Automated Scoring of Speaking Items in an Assessment for Teachers of English as a Foreign Language

Klaus Zechner, Keelan Evanini, Su-Youn Yoon, Lawrence Davis,
Xinhao Wang, Lei Chen, Chong Min Lee, Chee Wee Leong
Educational Testing Service (ETS)
Princeton, NJ 08541, USA
{kzechner,kevanini,syoon,ldavis,xwang002,lchen,clee001,cleong}@ets.org

Abstract

This paper describes an end-to-end prototype system for automated scoring of spoken responses in a novel assessment for teachers of English as a Foreign Language who are not native speakers of English. The 21 speaking items contained in the assessment elicit both restricted and moderately restricted responses, and their aim is to assess the essential speaking skills that English teachers need in order to be effective communicators in their classrooms. Our system consists of a state-of-the-art automatic speech recognizer; multiple feature generation modules addressing diverse aspects of speaking proficiency, such as fluency, pronunciation, prosody, grammatical accuracy, and content accuracy; a filter that identifies and flags problematic responses; and linear regression models that predict response scores based on subsets of the features. The automated speech scoring system was trained and evaluated on a data set involving about 1,400 test takers, and achieved a speaker-level correlation (when scores for all 21 responses of a speaker are aggregated) with human expert scores of 0.73.

1 Introduction

As English has become increasingly important as a language of international business, trade, science, and communication, efforts to promote teaching English as a Foreign Language (EFL) have seen substantially more emphasis in many non-English-speaking countries worldwide in recent years. In addition, the prevailing trend in English pedagogy has been to promote the use of spoken English in the classroom, as opposed to the respective native languages of the EFL learners. However, due to the high demand for EFL teachers in many countries, the training of these teachers has not always caught up with these high expectations, so there is a need for both governmental and private institutions involved in the employment and training of EFL teachers to assess their competence in the English language, as well as in English pedagogy.

Against this background, we developed a language assessment for EFL teachers who are not native speakers of English that addresses the four basic English language skills of Reading, Listening, Writing and Speaking. This paper focuses only on the speaking portion of the English assessment, and, in particular, on the system that we developed to automatically compute scores for test takers' spoken responses.

Several significant challenges needed to be addressed during the course of building this automated speech scoring system, including, but not limited to:

- The 21 Speaking items belong to 8 different task types with different characteristics; therefore, we had to select features and build scoring models for each task type separately.
- The test takers speak a variety of native languages, and thus have very different non-native accents in their spoken English. Furthermore, the test takers also exhibit a wide range of speaking proficiency levels, which contributes to the diversity of their spoken responses. Our speech recognizer therefore had to be trained and adapted to a large database of non-native speech.
- Since content accuracy is very important for the types of tasks contained in the test, even small error rates by the automatic speech recognition (ASR) system can lead to a noticeable impact on feature performance. This fact motivated the development of a set of

features that are robust to speech recognition errors.

- A significant amount of responses (more than 7%) exhibit issues that make them hard or impossible to score automatically, e.g., high noise levels, background speech, etc. We therefore implemented a filter to identify these non-scorable responses automatically.

The paper is organized as follows: Section 2 discusses related work; in Section 3, we present the data used for system training and evaluation; Section 4 describes the system architecture of the automated speech scoring system. We detail the methods we used to build our system in Section 5, followed by an overview of the results in Section 6. Section 7 discusses our findings; finally, Section 8 concludes the paper.

2 Related Work

Automated speech processing and scoring technology has been applied to a variety of domains over the course of the past two decades, including evaluation and tutoring of children's literacy skills (Mostow et al., 1994), preparation for high stakes English proficiency tests for institutions of higher education (Zechner et al., 2009), evaluation of English skills of foreign-based call center agents (Chandel et al., 2007), and evaluation of aviation English (Pearson Education, Inc., 2011), to name a few (for a comprehensive overview, see (Eskenazi, 2009)).

Most of these applications elicit restricted speech from the participants, and the most common item type by far is the Read Aloud, in which the speaker reads a sentence or collection of sentences out loud. Due to the constrained nature of this task, it is possible to develop ASR systems that are relatively accurate, even with heavily accented non-native speech. Several types of features related to a non-native speaker's ability to produce English sounds and speech patterns effectively have been extracted from these types of responses. Some of the best performing of these types of features include pronunciation features, such as a phone's spectral match to native speaker acoustic models (Witt, 1999) and a phone's duration compared to native speaker models (Neumeyer et al., 2000); fluency features, such as the rate of speech, mean pause length, and number of disfluencies (Cucchiarini et al., 2000); and prosody features, such as F0 and intensity slope (Hoenig, 2002).

