Automated Content Scoring of Spoken Responses in an Assessment for Teachers of English

Klaus Zechner, Xinhao Wang
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08541, USA
kzechner@ets.org, xwang002@ets.org

Abstract

This paper presents and evaluates approaches to automatically score the content correctness of spoken responses in a new language test for teachers of English as a foreign language who are non-native speakers of English. Most existing tests of English spoken proficiency elicit responses that are either very constrained (e.g., reading a passage aloud) or are of a predominantly spontaneous nature (e.g., stating an opinion on an issue). However, the assessment discussed in this paper focuses on essential speaking skills that English teachers need in order to be effective communicators in their classrooms and elicits mostly responses that fall in between these extremes and are moderately predictable. In order to automatically score the content accuracy of these spoken responses, we propose three categories of robust features, inspired from flexible text matching, *n*-grams, as well as string edit distance metrics. The experimental results indicate that even based on speech recognizer output, most of the feature correlations with human expert rater scores are in the range of $r = 0.4$ to $r = 0.5$, and further, that a scoring model for predicting human rater proficiency scores that includes our content features can significantly outperform a baseline without these features ($r = 0.56$ vs. $r = 0.33$).

1 Introduction

With the increased need for instruction of international learners of English as a foreign language (EFL), there is a concomitant rise in demand to assess the language competence of English teachers who are non-native speakers of English. This situation arises because it is neither possible nor affordable for countries where English is not spoken as a native language to employ only or even mostly native speakers of English as EFL teachers. Moreover, as the language of instruction increasingly becomes English in most classrooms, teachers' competence in the productive language modality of speaking becomes substantially more important than in the past. In order to meet this demand for assessing the English language proficiency of teachers of English, a new test, English Teachers Language Assessment (ETLA), was developed recently and piloted in 2012. The test comprises items for all four main language modalities: reading, listening, writing and speaking.

While reading and listening items use a multiple-choice paradigm, test items for speaking and writing elicit open responses. For cost and efficiency reasons, we aim to employ automated scoring of written and spoken responses in this test. This paper is concerned in particular with the conceptualization, implementation and evaluation of features that can assess one aspect of English speaking proficiency: the content correctness of a test taker's response. Our automated speech scoring system, SpeechRater^{SM} (Zechner et al., 2009), also has features addressing other aspects of speaking proficiency, such as fluency or pronunciation, but the details of these features will not be discussed as part of this paper.

The speaking items in ETLA range in complexity from reading a text passage aloud to more challenging tasks requiring multi-sentence responses related to typical teaching situations. The items, therefore, elicit speech in which predictability ranges from high (e.g., reading aloud) to medium (e.g., open responses based on teaching material).

While approaches to capture the content of mostly predictable speech have been widely used in the past (see, e.g., Alwan et al., 2007; Franco et al., 2010), this is not the case for responses that exhibit considerable variation but are still much shorter and more constrained than spontaneous items from other language tests, such as TOEFL iBT®.

Therefore, the goal of the study reported in this paper is to conceptualize, implement and evaluate features that can address the subset of ETLA speaking items where responses are not strongly predictable but are still fairly short and constrained by the context of the item stimulus and prompt.[1] One important aspect of any features used for content scoring is that they have to be robust with respect to speech recognition errors. Robustness is necessary because we are using an automatic speech recognition (ASR) system as a front end, and the average word error rate of the system is around 27% for moderately predictable item responses.

To illustrate what an ETLA speaking item may look like, we provide a relatively simple example here. Suppose the test taker (i.e., an English language teacher) is asked to request that the class open their textbooks on page 55. We could see a range of responses, from "perfect" (score level 3, e.g., "Please open your textbooks on page 55." or "Please open your textbooks and turn to page 55."), to "good" (score level 2, e.g., "Please open the books on the page 55.") and to "poor" (score level 1, e.g., "Open book page 55."). Again, note that for this paper we are not interested in potential issues with fluency, such as long pauses or speaking rate, nor with pronunciation or prosody. We just look at the content of the test takers' responses, either in idealized form by means of a human transcription of what a test taker actually said, or in a realistic operational scenario, where we look at the output of an ASR system. In both cases, we consider the sequence of words only (i.e., a textual representation of the test takers' spoken responses).

