Content-Based Automated Assessment of Non-Native Spoken Language Proficiency in a Simulated Conversation

Keelan Evanini
ETS Research
Princeton, NJ, 08534 USA
kevanini@ets.org

Sandeep Singh
ETS Research
Princeton, NJ, 08534 USA
ssingh016@ets.org

Anastassia Loukina
ETS Research
Princeton, NJ, 08534 USA
aloukina@ets.org

Xinhao Wang
ETS Research
San Francisco, CA, 94105 USA
xwang002@ets.org

Chong Min Lee
ETS Research
Princeton, NJ, 08534 USA
clee001@ets.org

Abstract

In this paper we present a task for assessing the English speaking proficiency of non-native speakers based on a simulated dialogic interaction with a computer interlocutor. In the task, the language learner is first presented with a set of stimulus materials and then participates in a simulated conversation by answering questions about the content of the materials. An automated speech scoring system based on features related to the language learner's delivery, grammar, and vocabulary is augmented with features that assess the appropriateness of the content in their responses. Experiments on a large corpus of spoken responses covering a range of L1 backgrounds demonstrate that the addition of the content features improves the performance of the automated scoring system when correlated with expert human ratings.

1 Introduction

As English continues to grow in dominance as the most commonly spoken international language for business and academic purposes, the need for automated language learning and assessment tools also continues to grow. While many automated spoken language assessment systems exist, both in the context of standardized language proficiency assessment for decision-making purposes as well as embedded in language learning applications that provide feedback to the learner, the majority of tasks that are presented to the user in these systems elicit restricted speech (such as reading a sentence out loud), not conversational speech. This is due to the difficulty in obtaining accurate ASR output to be used for scoring the responses, especially for non-native spoken responses that may contain pronunciation errors, large amounts of disfluencies, ungrammatical phrases, etc. However, in order for an assessment of English speaking proficiency to be valid, it should include a range of tasks that elicit the skills and abilities that are required for successful communication in an English-medium environment.

To address this, some automated assessment systems have also been developed to score spontaneous speech. In an early study, [1] developed a system to assess various aspects of a non-native speaker's fluency in spontaneous speech (such as articulation rate and average length of pauses) in the context of a standardized assessment of Dutch speaking proficiency. [2] developed a system to assess English speaking proficiency in the context of a practice test of English for academic purposes. This system also relied heavily on fluency features, but it also included features to assess pronunciation (acoustic model score) and grammar (language model score). In another study, [3] developed a system to score spontaneous speech in the context of an assessment of English for business purposes based on fluency features and additional features extracted from the audio signal (F0 and energy). However, these approaches to automated scoring of spontaneous speech do not fully cover all aspects of speaking proficiency, since they do not employ any sort of spoken language understanding techniques to determine the appropriateness of a response's content.

More recently, some automated scoring systems for spontaneous speech have expanded their coverage to include features that assess the content appropriateness of a response. The general approach taken in these systems is to build supervised models of content that is contained in responses at different score levels and then compare a new response to these models in order to determine which model it is most similar to. In one such study, [4] developed an automated scoring system for a test of non-native English in the K-12 domain that elicited several types of responses containing spontaneous speech (such as giving directions and providing instructions). This system used Latent Semantic Analysis (LSA) to compare the content in a spoken response to prompt-specific models trained on high-scoring responses; the results of this study showed that the LSA content features alone resulted in correlations with human ratings that approached the human-human agreement level of 0.822. In another study, [5] explored the use of a range of content similarity features in the context of an assessment of English for academic purposes, including LSA, Pointwise Mutual Information, and cosine similarity based on Content Vector Analysis (CVA). That study demonstrated that the highest-performing content features had correlations with human ratings around 0.55 - 0.60.

However, even though these systems incorporate spoken language understanding through the use of content features, they still elicit decontextualized individual utterances in isolation. In contrast, in order to be fully functional in an English-medium environment (such as in the workforce or a university) non-native speakers of English need to learn how to improve their ability to participate in interactive conversations. Therefore, automated language learning and assessment systems should also incorporate dialog-based tasks so that learners can be assessed on their ability to use interactive speech. In this paper, we present a novel approach to assessing a non-native speaker's interactive speaking ability given the constraints of state-of-the-art spoken dialog systems (SDS) and demonstrate how the inclusion of content features improves the performance of an automated scoring system designed to score the simulated conversations.

