THE INFLUENCE OF AUTOMATIC SPEECH RECOGNITION ACCURACY ON THE PERFORMANCE OF AN AUTOMATED SPEECH ASSESSMENT SYSTEM

Jidong Tao, Keelan Evanini, Xinhao Wang

Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
{jtao, kevanini, xwang002}@ets.org

ABSTRACT

The effectiveness of automated scoring systems for evaluating spoken language proficiency depends greatly on the quality of the automatic speech recognition (ASR) output that is used to calculate the features for the scoring model. In this paper, we examine the effects of ASR word error rate (WER) on the scores produced by a system for automated scoring of non-native English speaking proficiency, as well as on the scoring model features (especially content features) in order to demonstrate the impact of ASR improvements on the performance of the automated speech assessment system. Five different sets of transcriptions with varying degrees of WER ranging from 0% to 52% (including four sets of ASR hypotheses and manual transcriptions) were obtained for a dataset of spoken responses from a pilot administration of an assessment of non-native English speaking proficiency. The experimental results show that higher performing ASR leads to better performance in the automated assessment system; furthermore, the correlation between human and automated scores drops substantially with an increase in WER from 10.7% to 28.9%, whereas the correlation changes little within the following two ranges of WERs: 0% to 10.7% and 28.9% to 52%. A detailed analysis of the features used in the scoring model shows that the ASR errors have a bigger impact on the content features than the delivery and language use features.

Index Terms— automated scoring, English speaking proficiency, automatic speech recognition, English as a Foreign Language

1. INTRODUCTION

The rapid growth of English as a worldwide medium of communications has given rise to the need for automated methods of assessing English speaking proficiency, as well as computer-assisted language learning capabilities. These systems require the use of automated speech recognition (ASR) technology in order to process the English learner's spoken responses; however, achieving accurate ASR for non-native speech can be quite difficult, especially if the speech is heavily accented, disfluent, or ungrammatical, as is often the case with learners with lower English proficiency. To overcome this barrier, many automated English proficiency training and assessment systems rely primarily or exclusively on test questions that elicit restricted speech, such as reading a paragraph out loud. These types of test questions, however, only provide limited information about a non-native speaker's English proficiency; in order to be able to provide a valid assessment of a non-native speaker's proficiency, it is also necessary to elicit speech of a less constrained nature. However, it is exactly these types of test questions, namely ones that elicit speech of a more spontaneous nature, that can pose serious problems for the ASR system, and potentially have a negative impact on the overall performance of the automated scoring system. To date, however, little research has been done on examining the impact of ASR accuracy on automated speech scoring systems, or how the impact varies across test questions that elicit different types of spoken responses. This paper addresses this question in the context of a spoken language assessment for non-native speakers of English who are training to be English instructors in foreign countries. The assessment contains 8 different types of questions that elicit a range of different types of speech, and an automated scoring system is used with a variety of ASR configurations resulting in ASR hypotheses ranging in word error rate (WER) from 0% (for the manual transcriptions) to 52%. The results show that the performance of the ASR system has a different impact on different types of automated scoring features (in particular, features that rely on the content of the response vs. features that assess the speaker's delivery). These results have implications for the design and deployment of automated speech assessment and learning systems.

The remainder of this paper is organized as follows: Section 2 presents some related work in the field of automated speech assessment, and motivates the current study in the context of the literature; Section 3 describes the speaking items in the assessment used in the current study and presents the areas of speaking proficiency that are assessed. Section 4 presents the details of the experiments

that were conducted, followed by the results and analyses in Section 5; finally, Section 6 concludes the paper and points out directions for future work.