In order to investigate the effectiveness of candidate content features in a short-term development cycle before a larger amount of pilot data would be available, we first conducted a small scale in-house data collection effort focusing on the moderately predictable spoken items in ETLA. Based on the analysis of this mini-corpus, several different categories of promising features were selected for potential operational use and then evaluated on the pilot data.

The paper is organized as follows: Section 2 provides an overview on related work; Section 3 describes the in-house data set, the pilot data and the ASR system; the developed features are presented in Section 4; Section 5 presents our experiments; we then discuss our findings in Section 6 and we conclude the paper in Section 7.

2 Related Work

Related to the automated assessment of writing free-text, research to date has concentrated mainly on two tasks: (1) scoring of short answers (Mitchell et al., 2002; Leacock and Chodorow, 2003; Mohler and Mihalcea, 2009) and (2) scoring of essays (Foltz et al., 1999; Kanejiya et al., 2003; Attali and Burstein, 2006). For example, Leacock and Chodorow (2003) built an automated scoring system, c-rater™, to evaluate the short constructed or free-text responses, where the concepts given in test items were modeled, and the presence of these expected concepts in students' answers would be detected.

As for the evaluation of free-text essays, Attali and Burstein (2006) used a selected set of meaningful features to measure different constructed aspects of writing essays, such as grammar, usage, mechanics, style, organization, development, lexical complexity and prompt-specific vocabulary usage. In addition, the Intelligent Essay Assessor (Foltz et al., 1999) used Latent Semantic Analysis (LSA) to score students' answers by comparing them to domain-representative texts. Since LSA is based on the bag-of-words model, researchers have also tried to expand it by introducing additional information, such as part-of-speech (POS) tags (Kanejiya et al., 2003).

In addition, research efforts have also been made to evaluate the content relatedness and correctness for spoken responses. For example, Xie et al. (2012) used LSA and Pairwise Mutual Information approaches to evaluate the content correctness of unrestricted spontaneous spoken responses. Moreover, Chen and Zechner (2011) explored fea-

[1] A test item is a basic element of a test, consisting of stimulus material, such as text and/or visuals, and a prompt (test question) that elicits a response from the test taker.

tures related to grammatical complexity in an automated speech scoring system.

In order to address the moderately predictable speaking test items in the new ETLA, this paper presents several different types of features to score the content correctness of the elicited spoken responses. Following a series of experiments and comparisons, seven features from three content feature categories are selected and evaluated.

3 Data Sets and ASR System

This study conducts experiments and evaluations based on two different data sets: (1) a small scale in-house data collection effort, which was used for the design and development of content features; and (2) a larger-scale pilot data collection, which was used to further evaluate the features selected according to the in-house data and to build scoring models for the prediction of human proficiency scores.

3.1 In-house Data Collection

Twenty-two items from ETLA with moderately predictable responses were selected for the in-house data collection.[2] Firstly, 1,053 text responses in total for all three score levels (3 = high proficiency, 2 = medium proficiency, 1 = low proficiency) were drafted and collected by human experts. In order to simulate the operational scenario with an ASR system in place, a subset of responses was recorded by a small set of predominantly non-native speakers of English. For each test item, four responses were randomly selected from each score level, which resulted in 22 × 3 × 4 = 264 responses for voice recording. The remainder of 789 text responses comprised the set for feature development and training. In addition, about two thirds of the 264 text responses were randomly double-recorded by a second speaker, resulting in a speech corpus with 444 spoken responses in total, used as the evaluation set. Furthermore, all these spoken responses were manually transcribed to accommodate the errors introduced by reading, such as insertions of various speech disfluencies.

