Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 883–895
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Automatic scoring of non-native spontaneous speech in tests of spoken English

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

Educational Testing Service, Automated Scoring and NLP, Rosedale Road, MS 11-R, Princeton, NJ 08541, USA

Received 28 June 2008; received in revised form 14 April 2009; accepted 20 April 2009

Abstract

This paper presents the first version of the SpeechRater[SM] system for automatically scoring non-native spontaneous high-entropy speech in the context of an online practice test for prospective takers of the Test of English as a Foreign Language® internet-based test (TOEFL® iBT).

The system consists of a speech recognizer trained on non-native English speech data, a feature computation module, using speech recognizer output to compute a set of mostly fluency based features, and a multiple regression scoring model which predicts a speaking proficiency score for every test item response, using a subset of the features generated by the previous component. Experiments with classification and regression trees (CART) complement those performed with multiple regression. We evaluate the system both on TOEFL Practice data [TOEFL Practice Online (TPO)] as well as on Field Study data collected before the introduction of the TOEFL iBT.

Features are selected by test development experts based on both their empirical correlations with human scores as well as on their coverage of the concept of communicative competence.

We conclude that while the correlation between machine scores and human scores on TPO (of 0.57) still differs by 0.17 from the inter-human correlation (of 0.74) on complete sets of six items (Pearson *r* correlation coefficients), the correlation of 0.57 is still high enough to warrant the deployment of the system in a low-stakes practice environment, given its coverage of several important aspects of communicative competence such as fluency, vocabulary diversity, grammar, and pronunciation. Another reason why the deployment of the system in a low-stakes practice environment is warranted is that this system is an initial version of a long-term research and development program where features related to vocabulary, grammar, and content will be added in a later stage when automatic speech recognition performance improves, which can then be easily achieved without a re-design of the system.

Exact agreement on single TPO items between our system and human scores was 57.8%, essentially at par with inter-human agreement of 57.2%.

Our system has been in operational use to score TOEFL Practice Online Speaking tests since the Fall of 2006 and has since scored tens of thousands of tests.

© 2009 Elsevier B.V. All rights reserved.

Keywords: Speech scoring; Automatic scoring; Spoken language scoring; Scoring of spontaneous speech; Speaking assessment

1. Introduction

In testing language proficiency with tests such as the TOEFL® iBT, an important distinction can be made between the two receptive modalities of language (listening, reading) and the two productive modalities (speaking, writing). While the receptive modalities can not easily be tested directly, e.g., by observing how the listening and language understanding processes unfold in the test taker's brain, they can be quite conveniently tested by focusing on the comprehension aspect of both listening and reading. These tests have traditionally used a multiple-choice paradigm, a design that has flourished as the preferred item type for

* Corresponding author. Tel.: +1 609 734 1031; fax: +1 609 734 1090.
E-mail addresses: kzechner@ets.org (K. Zechner), dhiggins@ets.org (D. Higgins), xxi@ets.org (X. Xi), dmwilliamson@ets.org (D.M. Williamson).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.04.009

more than a generation, and with good reason. It is efficient to develop and administer, can be scored relatively unambiguously and swiftly, and is supported by a rich infrastructure of statistical methods and test theory. Furthermore, there is an increasing availability and lower cost of online administration and instantaneous scoring.

When looking at the productive modalities of language, speaking and writing, however, an argument for a multiple-choice design is much harder to make since it defies the purpose of the assessment, namely, how well a test taker can use the language for communicative purposes – how proficient he or she is in speaking and in writing. Although approximations have been used in the past, such as combining partial sentences, they can not really assess genuine productive language use.

For these reasons, so-called "constructed response" items are used in the TOEFL® iBT speaking and writing sections where candidates have to produce several samples of speech, relating to a pre-specified task, and have to write several essays on a given topic.

This different test design has generally necessitated the use of the slower and more costly human scoring that accompanies use of constructed response items. Despite research into automated scoring of complex tasks extending back more than 40 years (e.g., Page, 1966), only relatively recently (Clauser et al., 1997; Burstein et al., 1998; Williamson et al., 1999) has the ability for computerized delivery and automated scoring of constructed response items enabled the practical operational use of automated scoring for such items. Initially, such applications were primarily in automated scoring of essays (e.g., Burstein et al., 1998; Chodorow and Burstein, 2004; Attali and Burstein, 2005; Landauer and Dumais, 1997; Rudner et al., 2006), which has matured to a considerable degree. However, recent research in natural language processing and speech recognition capabilities has expanded the nature of constructed response tasks that are automatically scorable to include short answer tasks requiring factual information (e.g., Leacock, 2004; Leacock and Chodorow, 2003) and tasks eliciting highly predictable speech (e.g., Bernstein, 1999).

What has not been attempted so far, however, is to build an automatic scoring system for speech with high linguistic entropy, i.e., where the sequence of words is largely unpredictable. TOEFL iBT Speaking has items that elicit spoken responses about everyday familiar topics, as well as about campus life and academic situations, all of which require free, spontaneous, high-entropy speech responses by the test candidates.

The two major challenges for scoring high-entropy speech — aside from challenges related to natural language processing which are similar to those involved with scoring essays — are that (a) one can not use a pattern matching or forced-alignment approach as it is commonly used in low-entropy scoring paradigms; and (b) that non-native spontaneous speech is inherently hard to recognize and only moderate word accuracies are realistically achievable. TOEFL iBT is taken all over the world by speakers of more than 100 native languages, which poses additional challenges in speech recognition.

To address the first challenge, we are using speech features, derived from speech recognition output, to be used by a machine learning component that predicts scores for speaking proficiency (the "scoring model"). The second challenge is addressed by focusing on low-level[1] fluency features where word identities are not of critical importance.

At first glance, this approach may seem somewhat contradictory since one could raise the question of why using automatic speech recognition (ASR) technology is necessary if the features are mostly related to fluency and not so much to word identity and if word accuracy for this task is expected to be rather low. Our reasons for relying on ASR technology are twofold, however: (a) not *all* of the most relevant features can be computed without knowing word identities; and (b) this system is an initial version of a long-term research and development program where higher-level features (e.g., related to vocabulary, grammar, and content) will be added in a later stage when ASR performance improves. This we can then do seamlessly without a re-design of the system, which would be necessary had we started out without using ASR technology for this initial version.