2. RELATED WORK

ASR-based automated speech scoring systems have been widely used for assessing second language learners' speaking proficiency for a variety of tasks ranging from reading aloud [1], [2] to spontaneous speech [3], [4]. The types of features that are typically assessed range from pronunciation-specific [5] to more general fluency [1] and prosodic [6] aspects of non-native speech. These systems have primarily been developed for non-native speakers of English, but are increasingly being used for non-native speakers of other languages [1], [3], [7]. The architecture of the automated scoring system used in this paper differs in some respects from other automated speech scoring systems, but the features for measuring speaking proficiency are still mostly extracted from the output of an ASR system. A recent study demonstrates that more accurate ASR improved the quality of the pronunciation-related features that were derived from it for automated speech proficiency scoring [8], but it was unclear whether the same conclusion could be drawn on the content features that were produced from the ASR hypotheses due to the low word accuracy. More recently, Deep Neural Network (DNN) approaches have been shown to improve ASR performance greatly [9]; furthermore, an automated speech scoring system that adopted a DNN-based ASR component significantly outperformed a baseline system using the most popular Hidden Markov Model (HMM)-based ASR approach [10]. Although a great deal of research effort has been put into developing an ASR system with the lowest possible WER in the context of automated speech scoring, there is little research on analyzing the impact of ASR accuracy in the area of automated speech assessment. The goal of this paper is to illustrate the influence of improved ASR on the performance of automated speech scoring by analyzing the predicted scores and the scoring model features (especially content features) through a case study for assessing teachers of English as a Foreign Language (EFL) who are non-native English speakers.

3. ASSESSMENT DESIGN

The pilot version of the assessment that was used for this study contains 35 speaking items belonging to 8 different task types with different characteristics [4]. Table 1 describes these 8 constructed-response speaking task types, which are divided into two groups based on how constrained the test taker's response is. The spoken responses in Group 1 are highly predictable due to the fact that all of the expected linguistic content is presented in the test prompt, whereas the task types in Group 2 require the test taker to produce some spontaneous speech. Each test taker responded to 28 out of the 35 test items in the pilot study. The distribution of the number of responses for each task type is highly uneven, ranging from 5% for the least frequent to 30% for the most frequent task type.

Group 1: Restricted speaking item types	
Item Type	**Description**
Multiple Choice (MC)	Read aloud the selected choice
Read Aloud (RA)	Read aloud a given set of instructions
Repeat Aloud (RP)	Repeat a short utterance
Group 2: Semi-restricted speaking item types	
Item Type	**Description**
Incomplete Sentence (IS)	Complete a fragmented sentence
Key Words (KW)	Compose a sentence as instructed using given key words
Chart (CH)	Formulate a similar sentence to an example from a chart using a given grammatical pattern
Keyword Chart (KC)	Construct a sentence using given keywords and information in a chart
Visuals (VI)	Give instructions based on the graphical information in given two visuals

Table 1. Eight speaking item types included in the assessment

Holistic scores for each spoken response were provided by two trained human raters using the following scores: 0, 1, 2, 3, and TD (Technical Difficulty). A score of 0 means that a spoken response contains no speech, contains non-English speech, or is off topic, etc. Scores 1 to 3 correspond to low, medium, and high ratings of speaking proficiency and task completion, respectively. Finally, a response that encountered a technical difficulty such as a microphone problem, background noise, etc. was scored as TD. For the restricted speaking item types in Group 1, the human raters assessed the test taker's delivery (including characteristics of their speech related to pronunciation, fluency and prosody) and the test taker's ability to read/repeat the prompt accurately. For the semi-restricted item types in Group 2, the raters were also required to take into account the appropriateness of the language used (including vocabulary and grammar) in addition to the delivery and accuracy. Furthermore, a second set of trained human raters provided analytic scores specifically for the language delivery and content accuracy components of each response separately in order to investigate the two main constructed feature groups: delivery and content. These analytic scores were provided on the same 0-3 scale as the holistic scores.

4. METHODOLOGY

4.1. Data partitions

The data in this paper are drawn from a pilot version of a standardized assessment that was developed EFL teachers who are non-native speakers in order to assess their competence in using the English language for the purpose of classroom English instruction [4]. Assessment items for the four general language modalities (Reading, Listening, Writing and Speaking) are included in the assessment; this study focuses solely on the Speaking items. The data for the study was collected from 2308 test takers across 10 non-English speaking countries in a pilot administration of the assessment. The spoken responses were partitioned into the following five sets (without speaker or response overlap) for the purpose of developing an automated scoring system to assess spoken responses: three sets were used for ASR training, development, and evaluation; the remaining two sets were used to train and evaluate the scoring models. The uneven distribution of responses for the 8 task types mentioned above was similar across all partitions, and was due to the design of the assessment.

