Measuring the Power of Learning.™

Research Report
ETS RR–18-10

Automated Scoring of Nonnative Speech Using the *SpeechRater*^{SM} v. 5.0 Engine

Lei Chen
Klaus Zechner
Su-Youn Yoon
Keelan Evanini
Xinhao Wang
Anastassia Loukina
Jidong Tao
Lawrence Davis
Chong Min Lee
Min Ma
Robert Mundkowsky
Chi Lu
Chee Wee Leong
Binod Gyawali

December 2018

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Automated Scoring of Nonnative Speech Using the *SpeechRater*^{SM} v. 5.0 Engine

Lei Chen, Klaus Zechner, Su-Youn Yoon, Keelan Evanini, Xinhao Wang, Anastassia Loukina, Jidong Tao, Lawrence Davis, Chong Min Lee, Min Ma, Robert Mundkowsky, Chi Lu, Chee Wee Leong, & Binod Gyawali

Educational Testing Service, Princeton, NJ

This research report provides an overview of the R&D efforts at Educational Testing Service related to its capability for automated scoring of nonnative spontaneous speech with the *SpeechRater*^{SM} automated scoring service since its initial version was deployed in 2006. While most aspects of this R&D work have been published in various venues in recent years, no comprehensive account of the current state of SpeechRater has been provided since the initial publications following its first operational use in 2006. After a brief review of recent related work by other institutions, we summarize the main features and feature classes that have been developed and introduced into SpeechRater in the past 10 years, including features measuring aspects of pronunciation, prosody, vocabulary, grammar, content, and discourse. Furthermore, new types of filtering models for flagging nonscorable spoken responses are described, as is our new hybrid way of building linear regression scoring models with improved feature selection. Finally, empirical results for SpeechRater 5.0 (operationally deployed in 2016) are provided.

Keywords Automated speech scoring; automated speech recognition; scoring models; natural language processing; English assessments

doi:10.1002/ets2.12198

Automated scoring of speech can be seen as a task of artificial intelligence, whereby a computer system assigns a speaking proficiency score to a digitized spoken response produced in a language assessment by a test taker who is not a native speaker of the language being assessed or is still a learner. Essentially, the task involves generating a mapping function from the speech signal to a speaking proficiency score, whereby scores usually generated by human raters are used as the gold standard to train the system.

Most automated speech scoring systems contain three main components: an automatic speech recognition (ASR) system that generates word hypotheses for a given speech sample along with other information, such as the duration of pauses between words; a set of modules based on digital signal processing and natural language processing (NLP) technologies that compute a number of features measuring various aspects of speech considered relevant by language assessment experts (e.g., fluency, pronunciation, grammatical accuracy); and finally, a scoring model that maps features to a score using a supervised machine learning paradigm.

Numerous significant challenges need to be addressed in automated speech scoring, including but not limited to a large variation in the speech input characteristics owing to variations in speech proficiency as well as native language (resulting in substantially higher ASR word error rates [WERs] than what is observed for native speech) or issues with audio quality related to audio capture, transmission, and environmental effects such as noise or background talk.

In addition to these technical challenges due to data characteristics and ASR performance, it is also not easy to compute measures of speech that both assess valid aspects of speaking proficiency and exhibit a reasonably high empirical performance when used for speech scoring. Some aspects of spoken proficiency that are highly valued by human raters, such as content appropriateness and organization of discourse, are very hard to capture with current ASR and NLP technology; a similar issue is also a challenge for the automated scoring of essays (Quinlan, Higgins, & Wolff, 2009). Finally, in terms of building a high-performing scoring model that combines the features or measures of speaking proficiency, several aspects in addition to empirical performance need to be considered when following best practices of the educational measurement field. These include model interpretability (this usually means a preference toward linear models),

Corresponding author: K. Evanini, E-mail: kevanini@ets.org

feature independence (minimizing collinearities or feature intercorrelations), fairness of the assessment (the automated scoring system having similar performance across different groups of test takers), and construct relevance (maximizing the overlap between features and aspects of the speaking construct that should be considered in a scoring model as defined by assessment experts). The section Hybrid Method of Feature Selection in this report addresses the latter challenge.

Research into the automated scoring of nonnative speech at Educational Testing Service (ETS) started in the early 2000s and resulted in ETS's automated speech scoring capability, the *SpeechRater*SM automated scoring service, whose scores have been operationally deployed as the sole scores for the *TPO*™ practice test since 2006 (Zechner, Higgins, Xi, & Williamson, 2009). This system was the first capability ever used in operation for scoring open-ended, spontaneous nonnative English, compared to other systems developed around the same time, as well as in the 1990s, which focused on the automated scoring of highly predictable speech, such as passages read aloud or sentences presented as audio stimuli and repeated aloud (Bernstein, Moere, & Cheng, 2010).

In the intervening decade, research and development related to the capability of automated scoring of nonnative speech at ETS has been substantially expanded, with the main focus on adding features to SpeechRater that can measure aspects of the speaking proficiency construct that were not previously addressed, for example, features related to vocabulary diversity or grammatical complexity.

While SpeechRater was initially geared only toward scoring nonnative spontaneous speech, over the years, it has also been used for speaking tasks that are more restricted, for example, in the *TEFT*™ assessment, which contains highly predictable as well as moderately predictable tasks (Zechner et al., 2015). For these task types, additional features that can measure, for example, the accuracy of reading or repeating a sentence or a passage were developed.

Furthermore, the process of building scoring models has been for the most part automatized and streamlined, allowing for a substantially faster turnaround time when building new scoring models and also allowing for exploring a large number of model alternatives in a short amount of time. Recent research into using particular versions of linear regression models that perform shrinkage of feature vector dimensions and preserve construct coverage commensurate with that of human experts has further allowed us to increase empirical model performance while maintaining a high degree of model interpretability, an aspect of high importance in the educational measurement field (Loukina, Zechner, Chen, & Heilman, 2015).

Last, but not least, SpeechRater has grown substantially in functionality, complexity, and size over the years. Two major rounds of code refactoring were undertaken with the goal of increasing code maintainability and allowing for easier updates and changes to the code by multiple research scientists and research engineers engaged with the work of automated speech scoring.

The aim of this research report is first to provide, in the following section, an overview of recent developments in the area of automated speech scoring in the last decade, then to succinctly summarize and describe the most important features and feature classes that have been developed and added to SpeechRater in recent years (in the Innovations in Scoring Features section); to describe the current SpeechRater system, including aspects of its refactored code structure (in the SpeechRater 5.0 section); and finally, in the Applying SpeechRater 5.0 section, to provide a case study of the current scoring model building process using SpeechRater for TPO.

Figure 1 provides an overview of the major data-processing steps inside SpeechRater. In addition, we also present a concise summary of the research activities enabling these steps in the left column and applications in the right column.

Previous Research

Speech Scoring Research in a Nutshell

Because automated speech scoring technology has widespread applications in the language testing and education fields, many research groups and companies around the world have been conducting research or developing commercial products. There are already comprehensive surveys on this topic. For example, a special issue related to this topic was published in the journal *Speech Communication* in 2009 (Eskenazi, Alwan, & Strik, 2009). Eskenazi (2009) summarized various research endeavors in speech scoring and computer-assisted pronunciation training (CAPT).

In this section, we provide information about several academic research groups and companies that have been active in speech scoring research and in industry to provide an up-to-date snapshot of the landscape of the entire domain. Note

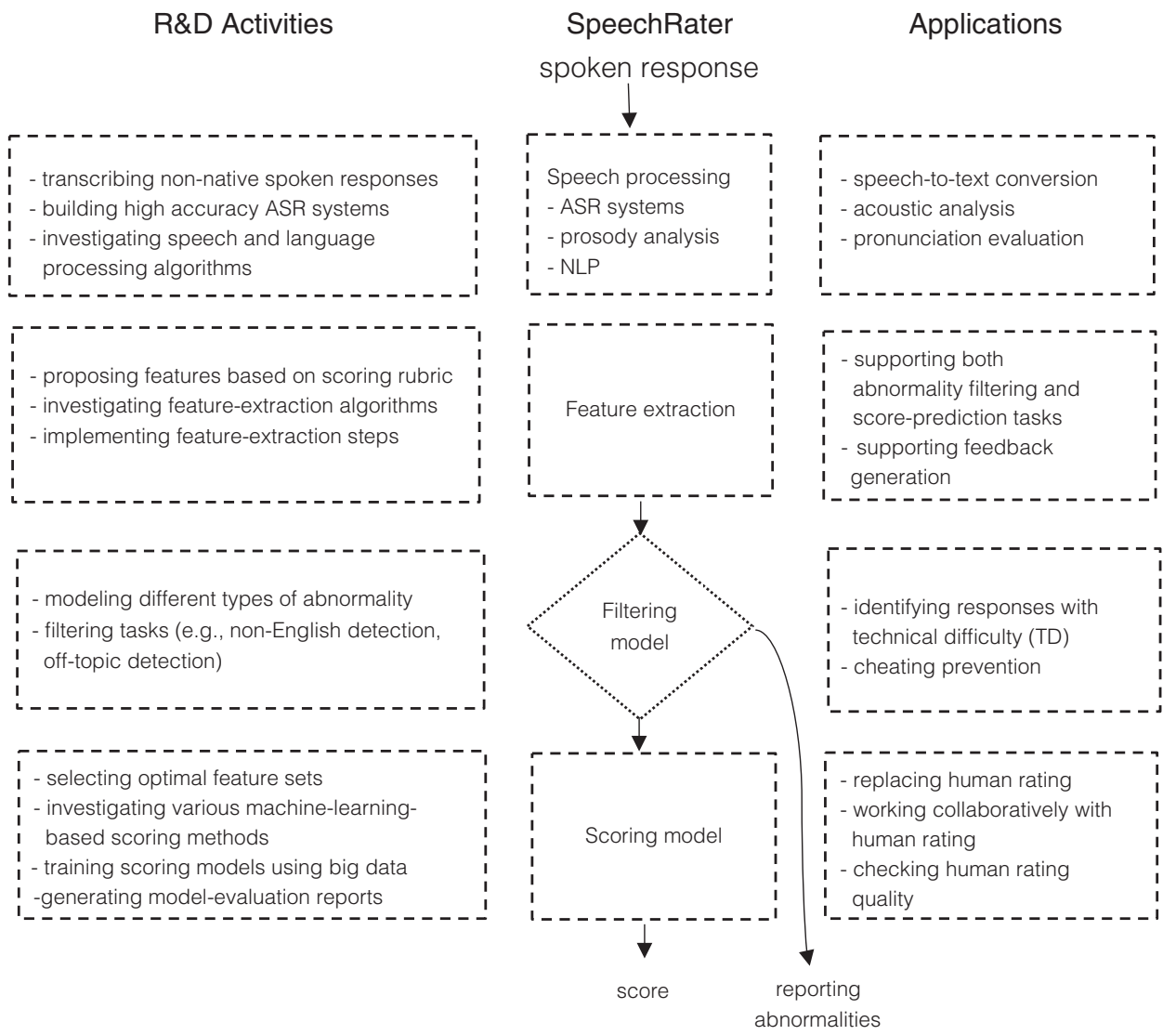

Figure 1 An overview diagram showing the major data-processing steps within SpeechRater (in the center column), the associated research activities (in the left column), and potential applications (in the right column).

that the brief survey presented in this section should be treated as a concise summary – in a nutshell – rather than as a comprehensive literature review.

First, we introduce several research groups in academia that have been consistently conducting research on speech scoring. The Center for Language and Speech Technology (CLST) at Radboud University in the Netherlands is well known for its long-term research into utilizing speech recognition technology to help language learning. For example, several papers were published in speech scoring, and these papers (e.g., Cucchiarini, Strik, & Boves, 2000; Strik & Cucchiarini, 1999; Strik, Truong, de Wet, & Cucchiarini, 2009) have received a considerable number of citations by other scholars. In recent years, CLST has studied utilizing its pronunciation training technology within a Web-based coaching system (Cucchiarini, Nejjari, & Strik, 2014) and has utilized a serious game setup for language learning (Strik, Palumbo, & de Wet, 2015).

The Institute for Automated Language Teaching and Assessment (ALTA) is another research agency that develops automated assessment technology. It is directly funded by Cambridge English Language Assessment with the goal of improving research in enhancing the linguistic proficiency of English learners and developing technologies to automatically rate both writing and speaking skills. In recent years, an increasing number of publications from ALTA have reported several advances in both essay and speaking scoring. For example, van Dalen, Knill, and Gales (2015) proposed using a Gaussian process model to score speaking proficiency.

Globally, utilizing speech technology to assist language learning has become a hot research topic. For example, the laboratory led by Dr. Meng at the Chinese University of Hong Kong has been actively working on this topic. In 2012, her group published the first paper to use deep neural network (DNN) acoustic models (AMs) for improving mispronunciation detection (Qian, Meng, & Soong, 2012).

We now introduce several industry agencies working on speech scoring and CAPT technologies. Besides ETS, several companies around the world have been carrying out research and product development related to automated speech scoring. Carnegie Speech provides English pronunciation training tools based on speech recognition technologies, such as Carnegie Speech Assessment and NativeAccent. The former is a solution for scoring general English speaking skills, whereas the latter focuses on providing pronunciation training for global enterprises and individual users.

SRI's eduSpeak SDK (Franco et al., 2010) provides a special speech recognition system for computer-based language learning and training applications. It can be utilized for building products for foreign language learning, English as a second language (ESL) training, reading comprehension, and interactive tutoring systems.