This paper presents the results of a research and development effort for SpeechRater℠ v1.0, an automated scoring system for the spontaneous speech of English language learners used operationally in the TOEFL Practice Online assessment. Its three main components are a state-of-the-art speech recognizer, generating word-level information (including inter-word pauses) as well as response-level scores (acoustic and language model scores), a feature computation module, operating on the output of the recognizer, and a scoring model which maps a subset of the speech features to a speaking proficiency score, trained on human rated test items.

Our two main success criteria for making a positive decision for operational use of SpeechRater in the TOEFL Practice Online Speaking test are (a) a reasonable coverage of the concept of communicative competence by our features (Condition 1); and (b) an only moderate drop in Pearson r correlation of machine–human score comparisons from human–human score comparisons (Condition 2).

These conditions would certainly be more stringent for a deployment in a high-stakes environment, but we feel that the advantage of the fast turnaround time compared to human scoring in conjunction with a fulfillment of the two stated conditions can warrant the implementation of our system for the TOEFL Practice Online Speaking test.

The organization of this paper is as follows: Section 2 presents related work, Section 3 describes the TOEFL Practice Online test, and Section 4 introduces the system

[1] With "low-level" we like to indicate the contrast to "high-level" linguistic features, such as those related to content or grammar. In other words, we consider the level of word-based linguistic complexity here.

architecture of SpeechRater. In Section 5 we describe the data used to develop and evaluate the system, and in Sections 6–8 we describe the development and evaluations of the three main components of SpeechRater, namely the speech recognizer, the feature computation module, and the scoring model. Next, we discuss our contributions in Section 9 and conclude in Section 10, with an outlook on future work.

2. Related work

There has been previous work to characterize aspects of communicative competence such as fluency, pronunciation, and prosody. Franco et al. (2000a,b) present a system for automatic evaluation of the pronunciation quality of both native and non-native speakers of English on the phone level and the sentence level (EduSpeak). Candidates read English texts and a forced alignment between the speech signal and the ideal path through the Hidden Markov Model (HMM) is computed. Next, the log posterior probabilities for pronouncing a certain phone at a certain position in the signal are computed to achieve a local pronunciation score. These scores are then combined with other automatically derived measures such as the rate of speech (number of words per second) or the duration of phonemes to yield global pronunciation scores.

Cucchiarini et al. (1997a,b, 2000a,b, 2002) describe a system for Dutch pronunciation scoring along similar lines. Their feature set, however, is more extensive and contains, in addition to log likelihood Hidden Markov Model scores, various duration scores, and information on pauses, word stress, syllable structure, and intonation. In an evaluation, correlations between four human scores and five machine scores range from 0.67 to 0.92.

Bernstein (1999) presents a test for spoken English (SET-10) that uses the following types of items: reading, sentence repetition, sentence building, opposites, short questions, and open-ended questions. All types except for the last are scored automatically and a score is reported that can be interpreted as an indicator of how native-like a speaker's speech is. In Bernstein et al. (2000), an experiment is performed to establish the generalizability of the SET-10 test. It is shown that the SET-10 test scores can predict different levels on the Oral Interaction Scale of the Council of Europe's Framework for describing oral proficiency of second/foreign language speakers with reasonable accuracy (North, 2000). This paper further reports on studies done to correlate the SET-10 automated scores with the human scores from two other tests of oral English communication skills. Correlations are found to be between 0.73 and 0.88.

Zechner and Bejar (2006) investigate the automated scoring of unrestricted, spontaneous speech of non-native speakers. They focus on exploring a number of different fluency features for the automated scoring of short (1 min) responses to test questions in a TOEFL-related program. They explore scoring models based on classification and regression trees (CART) as well as support vector machines (SVM). Their findings are that the SVM models are more useful for a quantitative analysis, whereas the CART models allow for a more transparent summary of the patterns that underlie the data.

In this paper, we compare using CART (Brieman et al., 1984) and multiple regression (MR) to build the scoring models for TOEFL Practice Online. Another major difference between previous work and the work reported in this paper is that we use feature normalization and transformation to obtain statistically more meaningful input variables for the scoring model. In addition, we do not use the whole set of features in an exploratory fashion. Instead, we have carefully selected a subset of features that are both good predictors of human scores and maximize the representation of the concept of communicative competence (Bachman, 1990; Bachman and Palmer, 1996), given the restrictions due to imperfect word accuracy by the ASR system which limits the feature space to mostly fluency-related features. Only with significantly higher word accuracy, could a more complete coverage of communicative competence be attempted by adding features related to grammatical accuracy, topical content, etc.

3. TOEFL Practice Online

The Speaking section of the TOEFL iBT is designed to measure the academic English speaking abilities of non-native speakers who plan to study at English-medium institutions for higher education. Using tasks that require language use typical of an academic environment, TOEFL iBT Speaking represents an important advancement in the large-scale assessment of productive skills. (Speaking was not a compulsory component of TOEFL prior to TOEFL iBT. Test of Spoken English, TSE®, was offered as a separate speaking test for institutions which required their applicants to submit speaking scores.) However, it poses challenges to learners in parts of the world where opportunities to practice speaking English are limited.

The TOEFL Practice Online (TPO) assessment is designed to help prospective TOEFL iBT examinees become familiar with and better prepared for the TOEFL iBT. It is designed to mirror the content and design characteristics of the TOEFL iBT to the extent possible in an economical practice environment. As such, the various sections of the TPO each have the same number of items, same mixture of content represented, and same item types that appear in the operational TOEFL. In fact, all of the items that are used for the TPO are retired operational TOEFL iBT items. However, unlike the TOEFL iBT test, the TPO allows users to customize their practice and take the test in a timed mode or untimed mode. The timed mode attempts to replicate the operational testing experience by using the same online delivery system and timing restrictions of the TOEFL iBT. In the untimed mode, users can progress at their own pace, starting or stopping the test whenever they like and revisiting items they have

completed if desired. Another important distinction between the TPO and the TOEFL iBT is that the former targets more immediate and cost-effective score feedback. This immediate feedback on students' performance is intended to inform their self-assessment of understanding of and comfort with the TOEFL iBT test administration. In early 2006 the users of TPO could instantly receive scores on reading and listening sections, both based on multiple-choice items, as well as the writing section, with automated writing scores provided by e-rater® (Attali and Burstein, 2004). The scores on speaking sections were produced by human raters within five business days. As a result of substantial interest in more immediate feedback from the speaking section of the TPO (hereafter called TOEFL iBT Speaking Practice test) a research agenda was launched to develop and deploy an automated scoring system for spontaneous speech. The immediate goal of this effort was to improve the scoring efficiency of the TOEFL iBT Speaking Practice test while maintaining quality comparable to that of trained human rater scoring for the TPO assessment. The long-term goal is to provide instructional and diagnostic feedback based on automated features beyond the score feedback provided by human scoring while maintaining a level of accuracy of scores nearly equivalent to that of human scoring.[2] The result of this effort was the release of SpeechRater v1.0 for operational use in the TPO in Fall of 2006.