All of the spoken responses in these five partitions were orthographically transcribed by trained human transcribers, and further received one of the 5 holistic human scores on the scale of 0, 1, 2, 3, and TD, as described in Section 3. For the experiments described in this paper, all responses receiving a score of 0 or TD were excluded from the scoring model evaluation partition. In addition, a small number of responses in the smEval partition did not receive two sets of holistic scores or analytic sub-scores for assessing delivery and content; these responses were also excluded from the study. In total, 5,301 responses in the smEval partition that obtained valid sets of double holistic and analytic scores were retained for the experiments. The numbers of responses in each type are as follows: 783 MC, 2115 RA, 272 RP, 148 IS, 451 KW, 1272 CH, 143 KC, and 117 VI.

ASR	True-Trans	OP-ItemSpec	OP-ID	HTK-ID	OP-OOD
WER	0	9.6	10.7	28.9	52

Table 2. WER for the 5 different ASR configurations

4.2. ASR configurations

To examine the effects of the WER on the automated scores and the scoring model features, five sets of transcriptions with varying WERs were obtained for the responses in the scoring model evaluation dataset using two different ASR systems. The first ASR system is a highly optimized, state-of-the-art speech recognizer (OP). The second ASR system is based on the open-source speech recognition toolkit HTK [11]. The acoustic models (AM) of both ASR systems consist of HMM-based crossword triphones. The OP system uses a 4-gram language model (LM), whereas the HTK system uses a bi-gram LM. In order to obtain transcriptions with five different WERs that mirror the types of ASR errors that would be encountered in a real-life operational deployment of an automated speech scoring system (instead of generating artificial transcriptions that simulate different levels of WER, but may not accurately reflect the types of errors that are made by actual ASR systems), five different ASR configurations were used with these two speech recognizers.

Firstly, an ASR system was configured using the OP speech recognizer trained on the in-domain (ID) ASR training data set with a single generic LM trained on the entire ASR training partition (OP-ID). Secondly, a system was configured using the same AM as OP-ID, but with 35 item-specific LMs that were trained using the transcripts in the ASR training partition for the 35 different items included in the study (OP-ItemSpecific). Thirdly, a configuration system was designed using the OP speech recognizer trained on out-of-domain (OOD) responses from a different assessment (OP-OOD). Next, an ASR system was configured using HTK with a single generic LM trained on the ASR training partition, as was done for OP-ID (HTK-ID). Finally, in order to simulate a "perfect" zero WER situation, another system was obtained by using the human transcriptions and running HTK forced alignment with the AM trained in the HTK-ID configuration (True-Trans). The 5 WERs obtained using these five ASR configurations for the responses in the scoring model evaluation partition are listed in increasing order in Table 2.

4.3. Scoring model features

The feature extraction and score prediction steps in this experiment were conducted using SpeechRater℠, an automated scoring engine for assessing non-native English speaking proficiency [12]. SpeechRater has four components that are connected sequentially as follows: an automatic speech recognizer, a feature computation module, a filtering model, and linear regression scoring models. In this process, the speech recognition component first decodes an input spoken response into a word-level transcript using an acoustic model trained on non-native speech, and forced-aligns the transcript using an acoustic model trained on native speech. Secondly, the feature extraction module uses the ASR output in combination with the speech signal to calculate speaking proficiency features to be used with the linear regression scoring models. Thirdly, the filtering model filters out spoken responses that cannot be given a regular score such as ones that contain no speech, non-English speech, noise, or have other sub-optimal audio characteristics [13]. Finally, the linear regression scoring models are used to predict a numerical score for the response.

SpeechRater calculates more than 100 scoring model features covering the three main construct areas of the assessment's scoring rubrics: delivery, language use, and content. The delivery group comprises features that assess a non-native English speaker's fluency, pronunciation, and prosody. The language use group is composed of features

that assess a speaker's vocabulary diversity and grammatical correctness. These features are only used for assessing responses to the semi-restricted item types, because language use is irrelevant for the highly predictable types. Finally, the content group of features measures the accuracy of the content of a speakers' response using the methods in natural language processing (NLP) such as string matching, n-grams, edit distance, and regular expressions [14].

A set of features were chosen for each of the eight item types (a total of 21 different features were used across the eight item types) based on the following criteria [4]: 1) high correlation with human scores, 2) construct relevance 3) construct coverage, and 4) feature independence. In the 21 selected features, 12 delivery features extract interruption points, short/long silences, repetitions, duration of vowels/words/clauses, and stressed syllables; 4 language use features calculate the global language model (LM) score, the part-of-speech (POS) based grammatical similarity scores, and number of word types; 5 content features estimate the number of correct read words, the read/repeat word error rate (WER), the response discrepancy from the high scoring responses in WER, the N-grams in response matching high scoring responses, and the matching of regular expression.