Another major player in the speech scoring arena is Pearson's speech scoring team, under its Knowledge Technology division. This group can be traced back to the Versant company before Pearson acquired it. Versant developed an automated speaking assessment deployed on the telephony platform; more details about the Versant speaking assessment can be found in Bernstein, Moere, et al. (2010). In recent years, Pearson's speech scoring team has improved their technical capabilities and applied automated speech scoring within several major Pearson English test products, for example, the Arizona English Language Learner Assessment, which is the domestic English language learning assessment for the state of Arizona (Metallinou & Cheng, 2014). Pearson embraced new high-accuracy DNN ASR technology in its speech scoring research with significantly increased recognition accuracy on nonnative speech. More details on Pearson's work utilizing a DNN ASR can be found in Cheng, Chen, and Metallinou (2015).

Leveraging its technical assets in advanced ASR and text-to-speech technologies, Microsoft Research Asia conducted research on pronunciation evaluation. Hu and colleagues (Hu, Qian, & Soong, 2014; Hu, Qian, Soong, & Wang, 2015) reported their work introducing the newest DNN ASR technology to provide higher quality pronunciation evaluation. This pronunciation evaluation technology has been used as an important learning module inside Microsoft's Bing Dictionary app.

Last, we mention several start-up companies that have been working on bringing speech scoring technology to new disruptive language learning and testing products on the mobile platform. Duolingo is a start-up company originating from CMU. It provides free and paid language learning services through its online and app-based platforms. Recently, on its Test Center platform, Duolingo pushed out the Duolingo English Test (DET), a computer-adaptive test of general English language ability with a testing fee of US$50. The Test Center platform is based on mobile devices and can utilize different sensors, such as the microphone and camera, to deliver advanced testing items and to conduct test proctoring. Using adaptive testing, the total duration of a DET test can be reduced by about half compared to the nonadaptive test, and it generally lasts approximately 10 – 25 minutes. The testing results are returned within 48 hours, and a median wait time is 18.5 hours. Wagner and Kunnan (2015) is an independent and rigorous validity study. It criticizes DET for its high-profile nature and the fact that there is little publicly available material about its assessment development or the model of language ability being measured. Similarly, Liulishuo is a start-up company in Shanghai, China. It provides a mobile app on both iOS and Android platforms to allow English learners (mostly from China) to use their spare time to improve English skills. ASR-based pronunciation training and feedback have been utilized in Liulishuo's English practice module.

From the preceding brief survey, we have the following observations. Developing more advanced automated speech scoring technology is a hot research area, in which many research groups around the world are active. Second, new technology, such as DNN-based ASR, has been introduced into the speech scoring research area very quickly. Last, new trends have become clear in this area. From the point of view of learning, game-based and mobile-based products (for utilizing learners' spare time) have received more attention. Also, a unified platform holding learning and testing in one place (mostly on a mobile platform) is an interesting new development.

A Summary of the Improvements Made for SpeechRater

Within ETS, automated speech scoring R&D work can be dated back to 2002. SpeechRater 1.0, released in 2006, was the first automated scoring system for spontaneous nonnative English spoken responses. Since then, several versions of the

SpeechRater system have been built to act as the sole scorer for the TPO test. Automated scoring technology allows the TPO test to provide consistent and timely scoring that is very close to the performance level of human rating.

Xi, Higgins, and Zechner (2008) summarized the SpeechRater 1.0 system. Since the publication of that research report, there have been many significant improvements. Prior to describing these improvements in the following sections, we list major improvements here to provide readers with a high-level overview of this report:

- *More construct-related features*. SpeechRater 1.0 provided the features mostly covering the fluency aspect of the *TOEFL*® test scoring rubric. Since then, new features have been developed to provide more comprehensive construct coverage, especially on pronunciation, rhythm, and high-level linguistic skills, for example, vocabulary and grammar; the section Innovations in Scoring Features in this report provides more detail.
- *More accurate automatic speech recognition technology*. SpeechRater 1.0 used an ASR system with a relatively high error rate; in contrast, a new ASR system with much improved recognition accuracy was introduced in SpeechRater 5.0. More details about the updates to the ASR can be found in the section Automatic Speech Recognition System.
- *More accurate and comprehensive abnormality detection*. When running an assessment, some abnormal spoken responses may show up for various reasons. For example, audio files may be missing due to technical issues throughout the entire audio recording, transferring, and storage pipeline. Test takers may provide off-topic responses or even memorized responses to intentionally cheat the scoring system. Therefore abnormality detection is important for maintaining the entire scoring system's accuracy and validity. However, only limited abnormality detection was utilized in SpeechRater 1.0. In contrast, SpeechRater 5.0 can detect more types of abnormalities with increased detection performance. More details on these updates can be found in the section Filtering Models.
- *Improved code implementation*. In recent years, consistent efforts have been made to systematically update the SpeechRater code repository. These efforts have improved code quality, and the updated code repository of SpeechRater 5.0 is easy to maintain and expand.
- *More advanced model building*. The scoring model plays an important role in the entire SpeechRater system. Therefore the methods and tools for building accurate scoring models meeting various psychometric criteria have become a focus of our team's research. As a result, many major improvements have been made in this direction. For example, in SpeechRater 1.0, feature selection and regression model parameters were determined manually by experts. This has been updated to a more efficient and accurate hybrid model-building process that jointly utilizes both a data-driven approach and human expertise. More details on these updates can be found in the section Hybrid Method of Feature Selection.
- *More data*. We have been utilizing the power embedded in ample-sized data sets. Compared to the data sets used in SpeechRater 1.0, a larger data set that was double-scored completely was used in SpeechRater 5.0. In particular, to train the scoring model, a sizable number of prescored responses is required, and test takers' profiles are expected to be close to what will be met in real tests. Therefore, instead of using the data set collected in 2006 for building SpeechRater 1.0, a larger and more recent set of TPO responses has been used for building scoring models. These responses were obtained in 2012, and double human ratings were provided by trained raters. This data set contains a total of 6,000 TPO responses from 1,000 test takers.

Innovations in Scoring Features

In this section, we introduce several groups of new features that have been investigated and proposed in recent years. These features help to increase the construct coverage compared to what we achieved in the SpeechRater 1.0 system, as described in Xi et al. (2008). Each group of new features is discussed in a separate subsection following a standard structure to ease the reader's understanding. The presentation structure contains (a) introduction and related work, (b) methods (describing the procedures of how to extract this group of features), (c) evaluation (recapping the experimental results in corresponding publications), and (d) optional discussion.

Note that we report on the evaluation results of the features in the following manner: The evaluation section recaps the findings with respect to these new features in their original publications; later, a more recent systematic evaluation using the tools and data sets for developing SpeechRater 5.0 is reported in the section Feature Correlations.

The main metric we use to evaluate feature performance is correlations with human holistic scores. While this is a good first indicator in terms of how "useful" a certain feature may be to measure spoken proficiency, additional, more detailed

evaluations of features, for example, comparing their internal representations and derived values with human annotations, may yield deeper insights into the strengths and weaknesses of particular speech features. Since holistic scores are evaluating a large set of areas related to speaking proficiency (such as fluency, intonation, vocabulary complexity, grammatical accuracy, content, and discourse) but individual features only measure a small aspect of the speaking construct, ideally, human scores based only on the same narrow aspect of speaking proficiency would be obtained and also used for feature evaluations. However, in practice, it is quite difficult for human raters to focus their scoring on a very narrow aspect of speech, and additionally, it turns out that oftentimes the measurement of various aspects of spoken proficiency are highly correlated with each other, resulting in little additional information obtained from this approach (Xi, 2007).

Utilizing Structural Events

Introduction and Related Work

Disfluencies have been considered an important key to understanding the sentence planning process, and researchers in psycholinguistic and second language acquisition (SLA) have actively investigated characteristics of disfluencies to understand L1 and L2 speakers' sentence planning. Boomer (1965) and Bock and Levelt (1994) found that disfluencies can be classified into groups according to their locations within utterances. There are two groups, and each group has a different function. Disfluencies that occurred at clause boundaries (hereinafter *boundary disfluencies*) serve as sentence planning time, whereas disfluencies that occurred within clauses (hereinafter *within-clause disfluencies*) occur when speakers have problems in sentence generation, such as failures in lexical retrieval.

In ESL research, Temple (2000) found a strong relationship between within-clause disfluencies and L2 speaker proficiency. Compared to L1 speakers, L2 speakers have reduced lexical and syntactic knowledge and must consciously control speech because it does not come automatically. Because of these issues, speakers with low proficiency have more problems during sentence generation, resulting in more frequent within-clause pauses than speakers with higher proficiency show. More recently, Mizera (2006) showed that the frequency of within-clause disfluencies is more strongly correlated with human proficiency scores than is the frequency of all disfluencies, including both within-clause and boundary disfluencies. A combination of utterance structure and disfluency profile can more accurately estimate speakers' proficiency levels.

Furthermore, ESL researchers have developed various quantitative measures based on the frequency and distribution of disfluencies and have used them in estimating L2 learners' oral proficiency. For instance, Lennon (1990) and Riggenbach (1991) found strong correlations between proficiency levels and features such as filled pauses per sentence and percentage of T-units followed by pauses. These features have been used since the beginning of automated speech scoring systems (e.g., Cucchiarini, Strik, & Boves, 2002; Zechner et al., 2009). In particular, Zechner et al. extracted these disfluency features in a fully automated manner. Given a spoken response, the automated scoring system calculated the frequency and duration of disfluencies (e.g., "uh," "um," and silent pauses) from the transcription created by a speech recognition system and generated multiple disfluency-related features. However, the features used in Zechner et al. were limited to relatively simple features; the study did not cover important aspects such as sentence structure. To address this gap, we developed a new set of features based on both disfluency profile and utterance structure.

Methods

First, we detected a set of structural events (SEs) from spoken responses. The SEs included two different types of events: clause boundaries and disfluencies. The disfluencies were further classified into edit disfluencies, silent pauses, and fillers. Silent pauses and fillers were detected from ASR output directly, while clause boundaries and edit disfluencies were detected using an automated model trained on the similar body of nonnative speakers' spoken responses (a total of 660 responses) with manual SE annotations. The model was based on a maximum entropy model using both lexical features (word bigrams, part of speech [POS] tag bigrams) and pause features. It was evaluated using the held-out annotation data comprising 330 spoken responses. The model achieved an F-score of .60 in clause boundary detection and an F-score of .30 in edit disfluency detection. The low accuracy in disfluency detection was expected because systems based on state-of-the-art speech technology, such as Liu (2004), achieved low accuracy even on native speakers' speech. The details of the experiments can be found in L. Chen and Yoon (2011, 2012).

Next, we generated the disfluency features using ASR output and the output of the automated SE detection model. All disfluencies were classified into two groups: within-clause pause/disfluency or clause-boundary pause/disfluency. Finally, the following five features were calculated: (a) *ipcount*, the number of edit disfluencies; (b) *clausecount*, the number of clause boundaries detected by the automated detection system; (c) *IPC*, the number of disfluencies per clause; (d) *IPW*, the number of disfluencies per word; and (e) *longSilRatio*, the proportion of long within-clause silence to all within-clause silence.

Evaluation

As reported in L. Chen and Yoon (2011), several features described in this section had significant correlations with holistic human scores. The best performing feature was longSilRatio, and the Pearson correlation coefficient with human scores was −.36. In the section Feature Correlations, the evaluation results of this group of features in SpeechRater 5.0 are presented.

Discussion

In L. Chen and Yoon (2011), SE detection using speech transcriptions shows generally good performance, and the features derived from the detected SEs show promising usefulness for predicting speaking proficiency levels. However, when using the noisy ASR output directly, as shown in L. Chen and Yoon (2012), disfluency detection was negatively impacted more than clause boundary detection was. Consequently, the features related to clause boundary and silent pauses show small reductions in correlations with human-judged scores, for example, longSilRatio.

Improved Pronunciation Features Measuring Spontaneous Nonnative Speech

Introduction and Related Work

Pronunciation measurement is one of the most important subtasks in the automated speech scoring task. A seminal approach, goodness of pronunciation, was proposed by Witt (1999) for measuring read-aloud pronunciation based on hidden Markov model (HMM) AM log likelihood. This approach and its extended version have been widely used in measuring read-aloud pronunciation (e.g., Cucchiarini, Strik, & Boves, 1997; Franco et al., 2000; Hacker et al., 2005; Neumeyer, Franco, Digalakis, & Weintraub, 2000). Moustroufas and Digalakis (2007) supported language teachers in using their own reading texts rather than the predefined ones. This improvement gives language teachers more freedom to use any reading text that helps students' learning.

A limited number of studies have been conducted on assessing speaking proficiency based on spontaneous speech. Zechner and Bejar (2006) presented a system to score nonnative spontaneous speech using features derived from the recognition results. The same technology was used in SpeechRater 1.0 for assessing pronunciation quality. However, there are some issues with the method of extracting pronunciation features in the previous research (Zechner & Bejar, 2006; Zechner, Higgins, & Xi, 2007). For example, the AM that was used to estimate the likelihood of a phoneme being spoken matches the acoustic properties of nonnative speech. However, for such measurements, an AM trained on native speech data needs to be utilized for more accurate and objective computation. Furthermore, other important aspects of pronunciation, for example, vowel duration, were not utilized as a feature in SpeechRater 1.0. In addition, likelihoods estimated on nonwords (such as silences and fillers) that were not central to the measurement of pronunciation were used in the feature extraction.