In the TOEFL iBT Speaking Practice test, as for the TOEFL iBT test, each test contains six tasks.[3] The first two tasks are *independent* tasks that ask candidates to provide information or opinions on familiar topics based on their personal experience or background knowledge, with 45 seconds of speaking time. The purpose of independent tasks is to measure the speaking ability of examinees independent of their ability to read or listen to English language. The remaining four tasks are *integrated* tasks that engage reading, listening and speaking skills in combination to mimic the kinds of communication expected of students in campus-based situations and in academic courses. Candidates read and/or listen to some stimulus materials and then respond to a question based on them. Each of the four integrated tasks has a speaking time of 60 seconds. The entire Speaking section of the test takes approximately 20 minutes. For each of the six tasks, after task stimulus materials and/or test questions are delivered, the examinees are allowed a short time to consider their response and then provide their responses in a spontaneous manner.

The scoring rubric used by human raters to evaluate the responses to the TOEFL iBT Speaking Practice test is identical to that used for the TOEFL iBT Speaking Test. (A rubric is a table where for each score level, typical characteristics of candidates' performances are listed to guide the human rater.) Similarly, the scoring of the practice test is conducted by raters who regularly score the operational TOEFL iBT Speaking test. The raters issue a holistic score for each response on a score scale from 1 to 4, four being the best, that is based on three key categories of performance: delivery (pronunciation, intonation, and fluency), language use (vocabulary and grammar), and topical development (content, coherence, and organization).

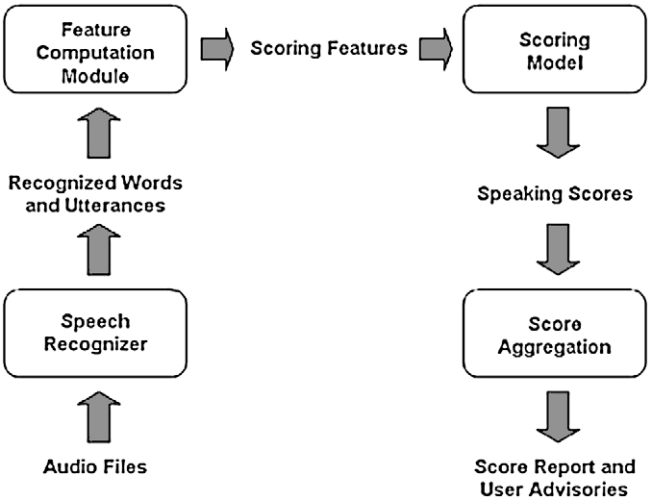

Fig. 1. Architecture of the SpeechRater automated speech scoring system.

4. SpeechRater system architecture

This section describes the architecture of our automated speech scoring system, which serves as a natural organizing structure for the remainder of the paper. An automated speech scoring system consists of three major components (see Fig. 1). First, the test taker's voice is recorded in Windows Media (wma) format (22 kHz, 16 bit, mono), then converted to wav format (16 bit PCM, 11 kHz, mono), and sent to the speech recognizer. Second, the feature computation module reads the output hypotheses from the speech recognizer and generates a feature vector for each recorded speech sample. Third, a pre-determined subset of the features are extracted and sent to the scoring model which then produces a score for every spoken response.[4] Finally, all six scores of a test are added and scaled to be sent back to the test taker.

5. Data

In building and evaluating the scoring models described in this report, we made use of two data sets: responses to

[2] We realize that these mentioned long-term goals are still far out of reach currently given the low word accuracy of the recognizer and more research needed in the area of useful feedback for (prospective) test candidates.

[3] In this paper, we use the terms "tasks" and "items" interchangeably. An item or task refers to a test question and the associated stimulus materials (if any).

[4] While this paper reports on two different scoring model approaches, multiple regression and CART trees, the operational system only uses multiple regression.

the TOEFL Practice Online assessment (the *TPO data set*), and responses from a TOEFL iBT Field Study (the *iBT data set*).

5.1. TOEFL iBT practice speaking data

In total, the TPO data contains 4162 spoken responses. These responses were double-scored by human raters. For the purposes of model building and analysis, we used the second set of human scores, because they were undertaken under more optimal rating conditions.

The TPO data contains responses from four distinct test forms, with each test form containing six distinct speaking tasks: two independent tasks and four integrated tasks (see Section 3). Each TPO response may be assigned a score in the range of 1–4 (with four indicating the highest proficiency), or 0 if the candidate makes no attempt to answer or produces only a few words unrelated to the topic. It may also be labeled as "technical difficulty" (TD) when technical issues may have degraded the audio quality so that a fair evaluation is not possible. These scoring rules are in accordance with the scoring of the operational TOEFL iBT test, and with the scoring of the iBT Field Study data described below.

We set aside a portion of the TPO data for the training of the speech recognizer (the **rec-train** set, about 1900 responses), and partitioned the remaining data into the scoring model train (**sm-train**, about 1300 responses) and evaluation (**sm-eval**, about 500 responses) sets to maximize its utility in evaluating the features and in building and evaluating the scoring models discussed in Section 8. (The remaining responses were TD or 0 which were excluded from this study.[5]) The scoring model train data were also used in evaluating the statistical properties of features.

Orthographic human transcriptions were made for the entire rec-train set and for portions of the scoring model sets; in all about 3000 responses were transcribed.

The partitioning of the TPO data was done in such a way that no overlap between speakers or tasks was allowed between the scoring model training and evaluation sets. The partitioning was also designed to minimize speaker and task overlap between the recognizer training set and all other sets, although this constraint could not be enforced absolutely. In order to ensure that all data partitions were of sufficient size for their intended purposes, while meeting our other constraints, we were forced to accept some speaker and task overlap between the **rec-train** partition and other partitions. The total proportion of responses with task and speaker overlap with the **rec-train** set amounts to 25% of the **sm-train** set, and 31% of the **sm-eval** set. Because there is still no overlap between the **sm-train** and **sm-eval** sets, it is unlikely that this will result in inflated estimates of scoring accuracy for SpeechRater.[6]

The partitioning process was also designed to ensure that the scoring model training and evaluation sets (a) contain a broad set of tasks, (b) contain similar proportions of responses from speakers of particular linguistic backgrounds, and (c) contain approximately the same proportion of responses to independent and integrated tasks.

