

Burstein, J., 2003. The e-rater® scoring engine: Automated essay scoring with natural language processing. In: Shermis, M.D., Burstein, J. (Eds.), Automated Essay Scoring: A Cross-disciplinary Perspective. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 113–121.

Callear, D., Jerrams-Smith, J., Soh, V., 2001. CAA of short non-MCQ answers. In: Proceedings of the 5th International CAA Conference, Loughborough University, Loughborough, UK.

Charney, D., 1984. The validity of using holistic scoring to evaluate writing: a critical overview. Research in the Teaching of English 18, 65–81.

Chevalier, S., 2007. Speech interaction with Saybot player, a CALL software to help Chinese learners of English. In: Proceedings of The International Speech Communication Association Special Interest Group on Speech and Language Technology in Education (SLaTE), Farmington, PA.

Clark, J.L.D., Swinton, S.S., 1979. An exploration of speaking proficiency measures in the TOEFL context. Tech. Re ETS-RR-04–79, Educational Testing Service, Princeton, NJ.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1), 37–46.

Cucchiarini, C., Strik, H., Boves, L., 1997a. Automatic evaluation of Dutch pronunciation by using speech recognition technology. In: IEEE Automatic Speech Recognition and Understanding Workshop, Santa Barbara, CA.

Cucchiarini, C., Strik, H., Boves, L., 1997b. Using speech recognition technology to assess foreign speakers' pronunciation of Dutch. In: Third International Symposium on the Acquisition of Second Language Speech: NEW SOUNDS 97. Klagenfurt, Austria.

Elliott, S., 2003. Intellimetric™: From here to validity. In: Shermis, M.D., Burstein, J. (Eds.), Automated Essay Scoring: A Cross-disciplinary Perspective. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 71–86.

Eskenazi, M., Kennedy, A., Ketchum, C., Olszewski, R., Pelton, G., 2007. The Native Accent™ pronunciation tutor: measuring success in the real world. In: Proceedings of The International Speech Communication Association Special Interest Group on Speech and Language Technology in Education (SLaTE), Farmington, PA.

Fayyad, U.M., Irani, K.B., 1993. Multi-interval discretization of continuousvalued attributes for classification learning. In: Proceedings of the International Joint Conference on Uncertainty in AI, pp. 1022–1027.

Feldt, L.S., Brennan, R.L., 1989. Reliability. In: Linn, R.L. (Ed.), Educational Measurement, American Council on Education, 3rd edition. Macmillan Publishing, Phoenix, AZ, pp. 105–146.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R., Cesari, F., 2000. The SRI EduSpeak system: recognition and pronunciation scoring for language learning. In: Proceedings of InSTILL (Integrating Speech Technology in Language Learning) 2000, Scotland, pp. 123–128.

Frank, E., Wang, Y.S., Inglis, G.H., Witten, I.H., 1998. Using model trees for classification. Machine Learning 32 (1), 63–76.

Garofolo, J.S., Fiscus, J.G., Fisher, W.M., 1997. Design and preparation of the 1996 HUB-4 broadcast news benchmark test corpora. In: Proceedings of DARPA Speech Recognition Workshop, Morgan Kaufmann, pp. 15–21.

Hall, M.A., 1998. Correlation-based feature selection for machine learning. Ph.D. thesis, University of Waikato, Hamilton, New Zealand.

Hanley, J., McNeil, B., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Hoyt, W.T., 2000. Rater bias in psychological research: when is it a problem and what can we do about it? Psychological Methods 5, 64–86.

Landauer, T.K., Laham, D., Foltz, P.W., 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor. In: Shermis, M.D., Burstein, J. (Eds.), Automated Essay Scoring: A Cross-disciplinary Perspective. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, pp. 87–112.
Leacock, C., Chodorow, M., 2003. C-rater: scoring of short-answer questions. Computers and the Humanities 37 (4), 389–405.
Messick, S., 1996. Validity and washback in language testing. Language Testing 13 (3), 241–256.
Mitchell, T., Russell, T., Broomhead, P., Aldridge, N., 2002. Towards robust computerized marking of free-text responses. In: Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, Loughborough, UK.
Moran, M.R., 1987. Options for written language assessment. Focus on Exceptional Children 19 (5), 1–10.
Murphy, K.R., Anhalt, R.L., 1992. Is halo error a property of the raters, ratees, or the specific behaviors observed? Journal of Applied Psychology 72, 494–500.
Page, E.B., 1966. The imminence of grading essays by computer. Phi Delta Kappan 48, 238–243.
Page, E.B., 1968. The use of the computer in analyzing student essays. International Review of Education 14 (2), 210–225.
Qualls-Payne, A.L., 1992. A comparison of score level estimates of the standard error of measurement. Journal of Educational Measurement 29 (3), 213–225.
Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1 (1), 81–106.
Sargeant, J., Wood, M.M., Anderson, S.M., 2004. A human–computer collaborative approach to the marking of free text answers. In: Proceedings of the 8th International CAA Conference, Loughborough University, Loughborough, UK.
Shohamy, E., 1983. Rater reliability of the oral interview speaking test. Foreign Language Annals 3, 219–222.
Stiggins, R.J., 1982. A comparison of direct and indirect writing assessment methods. Research in the Teaching of English 16 (2), 101–114.
Sukkarieh, J., Pulman, S., 2005. Information extraction and machine learning: Automarking short free text responses to science questions. In: Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED), Amsterdam, The Netherlands.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York.
Wainer, H., 1976. Estimating coefficients in linear models: it don't make no nevermind. Psychological Bulletin 83 (2), 213–217.
Wiggins, G., 1989. Teaching to the (authentic) test. Educational Leadership 46 (7), 41–47.
Wiggins, G., 1990. The case for authentic assessment. Practical Assessment, Research and Evaluation 2 (2).
Williamson, D.M., Bejar, I.I., Mislevy, R.J., 2006. Automated Scoring of Complex Tasks in Computer-based Testing: an Introduction. Lawrence Erlbaum Associates, Mahwah.
Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann.
Xi, X., Higgins, D., Zechner, K., Williamson, D., 2008. Automated scoring of spontaneous speech using SpeechRater v1.0. Tech. Re ETS-RR-08–62, Educational Testing Service, Princeton, NJ.
Yang, Y., Buckendahl, C., Juskiewicz, P.J., Bhola, D.S., 2002. A review of strategies for validating computer-automated scoring. Applied Measurement in Education 15 (4), 391–412.
Zechner, K., Bejar, I., 2006. Towards automatic scoring of non-native spontaneous speech. In: Proceedings of HLT-NAACL, New York, NY.
Zechner, K., Higgins, D., Xi, X., Williamson, D.M., 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication 51, 883–895.