Method

Figure 2 depicts our new method for extracting an expanded set of pronunciation features in a more meaningful way.

We used two different AMs for pronunciation feature extraction. First, we used an AM optimized for speech recognition (typically an AM adapted to nonnative speech so that it better fits nonnative speakers' acoustic patterns) to generate word hypotheses; then we used another AM optimized for pronunciation scoring (typically trained on native or near-native speech so that it is a good reference model reflecting the expected speech characteristics) to force-align the speech signal to the word hypotheses and to compute the likelihoods of individual words being spoken and the durations of phonemes; finally, new pronunciation features were extracted based on these measurements.

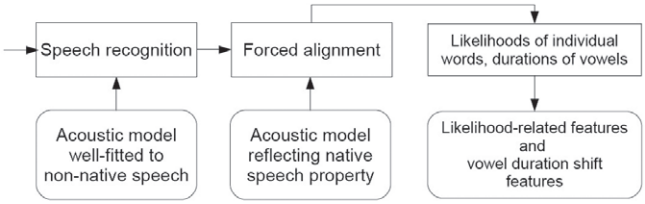

Figure 2 Two-stage pronunciation feature extraction.

Table 1 Notation Used for Pronunciation Feature Extraction

Variable	Meaning
$L(x_i)$	The likelihood of word x_i being spoken given the observed audio signal
t_i	The duration of word i in a response
T_s	The duration of the entire response
T	$\sum_{i=1}^{n} t_i$, the summation of the duration of all words, where $T \leq T_s$
n	The number of words in a response
m	The number of letters in a response
R	$\frac{m}{T_s}$, the frequency of letters (as the rate of speech)
v_i	Vowel i
N_v	The total number of vowels
P_{v_i}	The duration of vowel v_i
$\overline{P}$	The average vowel duration (across all vowels in the response being scored)
D_{v_i}	The standard average duration of vowel v_i (estimated on a native speech corpus)
$\overline{D}$	The average vowel duration (of all vowels in a native speech corpus)
S_{v_i}	$\mid P_{v_i} - D_{v_i} \mid$, the vowel duration deviation v_i (measured as the absolute value of the difference between the duration of vowel v_i and its standard value)
Sn_{v_i}	$\mid \frac{P_{v_i}}{\overline{P}} - \frac{D_{v_i}}{\overline{D}} \mid$, the normalized vowel duration deviation v_i (measured as the absolute value of the normalized difference between the duration of vowel v_i and its standard value)

Some notation used for computing the pronunciation features are listed in Table 1. On the basis of this notation, the proposed new pronunciation features are described in Table 2. To address the limitations of previous research on automated assessment of pronunciation, our proposed method has achieved the following improvements: (a) using the two-stage method to compute HMM likelihoods using a reference AM trained on native and near-native speech, (b) expanding the coverage of pronunciation features by using vowel duration deviations from native speakers' norms, and (c) using likelihoods on the audio portions that are recognized as words and applying various normalizations.

Evaluation

In the study described by L. Chen, Zechner, and Xi (2009), two AMs were created using the speech recognizer, utilized in SpeechRater 1.0, which is a gender-independent fully continuous HMM recognizer. The AM used in the recognition was trained on approximately 30 hours of nonnative speech from the TPO test. For language model (LM) training, a large corpus of nonnative speech (approximately 100 hours) was used and mixed with a large general-domain LM (trained from the Broadcast News corpus of the Linguistic Data Consortium [LDC]; Graff, Garofolo, Fiscus, Fisher, & Pallett, 1997). In the pronunciation feature extraction process depicted in Figure 2, this AM was used to recognize nonnative speech to generate the word hypotheses. The AM used in the forced alignment was trained on native speech and high-scoring non-native speech. It was trained as follows: Starting from a generic recognizer, which was trained on a large and varied native speech corpus, we adapted the AM using batch-mode MAP adaptation. The adaptation corpus contained approximately 2,000 responses with high scores in previous TPO tests and spoken responses to TOEFL questions collected from native speakers. In addition, this AM was used to estimate standard norms of vowels as described in Table 1.

Table 2 A List of Proposed Pronunciation Features

Feature	Formula	Meaning
L_1	$\sum_{i=1}^{n} L\left(x_i\right)$	Summation of likelihoods of all the individual words
L_2	L_1/n	Average likelihood across all words
L_3	L_1/m	Average likelihood across all letters
L_4	L_1/T	Average likelihood per second
L_5	$\frac{\sum_{i=1}^{n} \frac{L(x_i)}{t_i}}{n}$	Average likelihood density across all words
L_6	L_4/R	L_4 normalized by the rate of speech
L_7	L_5/R	L_5 normalized by the rate of speech
$\overline{S}$	$\frac{\sum_{i=1}^{N_v} S_{v_i}}{N_v}$	Average vowel duration deviations
$\overline{Sn}$	$\frac{\sum_{i=1}^{N_v} Sn_{v_i}}{N_v}$	Average normalized vowel duration deviations

The evaluation result in L. Chen et al. (2009) showed that new features provide promising measurement of pronunciation. The new pronunciation features, that is, L_6 and L_7, have $|r|$ values ranging around .44. In addition, $\overline{Sn}$, a new feature representing the vowel production aspect of pronunciation, shows a relatively high correlation with human holistic scores. This suggests that our new pronunciation feature set has an expanded coverage of pronunciation. Regarding the comprehensive evaluation results when using the updated, more accurate ASR in SpeechRater 5.0, please see the section Feature Correlations.

Discussion

When developing accurate and valid automated scoring systems, in addition to considering features that are highly correlated with human-rated scores, we need to pay attention to the features' construct relevance. This belief has guided us in designing this new group of features that measure pronunciation. The method of using one AM optimized for speech recognition and another AM optimized for pronunciation evaluation is well motivated theoretically (Witt, 1999; Xi et al., 2008). The results support the linkage of the features to the construct of pronunciation and their utility for use in a scoring model to predict human holistic judgments.

Rhythm Features

Introduction and Related Work

Several studies have investigated whether different languages can be classified into different groups (typically stress timed vs. syllable timed) based on rhythmic properties, such as variability of segmental and syllabic durations in an utterance (Dellwo, 2006; Grabe & Low, 2002; Ramus, Nespor, & Mehler, 2000). While more recent experimental studies suggest that there are not clear categorical differences between languages, there is evidence that different languages have different characteristic patterns (Loukina, Kochanski, Rosner, & Keane, 2011; White, Mattys, & Wiget, 2012). These findings have motivated investigations into rhythmic differences between native speech and nonnative speech under the hypothesis that rhythmic patterns from a speaker's L1 may carry over into his or her L2 speech. Several studies along these lines have found rhythmic differences between native English speech and L2 English produced by speakers from a variety of L1 backgrounds, including Spanish and Dutch (White & Mattys, 2007), Cantonese and Mandarin (Mok & Dellwo, 2008), French (Tortel & Hirst, 2010), and Japanese (Tepperman, Stanley, Hacioglu, & Pellom, 2010). In addition, some research studies have employed rhythm metrics to score the English speaking proficiency of nonnative speakers. Most of these studies have investigated read-aloud speech produced by speakers from a uniform L1 background, including Korean (Jang, 2009), Spanish (Nava, Tepperman, Goldstein, Zubizarreta, & Narayanan, 2009), and Mandarin (L. Chen & Zechner,

Table 3 Summary of Rhythm Metrics

Metric	Description
percentX	Percentage of speech consisting of X intervals
stddevX	Standard deviation of X intervals
varcoX	$\Delta X \times 100/\text{mean(X)}$
rpviX	Raw Pairwise Variability Index: $\sum_{k=1}^{n-1} \mid x_{k+1} - x_k \mid /n-1$
npviX	Normalized Pairwise Variability Index: $100 \times \sum_{k=1}^{n-1} \mid x_{k+1} - x_k / \left(x_{k+1} + x_k/2\right) \mid /n-1$, where k is the index of linguistic intervals ranging from 1 to $n-1$

Note. The metrics are calculated using durations of three different types of linguistic intervals: X ∈ {v = vowels, c = consonants, s = syllables} (except for %X, which is not meaningful for syllables).

2011). In addition, Lai, Evanini, and Zechner (2013) investigated spontaneous speech produced by speakers from a wide range of L1 backgrounds in the context of the TPO practice test. In general, these studies have demonstrated that nonnative speakers tend to have different rhythmic patterns of segmental and syllabic duration compared to native speakers and that these differences can be beneficial for automated speaking proficiency assessment. These findings motivated the addition of rhythm features to SpeechRater.

Method

Table 3 summarizes the rhythm metrics that were added to SpeechRater based on the findings from the studies described in the preceding introduction. These metrics are calculated using durations of consonantal (c), vocalic (v), and syllabic (s) intervals (except for percentX, which is defined for vowels and consonants but not for syllables).

Evaluation

As reported in Lai et al. (2013), several of the rhythm features described in this section had significant correlations with holistic human scores for spontaneous spoken responses captured in the TPO practice test. The highest performing feature was rpvis, with a correlation of −.44; this indicates that nonnative speakers who produce more uniform syllable durations throughout an utterance tend to receive higher scores. Additional performance results for the rhythm features are presented in the section Feature Correlations.

Discussion

Most studies of L2 rhythm have focused on patterns of duration across an utterance, which motivated the selection and addition to SpeechRater of duration-based rhythm features. However, some studies have also considered other acoustic properties that contribute to the perception of a nonnative speaker's rhythm; in particular, He (2012) and Selouani, Alotaibi, Cichocki, Gharsellaoui, and Kadi (2015) found systematic rhythmic differences between L1 and L2 speech based on patterns of average intensity values across linguistic intervals, similar to the duration-based rhythm metrics percentX, stddevX, varcoX, and rpviX. Future research will address whether these additional types of features can improve SpeechRater's automated assessment of nonnative rhythm in addition to the duration-based features.

Vocabulary Features

Introduction and Related Work

Vocabulary usage comprises two subconstructs: sophistication and precision. The vocabulary features described in this section are designed to measure lexical sophistication. Vocabulary sophistication features assess the degree to which a varied and large vocabulary is used (Laufer & Nation, 1995).

Researchers from SLA have developed many quantitative features to assess lexical sophistication (e.g., Daller, Van Hout, & Treffers-Daller, 2003; Vermeer, 2000). These features can be grouped into one of two groups: (a) quantitative or (b)

Table 4 List of Vocabulary Features

Feature name	No. features	Feature type	Description
TOP1 ... TOP6	6	List-rel	Relative frequency of word types in reference vocabulary list as a % of total types
avgRank	1	Rank	Average word rank ("rank" is the ordinal number of words in a list that is inverse sorted by word frequency)
avgFreq	1	Freq	Average word frequency
logFreq	1	Freq	Average log word frequency (the logarithm of the word frequency)

qualitative. The features in the first group merely assess the number of words known; they do not make any distinctions among them. The most representative feature in this group is type–token ratio. It has been widely applied but is unstable owing to its sensitivity to the length of language samples used in calculating the feature. The features in the second group take into account distinctions among words, such as their parts of speech or difficulty levels. The lexical frequency profile (LFP) as described in Laufer and Nation (1995) is a representative feature in this group. LFP uses a vocabulary profile for a given body of written text or spoken utterances and gives the percentage of words used at different vocabulary frequency levels (such as from the 1,000 most common words, the next 1,000 most common words, etc.), where the words themselves come from a vocabulary list that is precompiled based on frequencies of actual usage in corpora. Laufer and Nation have shown that LFP is a strong measure in assessing the written proficiency levels of ESL learners. However, limited studies have explored the relationship between vocabulary features and oral proficiency level from spoken responses.

Method

Table 4 summarizes LFP-based vocabulary features that were added to SpeechRater 5.0. First, the frequency of each vocabulary item was calculated from the TOEFL Academic Language Corpus covering the variety of language used in academic situations. We used frequency to estimate the difficulty of each vocabulary item; low-frequency items were considered to be difficult. Words were classified into seven groups based their frequency: top 100 words (TOP1), top 101–300 words (TOP2), top 301–700 words (TOP3), top 701–1,500 words (TOP4), top 1,501–3,000 words (TOP5), and over 3,001 words (TOP6). Next, we generated the four types of features shown in Table 4. The details of the feature generation process can be found in Yoon, Bhat, and Zechner (2012).

Evaluation

As shown by Yoon et al. (2012), the best performing features are avgFreq followed by TOP1. The feature evaluation result from SpeechRater 5.0 is presented in the section Feature Correlations.

Discussion

The empirical performance of LFP features was strongly influenced by the length of input. In particular, the proportion of low-frequency word types fluctuated largely even within the same speaker when each response was composed of a small amount of speech. This large variation within a speaker may decrease the correlation with oral proficiency scores. In addition, there was a strong impact on feature values by task types. Given two different task types — tasks that elicited opinions about familiar topics and tasks that elicited a summary or opinion about reading passages or listening stimuli — responses to the latter type tended to include more low-frequency words than responses to the former type. To address this task-type impact, a special approach (e.g., task-type-specific models) may be required in future research.

Grammatical Complexity Features

Introduction and Related Work

Grammatical complexity, the mastering of a variety of syntactic structures, is an important aspect of spoken proficiency assessed by language tests, in particular, when such tests elicit spontaneous rather than predictable speech. Research in

SLA has identified a number of different measures of syntactic complexity and has also looked at how well they correlate with oral proficiency scores by human raters (Broussard, 2001; Iwashita, 2010; Iwashita, Brown, Mcnamara, & Hagan, 2008; Lu, 2010; Ortega, 2003). However, the focus of such research has been predominantly on written production using manual annotation of such measures.