In addition to the large majority of applications that elicit restricted speech, a small number of applications have also investigated automated scoring of non-native spontaneous speech, in order to more fully evaluate a speaker's communicative competence (e.g., (Cucchiarini et al., 2002) and (Zechner et al., 2009)). In these systems, the same types of pronunciation, fluency, and prosody features can be extracted; furthermore, features related to additional aspects of a speaker's proficiency in the non-native language can be extracted, such as vocabulary usage (Yoon et al., 2012), syntactic complexity (Bernstein et al., 2010a; Chen and Zechner, 2011), and topical content (Xie et al., 2012).

As described in Section 1, the domain for the automated speaking assessment investigated in this study is teachers of EFL around the world. Based on the fact that many of the item types are designed to assess the test taker's ability to productively use English constructions and linguistic units that commonly recur in English teaching environments, several of the item types elicit semi-restricted speech (see Table 1 below for a description of the different item types). These types of responses fall somewhere between the heavily restricted speech elicited by a Read Aloud task and unconstrained spontaneous speech. In these semi-restricted responses, the test taker may be provided with a set of lexical items that should be used to form a sentence; in addition, the test taker is often asked to make the sentence conform to a given grammatical template. Thus, the responses provided for a given prompt of this type by multiple different speakers will often overlap with each other; however, it is not possible to specify a complete list of all possible responses. These types of items have only infrequently been examined in the context of automated speech scoring. Some related item types that have been explored previously include the Sentence Build and Short item types described in (Bernstein et al., 2010b); however, those item types typically elicited a much narrower range of responses than the semi-restricted ones in this study.

3 Data

The data used in this study was drawn from a pilot administration of a language assessment for teachers of English as a Foreign Language. This test is designed to assess the ability of a non-native teacher of English to use English in classroom settings. The language forms and functions included in this test are based on the materials included in a curriculum that the test takers studied prior to taking the assessment. The assessment includes items that cover the four language skills: Reading, Listening, Writing, and Speaking. There are a total of 8 different types of Speaking items included in the assessment. These can be divided into the following two categories, depending on how constrained the test taker's response is:

- *Restricted Speech*: In these item types, all of the linguistic content expected in the test taker's response is presented in the test prompt, and the test taker is asked to read or repeat it aloud.

- *Semi-restricted Speech*: In these item types, a portion of the linguistic content is presented in the prompt, and the test taker is required to provide the remaining content to formulate a complete response.

Sets of 7 Speaking items are presented to the test taker in thematic units, called "lessons", based on their instructional goals; in total, each test taker completed three lessons, and thus responded to 21 Speaking items. Table 1 presents descriptions of the 8 different item types included in the assessment.

The numbers of responses provided by the test takers to each type (along with their respective response durations) are as follows: four Multiple Choice (10 seconds each), six Read Aloud (four 40 second responses and two 60 second responses), two Repeat Aloud (15 seconds each), one Incomplete Sentence (20 seconds), one Key Words (15 seconds), five Chart (four 20 seconds and one 40 seconds), one Keyword Chart (15 seconds), and one Visuals (15 seconds). Thus, each test taker provided a total of approximately 9 minutes of audio.

The responses were all double-scored by trained human raters on a three-point scale (1 - 3). For the Restricted Speech items, the raters assessed the test taker's pronunciation, pacing, and intonation. For the Semi-restricted Speech items, the responses were also scored holistically on a 3-point scale, but raters were also asked to take into account the appropriateness of the language used (e.g., grammatical accuracy and content correctness) in addition to aspects of fluency and pronunciation. For some responses, the raters were not able to provide a score on the 1 - 3 scale, e.g., because the audio response contained no speech input, the test taker responded in their native language, etc. These responses are labeled NS for Non-Scoreable.

Restricted Speech	
Type	Description
Multiple Choice (MC)	The test taker selects the correct option and reads it aloud
Read Aloud (RA)	The test taker reads aloud a set of classroom instructions
Repeat Aloud (RP)	The test taker listens to a student utterance twice and then repeats it
Semi-restricted Speech	
Type	Description
Incomplete Sentence (IS)	The test taker is given a sentence fragment and completes the sentence according to the instructions
Key Words (KW)	The test taker uses the key words provided to speak a sentence as instructed
Chart (CH)	The test taker uses an example from a language chart and then formulates a similar sentence using a given grammatical pattern
Keyword Chart (KC)	The test taker constructs a sentence using keywords provided and information in a chart
Visuals (VI)	The test taker is given two visuals and is asked to give instructions to students based on the graphical information

Table 1: Types of speaking items included in the assessment

After receiving scores, all of the responses were transcribed using standard English orthography (disfluencies, such as filled pauses and partial words are also included in the transcriptions). Then, the responses were partitioned (with no speaker overlap) into five sets for the training and evaluation of the ASR system and the linear regression scoring models. The amount of data and human score distributions in each of these partitions are displayed in Table 2.