This resulted in the division of the TPO data scored in the range of 1–4 into three sets (Table 1).

The item-based agreement between human raters on rating scorable responses in the range 1–4 on the combined rec-train and sm-eval sets was fairly low.[7] Exact agreement was only 57.2%, with a quadratic-weighted κ of 0.54, and Pearson r of 0.55. The level of human agreement in terms of correlation and kappa improves somewhat as we aggregate scores; exact/adjacent agreement, correlation and kappa on summed pairs of scores, triples, and full sets of six is presented in Table 2. We use quadratically weighted kappa (Cohen, 1968) throughout the paper, where the weights correspond to the squares of score differences. E.g., if the human score for a certain item was 1 and the machine score 3, the weight for the kappa computation would be the square of 2, i.e., 4. That way, larger deviations from the gold standard are more heavily penalized than smaller ones.

5.2. TOEFL iBT Field Study data

The *TOEFL iBT Field Study* was a pilot study undertaken before the official roll-out of the TOEFL iBT test. While we were primarily interested in model performance on TPO data, we used the Field Study data in doing some evaluation runs for a number of reasons. First, the conditions under which the Field Study data were scored were closer to best practice than they were with the TPO data sets. Second, the partitioning of the Field Study data allows for better evaluation of the effects of item score aggregation, since the evaluation set contains more complete forms (sets of six tasks for a given examinee).

Third, the score distribution of the Field Study is more uniform and even, representing a situation encountered by a typical TOEFL iBT administration. Finally, evaluation on the Field Study data provides us with some idea of how our model generalizes across populations, file formats and speech recognizers. (The file format for TPO Speaking is .wma, but for the Field Study it is .au.)

The Field Study data contained 3502 responses from a single TOEFL iBT Speaking test form that were scored

[5] In operational mode, these responses are filtered out using a predictor module based on some of the speech features described below in Section 7, as well as some basic prosodic features derived from a pitch tracker and from frame-based power information. We built an automatic classifier for this filtering model that has an accuracy on TPO data of over 99%.

[6] While there can be an effect on word accuracy, we found in preliminary experiments that small changes in word accuracy have very little effect on features and scores.

[7] We combine these sets in order to be able to compute agreement on full sets of six items belonging to one test form. The scoring model evaluations done in Section 8 also refer to this data set combination.

Table 1
Summary statistics of TPO data scored in the range of 1–4 (TD and 0 scores excluded).

Data set	Responses	Speakers	Topics	Average score	SD of score	Score distribution (percent in brackets)			
						1	2	3	4
Rec-train	1907	320	24	2.81	0.72	52 (3%)	550 (29%)	1011 (53%)	294 (15%)
Sm-train	1257	263	15	2.74	0.77	58 (5%)	405 (32%)	603 (48%)	191 (15%)
Sm-eval	520	120	9	2.73	0.69	18 (3%)	159 (31%)	289 (56%)	54 (10%)

Table 2
Human agreement on aggregated TPO scores (TD and 0 scores excluded) for the TPO sm-eval + rec-train sets.

Number of scores	Exact agreement	Exact + adjacent agreement	Quadratic-weighted kappa	Pearson *r*
1	57.2%	97.5%	0.537	0.547
2	Not [a]computed	Not computed	0.614	0.632
3	Not computed	Not computed	0.656	0.679
6	Not computed	Not computed	0.710	0.742

[a] As a common practice, exact agreement and exact & adjacent agreement rates are only computed for item or task level scores.

Table 3
Summary statistics of iBT Field Study data sets.

Data sets	Responses	Speakers	Topics	Average score	SD of score	1	2	3	4
smFS-train	1750	311	6	2.44	1.02	366	573	482	329
smFS-eval	1752	315	6	2.48	1.00	339	553	542	318

Table 4
Human agreement on aggregated Field Study scores.

Number of scores	Exact agreement	Exact + adjacent agreement	Quadratic-weighted kappa	Pearson *r*
1	57.1%	98.3%	0.77	0.77
2	Not computed	Not computed	0.86	0.86
3	Not computed	Not computed	0.93	0.94

on the 1–4 scale (0 s and TDs were not included). Since we already had a non-native speech trained recognizer for this file format, and none of these data were transcribed, all of the data were used for the scoring model train (smFS-train) and evaluation (smFS-eval) sets. These two sets of data were constructed to maximize the number of examinees with six complete tasks in a set so that we could evaluate candidates' total scores on this section. This constraint prevented us from enforcing a ban on task overlap between the smFS-train and smFS-eval sets, but did allow us to prevent speaker overlap. Table 3 shows the properties of these two data sets.

Not all of the responses in these sets were double-scored, so we were forced to evaluate the level of human agreement on that subset of the data which had been double-scored. These results are provided in Table 4. (Note that we did not have enough double-scored responses to provide agreement results for sets of six tasks.)

One point to note is that the human–human agreement as indicated by the weighted kappa and the correlation was much higher for the Field Study data than for the TPO data. This reflects in part the fact that the Field Study scores were more varied and more evenly distributed across the four score levels than the TPO scores. In contrast, in the TPO data, the scores clustered around 3, with very few at the score level of 1.

6. Speech recognizers

For both data sets, we use a state-of-the-art gender-independent HMM recognizer bootstrapped on a native speech recognizer but trained on non-native speech data. Whereas the TPO recognizer makes use of genuine TPO data for its training, we did not have transcribed iBT data and had to use a recognizer trained on similar non-native data, from the TOEFL Academic Speaking Test (TAST).

Table 5 provides the information on the recognizers used for the TPO Speaking and the Field Study data sets. The word error rates were around 50%, which would be a low number for native speech, but not unreasonable for non-native speech with a large diversity of native language backgrounds and large variation in speaking proficiency. Word accuracies of individual responses vary considerably, typically between 10% and 80%, where speakers with higher scores are also typically better recognized – correlations between word accuracy and human scores are around 0.4 (Pearson *r* values are significant at the 0.01 level).

Table 5
Characteristics of the speech recognizers used for TPO Speaking and Field Study data sets.