In recent years, several research groups have started also to explore the use of measures for evaluating syntactic complexity in oral production (Bernstein, Cheng, & Suzuki, 2010; Bhat & Yoon, 2015; M. Chen & Zechner, 2011; Yoon & Bhat, 2012). Most measures used here are inspired by those used previously for analyzing the grammatical complexity of written language, but some more robust measures based on POS were added to this set that address specific issues of the automated scoring of spontaneous speech, for example, the issue of words incorrectly recognized by the ASR systems, which are the first step of automated speech scoring systems (Yoon & Bhat, 2012).

In general, computing grammatical measures based on nonnative spontaneous speech is very challenging, not only because of the aforementioned errors by ASR systems, but also because of speech disfluencies, such as hesitations, filled pauses, or false starts; various errors by the speaker; and the need to predict clause boundaries automatically. (ASR systems do not generate any interpunctuation in their output.) For these reasons, it is advantageous to use measures that are robust and not too complex to compute. We also note here that in contrast to the research on the automated scoring of essays, very little research has been done related to grammatical error detection in nonnative speech. The features we use in SpeechRater also focus, for the most part, on grammatical complexity rather than on grammatical accuracy.

Method

SpeechRater currently uses three groups of features measuring grammatical complexity:

- *Part-of-speech-based features*. This set of features assesses the range and sophistication of grammatical expressions based on their similarity with a corpus of learners' speech. The features are based on shallow processing (POS tagging) and are more robust against ASR errors. The features measure vector similarities (dot products) between samples of responses for each score level (1 – 4) and a particular ASR hypothesis generated from a spoken response (Bhat & Yoon, 2015; Yoon & Bhat, 2012). (These vectors contain frequencies of POS bigrams.) Aside from the similarity measures for each score level (poscva1, … , poscva4), a fifth feature (poscvamax) returns the score level with the maximum similarity score.
- *Clause-based features*. This set of features looks at the occurrence and frequency of certain clauses in the ASR hypothesis (M. Chen & Zechner, 2011). Clause-based features are related to the following two clause types: coordinate clauses (coord) and dependent clauses (dep).
- *Phrase-based features*. This set of features looks at the frequency of certain syntactic phrases in the ASR hypothesis (M. Chen & Zechner, 2011). In addition to basic syntactic constituent phrases — noun phrases (NPs), prepositional phrases (PPs), and verb phrases (VPs) — SpeechRater also computes statistics on coordinate phrases (coord), complex nominal phrases with embeddings (CX_Nominals), and dependent infinitives (Dep_Inf).

Clause-based and phrase-based features are computed using a pipeline approach, whereby the ASR hypothesis is first cleaned up (e.g., filler words are removed), then automatically segmented into clauses, then syntactically parsed using the Stanford parser (Klein & Manning, 2003), and finally processed using a script to extract syntactic structures from the output of the parser.

A similar process is used for computing POS-based features: The cleaned ASR hypothesis is tagged for POS. Next, features are generated by calculating similarity with a POS-based vector space model trained from a large collection of learners' spoken response transcriptions.

Evaluation

In our previous work (L. Chen & Zechner, 2011), we found a few features with correlations with human holistic scores in the range of .3 – .4, for example, features measuring the mean length of a sentence, the number of fragments, the number of dependent infinitives, or the number of prepositional phrases. As for POS-based features, the highest correlations previously reported were above $r = .4$ for a feature trained on responses for score level 4 using POS bigrams (Bhat & Yoon, 2015). The section Feature Correlations presents the performance of these features in the SpeechRater 5.0 system.

Discussion

As we pointed out earlier, measuring syntactic complexity in nonnative speech is very challenging because errors can be introduced at all stages of the feature computation process: the ASR system, removal of disfluencies from the ASR hypothesis, POS tagging, the assignment of clause boundaries, syntactic parsing, and structure extraction. However, despite all of these challenges, it is encouraging to see that at least for a subset of grammar features, reasonable performance (correlation with human scores) can be achieved.

For future work, we plan, on the one hand, to improve and refine the various components of the feature computation pipeline and, on the other hand, to explore additional features that exhibit promise for measuring the syntactic complexity of nonnative speech.

Content Features

Introduction and Related Work

The appropriateness of a response's content in completing the specified speaking task is typically an important component of the human scoring criteria for spontaneous speech, as exemplified by the following description of high-scoring responses in the topic development category of the TOEFL Speaking scoring rubrics for the integrated tasks: "The response … conveys the relevant information required by the task. It includes appropriate detail." However, features addressing content appropriateness were not included in early versions of SpeechRater, primarily because of the relatively low accuracy of the ASR engine and the associated difficulty in extracting meaningful content features. With the inclusion of a more accurate ASR engine (see the section Automatic Speech Recognition System for details), research was initiated to investigate the performance of different types of content features. This section describes features based on standard word-level vector space models (Salton, Wong, & Yang, 1975), which have been used to extract effective content appropriateness features—referred to as content vector analysis (CVA) features—in the context of automated essay scoring (Attali & Burstein, 2006). The basic motivation behind these features is to use a data set of human-scored responses to train CVA models at each score point and then compare the similarity between the content in a test response to these models to calculate the content features. After presenting this basic approach, the Discussion section briefly mentions some additional approaches that have been investigated recently.

Method

To develop the CVA features, lexical vectors containing term frequencies weighted by inverse document frequency values (tf-idf) were trained for a set of responses from each of the score points represented in the human scoring rubrics (this represents a 1–4 range for TOEFL Speaking). For each of the score points, s, the tf-idf value for each word, i, in the vector was therefore calculated as follows:

$$\text{tf-idf}_{i,s} = \text{tf}^{*}_{i,s} \log\left(N/N_i\right), \tag{1}$$

where $\text{tf}_{i,s}$ is the frequency of the word i across all responses at score point s, N is the total number of responses in the corpus used to calculate the idf values, and N_i is the total number of responses containing word i across all score points in the idf corpus. Then, for a given spoken response, the tf-idf value for each word in the vector was calculated as follows:

$$\text{tf-idf}_{i} = \text{tf}^{*}_{i} \log\left(N/N_i\right), \tag{2}$$

where tf_i is the frequency of the word i in the response. Then, to calculate the content features, the cosine similarity score between the vector for the response and each of the CVA models is computed. These cosine similarity scores are then used directly as features to predict proficiency scores and are referred to as follows: $\cos_s$ for $s \in 1, \ldots, 4$. An additional feature was calculated by comparing all of the cosine similarity scores to the models for the five score points for a given response and taking the score of the model that has the highest similarity; this feature is referred to as max_cos.

The CVA models can be trained either by using human transcriptions or the output of the ASR engine. Using the transcriptions provides the most accurate representation of the content of the response; however, this approach leads to a mismatch between the CVA models and the CVA vector for a test response, because the human transcription is not available in a live deployment. Therefore results using both approaches are presented in the following section.

Table 5 Correlations With Human Scores for Two Content Vector Analysis Content Features on Responses to TOEFL Integrated and Independent Prompts

Prompt type	CVA model source	max_cos	cos_4
TOEFL independent	Transcription	0.37	n.s.
	ASR	0.30	n.s.
TOEFL integrated	Transcription	0.50	0.51
	ASR	0.49	0.53

Note. CVA = content vector analysis.

Evaluation

Separate CVA models were trained using responses from a set of 8 TOEFL independent and 16 TOEFL integrated prompts and evaluated on responses to the same prompts (for further details about these experiments, see Xie et al., 2012). The results consistently showed that the cos_4 feature, in which the content of the test response is compared to the CVA model based on responses with the highest human score, outperformed the other cos_s features. Table 5 presents the performance of two CVA features, cos_4 and max_cos, in terms of correlations with human scores for CVA models trained using both human transcriptions and ASR output.

As shown in Table 5, the performance of the CVA features was higher on the responses to TOEFL integrated prompts than to TOEFL independent prompts, and the cos_4 feature does not even have a significant correlation with human scores for the TOEFL independent prompts. This result is not surprising, because the TOEFL independent prompts are not source based, and the content of high scoring responses is therefore expected to exhibit much more variation, thus reducing the effectiveness of the CVA models. The results in Table 5 also demonstrate that the CVA features are robust to ASR errors: The performance of the max_cos and cos_4 features changes very little on TOEFL integrated responses when ASR output is used to train the CVA models compared to when human transcriptions are used, despite the fact that the ASR WER on this set was 33%.

Discussion

In addition to the relatively straightforward method of using CVA models and cosine similarity calculations to produce the content features, additional approaches have been investigated for scoring spontaneous speech. Some of these include using latent semantic analysis (LSA; Metallinou & Cheng, 2014), pointwise mutual information (Xie, Evanini, & Zechner, 2012), and the ROUGE summarization evaluation metric (Lin & Rey, 2004; Loukina, Zechner, & Chen, 2014).

Finally, it should be emphasized that these approaches all assume the existence of human-scored responses to the same prompts that can be used to train the content models. In the absence of such data, for example, when new prompts are first deployed in an assessment, alternative approaches for assessing the content are required. One general approach has been to compare the content in the test response with elements from the stimulus materials presented to the test taker in the source-based task, such as a listening passage or an article. This approach has resulted in some features that have significant correlations with human scores but that do not perform as well as the features calculated using models trained on human-scored responses; Evanini, Xie, and Zechner (2013) have presented results using this type of prompt-based feature for spoken responses, and Beigman Klebanov, Madnani, Burstein, and Somasundaran (2014) have presented results using prompt-based features for essays.

Discourse Coherence Features

Introduction and Related Work

Discourse coherence related to topic development has always been used as a key metric in human scoring rubrics for various assessments of spoken language. However, very little research has been done to assess a speaker's coherence in automated speech scoring systems. To address this, we present a corpus of spoken responses that has been annotated for discourse coherence quality, and we explore a set of surface-based features to capture the use of nouns, pronouns, conjunctions, and discourse connectives in a spoken response.

Methods for automatically assessing discourse coherence in text documents have been widely studied in the context of applications such as natural language generation, document summarization, and assessment of text readability. For example, Foltz, Kintsch, and Landauer (1998) measured the overall coherence of a text by utilizing LSA to calculate the semantic relatedness between adjacent sentences. Barzilay and Lee (2004) introduced a model for the document-level analysis of topics and topic transitions based on HMMs. Barzilay and Lapata (2005, 2008) presented an approach for coherence modeling focused on the entities in the text and their grammatical transitions between adjacent sentences and calculated the entity transition probabilities on the document level. Pitler, Louis, and Nenkova (2010) provided a summary of the performance of several different types of features for automated coherence evaluation, including features based on cohesive devices, measurements of adjacent sentence similarity, Coh – Metrix features (Graesser, McNamara, Louwerse, & Cai, 2004), word co-occurrence patterns, and entity grids (Barzilay & Lapata, 2008).

In addition to studies on well-formed text, researchers have also addressed coherence modeling on text produced by language learners, which may contain multiple spelling, vocabulary, and grammar errors. Utilizing LSA and random indexing methods, Higgins, Burstein, Marcu, and Gentile (2004) measured the global coherence of students' essays by calculating the semantic relatedness between sentences and the corresponding prompts. In addition, Burstein, Tetreault, and Andreyev (2010) combined entity-grid features with writing quality features produced by an automated essay assessment system to predict the coherence scores of student essays. Recently, Yannakoudakis and Briscoe (2012) systematically analyzed a variety of coherence modeling methods within the framework of an automated assessment system for nonnative free text responses and indicated that features based on incremental semantic analysis, local histograms of words, POS co-occurrence patterns in adjacent sentences, and word length were the most effective.

In contrast to these previous studies on written texts, Hassanali, Liu, and Solorio (2012) investigated coherence modeling for spoken language in the context of a story retelling task for the automated diagnosis of children with language impairment. They annotated transcriptions of children's narratives with coherence scores as well as markers of narrative structure and narrative quality; furthermore, they built models to predict the coherence scores based on Coh – Metrix features and the manually annotated narrative features. The study of Wang, Evanini, and Zechner (2013) differed from this one in that it dealt with free spontaneous spoken responses provided by students at a university level; these responses therefore contained more varied and more complex information than the child narratives did.

Data and Annotation

The data used in this study were drawn from the *TOEFL iBT*®test and comprised 1,440 spoken responses from one test form (240 responses from each item). The spoken responses were all manually transcribed, and the average number of words per response was 113.8 ($SD = 33.6$), and the average number of sentences was 4.8 ($SD = 2.1$).

The coherence annotation guidelines used for the spoken responses in this study were modified based on the annotation guidelines developed for written essays described by Burstein et al. (2010). According to these guidelines, expert annotators provided each response with a score on a scale of 1 – 3. The three score points were defined as follows: 3 = highly coherent (contains no instances of confusing arguments or examples), 2 = somewhat coherent (contains some awkward points in which the speaker's line of argument is unclear), 1 = barely coherent (the entire response was confusing and hard to follow; it was intuitively incoherent as a whole, and the annotators had difficulty identifying specific weak points). For responses receiving a coherence score of 2, the annotators were requested to highlight the specific awkward points in the response. In addition, the annotators were specifically required to ignore disfluencies and grammatical errors as much as possible; thus they were instructed not to label sentences or clauses as awkward solely because of the presence of disfluent or ungrammatical speech.