[2] We decided to focus our efforts only on the moderately predictable items since scoring of highly predictable item types has been extensively studied in previous research already.

3.2 Pilot Data Collection

This study uses data from a 2012 pilot administration of the ETLA assessment. In particular, we focus on 14 moderately predictable items from the pilot, covering 2,308 test takers. In order to build the automatic speech recognizer and the scoring models, the pilot data were partitioned into five different subsets without any speaker and response overlaps. The first three data partitions were used for training, development and evaluation of the speech recognition system (hereafter, "asrTrain", "asrDev" and "asrEval"), which included spoken responses from both the moderately and highly predictable items. The asrTrain partition was further used to develop and train the content features described below. The remaining two partitions were used for training and evaluation of scoring models that predicted item scores based on a set of features (hereafter, "smTrain" and "smEval"), where only the spoken responses from 14 moderately predictable items from one pilot form were included.

The detailed partition information is listed in Table 1. All these spoken responses have been manually transcribed and scored with holistic scores from 1 to 3 by trained human expert raters. For the smTrain and smEval partitions, there were 6,367 responses receiving double annotation, and the inter-rater correlation was 0.73. Furthermore, the average length of responses from smTrain and smEval sets was 10.5 words, and the corresponding vocabulary size was 855 (not including partial words).

Partitions	# Speakers	# Responses
asrTrain	1,658	27,604
asrDev	25	700
asrEval	25	700
smTrain	300	3,452
smEval	300	3,466

Table 1. Number of speakers and number of responses included within each data partition.

3.3 System Architecture

Our automated speech scoring system, SpeechRater (Zechner et al., 2009), consists of an ASR system described below which generates a word hypothesis for every response by a test taker, including information about timing, energy and pitch, and other information from the input audio

file. Next, the feature computation modules take the outputs of the ASR system and compute a set of features, related to fluency, pronunciation, prosody, as well as content, the focus of this paper. Finally, a scoring model (linear regression model) is trained based on the smTrain set to predict scores and then evaluated on unseen data (smEval set).

3.4 ASR System

In this study, a state-of-the-art gender-independent Hidden Markov Model speech recognition system trained on about 800 hours of non-native speech is taken as the baseline recognizer, and its language model (LM) is then further adapted using the transcriptions from the asrTrain data partition. The language model adaptation weights are tuned on the asrDev set, and the resulting word error rate (WER) on the asrEval set (with both moderately and highly predictable responses) is 11.7%, and its WER on the subset of 264 moderately predictable responses is 19.7%. This speech recognizer is further evaluated on both smTrain and smEval sets as shown in Table 2, only including moderately predictable responses.

Partition	WER (%)
smTrain	26.7
smEval	26.9

Table 2. Word error rates (WER) of the speech recognizer on smTrain and smEval[3] data sets.

4 Content Features

Following a careful inspection and analysis of the collected in-house data (described in Section 3.1 above), several different categories of content features were designed and developed. The initial data analysis showed that features need to be able to capture very narrow ranges of expressions with minor variations, but also should be able to capture something like the "overall accuracy" of expression, where local word sequences or phrases should conform to the expectations of the item design without requiring that a response follows a confined pattern in its entirety. For the former situation, features like regular expression matches seem appropriate to be a good match, whereas for the latter, more flexible approaches such as *n*-gram models or string edit distance metrics may be more appropriate. We list and describe our proposed content features in the following section.

A. Flexible String Matching Metrics

AI. Regular Expressions

Since many responses in ETLA are expected to follow certain patterns, it is intuitive to construct limited regular expressions (RegEx) to match gold standard responses for candidates with high proficiency score levels. Accordingly, one type of regular expression related features, *re_match*, can be extracted to detect whether the test response can be matched by any of the pre-built regular expressions. This feature can obtain the values of 0 (does not match), 1 (partially matches) and 2 (exactly matches). Here, a partial match indicates that a RegEx can be matched within a test response that also has other spoken material, which is useful when the speaker repeats or corrects the answer multiple times in a single item response, and the compiled RegEx can still be used to match parts of the test response.