4 System Architecture

The automated scoring system used for the teachers' spoken language assessment consists of the following four components, which are invoked one after the other in a pipeline fashion (ETS SpeechRater^{SM}, (Zechner et al., 2009; Higgins et al., 2011)):

- an automated speech recognizer, generating word hypotheses from input audio recordings of the test takers' responses
- a feature computation module that generates features based on the ASR output, e.g., measuring fluency, pronunciation, prosody, and content accuracy
- a filtering model that flags responses that should not be scored automatically due to issues with audio quality, empty responses, etc.
- linear regression scoring models that predict the score for each response based on a set of selected features

Furthermore, we use Praat (Boersma and Weenick, 2012) to extract power and pitch from the speech signal; this information is used for some of the feature computation modules, as well as for the filtering model.

The ASR is an HMM-based triphone system trained on approximately 800 hours of non-native speech from a different data set; a background Language Model (LM) was also trained on the same data set. Subsequently, 8 adapted LMs were trained (with an interpolation weight of 0.9 for the in-domain data) using the responses in the ASR Training partition for the 8 different item types listed in Table 1. The ASR system obtained an overall word error rate (WER) of 13.0% on the ASR Evaluation partition and 15.6% on the Model Evaluation partition. As would be expected, the ASR system performed best on the responses that were most restricted by the test item and performed worse on the responses that were less restricted. The WER ranged from 11.4% for the RA responses to 41.4% for the IS responses in the Model Evaluation partition.

5 Methodology

5.1 Speech features

The feature computation components of our speech scoring system compute more than 100 features based on a speaker's response. They belong to the following broad dimensions of speaking proficiency: fluency, pronunciation, prosody, vocabulary usage, grammatical complexity and accuracy, and content accuracy (Zechner et al., 2009; Chen and Yoon, 2012; Chen et al., 2009; Zechner et al., 2011; Yoon et al., 2012; Yoon and Bhat, 2012; Zechner and Wang, 2013).

After initial feature generation, we selected a set of about 10 features for each of the 8 item types, based on the following considerations[1] (Zechner et al., 2009; Xi et al., 2008):

- empirical performance, i.e., feature correlation with human scores
- construct[2] relevance, i.e., to what extent the feature measures aspects of speaking proficiency that are considered to be relevant and important by content experts
- overall construct coverage, i.e., the feature set should include features from all relevant construct dimensions
- feature independence, i.e., the intercorrelation between any two features of the set should be low

Furthermore, some features were transformed (e.g., by applying the inverse or log function), in order to increase the normality of their distributions (an assumption of linear regression classifiers). All feature values that exceeded a threshold of 4 standard deviations from the mean were replaced by the respective threshold (outlier truncation).

The composition of feature sets is slightly different for the two item type categories: for the 3 restricted item types, features related to fluency, pronunciation, prosody and read/repeat accuracy were chosen, whereas for the 5 semi-restricted item types, vocabulary and grammar features were also added to the set. Further, while accuracy

[1] While automated feature selection is conceivable in principle, in our experience it typically does not result in a feature set that meets all of these criteria well.

[2] A construct is the set of knowledge, skills, and abilities measured by a test.

Partition	Spk.	Resp.	Dur.	1	2	3	NS
ASR Training	773	16,049	116.7	1,587 (9.9)	4,086 (25.5)	8,796 (54.8)	1,580 (9.8)
ASR Development	25	525	3.8	53 (10.1)	133 (25.3)	327 (62.3)	12 (2.3)
ASR Evaluation	25	525	3.8	31 (5.9)	114 (21.7)	326 (62.1)	54 (10.3)
Model Training	300	6,300	45.8	675 (10.7)	1,715 (27.2)	3,577 (56.8)	333 (5.3)
Model Evaluation	300	6,300	45.7	647 (10.3)	1,637 (26.0)	3,487 (55.3)	529 (8.4)
Total	1,423	29,699	215.8	2,993 (9.38)	7,685 (25.14)	16,513 (58.26)	2,508 (7.22)

Table 2: Amount of data contained in each partition (speakers, responses, hours of speech) and distribution of human scores (percentages of scores per partition in brackets).

features for the restricted items were based only on string alignment measures, content accuracy features for the semi-restricted items were more diverse, e.g., based on regular expressions, keywords, and language model scores (Zechner and Wang, 2013). Table 3 lists the features that were used in the scoring models for restricted and semi-restricted item types, along with sub-constructs they measure and their description.