Eight linear regression scoring models (one for each item type) were constructed with the associated selected features in each type as predictors and the average holistic scores provided by human raters as the dependent variable. The scoring models' performance was eventually evaluated in terms of the ability of each model to predict first human rater (H1) scores in the scoring model evaluation data set.

5. RESULTS

5.1. The effect of WER on automated scores

For the scoring model evaluation data set, the inter-rater agreement calculated in terms of both the Pearson correlation coefficient (r) and quadratic weighted kappa (κ) is 0.65. Table 3 lists the correlations between SpeechRater scores and first human rater (S-H1) across the 5 different WERs. The correlations in each column of the table illustrate that the performance of the automated scoring system decreases nearly monotonically as the WER increases; furthermore, the score correlations between humans and SpeechRater drop substantially with the increase in WER from 10.7% to 28.9%, whereas the correlation changes little within the following two ranges of WERs: 0% to 10.7% and 28.9% to 52%. The True-Trans condition ASR has the highest correlations among all the five configurations, but it is still lower than human-human agreement. This is because 1) human transcription of non-native speech is imperfect and can lead to high levels disagreement among transcribers [15]; 2) while the scoring model features in SpeechRater represent the three major areas of the scoring rubrics, they are still not a perfect match for the information used by human raters during the scoring process. Although the OP-OOD ASR nearly doubles the WER of the HTK-ID, the correlations of the OP-OOD are even slightly higher than the HTK-ID.

Configuration	WER (%)	r	κ
True-Trans	0	0.616	0.523
OP-ItemSpec	9.6	0.602	0.508
OP-ID	10.7	0.601	0.505
HTK-ID	28.9	0.461	0.385
OP-OOD	52	0.47	0.394

Table 3. Pearson correlation coefficient (r) and quadratic weighted kappa (κ) between SpeechRater and first set of human ratings (S-H1) across 5 WERs; inter-rater agreement (H1-H2) $r = \kappa = 0.65$.

5.2. The effect of WER on scoring model features

The results in Section 5.1 demonstrate the overall degree of degradation in performance of the automated scoring system with the increase in WER of the ASR component. In this section, the differential effect of WER on specific scoring model features is interpreted as the extent of the degradation by calculating the slope of the five correlations (one for each WER level) between each feature included in the scoring models and both the holistic human scores (H1) and the analytic sub-scores for delivery and content. The reliability of the analytic sub-scores in terms of the human-human agreement was $r = 0.48$ and $r = 0.77$ for delivery and content, respectively.

5.2.1. Correlations between individual features and holistic scores

The range of correlations of the True-Trans calculated between the scoring model features and H1 scores by item type for the 3 main constructs is shown in Table 4, which is the starting point to examine the degradation in correlation for the rest of ASR systems. The range of degradation in correlation between the scoring model features and H1 scores across the eight item types for the three main constructs is shown in Figure 1. The results for each of the eight item types are separated by dotted lines in the figure and are distributed from left to right corresponding to the two groups of speaking item types defined in Table 1: restricted (MC, RA, RP) and semi-restricted (IS, KW, CH, KC, VI). Each box plot in the figure represents a range of correlation degradation for one of the three main speaking proficiency constructs: delivery, grammar, and content; the results for each of the three constructs are listed in that order from left to right for each of the eight item types in the figures. Note that the restricted speaking types were only measured by features from the delivery and content constructs, whereas the semi-restricted types were covered by all three constructs. In each box plot, the central mark is the median, the edges of the box are the lower hinge (defined as the 25th percentile) and the upper hinge (the 75th percentile), and the whiskers extend to the most

extreme data points not considered outliers. In Figure 1, it can be seen that the content features have a range of degradation that is higher than the delivery and language use features for all but one restricted speaking item type (RA).