2 Data and Methodology

In this section we present details of the dialog-based task that language learners participated in, the human ratings that were given to each learner's responses in the simulated conversation, the data that was collected in a pilot study that included the dialog-based tasks, and the approach taken to develop an automated scoring system for the responses.

2.1 Task Design

In order to elicit evidence of a language learner's ability to participate in an interactive dialog about topics related to university life, we developed tasks that incorporated simulated conversations that we refer to as *pseudodialogs*. In these tasks, the language learner is presented with a set of stimulus materials, such as a course schedule, an advertisement for a job on campus, an email about a group meeting, etc., and is then presented with a series of spoken prompts from a computer-based interlocutor. After each prompt, the language learner is given a fixed amount of time to provide a spoken response. After each language learner's response, the subsequent prompt from the computer-based interlocutor is played, until the final prompt has been reached. Regardless of the content of the response provided by the language learner, a single, fixed order of system prompts is used. Thus, the system is not truly interactive (in the sense that the system's responses do not vary based on the user's input), but the sequence of turns is designed to simulatate an actual conversation, hence the term *pseudodialog*. In the context of standardized assessment, this format is beneficial for psychometric analyses (compared to a branching dialog system), since it means that all test takers provide comparable data in the assessment.

Table 1 presents the stimulus materials from a sample task incorporating a pseudodialog. In this task, the language learner is presented with information about two jobs on campus and then participates in a pseudodialog with a friend who asks questions about the specifics of the two job postings.

	Laboratory Assistant	Student Manager
Department	Chemistry	Dining Services
Hours	10 hrs. / week; weekdays (flexible)	12–19 hrs. / week; weekdays, weekends, evenings
Pay Rate	$12 / hour	$ 12 / hour
Responsibilities	Assist researchers with lab experiments; maintain and clean laboratory area and equipment; order and stock supplies	Schedule, train, and supervise student workers; enforce Dining Services safety rules; communicate with food preparation staff
Qualifications	Good communicator; responsible; must have completed introductory Chemistry courses	Two or more semesters of dining service experience; excellent customer service skills

Table 1: Sample stimulus materials for a dialog-based language assessment task

Table 2 provides selected system prompts from this pseudodialog as well as sample responses from two language learners, one who received a high score for the conversation (5 on the scoring rubrics described in Section 2.2), and one who received a low score (2). As shown in Table 2, the system's prompts are identical at all turns in the conversation, regardless of the responses provided by the speaker.

2.2 Data Collection

The dialog-based tasks were administered to English language learners around the world in a pilot study that included a variety of additional English proficiency assessment tasks. In total, 1825 test takers participated from the following 9 countries: Brazil, China, France, Germany, India, Japan, Mexico, South Korea, and the United Arab Emirates. Three different versions of the dialog-based assessment task (each with different stimulus materials and conversational prompts) were included in the pilot, and each participant provided responses to one of the three versions. The three versions of the task varied in the number of turns and responses elicited from the language learner: one version elicited 4 responses, another elicited 5, and the third elicited 7. The 1825 simulated conversations that were collected correspond to 9715 distinct responses and were divided into the following three sets for conducting the experiments described in this paper: ASR Training, Scoring Model Training, and Scoring Model Evaluation. Table 3 presents the number of conversations and distinct responses contained in each of these four partitions.[1]

Expert human raters then provided proficiency ratings for each language learner's performance in the entire simulated conversation; i.e., a rater listened to all responses provided by the learner in the conversation and then provided a single score for the entire conversation. The scores were given on a scale of 1–5 and the scoring rubrics encompassed a range of characteristics of the response based on the language learner's spoken English proficiency and how well the task was completed. Table 4 presents the detailed scoring rubrics for a high-scoring response (i.e., score level 5) and thus indicates the specific linguistic aspects of a response that the raters took into account when providing their scores.

2.3 Automated Scoring Methodology

2.3.1 Baseline Features

Baseline features for assessing a language learner's English speaking proficiency were extracted using the SpeechRater automated speech scoring system [2], which employs a two-pass approach that

[1] Responses that were not able to receive a valid human score, due to poor audio quality in the response or other technical difficulties, were removed from the Scoring Model Training and Evaluation partitions. This resulted in the removal of approximately 3% of the data from each partition before the modeling experiments were conducted.