5.2 Filtering model

In order to automatically identify responses that have technical issues (e.g., loud background noise) or are otherwise not scorable (e.g., empty responses), a decision tree-based filtering model was developed using a combination of features derived from ASR output and from pitch and energy information (Yoon et al., 2011; Jeon and Yoon, 2012). The filtering model was tested on the scoring model evaluation data, and obtained an accuracy rate (the exact agreement between the filtering model and a human rater concerning the distinction between scorable and non-scorable responses) of 97%; it correctly identified 90% of the non-scorable responses in the data set with a false positive rate of 21% (recall=0.90, precision=0.79, F-score=0.84).

5.3 Scoring models

We used the Model Training set to train 8 linear regression models for the 8 different item types, using the previously determined feature sets. We used the features as independent variables in these models and the summed scores of two human raters as the dependent variable. These trained scoring models were then employed to score responses of the Model Evaluation data (excluding responses marked as non-scorable by human raters) and rounded to the nearest integer to predict the final scores for each response. These scores were then evaluated against the first human rater score (H1).

Item	N	S-H1	H1-H2	WER (%)
RA	1653	0.34	0.51	11.4
RP	543	0.41	0.73	21.8
MC	1036	0.67	0.83	17.1
CH	1372	0.44	0.67	26.3
KW	275	0.45	0.67	28.7
KC	274	0.57	0.74	28.8
IS	260	0.46	0.69	41.4
VI	272	0.43	0.80	30.4

Table 4: Correlations between system and first human rater (S-H1) and between two human raters (H1-H2), for all responses of each item type in the Model Evaluation partition (N). The last column provides the average ASR word error rate (WER) in percent.

Additionally, for responses flagged as non-scorable by the automatic filtering model, the second human rater score (H2) was used as final item score in order to mimic the operational scenario where human raters score responses that are flagged by the filtering model.

We also compute the agreement between system and human raters based on a set of all 21 responses of a speaker. Score imputation was used for responses that were labeled as non-scorable by both the system and H2; in this case, the response was given the mean score of the total scorable responses from the same speaker. Similarly, the same score imputation rule was applied to the H1 scores.

6 Results

Table 4 presents the Pearson correlation coefficients between human and automated scores for the responses from the 8 different item types along with the human-human correlation for each item type. Furthermore, we also provide the word error rates of the ASR system for the same 8 item types in the last column of the table.

Feature	Sub-construct	Description
Content_Ed1	Read/repeat accuracy / Fluency	Correctly read words per minute
Content_Ed2	Read/repeat accuracy	Read/repeat word error rate
Content_RegEx	Content accuracy	Matching of regular expressions
Content_WER	Content accuracy	Response discrepancy from high scoring responses
Content_NGram	Content accuracy	N-grams in response matching high scoring response n-grams
Fluency_Rate	Fluency	Speaking rate
Fluency_Chunk	Fluency	Average length of contiguous word chunks
Fluency_Sil1	Fluency	Frequency of long silences
Fluency_Sil2	Fluency / Grammar	Proportion of long within-clause-silences to all within-clause-silences
Fluency_Sil3	Fluency	Mean length of silences within a clause
Fluency_Disfl1	Fluency	Frequency of interruption points (repair, repetition, false start)
Fluency_Disfl2	Fluency	Number of disfluencies per second
Fluency_Disfl3	Fluency	Frequency of repetitions
Pron_Vowels	Pronunciation	Average vowel duration differences relative to a native-speaker model
Prosody1	Prosody	Percentage of stressed syllables
Prosody2	Prosody	Mean deviation of time intervals between stressed syllables
Prosody3	Prosody	Mean distance between stressed syllables
Vocab1	Vocabulary / Fluency	Number of word types divided by utterance duration
Grammar_POS	Grammar	Part-of-speech based distributional similarity score between a response and responses with different score levels
Grammar_LM	Grammar	Global language model score (normalized by response length)

Table 3: List of features used for item type scoring models, with the sub-constructs they represent and descriptions.