Two annotators first made independent coherence annotations for 600 spoken responses, including 25 samples from each of the four score levels of speaking proficiency for each of the six test questions. The two annotators achieved a moderate interannotator agreement (Landis & Koch, 1977) of $\kappa = .68$ on the 3-point scale of coherence scores. Subsequently, the same two annotators provided coherence annotations for the remaining 840 responses in the corpus using the following approach: Each annotator provided a single annotation for 420 responses from three test questions, that is, 35 responses from each score level for each test question.

To verify the effectiveness of the proposed coherence cues in the assessment of speaking proficiency, we extracted two types of features based on the human annotations, including the coherence scores and the number of awkward points identified in responses. These two features are correlated with the holistic proficiency scores for evaluation. The double

Table 6 Pearson Correlation Coefficient *r* of Human Coherence Scores and Number of Awkward Points With Holistic Human Scores

	r with coherence scores	*r* with awkward points
Double annotation	.656	−.626
Single annotation	.615	−.597

annotated set received a special treatment: The average coherence scores from two annotators were used, and the union set of awkward points identified by either annotator on each response was counted. As shown in Table 6, on the 600 double-annotated responses, the average coherence scores correlate with the proficiency scores at $r = .656$, and the number of the union set of identified awkward points correlates at $r = -.626$, indicating that the assessment of speaking proficiency can greatly benefit from modeling the coherence cues proposed in this study.

Features and Evaluation

This work[1] explored a set of simple features that were designed to capture the use of nouns, pronouns, conjunctions, and discourse connectives in a test taker's spoken response, henceforth referred to as surface-based features. For this purpose, the discourse connective list from the Penn Discourse Treebank (Prasad et al., 2008) was used. Various basic features were counted, such as the numbers of nouns, pronouns, conjunctions, and discourse connectives (counted based on both word types and word tokens). The ratios between these counts were also extracted. An evaluation was conducted to examine the correlations of these features with the averaged coherence scores on the 600 double-annotated responses, and only features with absolute correlations greater than .1 on both the human transcriptions and the ASR outputs were adopted.

As shown in Table 7, besides the first five features based on word counts, two additional features were designed to capture the global coherence, which represented the use of conjunctions and discourse connectives across a test response. To obtain these features, a reference corpus with high-proficiency responses was collected, and then a connective chain was extracted from each reference response, where only the pronouns, conjunctions, and discourse connectives were retained and all other words were removed from the response. Given a test response, a similar connective chain can be also extracted. Then, by comparing the similarity of the test chain with each of the reference chains, the maximum similarity or the minimum distance can be extracted as a feature to measure the proper use of the connective sequence in a test response. The following three evaluation metrics were investigated to evaluate the similarity between two chains: BLEU score (Papineni, Roukos, Ward, & Zhu, 2002), edit distance, and WER.

The reference chains can be built as either item-specific or generic ones: The item-specific references indicate that they were elicited with the same test question used to get the test response; conversely, generic references were elicited across multiple different test questions. In this work, the reference samples were extracted from a corpus that was used to train the speech recognizer in SpeechRater. Approximately 200–260 responses with the highest speaking proficiency scores were obtained for each test question, and in total, 1,395 responses across six test questions were collected as references. A preliminary experiment indicated that the BLEU similarity with the item-specific models, that is, connective_chain_bleu, and the edit distance with the generic models, that is, connective_chain_ed, achieve relatively higher correlations, as shown in Table 7.

Table 7 Pearson Correlation Coefficients *r* of Surface-Based Features With the Averaged Coherence Scores, Extracted From the Human Transcriptions and the Automatic Speech Recognition Outputs Separately

Features	Transcription	ASR
num_pronouns	.204	.186
ratio_pronoun_nountype	−.128	−.106
num_conjunctions	.174	.209
num_connective_types	.381	.352
num_connective_tokens	.337	.330
connective_chain_bleu	.068	.155
connective_chain_ed	.282	.268

Note. ASR = automatic speech recognition.

Discussion

In Wang et al. (2013), we presented a corpus of coherence annotations for spontaneous spoken responses, and the analysis of these annotations shows that an automated speech scoring system can benefit from modeling the coherence of spoken responses. On the basis of this finding, one set of surface-based features was employed to model the discourse coherence of spontaneous speech. In the future, we will continue the work on discourse coherence modeling of spoken responses, either by predicting the coherence quality scores or by identifying the awkward points. More importantly, we will attempt to develop more effective discourse-related features that are robust against recognition errors.

SpeechRater 5.0

Automatic Speech Recognition System

ASR is an essential component in the SpeechRater system. It is used as the first step in SpeechRater for generating information used for extracting a large variety of linguistic features that are then combined in a scoring model for a spoken language assessment. A systematic investigation (Tao, Evanini, & Wang, 2014) showed that the ASR module used within SpeechRater plays an extremely important role in achieving high performance in the scoring task. The primary ASR we are currently using within SpeechRater 5.0 is provided by an external vendor. Compared to the ASR system used in SpeechRater 1.0, the current system has many significant improvements on the ASR technology itself and the size of the training data.

Because the spoken responses are nonnative spontaneous speech, a two-stage ASR method is developed by which the spoken responses are first recognized based on the acoustic properties of nonnative speech, and then the speech and the recognized responses are force-aligned using an AM that reflects the properties of native speech (L. Chen et al., 2009). In the initial stage, the recognizer uses a highly optimized speaker-independent cross-word triphone HMM with Gaussian mixture for each state as the AM and a four-gram statistical LM. To obtain the most accurate word hypotheses from nonnative spontaneous speech, the AM and the LM were trained using a corpus of approximately 800 hours of nonnative spontaneous English speech collected from the TOEFL iBT assessment. This ASR engine achieved a WER of 28.5% on the evaluation partition of the corpus. In the forced-alignment stage, the nonnative spontaneous speech is force-aligned to the word hypotheses recognized from the previous stage using another AM trained on native English speech. To accommodate both North American and British accents, two LDC speech corpora, Broadcast News (Graff et al., 1997) and Cambridge Read News (Robinson et al., 1995), were combined to train the AM. This two-stage ASR system not only generates word hypotheses but also computes word confidence scores, LM likelihood, and word and phone acoustic likelihood.

Although the external vendor–provided ASR achieved a 28.5% WER on the evaluation partition of this corpus for the purpose of building the nonnative spontaneous English ASR system, the WER on the TPO scoring evaluation partition is 38.5%, owing to a vocabulary mismatch caused by the fact that many of the prompts in the TPO data set were not contained in the ASR training set.

Filtering Models

In a large-scale language proficiency assessment including a speaking test, some spoken responses have suboptimal characteristics that make it difficult for the automated scoring system to provide a valid score. Hereinafter we call these problematic responses nonscorable responses. These nonscorable responses can be classified into the following two groups: (a) TD,[2] responses with serious audio quality problems that make it impossible to assign fair scores, for example, responses with a high level of noise; and (b) 0,[3] responses from uncooperative test takers, for example, responses in which the test taker does not speak.

They are likely to cause problems in automated speech recognition or in extracting the linguistic features used in generating an automated score. As a result, they may cause failure in score generation or reduce the validity of the automated scores. To address these issues, we used a two-step approach: These problematic responses were filtered out by a "filtering model," and only the remaining responses were scored using the automated scoring model. By filtering out these responses, the robustness of the automated scoring system can be improved. For SpeechRater 5.0, we focused on

developing new capabilities to detect the responses that receive a score of 0 (group 2 from the above) because we already achieved a high accuracy of detecting TD responses (e.g., Higgins, Xi, Zechner, & Williamson, 2011).

Recently, a few researchers have investigated the filtering of nonscorable responses for automated speech scoring, but most studies have focused on restricted speech. van Doremalen, Strik, and Cucchiarini (2009) and Lo, Harrison, and Meng (2010) used normalized confidence scores of a speech recognizer in recasting speech. They identified nonscorable responses with promising performance (equal error rates ranged from 10% to 20%). Cheng and Shen (2011) extended these studies and combined an AM score, a LM score, and a garbage model score with confidence scores. They applied this new filter to less constrained items (e.g., picture description) and identified off-topic responses with an accuracy rate of 90% and with a false positive rate of 5%.

Although these models achieved promising performance in restricted speech, they are not appropriate for unconstrained speech. The types of nonscorable responses that arise in a speaking test that elicits unconstrained spontaneous speech may be different from what is encountered in restricted speech. To better understand this issue, first we analyzed the types of nonscorable responses using a large collection of spoken responses from the TOEFL iBT. For the responses that received a score of 0, we found the following subcategories: (a) no-speech, or responses that do not have any speech but have clear evidence of the speaker's presence, for example, breathing, coughing; (b) non-English response, or responses spoken entirely in another language; (c) off-topic, or responses that are entirely irrelevant to the task; (d) generic responses, or responses that only include filler words or generic responses, such as "I don't know," "it is too difficult to answer," or "well"; (e) question-copy, or full or partial repetition of the question or reading/listening stimuli; (f) canned responses, or responses only including memorized segments provided by external sources (often Web sites); and (g) other, or responses from speakers who do not attempt to respond in a way not otherwise covered by the preceding categories. In addition, there is a "complicated responses" category. These are responses that belong to more than one type, for instance, a response that comprises both off-topic sentences and sentences in a speaker's native language.

Responses that belong to non-English, off-topic, and canned responses are likely to be associated with test takers who try to game the automated system. By speaking in their native languages, citing memorized responses for unrelated topics, or reading questions or parts of questions, test takers can generate fluent speech, and the automated proficiency scoring system, which utilizes fluency as one of the important factors, may assign a high score. The proportion of these gaming responses was extremely low (less than 0.5%) in both TOEFL iBT, for which all responses are scored by human raters, and TPO, for which students may have low motivation owing to the low-stakes nature of the test. Because of this skewed distribution, it was not easy to train the filtering models for gaming responses. To address this issue, we trained two new filtering models — a filtering model for non-English responses and a filtering model for responses with topicality issues — using different data sets.

Filtering Model for Non-English Responses

The non-English filtering model was developed based on speech-based language identification (LID) technology and fluency features (e.g., speaking rate) from SpeechRater. The LID technology used in this study was based on the output of multiple language-dependent phone recognizers, as in Zissman (1996). The frequencies of phones and phone sequences differ according to languages, and some phone sequences occur only in certain languages. The system first ran 10 language-dependent phone recognizers[4] and created a set of features, such as the language with the highest phone recognizer score, the normalized phone recognizer score, and the difference between normalized English phone recognizer score. Details of the implementation are presented in Yoon and Higgins (2011).

We trained a non-English filtering model using LID technology and a corpus comprising 3,021 English responses from TPO and 158 non-English audio files from the OGI Multi-Language corpus (Muthusamy, Cole, & Oshika, 1992). The OGI Multi-Language corpus is a standard language identification development data set including speech in 10 different languages. A decision tree model was trained to predict binary values (0 for English and 1 for non-English) using the J48 algorithm (WEKA implementation of C4.5) of the WEKA machine learning tool kit (Hall et al., 2009). We performed threefold cross-validation during the non-English filtering model training and evaluation. The model achieved high performance: The accuracy was .98, and the *F*-score was .82. The accuracy was .03 higher than the accuracy of the baseline using majority voting (in which all responses were classified as English, the majority class), and there was a 35% relative error reduction.

Filtering Model for Responses With Topicality Issues

The topicality filtering model was developed to filter out responses with topicality problems: off-topic responses, question-copy, and generic responses. For this purpose, in addition to SpeechRater features, we developed a set of new features based on the metrics frequently used to identify documents with relevant topics (e.g., Hoad & Zobel, 2003; Sanderson, 1997). The new features were classified into the following two subgroups:

1. *Response-based features*. These were features based on CVA models using cosine similarity and term frequency-inverse document frequency (tf-idf). Two different CVAs (one trained on samples with the highest proficiency scores and one trained on the test prompts) were used to detect both off-topic responses and question-copy responses.
2. *Sentence-based features*. These were similarity scores between a prompt question and a response at each sentence level, similar to Metzler, Bernstein, Croft, Moffat, and Zobel (2005). The response was first split into the sentences, and the proportion of word overlap with the prompt question was calculated. Finally, the response was determined as nonscorable or not based on aggregated sentence-level features.

We used 11,560 spoken responses from TOEFL iBT, and the proportion of nonscorable responses in this data set was 13% (for a total of 1,560 responses).[5] We performed 10-fold cross-validation to train and evaluate the filtering model. In addition to the newly developed features, we used SpeechRater features, and 30 features were selected using the WEKA feature selection algorithm.[6] Finally, the filtering model was trained using the support vector machine algorithm with the radial basis function of the WEKA machine learning tool kit (Hall et al., 2009).

The model achieved extremely high performance; the accuracy was .98, and *F*-score was .91. There was a substantial improvement over the majority class baseline (classifying all responses as scorable responses); the accuracy of this system was .87. However, there was only slight improvement over the model that was based exclusively on SpeechRater features; the accuracy and *F*-score were .98 and .90, respectively, and the improvement was .01.

Because the majority of nonscorable responses were no-speech, the new features designed for responses with topicality issues may not have a strong impact on the results. To evaluate the impact of the new features on the target nonscorable types, we created a new data set by removing the responses that were nonscorable for reasons other than topicality. We removed 1,246 nonscorable responses, and the data set included only 314 nonscorable responses (4%).