	TPO recognizer	Field Study recognizer
Audio format	WAV (from WMA)	AU
Sampling rate	11 kHz	8 kHz
Resolution	16 bit	8 bit
Channels	Mono	Mono
AM training – non-native responses (approx.)	1900	650
LM training – non-native responses (approx.)	600,000 words	80,000 words
LM training – native speech Corpora	Broadcast News (Linguistic Data Consortium, 1997)	Broadcast News (Linguistic Data Consortium, 1997)
Evaluation set size in responses	645	150
Average word accuracy	49.6%[a]	52.6%[a]
Range of word accuracy on evaluation set	9.5–83.1%	4.2–82.6%
Correlation between word accuracy and score	0.39	0.47

[a] We use an unbiased formula for word accuracy computation: word accuracy = 100.0 * 0. 5 * (correct/(correct + deletions + substitutions) + correct/(correct + insertions + substitutions)). Unlike many other formulas for word accuracy and word error rate, this one is symmetrical between hypothesis and reference.

7. Scoring features

The output of the recognizer is a list of words with timing and confidence score information. All features are computed based on this word list and the associated timing information by the feature computation module, except for amscore and lmscore (see Table 6) which were drawn from the recognizer's feedback on the entire utterance as opposed to individual words.

Based on suggestions from the literature (e.g., Cucchiarini et al., 1997b, 2002), and from experts in test development and training of human raters, we put together a list of 29 initial features (Table 6).

They cover different aspects of communicative competence as denoted in the TOEFL iBT rubric; the focus is on fluency, with pronunciation, vocabulary diversity and grammatical accuracy added to the mix.

Based on a thorough review by test development experts and human rater training specialists, a set of 13 features was selected from the initial list for use in the final component, the scoring model. The main criteria used in selection are (a) relevance to the construct of "speaking" (i.e., communicative competence), (b) coverage of all the features combined in representing the speaking construct (the overall quality of speech) and (c) good empirical evidence that a feature correlates well with human scores.

Due to high inter-correlations between some of these features, two more features, which were essentially representing redundant information, were removed from the list to yield 11 final feature candidates for the scoring models (marked with a * in Table 6).

Note that the final multiple regression scoring model only uses five of those 11 features. The feature set of 11 candidates was reduced to five features by means of building a regression model and keeping only the top performing features, while also considering maximal construct coverage (best representation of "communicative competence") (see Section 8.1).

Since one of our scoring model methods is multiple regression, we have to address the fact that the features we have developed for speech scoring may not conform to this model's assumptions, notably the assumption of a linear relationship between the features and the score, and the assumption that the error term in the regression equation be normally distributed. To address this possibility, we examined the distribution of each of our features, and considered transformations of the features which might improve the correlation between the feature and the item score to be predicted, as well as making the feature's distribution more normal. We limited ourselves to basic transformations such as "inverse", "square", "square root", or "logarithm". (Since CART trees are not making any assumptions on normality of feature distributions, only features used in multiple regression scoring models were transformed.)

Table 7 shows the change in correlations between the features, which were transformed, and human scores before and after these transformations.

Finally, also only for multiple regression, outliers were defined as feature values more than four standard deviations from the mean; these values were mapped to this boundary (+-4 standard deviations from the mean). The same procedures (transformations and outlier handling) were also applied to the test data for the evaluation of the scoring model.

For the regression model, all features were further normalized to a standard Gaussian distribution with mean of 0 and standard deviation of 1.

8. Scoring models

We explored two types of scoring models: classification and regression trees (CART) and multiple regression (MR).

A main advantage of CART is that trained trees can be seen as mirroring the decision behavior of human raters to some extent, whose scoring results can be directly reconstructed as a sequence of decisions. Also, CART is not sensitive to feature outliers or non-normal feature distributions.

The main advantages of MR, on the other hand, are its much simpler design and much longer history of being used

Table 6
Candidate features for the development of the scoring models. Features marked with a * were among the 11 selected features for scoring model training.

Feature number	Feature name	Feature class	Dimension	Description
1	numwds	Length	NA	Number of words
2	Numtok	Length	NA	Number of tokens [numwds + numdff]
3	globsegdur	Length	NA	Duration of entire transcribed segment, including all pauses
4	segdur	Length	NA	Total duration of segment without disfluencies & pauses
5	uttsegdur	Length	NA	Duration of entire transcribed segment but without inter-utterance pauses
*6	wdpchk	Fluency	Delivery	Average chunk length in words; a chunk is a segment of contiguous words
7	secpchk	Fluency	Delivery	Average chunk length in seconds
*8	wpsec	Fluency	Delivery	Articulation rate
9	Wpsecutt	Fluency	Delivery	Speaking rate
10	secpchkmeandev	Fluency	Delivery	Mean deviation of chunks in seconds
*11	wdpchkmeandev	Fluency	Delivery	Mean deviation of chunks in words
12	numsil	Fluency	Delivery	Number of silences
*13	silpwd	Fluency	Delivery	Duration of silences per word: total duration of silences divided by # of words
14	silpsec	Fluency	Delivery	Duration of silences per second: total duration of silences divided by total duration of response without disfluencies & pauses
*15	silmean	Fluency	Delivery	Mean of silence duration (in seconds)
16	silmeandev	Fluency	Delivery	Mean deviation of silences
17	longpfreq	Fluency	Delivery	Frequency of longer pauses ($\geqslant 0.5$ s)
*18	longpmn	Fluency	Delivery	Mean duration of long pauses
*19	longpwd	Fluency	Delivery	Frequency of longer pauses divided by number of words
20	longpmeandev	Fluency	Delivery	Mean deviation of long pauses
21	silstddev	Fluency	Delivery	Standard deviation of silence duration
22	longpstddev	Fluency	Delivery	Standard deviation of long pauses
23	numdff	Fluency	Delivery	Number of disfluencies (filled pauses)
24	dpsec	Fluency	Delivery	Disfluencies per second
25	repfreq	Fluency	Delivery	Number of repetitions divided by number of words
*26	tpsec	Fluency & Vocabulary diversity	Delivery & Language use	Types per second (types are unique words)
*27	tpsecutt	Fluency & Vocabulary diversity	Delivery & Language use	Types divided by uttsegdur
*28	amscore	Pronunciation	Delivery	Global HMM acoustic model score (normalized)
*29	lmscore	Grammatical accuracy	Language use	Global language model score (normalized)

Notes: (1) Mean deviation is computed as the mean of the absolute differences between feature values and the mean of all feature values.
(2) The terms "pauses" and "silences" are synonymous here.
(3) In all cases where the denominator would be zero (0.0), the respective value of a feature or component of a feature is also set to zero (0.0).

Table 7
Changes in correlation before and after the transformation.