Comparison	Pearson r
S-H1	0.725
S-H2	0.742
H1-H2	0.934

Table 5: Speaker-level performance (Pearson r correlations) computed over the sum of all 21 scores from each speaker, N=272

Sub-construct	Restricted	Semi-restricted
Content	0.33–0.67	0.34–0.61
Fluency	0.19–0.33	0.20–0.33
Pronunciation	0.20–0.22	0.13–0.31
Prosody	0.18–0.24	0.12–0.27
Grammar	–	0.23–0.49
Vocabulary	–	0.21–0.32

Table 6: Range of Pearson r correlations for different features with human scores (H1) by sub-construct for restricted and semi-restricted item types.

Table 5 presents the Pearson correlation coefficients between the speaker-level scores produced by the automated scoring system (S) and the two sets of human scores (H1 and H2). These speaker-level scores were computed based on the sum of all 21 scores from each speaker in the Model Evaluation partition. Responses that received a non-scorable rating from the human raters were imputed, as described above. Furthermore, 28 speakers were excluded from this analysis because they had more than 7 non-scorable responses each.[3]

Finally, Table 6 provides an overview of Pearson correlation ranges with human rater scores (H1) for the different features used in the scoring models, summarized by the sub-constructs that the features represent.

7 Discussion

When looking at Table 4, we see that the inter-rater reliability for human raters ranges between 0.51 (for RA items) and 0.83 (for MC items). Inter-rater reliability varies less for the 5 semi-restricted item types (0.67–0.80), compared to the 3 restricted item types (0.51–0.83). As for automated score correlations with human raters, the Pearson r coefficients range from 0.34 (RA) to 0.67 (MC). Again, the variability of Pearson r coefficients is larger for the 3 restricted item types (0.34–0.67) than for the 5 semi-restricted item types (0.43–0.57). The degradation in correlation between the inter-human results and the machine-human results varies from 0.16 (MC) to 0.37 (VI).

Speech recognition word error rate does not seem to have a strong influence on model performance (RA items have the lowest WER with S-H1 r=0.34, but r=0.46 for IS items that have the highest WER). However, we found other factors that affect model performance negatively; for example, multiple repeats of responses by test takers contribute to the large performance difference between S-H1 and H1-H2 for the RP items. In general, we conjecture that using features for a larger set of sub-construct areas—in the case of semi-restricted item types—may contribute to the lower variation of scoring model performance for this subset of the data.

As for speaker-level results (Table 5), the overall degradation between the inter-human correlation and the system-human correlations is of a similar magnitude (around 0.2) as observed for most of the individual item types. Still, the speaker-level correlation of 0.73 is 0.26 higher than the average item type correlation between the system and H1.

When we look into more detail at the Pearson r correlations between individual features used in the item type scoring models and human scores (Table 6), we can see that features related to content accuracy exhibit a substantially stronger performance (r=0.33–0.67) than features related to most other sub-constructs of speaking proficiency, namely fluency, pronunciation, prosody, and vocabulary ($r \sim 0.2$). One exception is features related to grammar, where correlations with human scores are as high as 0.49. Since related work on scoring speech using features indicative of fluency, pronunciation, etc. showed higher correlations (e.g., (Cucchiarini et al., 1997; Franco et al., 2000; Zechner et al., 2009)), we conjecture that the reason behind this difference is likely to be found in the fact that the responses in this assessment for teachers of English are quite short (6–14 words on average for all items except for Read Aloud items that are about 46 words on average). Since content features are less reliant on longer stretches of speech, they still work fairly well for most items in our corpus.

[3] In an operational setting, these test takers would not receive a test score; instead, they would have the opportunity to take the test again.

Finally, while the proportion of words contained in responses in restricted items is much larger than those contained in responses in semi-restricted items, these two item type categories are more evenly distributed over the whole test, i.e., each test taker responds to 9 semi-restricted and 12 restricted items, and the item scores are then aggregated for a final score with equal weight given to each item score.

8 Conclusion

This paper presented an overview of an automated speech scoring system that was developed for a language assessment for teachers of English as a Foreign Language (EFL) whose native language is not English. We described the main components of this prototype system and their performance: the ASR system, features generated from ASR output, a filtering model to flag non-scorable responses, and finally a set of linear regression models, one for each of 8 different types of test items.

We found that overall, the correlation between our speech scoring system's predicted scores and human rater scores range between 0.34 and 0.67, evaluated on responses from 8 item types. Furthermore, we found that correlations based on complete sets of 21 spoken responses per test taker improve to around $r = 0.73$.