The accuracy was .98, and the *F*-score was .66. Furthermore, there was a substantial improvement over the model based exclusively on SpeechRater features. The *F*-score of the model using all features was .19 higher than the *F*-score of the model based only on SpeechRater features (.47).

Applying SpeechRater 5.0 to Building an Automated Scoring Model for the TPO Version 5

Feature Correlations

In the section Innovations in Scoring Features, a series of the new features that have been proposed by the speech scoring research team was presented. In the past several years, through collaboration between research scientists and engineers, most of these feature extraction methods have been implemented into SpeechRater 5.0. Therefore, when building the fifth version of the TPO scoring models (TPO V5), these new features can be computed directly and can be considered with other preexisting features during the model-building process. In addition, the implementation of these features into SpeechRater 5.0 also provides us an opportunity to evaluate these features on a common data set. This is useful, because various data sets were used in the evaluations that took place over the course of development of these features, as shown in previous publications.

When evaluating the features described in this section, we used the new larger standard data set collected in 2012. Two thirds of this data set (667 test takers; 4,002 responses) were allocated to the model training partition, and one-third (333 test takers; 1,998 responses) was allocated to the model evaluation partition. Table 8 presents the interrater agreement for the two partitions, in terms of both Pearson correlation (*r*) and quadratic weighted κ, between two raters' holistic scores.

On this new data set, we applied SpeechRater to extract speech features. To measure the features' usefulness for predicting human-rated scores, we computed the Pearson correlations (*r*) between the features and human rated scores. The following tables report on the correlation analysis results regarding the features introduced in the section Innovations in Scoring Features.

Table 8 Interrater Agreement on Model Training (sm-train) Partition and Model Evaluation (sm-eval) Partition

Data set	Correlation *(r)*	Quadratic weighted κ
sm-train	.61	.61
sm-eval	.59	.59

Table 9 Pearson Correlation *r* Between Human-Rated Scores and the Structural-Event Features Described in the Section Utilizing Structural Events

Feature	*r* to human-rated scores
ipcount	.141
clausecount	.199
IPC	.237
IPW	.279
longSilRatio	−.356

Table 10 Pearson Correlation *r* Between Human-Rated Scores and the Pronunciation Features Described in the Section Improved Pronunciation Features Measuring Spontaneous Nonnative Speech

Feature	*r* to human-rated scores
L1	.307
L2	.187
L3	.264
L4	.125
L5	.121
L6	.398
L7	.396
phn_{shift}	.430

Table 9 reports on the Pearson correlations between the features related to SEs, which were described in the section Utilizing Structural Events, and human-rated scores. On the SpeechRater 5.0 system, the feature with the highest $|r|$ is longSilRatio ($r = -.356$). It is interesting to see that the IPW feature now has a high $|r|$ ($r = .280$) after implementing a more accurate ASR in the SpeechRater 5.0 system. Clearly the features related to the internal structure of spoken responses (i.e., clauses and disfluencies) provide additional measurements over the feature set used in SpeechRater 1.0.

Table 10 reports the correlation analysis results on the new group of pronunciation features described in the section Improved Pronunciation Features Measuring Spontaneous Nonnative Speech. Both L6 and L7 features show very high correlation (*r* values around .4), quite close to the *r* value from the amscore feature ($r = .404$). Because the new pronunciation features are more relevant to the construct, these new features help SpeechRater 5.0 increase its construct validity. In addition, the new feature related to phoneme durations, phn_{shift}, shows quite a high *r* (.430) and will help SpeechRater 5.0 reach higher scoring accuracy with respect to pronunciation.

Table 11 reports the correlation analysis results on the features measuring rhythm described in the section Rhythm Features. The rpvis feature shows a high correlation ($r = .357$). Both npvis and stddevs show a correlation higher than .25. This suggests that new rhythm-related features provide additional support for more accurate scoring.

Table 12 reports the correlation analysis results on the vocabulary features described in the section Vocabulary Features. Consistent with the evaluation result reported in Yoon et al. (2012), the best performing vocabulary profile features are avgFreq ($r = .33$) and TOP1 ($r = .20$).

Table 13 shows correlations between the grammatical complexity features described in the section Grammatical Complexity Features. Features counting occurrences of traditional constituent phrases, for example, NPs and PPs, show fairly strong correlations with human holistic proficiency scores (i.e., compared to SpeechRater features from other construct areas), almost reaching .4 for NPs and PPs. A similar level of performance is also achieved by POS-based features, where

Table 11 Pearson Correlation *r* Between Human-Rated Scores and the Rhythm Features Described in the Section Rhythm Features

Feature	*r* to human-rated scores
percentv	.171
rpvic	.216
rpvis	.357
rpviv	.243
npvic	.131
npvis	.268
npviv	.121
varcoc	.150
varcos	.248
varcov	.084
stddevc	.216
stddevs	.296
stddevv	.216

Table 12 Pearson Correlation *r* Between Human-Rated Scores and the Vocabulary Features Described in the Section Vocabulary Features

Feature	*r* to human-rated scores
TOP1	.201
TOP2	.065
TOP3	.141
TOP4	.114
TOP5	.146
TOP6	.161
avgRank	−.069
avgFreq	.326
logFreq	.165

POS sequences of training data are compared to those observed in the ASR hypothesis for a test taker's response. In contrast, clause-based features have correlations below .3 in the evaluation result on SpeechRater 5.0. These results indicate that grammatical features based on shorter spans, such as POS sequences, NPs, or PPs, can be more accurately computed than those using longer spans of information, such as coordinate phrases, VPs, and various clause types.

The cvamax feature described in the section Content Features shows a correlation of .311 with human-rated scores and will provide an important measurement regarding content relevance.

Hybrid Method of Feature Selection

Building automated scoring models for constructed responses is a complex endeavor because such models need to balance good empirical performance with the validity and interpretability of the scoring models (cf. Bernstein, Moere et al., 2010; Ramineni & Williamson, 2013; Williamson, Xi, & Breyer, 2012). One very important aspect of validity is the extent to which the automated scoring model reflects important dimensions of the construct measured by the test. Furthermore, relative contributions by features to each construct dimension should be transparent to the test taker and the score user. Finally, the contribution of each feature to the final score should be consistent with expectations: If all of the features in the model are designed to be positively correlated with a criterion score, the coefficients of all such features in the final model should be positive as well.

Fulfilling all of these requirements when building automated scoring models is not trivial, and therefore, in previous versions of SpeechRater, the scoring models for constructed responses were built using human experts who selected features based on these criteria in an iterative fashion, training and evaluating scoring models after each feature set was chosen. However, there are certain limitations to this manual process of building scoring models, not the least of which is the aspect of the time it takes to build models with iterative evaluations and changes in the feature set composition.

Table 13 Pearson Correlation *r* Between Human-Rated Scores and the Grammatical Complexity Features Described in the Section Grammatical Complexity Features

Feature	*r* to human-rated scores
poscva1	−.316
poscva2	.344
poscva3	.379
poscva4	.401
poscvamax	.261
coord_clauses	.075
coord_clauses_per_clause	.049
dep_clauses	.273
dep_clauses_per_clause	.134
coor_phrases	.161
coor_phrases_per_clause	.084
cx_nominals	.275
cx_nominals_per_np	.112
dep_inf	.217
dep_inf_per_clause	.096
np	.388
pp	.304
vp	.379

To solve this problem, Loukina et al. (2015) introduced an automatic method of feature selection based on penalized linear models. In this approach, the feature selection is done using Lasso regression (Tibshirani, 1996) constrained to positive-only coefficients (Goeman, 2010) and fine-tuned to enforce more aggressive feature selection. They showed that this method allowed them to achieve simultaneously satisfying construct coverage, maximal interpretability of the resulting scoring model, and good empirical performance.

As discussed in Loukina et al. (2015), sometimes the feature set selected by the fully automated method may not result in optimal construct coverage. Therefore the scoring models for SpeechRater 5.0 are built using a hybrid method of feature selection, which includes the following steps:

1. An expert identifies a subset of SpeechRater features applicable to a particular assessment or item type.
2. The Lasso-based method described is used to select the initial set of features.
3. The final set is reviewed by an expert and adjusted as necessary to ensure optimal construct coverage.

Scoring Model for TPO Version 5

The scoring model for TPO Version 5 was built using the hybrid method of feature selection described in the previous section: We identified 102 features applicable to TPO items and used a combination of Lasso regression and expert judgment to identify the optimal set of features. This fine-tuning was done using 10-fold cross-validation on the training set. The coefficients for the selected features were then estimated using ordinary least squares linear regression. The final model was then evaluated on the evaluation set.

The scoring model included 20 features that covered all constructs currently represented in SpeechRater. Though only 20 of 102 features were used in the scoring model, it is worth noting that the features not appearing in the scoring model could be used in other nonscoring scenarios, such as being used for providing constructive feedback to English learners. Table 14 shows the features included in the final hybrid model, the corresponding constructs and subconstructs, and the relative contribution of each feature to the final score. The construct attributions for features were determined based on a theoretical understanding of different constructs and an understanding of the phenomena addressed by different features.

Table 15 shows the model performance at the level of individual response in comparison to agreement between two human raters. Although the correlation between system and human scores falls below the .7 threshold recommended by Williamson et al. (2012), the degradation from the human–human score agreement was substantially below the recommended threshold of .1. Similarly, the standardized mean score difference between machine and human scores was −.02 and did not exceed the recommended threshold of .15. As also noted by Williamson et al., low interrater agreement

Table 14 List of Features Included in the Final Scoring Model for TOEFL Practice Online Version 5 and Their Relative Contributions to the Final Score

Construct	Subconstruct	Feature	Description	Relative coefficient
Delivery	Fluency	silmean	Mean silence duration	.119
	Fluency	wpsec	Speaking rate in words per second	.097
	Fluency	secpchk	Average of chunk length in seconds	.066
	Fluency	numrep	No. repetitions	.061
	Fluency	numdff	No. disfluencies	.056
	Fluency	silpsecutt	No. silences per second	.056
	Fluency	IPC	No. interruption points per clause	.012
	Fluency	withinClauseSilMean	Average duration of all within-clause silences	.008
	Pronunciation	L1	Total acoustic model score for all words with model trained on native data	.081
	Pronunciation	amscore	Total acoustic model score with model trained on nonnative data	.038
	Prosody	powstddev	*SD* of power	.057
	Prosody	pitdeltanorm	Range of normalized pitch	.028
	Prosody	phn_shift	Mean of absolute shifts of the normalized vowel durations compared to standard normalized vowel durations estimated on a native speech corpus	.014
	Rhythm	rpvic	Raw Pairwise Variability Index for consonants	.028
	Rhythm	stresyllmdev	Mean deviation of distances between stressed syllables in syllables	.014
Language use	Grammar	poscvamax	Score point with the highest grammatical similarity score	.062
	Grammar	dep_clauses_per_clause	Mean no. dependent clauses per clause	.001
	Vocabulary [a]	cvamax	Score point with the highest word CVA similarity score	.099
	Vocabulary	types	Total no. different lexical types	.061
	Vocabulary	logFreq	Average of log frequency of word types in the response	.042

Note. CVA = content vector analysis.
[a]When using generic CVA models trained on responses from a range of prompts, the CVA-related features do not measure content appropriateness directly; rather they provide a measure of general vocabulary usage.

Table 15 Human–Human and System–Human Agreement at Item Level

	N	*R*	R^2	*SMD*
Human–human	1,756	.591	.350	.032
System–human	1,775	.557	.308	−.019

Note. All metrics are computed on the evaluation set (total *N* responses, Pearson's correlation, R^2, and standardized mean difference [SMD]).

of independent human raters is likely to result in lower performance of the automated scoring "because of the inherent unreliability of the human scoring upon which it is both modeled and evaluated" (p. 7).

While item-level performance is commonly used to evaluate the performance of the automated scoring system, the speakers only receive a report of the final score aggregated across the six items. Table 16 shows the agreement for such speaker-level scores. The table only includes the speakers who received six numeric nonzero scores from both the first rater and the system.

The model-building process uses a data-driven approach. Consequently, the training data set's properties (e.g., test takers' L1 distributions, gender distributions) will influence the obtained scoring models. For tests that aid in making high-stakes decisions with a high standard on testing fairness, more careful controls on the training data sets are highly

Table 16 Human – Human and System – Human Agreement at Speaker Level

	N	*R*	R^2	*SMD*
Human-human	263	.88	.751	.036
System-human	264	.770	.584	−.061

Note. All metrics are computed on the evaluation set (total *N* speakers, Pearson's correlation, R^2, and standardized mean difference [SMD]).

needed. As for this line of research and practice in assessment, readers can refer to Dorans and Holland (2000), Dorans and Cook (2016), and Zhang, Dorans, Li, and Rupp (2017).

Conclusion

More than 10 years have passed since the first version of ETS's automated speech scoring capability, SpeechRater, was first used operationally as the sole score for the TPO Speaking section; many aspects of this early system were described in detail by Zechner and colleagues (Zechner et al., 2009; Zechner, Higgins, & Xi, 2007), Higgins et al. (2011), and Xi et al. (2008). In the meantime, research and development related to automated speech scoring at ETS has made substantial progress in many areas, including improved automated speech recognition, a substantially expanded set of linguistically based features to evaluate spoken proficiency, improved methods of feature selection and model building, and additional methods for flagging nonscorable responses. While these individual R&D efforts have been published in journals and in conference and workshop proceedings in previous years, no monograph has summarized and captured the major developments in the area of automated scoring of spontaneous speech at ETS since the initial publications of 2007 – 2011 mentioned earlier. This research report was meant to summarize these R&D efforts as comprehensively, and also as succinctly, as possible. Research efforts related to the automated scoring of predictable speech also undertaken at ETS in these years (e.g., Zechner et al., 2015; Zechner & Xi, 2008) were not in the scope of this report, as these types of spoken responses have received more attention by other research groups and for a longer time. This report focuses only on the automated scoring of spontaneous speech, even more precisely, on speech elicited by TOEFL iBT items.