Item Type	Delivery	Language Use	Content
MC	0.155-0.307	N/A	0.596-0.779
RA	0.227-0.39	N/A	0.346-0.37
RP	0.247-0.36	N/A	0.571-0.572
IS	0.08-0.247	0.048-0.426	0.604-0.684
KW	0.277-0.338	0.301-0.41	0.561-0.637
CH	0.158-0.303	0.271-0.434	0.432-0.53
KC	0.121-0.157	0.164-0.463	0.667-0.721
VI	0.06-0.255	0.248-0.429	0.4-0.489

Table 4. Range of correlations between the scoring model features and H1 scores by item type for three constructs

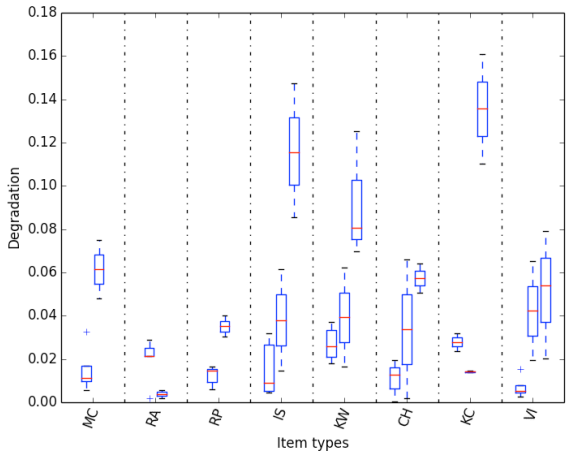

Figure 1. Range of degradations in correlation between the scoring model features and H1 scores for delivery, language use, and content, from left to right for each item type

5.2.2. Correlations between individual features and analytic delivery scores

To further analyze the degradation in correlation between the scoring model features and the human analytic delivery scores, the range of degradations in Figure 2 is illustrated in the same manner as in Figure 1. It can be seen that the entire range of correlation degradation is restricted to the range 0 - 0.06, whereas the range in Figure 1 is 0 - 0.16. It is also hard to differentiate the degradations among the three sets of feature constructs.

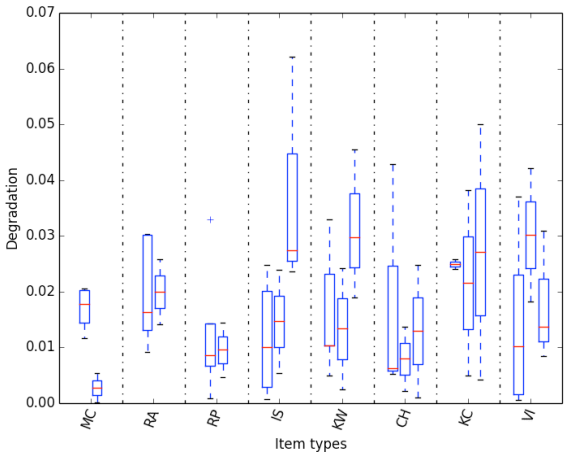

Figure 2. Range of degradations in correlation between the scoring model features and delivery analytic scores

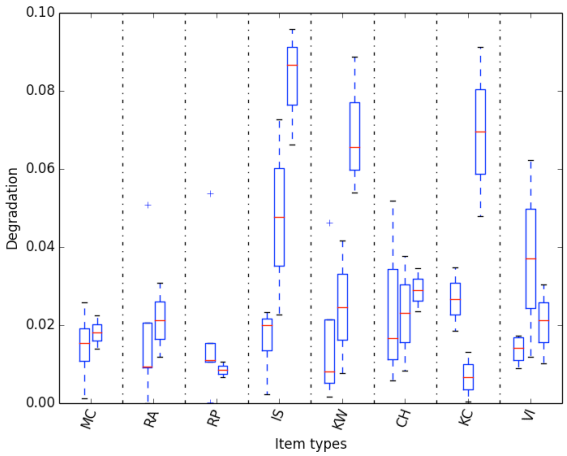

Figure 3. Range of degradations in correlation between the scoring model features and content analytic scores

5.2.3. Correlations between individual features and analytic content scores

In Figure 3, the degradations are computed using the slope of the five correlations between each feature and the human analytic scores for content in the same manner as in Figure 1. The entire range of degradation falls between the ranges depicted in Figure 1 and Figure **2** Similar to the behavior of the correlation degradation found in Section 5.2.1, the content features for all item types but RP and VI have higher ranges of degradation than the delivery and language use features.