Feature	Transformation performed	Correlations with human scores	
		Original	Transformed
Wdpchk	Natural log (wdpchk + 1)	0.106	0.222
Amscore	Inverse	−0.445	0.510
Lmscore	Inverse	−0.295	0.282
wdpchkmeandev	Inverse	0.097	−0.248

for automated scoring purposes [such as, e.g., in e-rater® (Attali and Burstein, 2005)]. The relationships between the features and the predicted scores are straightforward, as well as the relative weights of the features, and so it may be more perspicuous to an outsider than a CART model.

8.1. Multiple regression

Our aim in developing the SpeechRater multiple regression model was to produce a model with high agreement with human raters, but also to structure the model so that its use of our predictive features is in conformance with our understanding of the concept of communicative competence – the selected features should cover a wide range of aspects of communicative competence. Further, the feature weights should reflect the relative importance of different aspects of communicative competence and also should have a correct direction of association, i.e., if higher feature values correspond to higher proficiency, the feature weight has to be positive.

In a first step, we built a preliminary MR model using the 11 selected features described above. Then we determined a subset of these features in consultation with the content advisory committee (CAC) who also assisted in determining the feature-specific weights (see Eq. (1)), using the weights of the initial MR model as guidance. The CAC is a group of content-area specialists convened to ensure the appropriateness of our scoring model design for the concept of communicative competence.

We standardized the feature values (to zero mean and unit variance) so that the CAC weights assigned were comparable across all features. These standardization parameters (the mean and variance of the feature as observed in the training data) were retained for scaling of the features in the test samples as well.

Based on our initial MR model, the CAC agreed on the use of a model with the features *amscore*, *wpsec*, *tpsecutt*, *wdpchk*, and *lmscore*. This set of features was deemed to provide the widest range of coverage of the different aspects of the speaking construct, and could be weighted in such a way that the relative importance of each of these measures was represented.

To compute the final MR model, we wanted the feature weights to be fixed a priori to the values chosen by the CAC and so we restructured the standard regression equation, shown here (Eq. (1)):

$$Score = \sum_i \alpha_i f_i + \beta. \quad (1)$$

This original equation has a set of free parameters α_i associated with each scoring feature f_i. Our modified equation still has a parameter α'_i associated with each feature, but these parameters are not allowed to vary in the optimization of the model for a given training set. The only two parameters which need to be learned from the data are the slope parameter μ, and the intercept β (Eq. (2)):

$$Score = \mu \sum_i \alpha'_i f_i + \beta. \quad (2)$$

Table 8 provides the feature class and dimension represented by each of the features used, together with their weights in the regression model.

We fixed the weights α'_i of the standardized features to the CAC-defined values shown in Table 8, and trained the model using the TPO scoring model training data (**sm-train**). This involved setting the model slope parameter μ and intercept β to minimize the least-squares error on this training data.

As mentioned earlier, one shortcoming of just using the TPO **sm-eval** set for evaluation is that it does not allow a direct estimation of the correlation of predicted scores with human-assigned scores on a full complement of six tasks, which is the level of the score we wish to report (since there are six tasks in the TOEFL iBT Practice Speaking test). There were only 58 candidates with complete sets of six task scores in this evaluation set. To address this deficiency, we performed the evaluation run on the combined data from the TPO **sm-eval** and **rec-train** sets. This combined set of evaluation data contained many more (308) complete sets of six tasks per candidate than the **sm-eval** set alone.[8]

8.2. CART trees

CART 5.0 (Steinberg and Colla, 1995) was used to build the classification trees. We used all 11 selected features described above and explored different model configurations, i.e., different combinations of priors and splitting rules. For each combination, a 10-fold cross-validation was conducted. Subsequently, the optimal sub tree that was a relatively small tree with the highest or near-highest agreement with the human scores (weighted kappa) on the cross-validation sample was identified. Mixed priors (average of equal priors and training sample priors) with the Gini splitting rule[9] gave comparable weighted kappas with the human scores on the cross-validation sample, among all the combinations. Then the **sm-eval** sample cases were dropped down the best tree to obtain the classification rates. The best tree contained only five features, which were automatically selected by the CART learning procedure.

8.3. Model performance

In Table 9, the agreement results between MR and CART scoring models with human scores are broken down into four sections – groups of one, two, three and six task scores (grouped always from one examinee).

This table also includes results for the iBT Field Study data set (**smFS-eval**) as a comparison. This data set has a more even score distribution than TPO, but does have overlap in items between **smFS-train** and **smFS-eval**. Also, hypotheses were generated with a different recognizer due to a different audio format and no corpus-specific adaptation could be performed due to a lack of transcribed data.

8.3.1. MR analysis

While different measures for scoring quality are presented in Table 9, we choose the Pearson r correlation of the predicted scores with the human-assigned scores as our main evaluation metric. While exact and adjacent agreement measures can provide important information

[8] Strictly speaking, however, there is a methodological issue with doing the evaluation this way. Since the data from the **rec-train** set was used to train the speech recognizer, it is possible that some of the learning from this stage (relative probabilities of word sequences and pronunciation variants) might cause the scoring model to perform uncharacteristically on this particular set of data. In practice, however, this seems unlikely, given that our feature set abstracts away from the actual hypothesized word sequence returned by the recognizer. While the *lmscore* and *amscore* features do use information about the internal state of the recognizer, and therefore could be affected by the use of a particular response in recognizer training, we expect this effect to be small.

[9] The Gini splitting rule aims to get pure terminal nodes as soon as it can. It is the default splitting rule implemented in CART since it typically performs the best. However, given specific circumstances and data characteristics, other splitting rules can generate more accurate models.

Table 8
Features used in CAC regression model.

Feature	Weight	Feature class	Dimension
Amscore	4	Pronunciation	Delivery
Wpsec	2	Fluency	Delivery
Tpsecutt	2	Vocabulary, Fluency	Delivery & Language use
Wdpchk	1	Fluency	Delivery
Lmscore	1	Grammar	Language use

Table 9
CAC regression model and CART performances on TPO Evaluation + Recognizer train set, and iBT Field Study data set.