In what follows, we summarize where we currently are in the principal areas of the automated scoring of spontaneous speech and where we still need to make progress moving forward.

Automatic Speech Recognition

Although switching to a state-of-the-art ASR system in recent years has enabled us to achieve substantially lower WERs on spoken responses by test takers, recent developments in the field of speech science and ASR, for example, related to DNN technologies (Metallinou & Cheng, 2014), suggest that even more substantial improvements are possible in this area in the near term. Still, it needs to be pointed out that studies on human speech transcribers have shown that for nonnative spontaneous speech, it is difficult to reach agreement above 85% (Zechner, 2009); therefore WERs by automated speech recognition engines for spontaneous nonnative speech may bottom out at approximately 10% – 20%, compared to 5% or less for spontaneous native speech.

Speech Features

Comparing Version 1 and Version 5 of SpeechRater, the feature set was expanded from less than 50 to more than 100 features, specifically addressing spontaneous speech. More importantly, the features' construct coverage has been substantially increased to allow the automated scoring method to use rich information, as human raters do. As for the TOEFL iBT Speaking construct (as delineated in the TOEFL iBT Speaking rubrics), virtually all of the various subareas of the construct are currently addressed by some of the SpeechRater features. Still, the coverage of the topic development area is somewhat limited, and more features addressing more detailed aspects of content and discourse still need to be developed — not an insignificant challenge given that the WER of speech recognition is still substantial.

Filtering Model

Whereas the initial filtering model in SpeechRater only focused on detecting responses with no speech or responses with too much noise, we now have several additional components that can flag nonscorable responses, including detection of non-English speech or off-topic speech. All of these additional filtering components are important, in particular, if (and when) SpeechRater may be used in some way in the operational scoring of an assessment used to make high-stakes decisions, where test takers may be inclined to use certain strategies of gaming the system to inflate the scores they receive from the automated system. Developing additional and more effective filtering models is a high priority moving forward; one of the major challenges here is that very limited "real-life" data are available to train these models, and furthermore, it is not straightforward to anticipate all of the different ways test takers may try to artificially inflate their scores before automated scoring is actually introduced.

Scoring Model

In the first version of SpeechRater (as well as in several subsequent versions), scoring models were mostly handcrafted by content experts, using information about feature correlation with human scores, feature collinearity, feature normality, overall construct coverage, and so on, as input in their model design. Recently, as described in the section Hybrid Method of Feature Selection, we switched to a hybrid model of feature selection, whereby the initial step is automatic and determines a subset of the overall feature set to be used in the model. In a second step, a content expert further refines this feature set according to considerations similar to those listed earlier. This hybrid approach results in both improved empirical model performance and more balanced construct coverage, along with a substantial reduction in model-building time. Still, we are currently only using linear regression as the machine learning method to compute scores; research into more sophisticated and complex machine learning algorithms is ongoing and may yield scoring models with still improved empirical performance.

In summary, in this report, we have provided a comprehensive view of the major R&D efforts and results related to the automated scoring of spontaneous nonnative speech at ETS in the past 10 years, since the introduction of the first version of SpeechRater in 2006. R&D efforts are still continuing in all areas of this technology, and we may be nearing the time when our SpeechRater capability may be used, in conjunction with human raters, in more consequential assessments.

Acknowledgments

The authors thank previous members of the Speech Team and other collaborators at ETS for their contributions, including Kalliroi Georgila, Shasha Xie, Xiaoming Xi, and Pamela Mollaun. We thank many research assistants and external contractors for assisting with annotation work on multiple speech corpora. We also thank many previous summer interns for their hard work inside the Speech Team. Finally, we thank our colleagues at ETS for their valuable comments and suggestions on previous drafts of this report.

Notes

1 The code for extracting discourse coherence features had not been integrated into SpeechRater 5.0. Therefore we could not evaluate this group of features with other feature groups in the section Feature Correlations. Instead, we report the evaluation result presented in Wang et al. (2013) here.
2 TD stands for *technical difficulty*.
3 These responses are assigned a score of 0 by human raters.
4 The languages were English, Farsi, French, German, Japanese, Korean, Mandarin Chinese, Spanish, Tamil, and Vietnamese.
5 Owing to the extremely skewed distribution of NS responses (2% in the ASR set), it was not easy to train and evaluate the filtering model. To address this issue, we modified the distribution of NS responses in the FM set. Initially, we collected 90,000 responses, including 1,560 NS responses. While maintaining all NS responses, we downsampled the scorable responses in the FM set to include 10,000 responses. Finally, the proportion of NS responses was 6 times higher in the FM set (13%) than in the ASR set.
6 In addition to topicality problems, many different issues make responses nonscorable. A subset of SpeechRater features that measure spectrum characteristics, energy level, ASR confidence scores, and length of speech section were selected to detect various other issues, such as problems with audio quality and abberant behavior of the ASR system.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 3 – 30.

Barzilay, R., & Lapata, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd annual meeting of the ACL* (pp. 141 – 148). Stroudsburg, PA: Association for Computational Linguistics.

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics, 34*(1), 1 – 34. https://doi.org/10.1162/coli.2008.34.1.1

Barzilay, R., & Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. In D. M. Susan Dumais & S. Roukos (Eds.), *HLT-NAACL 2004: Main proceedings* (pp. 113 – 120). Stroudsburg, PA: Association for Computational Linguistics.

Beigman Klebanov, B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content importance models for scoring writing from sources. In *Proceedings of the 52nd annual meeting of the Association of Computational Linguistics* (pp. 247 – 252). Stroudsburg, PA: Association for Computational Linguistics.

Bernstein, J., Cheng, J., & Suzuki, M. (2010). Fluency and structural complexity as predictors of L2 oral proficiency. In T. Kobayashi, K. Hirose, & S. Nakamura (Eds.), *Proceedings of the ISCA Interspeech Conference* (pp. 1241 – 1244). Retrieved from http://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1241.pdf

Bernstein, J., Moere, A. V., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*, 355 – 377. https://doi.org/10.1177/0265532210364404

Bhat, S., & Yoon, S. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication, 67*, 42 – 57. https://doi.org/10.1016/j.specom.2014.09.005

Bock, K., & Levelt, M. (1994). Language production: Grammatical encoding. In M. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945 – 984). San Diego, CA: Academic Press.

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech, 8*, 148 – 158. https://doi.org/10.1177/002383096500800302

Broussard, K. M. (2001). Review of the technical report *Second language development in writing: Measures of fluency, accuracy and complexity* by K. Wolfe-Quintero, S, Inagaki, S., & H.-Y. Kim. *TESOL Quarterly, 35*, 342 – 343. https://doi.org/10.2307/3587656

Burstein, J., Tetreault, J., & Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics* (pp. 681 – 684). Stroudsburg, PA: Association for Computational Linguistics.

Chen, L., & Yoon, S.-Y. (2011). Detecting structural events for assessing non-native speech. In *6th workshop on innovative use of NLP for building educational applications* (pp. 38 – 45). Stroudsburg, PA: Association for Computational Linguistics.

Chen, L., & Yoon, S.-Y. (2012). Application of structural events detected on ASR outputs for automated speaking assessment. In *Interspeech 2012, 13th annual conference of the International Speech Communication Association, Portland, OR, USA, September 9 – 13, 2012* (pp. 767 – 770). Retrieved from http://www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_0767.pdf

Chen, L., & Zechner, K. (2011). Applying rhythm features to automatically assess non-native speech. In P. Cosi, R. D. Mori, G. D. Fabbrizio, & R. Pieraccini (Eds.), *Interspeech 2011, 12th annual conference of the International Speech Communication Association, Florence, Italy, August 27 – 31, 2011* (pp. 1861 – 1864). Retrieved from http://www.isca-speech.org/archive/archive_papers/interspeech_2011/i11_1861.pdf

Chen, L., Zechner, K., & Xi, X. (2009). Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of Human Language Technologies: The 2009 annual conference of the North American chapter of the Association for Computational Linguistics* (pp. 442 – 449). Stroudsburg, PA: Association for Computational Linguistics.

Chen, M., & Zechner, K. (2011, June). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 722 – 731). Stroudsburg, PA: Association for Computational Linguistics.

Cheng, J., Chen, X., & Metallinou, A. (2015). Deep neural network acoustic models for spoken assessment applications. *Speech Communication, 73*, 14 – 27. https://doi.org/10.1016/j.specom.2015.07.006

Cheng, J., & Shen, J. (2011). Off-topic detection in automated speech assessment applications. In P. Cosi, R. D. Mori, G. D. Fabbrizio, & R. Pieraccini (Eds.), *Interspeech 2011, 12th annual conference of the International Speech Communication Association, Florence, Italy, August 27 – 31, 2011* (pp. 1597 – 1600). Retrieved from http://www.isca-speech.org/archive/archive_papers/interspeech_2011/i11_1597.pdf

Cucchiarini, C., Nejjari, W., & Strik, H. (2014). My pronunciation coach: Computer-assisted English pronunciation training. In R. van den Doel & L. Rupp (Eds.), *Pronunciation matters: Accents of English in the Netherlands and elsewhere* (pp. 45 – 68). Amsterdam, the Netherlands: VU uitgeverij.

Cucchiarini, C., Strik, H., & Boves, L. (1997). Automatic evaluation of Dutch pronunciation by using speech recognition technology. In *1997 IEEE workshop on automatic speech recognition and understanding proceedings* (pp. 622 – 629). New York, NY: IEEE. https://doi.org/10.1109/ASRU.1997.659144

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America, 107*, 989 – 999. https://doi.org/10.1121/1.428279

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*, 2862 – 2873. https://doi.org/10.1121/1.1471894

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*, 197 – 222. https://doi.org/10.1093/applin/24.2.197

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ΔC. In P. Karnowski & I. Szigeti (Eds.), *Language and language processing* (pp. 231 – 241). Frankfurt, Germany: Peter Lang.

Dorans, N. J., & Cook, L. L. (2016). *Fairness in educational assessment and measurement*. New York, NY: Routledge.

Dorans, N. J., & Holland, P. W. (2000). *Population invariance and the equatability of tests: Basic theory and the linear case* (Research Report No. RR-00-19). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01842.x

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech Communication, 51*, 832 – 844. https://doi.org/10.1016/j.specom.2009.04.005

Eskenazi, M., Alwan, A., & Strik, H. (Eds.). (2009). Spoken language technology for education [Special issue] . *Speech Communication, 51*(10).

Evanini, K., Xie, S., & Zechner, K. (2013). Prompt-based content scoring for automated spoken language assessment. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 157 – 162). Stroudsburg, PA: Association for Computational Linguistics.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Textual coherence using latent semantic analysis. *Discourse Processes, 25*, 285 – 307. https://doi.org/10.1080/01638539809545029

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R.,Butzberger, J., ... Cesari, R. (2000). The SRI EduSpeak system: Recognition and pronunciation scoring for language learning. In P. Delcloque (Ed.), *Proceedings of InSTiLL (intelligent speech technology in language learning)* (pp. 123 – 128). Dundee, Scotland: University of Abertay.

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing, 27*, 401 – 418. https://doi.org/10.1177/0265532210364408

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal, 52*(1), 70 – 84.

Grabe, E., & Low, E. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology, 7*, 515 – 546. https://doi.org/10.1515/9783110197105.515

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, 36*, 193 – 202. https://doi.org/10.3758/BF03195564

Graff, D., Garofolo, J., Fiscus, J., Fisher, W., & Pallett, D. (1997). *1996 English Broadcast News Speech (HUB4) LDC97S44*. Philadelphia, PA: Linguistic Data Consortium.

Hacker, C., Cincarek, T., Gruhn, R., Steidl, S., Nöth, E., & Niemann, H. (2005). Pronunciation feature extraction. In W. G. Kropatsch, R. Sablatnig, & A. Hanbury (Eds.), *Pattern recognition: 27th DAGM symposium, Vienna, Austria, August 31 – September 2, 2005, Proceedings* (pp. 141 – 148). Berlin, Germany: Springer.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witte, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1), 10 – 18. https://doi.org/10.1145/1656274.1656278

Hassanali, K., Liu, Y., & Solorio, T. (2012). Coherence in child language narratives: A case study of annotation and automatic prediction of coherence. In *WOCCI 2014 Workshop on Child – Computer Interaction* (pp. 7 – 12). Retrieved from http://www.isca-speech.org/archive/wocci_2012/papers/wc12_007.pdf

He, L. (2012). Syllabic intensity variations as quantification of speech rhythm: Evidence from both L1 and L2. In Q. Ma, H. Ding, & D. Hirst (Eds.), *Proceedings of the 6th International Conference on Speech Prosody* (pp. 466 – 469). Shanghai, China: Tongji University Press.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004, May 2 – 7). Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL 2004: Human language technology conference of the North American Chapter of the Association for Computational Linguistics* (pp. 185 – 192). Stroudsburg, PA: Association for Computational Linguistics.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language, 25*, 282 – 306. https://doi.org/10.1016/j.csl.2010.06.001

Hoad, T. C., & Zobel, J. (2003). Methods for identifying versioned and plagiarized documents. *Journal of the American Society for Information Science and Technology, 54*, 203 – 215. https://doi.org/10.1002/asi.10170

Hu, W., Qian, Y., & Soong, F. K. (2014). A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training. In *ICASSP, IEEE international conference on acoustics, speech and signal processing, proceedings* (pp. 3206–3210). New York, NY: IEEE. https://doi.org/10.1109/ICASSP.2014.6854192

Hu, W., Qian, Y., Soong, F. K., & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication, 67*, 154–166. https://doi.org/10.1016/j.specom.2014.12.008

Iwashita, N. (2010). Features of oral proficiency in task performance by EFL and JFL learners. In *Selected proceedings of the 2008 Second Language Research Forum* (pp. 32–47). Somerville, MA: Cascadilla Proceedings Project.