System prompt	High-scoring conversation	Low-scoring conversation
Hi, it's Kathy. I got your message about the jobs you saw in the campus newspaper. What kind of jobs are they?	You can choose from two different jobs uh the first of all you can be a Laboratory Assistant which belongs to the department uh Chemistry and the other one is the stude- uh you can be a Student Manager which um is in the department of Dining Services so they are completely different to each other.	Hi, um I have two messages about the campus, campus jobs choice. One one is, one is the Laboratory A-, Laboratory Assistant and another is Student Manager.
Is the pay different for the two jobs?	The payment is not different. You will always receive twelve dollars per hour so there is nothing to choose from actually.	The the library Laboratory Assistant uh, um provi-, provide you um provide you twelve hou- twelve dollars a wee-.
...	...	...
They both sound like really good options. I'm not sure what to choose. Tell me what you would decide and why.	If I had to decide which job I choose, I would take the um Student Manager um job um in the Dining Services area because you can work more hours per week and you can make more money. Um it depends on how your studies are going and how much time you have, but if you really like Chemistry and you've already have really a lot experience in this area um you can choose Laboratory Assistant job too which is like less um.	Uh in my opinions, in my opinions the laboratory, the laboratory assistants and have you, have you do some research and have you experience the sums of Chemistry Chemistry experiments and you and you and the you the b-, and the you the better, uh if you, if you enjoys um chemistries I think, I think uh is is a good idea, is a good idea to enter the enter laboratory assistant to enforce.

Table 2: Sample excerpts from high-scoring and low-scoring conversations for a dialog-based language assessment task

	ASR Training	Scoring Model Training	Scoring Model Evaluation
# Conversations	612	911	302
# Responses	3292	4834	1589

Table 3: Number of conversations and distinct responses contained in each data partition

first conducts ASR on the spoken response using ASR models trained on non-native speech and then conducts forced alignment of the spoken response to the ASR output using an acoustic model trained on native speech (in order to calculate pronunciation features). The non-native acoustic model used for recognition was trained on over 800 hours of non-native spontaneous speech obtained in the context of a global English proficiency assessment; the language model used for these experiments was trained on the ASR Training partition. The SpeechRater system extracts a total of 135 features that cover a range of linguistic characteristics of the spoken response, such as fluency, intonation, stress, rhythm, pronunciation, vocabulary, and grammar. However, this baseline system does not contain any features that address the appropriateness of the content in a language learner's response.

2.3.2 Content Features

In order to be able to determine whether the content of the test taker's response is appropriate to the prompt, we employed standard features based on CVA models [5, 6]. To develop these features, lexical vectors containing term frequencies weighted by IDF values were trained for a set of responses from each of the score points in the 1-5 range. For each of the score points, s, the $tfidf$ value for each word, i, in the vector was therefore calculated as follows:

A response at this level demonstrates an ability to maintain a conversation by responding appropriately to the interlocutor. Furthermore, a response at this level is clear and easy to understand and demonstrates effective language usage. The response is characterized by the following:

- Responds appropriately and provides relevant and detailed explanation and support.
- Conveys relevant and accurate information from the reading material to support response (as needed).
- Conveys meaning clearly and efficiently through effective word choice.
- Uses a range of linguistic forms effectively with only minor grammatical errors that don't interfere with meaning.
- Speech is fluent with only minor hesitation. Delivery is mostly clear and effective (pronunciation, intonation, rate of speech, etc.) and requires little, if any, listener effort to follow.

Table 4: Scoring rubrics indicating linguistic characteristics of a high-scoring conversation

$$tfidf_{i,s} = tf_{i,s} * log(N/N_i) \tag{1}$$

where $tf_{i,s}$ is the frequency of the word i at score point s, N is the total number of responses in the ASR Training partition, and N_i is the total number of responses containing word i across all score points in the ASR Training partition. Then, for a given response in the Scoring Model Training and Evaluation partitions, the $tfidf$ value for each word in the vector was calculated as follows:

$$tfidf_i = tf_i * log(N/N_i) \tag{2}$$

where tf_i is the frequency of the word i in the response. Then, to calculate the content features, the cosine similarity scores between the vector for the response and the 5 CVA models are computed. These cosine similarity scores are then used directly as features to predict proficiency scores, and are referred to as follows: cos_s for $s \in 1, ..., 5$. An additional feature was calculated by comparing all of the cosine similarity scores to the models for the 5 score points for a given response and taking the score of the model which has the highest similarity; this feature is referred to as max_cos.

The CVA models (both the term frequencies and the IDF values) were trained using the responses from the ASR Training partition and separate models were trained based on human transcriptions of the responses and ASR output.[2] Since each response in the pseudodialogs corresponds to specific content in the stimulus materials that would be expected in a high-scoring response, separate CVA models were trained for each of the individual system prompts. In order to do this, the human rating that was given to the entire pseudodialog was used as the score point for each individual response contained in it.