6. CONCLUSIONS AND FUTURE WORK

In order to investigate the impact of ASR accuracy on the performance of an automated speech scoring system, transcriptions with five different WERs produced by 5 different ASR configurations were obtained. The correlations between the holistic human scores and

automated scores showed that the higher performing ASR systems lead to better automated scoring performance, as expected; furthermore, the correlation between the human scores and automated scores drops substantially with an increase in WER from 10.7% to 28.9%, whereas the correlation changes little within the following two WER ranges: 0% to 10.7% and 28.9% to 52%. This finding could indicate that a WER of 10% is a satisfactory goal for an automated scoring system, and that it may not be effective to expend additional effort pursuing a "perfect" ASR. In order to validate this finding, a follow-up study will explore the detailed effect of different WERs on performance at the item level and the speaker level. To further investigate the sensitive range of WER with the largest correlation drop from 10.7% to 28.9%, additional systems with different LMs (e.g., tri-gram) will be studied to fill the gap.

The more detailed scoring model feature analyses across the eight item types for the three feature constructs show that the ranges of degradation in correlation for the content-related features generally vary more than the language- and especially the delivery-related features. This observation indicates that ASR errors have a larger impact on the content features included in the scoring model than on the delivery and language use features.

The conclusions in this paper were drawn using a pilot dataset of teachers of EFL which has highly skewed score distribution: the average score across all items is approximately 2.5 on a 3-point scale. Furthermore, the eight item types are unequally represented in the 28 items that each test taker responded to (some of the item types are only represented by a single item) and the item types are associated with different constructs and response durations. Since these characteristics of the data set could potentially influence the robustness of the results obtained in this study, future work will validate these conclusions by extending the analysis to other non-native speaking proficiency tests, in particular ones that elicit completely unconstrained spontaneous speech.

7. REFERENCES

[1] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 989–999, 2000.

[2] D. Bolaños, R. A. Cole, W. H. Ward, G. A. Tindal, J. Hasbrouck, and P. J. Schwanenflugel, "Human and Automated Assessment of Oral Reading Fluency.," *J. Educ. Psychol.*, vol. 105(4), pp. 1142–1151, Nov. 2013.

[3] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," *J. Acoust. Soc. Am.*, vol. 111, no. 6, pp. 2862–2873, 2002.

[4] K. Zechner, K. Evanini, S.-Y. Yoon, L. Davis, X. Wang, L. Chen, C. M. Lee, and C. W. Leong, "Automated scoring of speaking items in an assessment for teachers of English as a Foreign Language," in *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, Baltimore, Maryland, 2014, pp. 134–142.

[5] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2–3, pp. 95 – 108, 2000.

[6] F. Hönig, T. Bocklet, K. Riedhammer, A. Batliner, and E. Nöth, "The Automatic Assessment of Non-Native Prosody: Combining Classical Prosodic Analysis with Acoustic Modelling," in *Proceedings of 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 2012.

[7] T. Lustyk, P. Bergl, and R. Cmejla, "Evaluation of disfluent speech by means of automatic acoustic measurements," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1457–1468, 2014.

[8] D. Higgins, L. Chen, K. Zechner, K. Evanini, and S.-Y. Yoon, "The impact of ASR accuracy on the performance of an automated scoring engine for spoken responses," presented at the National Council on Measurement in Education meeting, New Orleans, LA, 2011.

[9] G. Hinton, L. Deng, D. Yu, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[10] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Proceedings of 14th Annual Conference of the International Speech Communication Association*, Lyon, France, 2013, pp. 1886–1890.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book, version 3.4*. Cambridge University Engineering Department, 2006.

[12] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Spok. Lang. Technol. Educ. Spok. Lang.*, vol. 51, no. 10, pp. 883–895, Oct. 2009.

[13] J. H. Jeon and S.-Y. Yoon, "Acoustic Feature-based Non-scorable Response Detection for an Automated Speaking Proficiency Assessment," in *Proceedings of 13th Annual Conference of the International Speech Communication Association*, Portland, Oregon, USA, 2012.

[14] K. Zechner and X. Wang, "Automated Content Scoring of Spoken Responses in an Assessment for Teachers of English," in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, Georgia, 2013, pp. 73–81.

[15] K. Evanini, D. Higgins, and K. Zechner, "Using Amazon Mechanical Turk for Transcription of Non-Native Speech," in *Proc. NAACL HLT, Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk.*, 2010.