Iwashita, N., Brown, A., Mcnamara, T., & Hagan, S. O. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Jang, T.-Y. (2009). Automatic assessment of non-native prosody using rhythm metrics: Focusing on Korean speakers' English pronunciation. In J. Brooke, G. Coppola, E. Görgülü, M. Mameni, E. Mileva, S. Morton, & A. Rimrott (Eds.), *Simon Fraser University working papers in linguistics: Vol. 2: Proceedings of the 2nd International Conference on East Asian Linguistics*. Retrieved from https://www.sfu.ca/content/dam/sfu/linguistics/Gradlings/SFUWPL/Jang_T.pdf

Klein, D., & Manning, C. D. (2003, July). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics* (pp. 423–430). Stroudsburg, PA: Association for Computational Linguistics.

Lai, C., Evanini, K., & Zechner, K. (2013). Applying rhythm metrics to non-native spontaneous speech. In P. Badin, T. Hueber, G. Bailly, D. Demolin, & F. Raby (Eds.), *Proceedings of the ISCA workshop on speech and language technology in education (SLaTE)* (pp. 159–163). Retrieved from http://www.isca-speech.org/archive/slate_2013/papers/sl13_159.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307–322. https://doi.org/10.1093/applin/16.3.307

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*, 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Lin, C.-Y., & Rey, M. (2004). ROUGE: A package for automatic evaluation of summaries. In S. Szpakowicz & M.-F. Moens (Eds.), *Text summarization branches out: Proceedings of the ACL-04 Workshop* (pp. 74—81). Stroudsburg, PA: Association for Computational Linguistics.

Liu, Y. (2004). *Structural event detection for rich transcription of speech* (Unpublished doctoral dissertation). Purdue University, West Lafayette, IN.

Lo, W.-K., Harrison, A. M., & Meng, H. (2010). Statistical phone duration modeling to filter for intact utterances in a computer-assisted pronunciation training system. In *ICASSP, IEEE international conference on acoustics, speech and signal processing, proceedings* (pp. 5238–5241). New York, NY: IEEE. https://doi.org/10.1109/ICASSP.2010.5494988

Loukina, A., Kochanski, G., Rosner, B., & Keane, E. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America, 129*, 3258–3270. https://doi.org/10.1121/1.3559709

Loukina, A., Zechner, K., & Chen, L. (2014). Automatic evaluation of spoken summaries: The case of language assessment. In *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 68–78). Stroudsburg, PA: Association for Computational Linguistics.

Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. In *Proceedings of the 10th workshop on innovative use of NLP for building educational applications* (pp. 12–19). Stroudsberg, PA: Association for Computational Linguistics.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*, 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Metallinou, A., & Cheng, J. (2014). Using deep neural networks to improve proficiency assessment for children English language learners. In H. Li, H. Meng, B. Ma, E. S. Chng,& L. Xie (Eds.), *INTERSPEECH 2014, Proceedings of the 15th annual conference of the International Speech Communication Association: Celebrating the diversity of spoken languages* (pp. 1468–1472). Retrieved from http://www.isca-speech.org/archive/interspeech_2014/i14_1468.html

Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A., & Zobel, J. (2005). Similarity measures for tracking information flow. In *CIKM '05: Proceedings of the 14th ACM international conference on information and knowledge management* (pp. 517–524). New York, NY: ACM. https://doi.org/10.1145/1099554.1099695

Mizera, G. J. (2006). *Working memory and L2 oral fluency* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Mok, P., & Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the Speech Prosody 2008 conference* (pp. 423–426). Retrieved from http://www.isca-speech.org/archive/sp2008/papers/sp08_423.pdf

Moustroufas, N., & Digalakis, V. (2007). Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech & Language, 21*, 219 – 230. https://doi.org/10.1016/j.csl.2006.04.001

Muthusamy, Y. K., Cole, R. A., & Oshika, B. T. (1992). The OGI multi-language telephone speech corpus. In *Proceedings of the second international conference on spoken language processing (ICSLP'92)* (pp. 895 – 898). Retrieved from http://www.isca-speech.org/archive/archive_papers/icslp_1992/i92_0895.pdf

Nava, E., Tepperman, J., Goldstein, L., Zubizarreta, M., & Narayanan, S. (2009). Connecting rhythm and prominence in automatic ESL pronunciation scoring. In M. Uther, R. Moore, & S. Cox (Eds.), *Interspeech 2009 — 10th annual conference of the International Speech Communication Association* (pp. 684 – 687). Retrieved from http://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_0684.pdf

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication, 6*, 83 – 93. https://doi.org/10.1016/S0167-6393(99)00046-1

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*, 492 – 518. https://doi.org/10.1093/applin/24.4.492

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311 – 318). Stroudsburg, PA: Association for Computational Linguistics.

Pitler, E., Louis, A., & Nenkova, A. (2010, July). Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 544 – 554). Stroudsburg, PA: Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, & D. Tapias (Eds.), *Proceedings of the sixth international conference on language resources and evaluation (LREC'08)* (pp. 2961 – 2968). Marrakech, Morocco: European Language Resources Association.

Qian, X., Meng, H., & Soong, F. (2012). The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. In *Interspeech 2012: 13th annual conference of the International Speech Communication Association* (pp. 775 – 778). Retrieved from http://www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_0775.pdf

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02158.x

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(1), 25 – 39. https://doi.org/10.1016/j.asw.2012.10.004

Ramus, F., Nespor, M., & Mehler, J. (2000). Correlates of linguistic rhythm in the speech signal. *Cognition, 75*, 265 – 292. https://doi.org/10.1016/S0010-0277(00)00101-3

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14*, 423 – 441. https://doi.org/10.1080/01638539109544795

Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S., Woodland, P., & Young, S. (1995). *WSJCAM0 Cambridge Read News LDC95S24*. Philadelphia, PA: Linguistic Data Consortium.

Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*, 613 – 620. https://doi.org/10.1145/361219.361220

Sanderson, M. (1997). *Duplicate detection in the Reuters collection* (Technical Report No. TR-1997-5). Glasgow, UK: University of Glasgow.

Selouani, S.-A., Alotaibi, Y., Cichocki, W., Gharsellaoui, S., & Kadi, K. (2015). Native and nonnative class discrimination using speech rhythm- and auditory-based cues. *Computer Speech & Language, 31*, 28 – 48. https://doi.org/10.1016/j.csl.2014.11.003

Strik, H., & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication, 29*, 225 – 246. https://doi.org/10.1016/S0167-6393(99)00038-2

Strik, H., Palumbo, L., & de Wet, F. (2015). Web-based mini-games for language learning that support spoken interaction. In S. Steidl, A. Batliner, & O. Jokisch (Eds.), *SLaTE 2015 workshop on speech and language technology in education* (pp. 137 – 142). Retrieved from http://www.isca-speech.org/archive/slate_2015/papers/sl15_137.pdf

Strik, H., Truong, K., de Wet, F. D., & Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication, 51*, 845 – 852. https://doi.org/10.1016/j.specom.2009.05.007

Tao, J., Evanini, K., & Wang, X. (2014). The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system. In *IEEE spoken language technology workshop (SLT)* (pp. 294 – 299). New York, NY: IEEE. https://doi.org/https://doi.org/10.1109/SLT.2014.7078590

Temple, L. (2000). Second language learner speech production. *Studia Linguistica, 54*, 288 – 297. https://doi.org/10.1111/1467-9582.00068

Tepperman, J., Stanley, T., Hacioglu, K., & Pellom, B. (2010, May). *Testing suprasegmental English through parroting*. Paper presented at Speech Prosody, Chicago, IL.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Tortel, A., & Hirst, D. (2010, May). *Rhythm metrics and the production of English L1/L2*. Paper presented at Speech Prosody, Chicago, IL.

van Dalen, R., Knill, K., & Gales, M. J. F. (2015). Automatically grading learners' English using a Gaussian process. In S. Steidl, A. Batliner, & O. Jokisch (Eds.), *SLaTE 2015 workshop on speech and language technology in education* (pp. 7–12). Retrieved from http://www.isca-speech.org/archive/slate_2015/papers/sl15_007.pdf

van Doremalen, J., Strik, H., & Cucchiarini, C. (2009). Utterance verification in language learning applications. In *ISCA international workshop on speech and language technology in education (SLaTE 2009)* (pp. 13–16). Retrieved from http://www.isca-speech.org/archive/slate_2009/papers/sla9_013.pdf

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65–83. https://doi.org/10.1177/026553220001700103

Wagner, E., & Kunnan, A. J. (2015). The Duolingo English test. *Language Assessment Quarterly, 12*, 320–331. https://doi.org/10.1080/15434303.2015.1061530

Wang, X., Evanini, K., & Zechner, K. (2013). Coherence modeling for the automated assessment of spontaneous spoken responses. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 814–819). Stroudsburg, PA: Association for Computational Linguistics.

White, L., & Mattys, S. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics, 35*, 501–522. https://doi.org/10.1016/j.wocn.2007.02.003

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language, 66*, 665–679. https://doi.org/10.1016/j.jml.2011.12.010

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Witt, S. M. (1999). *Use of speech recognition in computer-assisted language learning* (Unpublished doctoral dissertation). University of Cambridge, Cambridge, England.

Xi, X. (2007). Evaluating analytic scoring for the TOEFLR academic speaking test (TAST) for operational use. *Language Testing, 24*, 251–286. https://doi.org/10.1177/0265532207076365

Xi, X., Higgins, D., & Zechner, K. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (Research Report No. RR-08-62). Princeton, NJ: Educational Testing Service. Retrieved from https://doi.org/10.1002/j.2333-8504.2008.tb02148.x

Xie, S., Evanini, K., & Zechner, K. (2012). Exploring content features for automated speech scoring. In *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 103–111). Stroudsburg, PA: Association for Computational Linguistics.

Yannakoudakis, H., & Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In *The 7th workshop on the innovative use of NLP for building educational applications* (pp. 33–43). Stroudsburg, PA: Association for Computational Linguistics.

Yoon, S.-Y., & Bhat, S. (2012). Assessment of ESL learners' syntactic competence based on similarity measures. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 600–608). Stroudsburg, PA: Association for Computational Linguistics.

Yoon, S.-Y., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the seventh workshop on building educational applications using NLP* (pp. 180–189). Stroudsburg, PA: Association for Computational Linguistics.

Yoon, S.-Y., & Higgins, D. (2011). Non-English response detection method for automated proficiency scoring system. In *Proceedings of the 6th workshop on innovative use of NLP for building educational applications* (pp. 161–169). Stroudsburg, PA: Association for Computational Linguistics.

Zechner, K. (2009). What did they actually say? Agreement and disagreement among transcribers of non-native spontaneous speech responses in an English proficiency test. In *ISCA international workshop on speech and language technology in education (SLaTE 2009)* (pp. 25–28). Retrieved from http://www.isca-speech.org/archive/slate_2009/papers/sla9_025.pdf

Zechner, K., & Bejar, I. (2006, June). Towards automatic scoring of non-native spontaneous speech. In *Proceedings of the Human Language Technology Conference of the NAACL, main conference* (pp. 216–223). Stroudsburg, PA: Association for Computational Linguistics.

Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C. M., Leong, C. W., Wang, X., & Yoon, S.-Y. (2015). *Automated scoring of speaking tasks in the Test of English-for-Teaching (TEFT)* (Research Report No. RR-15-31). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12080

Zechner, K., Higgins, D., & Xi, X. (2007). SpeechRater: A construct-driven approach to scoring spontaneous non-native speech. In *Proceedings of the SLaTE workshop on speech and language technology in education* (pp. 128–131). Retrieved from http://www.isca-speech.org/archive_open/archive_papers/slate_2007/sle7_128.pdf

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. https://doi.org/10.1016/j.specom.2009.04.009

Zechner, K., & Xi, X. (2008). Towards automatic scoring of a test of spoken language with heterogeneous task types. In *Proceedings of the third workshop on innovative use of NLP for building educational applications* (pp. 98 – 106). Stroudsburg, PA: Association for Computational Linguistics.

Zhang, M., Dorans, N., Li, C. & Rupp, A. (2017). Differential feature functioning in automated essay scoring. In H. Jiao & R. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (p. 185 – 208). Charlotte, NC: Information Age.

Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing, 4*, 31 – 44. https://doi.org/10.1109/TSA.1996.481450

Suggested citation:

Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., … Gyawali, B. (2018). *Automated scoring of nonnative speech using the SpeechRater*^{SM} *v. 5.0 engine* (Research Report No. RR-18-10). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12198

Action Editor: Beata Beigman Klebanov

Reviewers: Aoife Cahill and Mo Zhang