2.3.3 Scoring Model Building

Separate linear regression scoring models were trained on the responses in the Scoring Model Training partition using the scores from the two different scoring rubrics as the dependent variables in the following three conditions: the Baseline models included the 135 original features described in section 2.3.1; the Transcription models include the Baseline features plus the content features described in Section 2.3.2 calcualted using the transcription-based CVA models; the ASR models include the Baseline features plus the content features calculated using the ASR-based CVA models. Since the human ratings that are predicted by the scoring model correspond to entire pseudodialogs that consist of multiple responses from a language learner, the scoring features (including the CVA

[2] While it is sub-optimal to train the CVA models based on ASR output on the ASR Training partition, since the performance of the ASR system will be inflated compared to unseen test responses, this was necessary due to the limited amount of data available. This approach, however, is preferable to using the Scoring Model Training partition for training the CVA models, since the similarity scores used as features to train the model would then be artificially high, and would result in scoring models that do not generalize to unseen responses.

	cos_1	cos_2	cos_3	cos_4	cos_5	max_cos
Transcription	-0.130	0.309	0.346	0.375	0.389	0.416
ASR	-0.089	0.306	0.345	0.376	0.389	0.432

Table 5: Correlations of individual content features with human scores in the dialog-based task

Baseline	Transcription	ASR
0.632	0.743	0.748

Table 6: Correlations between automated scores and human scores for three conditions (Baseline = no content features; Transcription = inclusion of content features using transcription-based CVA models; ASR = inclusion of content features using ASR-based CVA models)

features) were first extracted for each of the individual responses in a pseudodialog. Then, the mean of each feature across the individual responses was used for training the scoring model.

3 Results

First, Table 5 presents the correlations between the content features (i.e., the mean values of the content features across all responses in a pseudodialog) and the human scores. As the table shows, the cos_5 feature (which measures the similarity between the language learner's response and the CVA model trained on responses that received a score of 5) and the max_cos feature (which provides the score point corresponding to the model that the response was most similar to) consistenly result in the highest correlations with human scores. Compared to the other non-content-based features in the scoring models, these correlation values are relatively high; for example, the max_cos feature typically falls within the top-ten performing features.

Next, Table 6 presents the results of the scoring model experiments in terms of correlations with the human scores. As the table shows, the performance of the baseline automated scoring system was 0.632. This baseline performance improved substantially when content features were added, with an increase in correlation of 0.11 for the CVA models based on transcriptions and an increase of 0.12 using the CVA models based on ASR output.

4 Discussion and Conclusion

This study examined the use of content features in the context of automated English speaking proficiency assessment for a simulated interactive conversation. The results demonstrated that the inclusion of content features into a linear regression scoring model substantially improves the prediction accuracy of the model, thereby improving the validity of the scores produced by the system. This represents an important step towards developing interactive spoken language learning and assessment systems that can provide real-time feedback to the language learner about a wide range of language speaking proficiency characteristics.

Since this system used simulated interactive conversations (*psuedodialogs*) and not actual interactive conversations using an SDS, future work will annotate the responses collected in this study based on the presence or absence of key content from the stimulus materials. Then, these annotations will be used to train the language understanding component of an SDS and a user study will be conducted to determine the extent to which the proficiency scores obtained by using the pseudodialogs compare to the ones obtained using truly interactive versions in an SDS.

References

[1] Catia Cucchiarini and Helmer Strik, "Automatic assessment of second language learners' fluency," in *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999.

[2] Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[3] Roger C. van Dalen, Kate M. Knill, and Mark J.F. Gales, "Automatically grading learners' English using a Gaussian process," in *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, Leipzig, Germany, 2015, pp. 7–12.

[4] Angeliki Metallinou and Jian Cheng, "Using Deep Neural Networks to improve proficiency assessment for children English language learners," in *Proceedings of Interspeech*, 2014, pp. 1468–1472.

[5] Shasha Xie, Keelan Evanini, and Klaus Zechner, "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*, Montréal, Canada, 2012, pp. 103–111, Association for Computational Linguistics.

[6] Yigal Attali and Jill Burstein, "Automated essay scoring with e-rater® V.2," *The Journal of Technology, Learning, and Assessment*, vol. 4, no. 3, pp. 3–30, 2006.