Given the many significant challenges of this work, including 8 different item types in the assessment, responses from speakers from different native languages and speaking proficiency levels, sub-optimal audio conditions for a part of the data, and a relatively small data set for both ASR system adaptation and linear regression model training, we find that the overall performance achieved by our automated speech scoring system was a good starting point for an eventual deployment in a low-stakes assessment context.

Future work will aim at improving the performance of the prediction models by the addition of more features addressing different aspects of the construct as well as an improved filtering model for flagging the different types of problematic responses. Furthermore, agreement between human raters, in particular for read-aloud items, could be improved by refining rater rubrics and additional rater training and monitoring.

Acknowledgments

The authors would like to thank Anastassia Loukina and Jidong Tao for their comments on an earlier version of this paper, and are also indebted to the anonymous reviewers of BEA-9 and ASRU 2013 for their valuable comments and suggestions.

References

Jared Bernstein, Jian Cheng, and Masanori Suzuki. 2010a. Fluency and structural complexity as predictors of L2 oral proficiency. In *Proceedings of Interspeech*.

Jared Bernstein, Alistair Van Moere, and Jian Cheng. 2010b. Validating automated speaking tests. *Language Testing*, 27(3):355–377.

Paul Boersma and David Weenick. 2012. Praat: Doing phonetics by computer, version 5.3.32. `http://www.praat.org`.

Abhishek Chandel, Abhinav Parate, Maymon Madathingal, Himanshu Pant, Nitendra Rajput, Shajith Ikbal, Om Deshmuck, and Ashish Verma. 2007. Sensei: Spoken language assessment for call center agents. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Lei Chen and Su-Youn Yoon. 2012. Application of structural events detected on ASR outputs for automated speaking assessment. In *Proceedings of Interspeech*.

Miao Chen and Klaus Zechner. 2011. 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*, pages 722–731.

Lei Chen, Klaus Zechner, and Xiaoming Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of NAACL-HLT*.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 1997. Automatic evaluation of Dutch pronunciation by using speech recognition technology. In *Proceedings of the IEEE Workshop on Auotmatic Speech Recognition and Understanding (ASRU)*.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America*, 107(2):989–999.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111(6):2862–2873.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.

Horacio Franco, Leonardo Neumeyer, Vassilios Digalakis, and Orith Ronen. 2000. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30(1-2):121–130.

Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David M. Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25(2):282–306.

Florian Hoenig. 2002. Automatic assessment of non-native prosody – Annotation, modelling, and evaluation. In *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*, pages 21–30, Stockholm, Sweden.

Je Hun Jeon and Su-Youn Yoon. 2012. Acoustic feature-based non-scorable response detection for an automated speaking proficiency assessment. In *Proceedings of Interspeech*.

Jack Mostow, Steven F. Roth, Alexander G. Hauptmann, and Matthew Kane. 1994. A prototype reading coach that listens. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*.

Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub. 2000. Automatic scoring of pronunciation quality. *Speech Communication*, 30:83–93.

Pearson Education, Inc. 2011. Versant^{TM} Aviation English Test. `http://www.versanttest.com/technology/VersantAviationEnglishTestValidation.pdf`.

Silke Witt. 1999. *Use of speech recognition in computer-assisted language learning*. Ph.D. thesis, Cambridge University.

Xiaoming Xi, Derrick Higgins, Klaus Zechner, and David M. Williamson. 2008. Automated scoring of spontaneous speech using SpeechRater v1.0. *Educational Testing Service Research Report RR-08-62*.

Shasha Xie, Keelan Evanini, and Klaus Zechner. 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*, pages 103–111, Montréal, Canada. Association for Computational Linguistics.

Su-Youn Yoon and Suma Bhat. 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*, pages 600–608, Jeju Island, Korea. Association for Computational Linguistics.

Su-Youn Yoon, Keelan Evanini, and Klaus Zechner. 2011. Non-scorable response detection for automated speaking proficiency assessment. In *Proceedings of NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*.

Su-Youn Yoon, Suma Bhat, and Klaus Zechner. 2012. Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT*, Montréal, Canada. Association for Computational Linguistics.

Klaus Zechner and Xinhao Wang. 2013. Automated content scoring of spoken responses in an assessment for teachers of english. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT*, Atlanta. Association for Computational Linguistics.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.

Klaus Zechner, Xiaoming Xi, and Lei Chen. 2011. Evaluating prosodic features for automated scoring of non-native read speech. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.