Sets of scores	Evaluation method	Multiple regression model (CAC weights)		CART model (mixed priors, Gini splitting)	
		TPO Eval + Recognizer training set	iBT Field Study test set	TPO Eval + Recognizer training set	iBT Field Study test set
Single score	Weighted κ	0.33	0.51	0.43	0.59
	Exact Agreement	57.8%	44.2%	50.5%	50.6%
	Exact + adjacent agreement	98.4%	95.1%	94.8%	93.3%
	Mean (SD) of predicted score	2.79 (0.37)	2.45 (0.61)	2.88 (0.80)	2.47 (0.92)
	Correlation (unrounded)	0.47	0.61	NA[a]	NA
	Correlation (rounded)	0.37	0.55	0.44	0.62
Pairs of scores	Weighted κ	0.45	0.58	0.49	0.66
	Correlation (unrounded)	0.53	0.66	NA	NA
	Correlation (rounded)	0.50	0.64	0.50	0.66
Triples of scores	Weighted κ	0.48	0.61	0.53	0.68
	Correlation (unrounded)	0.56	0.68	NA	NA
	Correlation (rounded)	0.54	0.67	0.54	0.68
Sets of six scores	Weighted κ	0.51	0.61	0.55	0.69
	Correlation (unrounded)	0.57	0.68	NA	NA
	Correlation (rounded)	0.57	0.68	0.57	0.70

[a] CART predicts only integer scores by design; therefore, no "unrounded" scores exist that a correlation can be computed on.

on the score overlap between automatic and human scores, it does not tell anything about the magnitude of errors outside of the agreement window. Furthermore, in cases with scales of very few score points such as in TPO (only four points), a high agreement is easily achievable. In particular, exact agreement is highly sensitive to the distribution of human ratings (the majority class baseline). If these are highly skewed, a high agreement figure might not reflect any sophistication on the part of the automated scoring system, but only a statistical artifact of the composition of the data.

In contrast, correlation is a more global metric which looks at distributional differences between all scores while ignoring consistent differences in the means between automatic and human scores, unlike for the exact agreement metric.

In our MR evaluations, Pearson r correlation ranges from 0.37 for single items to 0.57 for sets of six tasks for TPO and from 0.55 to 0.68 for the iBT Field Study. (We report correlations for rounded scores for better comparison with CART where all scores are integer in the first place.) As a comparison, the inter-human correlations are 0.55 for one item and 0.74 for six items for TPO data and 0.77 for one item and 0.94 for three items for the Field Study data.

Note also that there is less variation in SpeechRater's score estimates – the standard deviation of predicted scores is considerably lower than the standard deviation of human-assigned scores. This is partially due to the uneven distribution of task scores in the training data (with almost half the tasks receiving a score of 3), but may also have to do with inconsistency in the human scoring of these responses, and with the limited range of coverage of the concept of communicative competence in our feature set (emphasis on fluency features).

8.3.2. CART analysis

The optimal tree using the mixed priors and the Gini splitting rule contains features that partition the **sm-train** cases into different score classes at certain splitting values. Conceptually, the splitting features and values define the boundaries of different score classes. This is analogous to using responses that represent the lower and upper ends of a score class as range finders typically used in rater training.

Five features were present in the tree: *amscore, wpsec, wdpchk, silmean and lmscore.* The first one was a pronunciation feature (Delivery), the second through the fourth fluency features (Delivery) and the last one a grammar feature (Language Use). Not too surprisingly, we see a strong similarity between these CART features and those used in the multiple regression scoring model – *amscore, wpsec, wdpchk*, and *lmscore* occur in both models, whereas *tpsecutt* only occurs in MR and *silmean* only in CART.

Table 10
CART decision rules for different score classes (1–4).

Score class	Rule #	Rule
Score class 1	Rule 1	440.7 < Amscore ⩽ = 748.4, Longpmn > 1.02, Wdpchk ⩽ 4.6
	Rule 2	440.7 < Amscore ⩽ 748.4, Longpmn > 1.02, Wdpchk ⩾ 4.6, Silmean > 1.1
	Rule 3	Amscore > 440.7
Score class 2	Rule 1	392.9 < Amscore ⩽ 440.7, Wpsec ⩽ 2.8, Lmscore > 69.2
	Rule 2	440.7 < Amscore ⩽ 748.4, Longpmn ⩽ 1.02
	Rule 3	440.7 < Amscore ⩽ 748.4, Longpmn > 1.02, Wdpchk ⩾ 4.6, Silmean ⩽ 1.1
Score class 3	Rule 1	Amscore ⩽ 440.7, Wpsec ⩽ 2.8
	Rule 2	392.9 < Amscore ⩽ 440.7, Wpsec > 2.8, Lmscore ⩽ 69.2
Score class 4	Rule 1	Amscore ⩽ 392.9, Wpsec > 2.8

The decision rules that led to the terminal nodes are summarized above in Table 10.

These decision rules were considered by the CAC members to be reasonable. The different scoring rules for each score class were also deemed to be consistent with some of the typical profiles of students at a particular score level.

The CAC members also noted that some other features, such as vocabulary sophistication and precision features, and coherence and content relevance features, if available, may improve the accuracy in partitioning the cases into the right score classes.

8.4. Model selection

While the results we observed do not clearly favor either CART or MR, the final decision was made in favor of the MR model due to its greater simplicity, parsimony, and stability and (hence) easier defensibility towards the outside world. MR needs fewer training instances than CART for a stable model and is more easily modifiable in terms of adapting it to shifting means and variances due to changes in the test population.

9. Discussion

We have presented a system called SpeechRater for scoring non-native spontaneous speech in the context of an online English language testing program (TOEFL Practice Online Speaking). The system consists of three major components, the speech recognizer, the feature computation module, and the scoring model. The paper discusses each of these three stages of the system and provides motivations for the design decisions we made.

An important aspect of this work is that the features selected for the scoring model should not only have good statistical properties in terms of having high correlations with human rater scores, but at the same time should achieve as broad as possible a coverage of the concept of communicative competence. We see this work as a first step in a long-term process, where, over time, more features will be developed and added to the scoring model such that eventually all dimensions of communicative competence (delivery, language use, and topical development) are reasonably covered. One of the main challenges in this endeavor, as we move from low-level fluency features to higher-level features such as grammar, vocabulary, or content, is that those features rely on correct word identities to a much larger extent than the features we are using so far. This means that a substantial increase in the word accuracy of our recognizer will be mandatory for acceptable performance of these features and the overall system. We envision different ways of both acoustic and language model adaptation, e.g., adaptation to examinee responses to a particular task, which could aid in this effort of improving the recognizer's performance. Preliminary work on a related corpus with different methods of adaptation and other optimizations has already shown promising results.

Other challenges are much harder to address, e.g., the skewed score distribution in the TPO data set, which is likely one of the main reasons why both human agreement and automatic score predictions are worse than for the more evenly distributed iBT Field Study data set. We believe that the reason for the data skewedness is that speakers with a very high proficiency are likely aware of this and see no need in taking a practice test whereas speakers with a very low or moderate-low proficiency do not think that they can reasonably profit from such an exercise.

In terms of score prediction performance, the two models we evaluated, CART and MR yield comparable results overall. The gap between inter-human correlation and automatic performance in terms of Pearson r for all six items of a test is about 0.17 for the TPO data set and still higher for the iBT Field Study set. As this difference is of moderate magnitude for TPO data, Condition 2 of Section 1 was considered to be met. Condition 1, the appropriateness of the features in the scoring model in terms of covering the concept of communicative competence, was judged to be moderate, but adequate for a low-stakes practice test, such as the TPO, particularly in light of the known high correlations between the three dimensions of the TOEFL Speaking rubrics (delivery, language use, and topical development) (Xi and Mollaun, 2006). The five features of the

multiple regression scoring model cover mostly aspects of delivery (fluency, pronunciation), but also aspects of language use (grammatical accuracy and vocabulary diversity).

Future work will focus on developing additional features and improving the speech recognizer to close this performance gap between machine scores and human agreements.

10. Summary and future work

This paper has demonstrated the feasibility of automatic scoring of spontaneous non-native speech in a practice environment (TOEFL Practice Online Speaking), where the system used for scoring is composed of three main components, the speech recognizer, the feature computation module, and the scoring model. Two different corpora from related tests of English were used for system development and testing: the TOEFL Practice Online Speaking corpus and the TOEFL iBT Field Study corpus. From a set of 29 initial features, five were eventually used by our two scoring models, multiple regression and CART trees. (The feature sets were almost identical with an overlap of four of five features). Since we are reporting scores for overall performance of a complete test, results for grouping all six tasks of a form together were most interesting to us. The multiple regression correlations for TPO were 0.57 and for iBT around 0.68, compared to inter-human agreement rates of 0.74 and 0.94 (on three tasks), respectively. We decided to pick the multiple regression scoring model for the operational system since it is simpler and more perspicuous than CART tree models. The operational system has been online since the fall of 2006 and has scored tens of thousands of practice tests to date.

Future work will focus predominantly on two areas: (a) extending the feature set to allow for broader coverage of the concept of communicative competence; and (b) substantially improving the word accuracy of the speech recognizer by various methods of adaptation.

References

Attali, Y., Burstein, J., 2004. Automated essay scoring with e-rater V.2.0. Presented at the Annual Meeting of the International Association for Educational Assessment, Philadelphia, PA.

Attali, Y., Burstein, J., 2005. Automated essay scoring with e-rater V.2.0 (ETS RR-04-45). Educational Testing Service, Princeton, NJ.

Bachman, L.F., 1990. Fundamental Considerations in Language Testing. Oxford University Press, Oxford.

Bachman, L.F., Palmer, A., 1996. Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford University, Oxford.

Bernstein, J., 1999. PhonePass Testing: Structure and Construct. Ordinate Corporation, Menlo Park, CA.

Bernstein, J., DeJong, J., Pisoni, D., Townshend, B., 2000. Two experiments in automatic scoring of spoken language proficiency. In: Proceedings of InSTILL2000, Dundee, Scotland.

Brieman, L., Jerome, F., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth, Pacific Grove.

Burstein, J., Kukich, K., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Lu, C., Nolan, J., Rock, D., Wolff, S., 1998. Computer analysis of essay content for automated score prediction: a prototype automated scoring system for GMAT analytical writing assessment (ETS RR-98-15). Educational Testing Service, Princeton, NJ.

Chodorow, M., Burstein, J., 2004. Beyond essay length: evaluating e-rater's performance on TOEFL essays (TOEFL Research Report No. RR-73, ETS RR-04-04). Educational Testing Service, Princeton, NJ.

Clauser, B.E., Margolis, M.J., Clyman, S.G., Ross, L.P., 1997. Development of automated scoring algorithms for complex performance assessments: a comparison of two approaches. Journal of Educational Measurement 34, 141–161.

Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70, 213–220.

Cucchiarini, C., Strik, H., Boves, L., 1997a. Automatic evaluation of Dutch pronunciation by using speech recognition technology. Paper Presented at the 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. Paper Presented at the Third International Symposium on the Acquisition of Second Language Speech: NEW SOUNDS 97, Klagenfurt, Austria.

Cucchiarini, C., Strik, H., Boves, L., 2000a. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. Speech Communication 30 (2–3), 109–119.

Cucchiarini, C., Strik, H., Boves, L., 2000b. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. Journal of the Acoustical Society of America 107, 989–999.

Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. Journal of the Acoustical Society of America 111 (6), 2862–2873.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., 2000a. The SRI EduSpeak system: recognition and pronunciation scoring for language learning. In: Proceedings of InSTiLL-2000 (Intelligent Speech Technology in Language Learning), Dundee, Scotland.

Franco, H., Neumeyer, L., Digalakis, V., Ronen, O., 2000b. Combination of machine scores for automatic grading of pronunciation quality. Speech Communication 30, 121–130.

Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review 104, 211–240.

Leacock, C., 2004. Scoring free-responses automatically: a case study of a large-scale assessment. Examens 1 (3).

Leacock, C., Chodorow, M., 2003. C-rater: scoring of short-answer questions. Computers and the Humanities 37 (4), 389–405.

Linguistic Data Consortium (LDC), 1997. HUB-4 Broadcast News corpus (English).

North, B., 2000. The Development of a Common Framework Scale of Language Proficiency. Peter Lang, New York, NY.

Page, E.B., 1966. The imminence of grading essays by computer. Phi Delta Kappan 47, 238–243.

Rudner, L.M., Garcia, V., Welch, C., 2006. An evaluation of the Intellimetric essay scoring system. Journal of Technology, Learning and Assessment 4 (4). Retrieved from <http://www.jtla.org>.

Steinberg, D., Colla, P., 1995. CART: Tree-Structured Non-Parametric Data Analysis. Salford Systems, San Diego, CA.

Williamson, D.M., Bejar, I.I., Hone, A.S., 1999. "Mental model" comparison of automated and human scoring. Journal of Educational Measurement 36, 158–184.

Xi, X., Mollaun, P., 2006. Investigating the Utility of Analytic Scoring for the TOEFL® Academic Speaking Test (TAST). TOEFL iBT Research Report No. TOEFLiBT-01.

Zechner, K., Bejar, I., 2006. Towards automatic scoring of non-native spontaneous speech. In: Proceedings of the 2006 Conference on Human Language Technology and the North-American Association for Computational Linguistics (HLT-NAACL-06), New York, NY.