This content feature has the advantage of high precision, as it can precisely examine the content correctness of the test responses. Thus, the RegEx should be compiled to match all the example responses at the highest score level 3 from the training set. For some test items with relatively short and fixed answer patterns, this feature is quite useful; however, it is very time-consuming and difficult to manually build regular expressions for items with longer and more flexible expressions. Meanwhile, the mechanism of exact matching can make this feature fail in very small variations of expression. Especially when applying this feature on ASR output, it is difficult to successfully match some content-correct responses that have disfluencies or recognition errors.

Therefore, in order to improve the robustness of RegEx, another regular expression related feature is proposed. In general, for each item in ETLA, some pieces of specific expressions are required in a test response to represent its content correctness. Accordingly, we can segment the reference responses into several fragments and identify some pieces as key fragments. For example, when looking at the reference response "Please open your

[3] The calculation of WER is based on only the recognized outputs with more than one word. Thus, the number of actually recognized responses is less than that in Table 1, i.e., 3,264 responses for smTrain and 3,255 responses for smEval.

text books and turn to page 55." two key fragments can be extracted with "Please open your text books" and "turn to page 55." We group versions of these key fragments from the training corpus together and construct regular expressions to match each group. Afterwards, a feature can be defined to count how many key fragments can be matched by a test response, namely *num_fragments*.

AII. Keyword Detection

For moderately predictable items on ETLA, keyword lists can be extracted from the stimulus material and the item prompt, containing the words that need to be included in a test response by test takers. Then a feature, *num_keywords,* can be used to examine how many keywords appear in a test response, which can be further normalized by the number of predefined keywords for each item, i.e., *percent_keywords*. In addition, as some keywords may be a phrase with multiple words, such as "page 55," we can split all the keywords into single words and get another sub-keywords list. Then two corresponding features can be extracted as *num_sub_keywords* and *percent_sub_keywords*.

B. *N*-grams

BI. Word *N*-grams

The word *n*-gram model is introduced here to capture the similarity of word usage between the test and the reference responses. Based on the collected training samples, trigrams are trained using the text responses from the highest score level 3. Then, the LM can be used to score a test response, and the resulting probability can be taken as feature, called *lm_3*.

BII. POS Similarity

This feature measures the syntactic complexity of test responses based on the distribution of POS tags. First, all the responses from the training data set are assigned with POS tag sequences via an automatic POS tagger. Then, a POS vector according to each score level can be obtained by gathering the POS unigram, bigram or trigram statistics from the same score level.

Given a test response, its corresponding POS sequence can be determined by the same POS tagger, and the cosine similarities between the test POS *n*-gram vector and the POS vectors from three different score levels can be calculated as *pos_1*, *pos_2* and *pos_3*, where *pos_3* is used as a feature in our experiments below. Furthermore, by comparing these three cosine similarities, the score category with the highest similarity can be extracted as another feature, i.e., *pos_score*.

BIII. Machine Translation Evaluation Metric (BLEU)

BLEU (Papineni et al., 2002) is one of the most popular metrics for automatic evaluation of machine translation, where the score is calculated based on the modified *n*-gram precision. In this study, the BLEU score is introduced to evaluate the content quality of a test response, where three different gold standard reference corpora are extracted from the training set according to each score level. Similar to the edit distance and WER features described below, three BLEU scores are calculated by comparing them with reference responses from each score level (i.e., *bleu_1*, *bleu_2* and *bleu_3*). We decide to use the following two features for our experiments below: *bleu_3* and *bleu_score*, the score level which receives the maximum BLEU score.

C. String Edit Distance Metrics

CI. String Edit Distance

As the edit distance is an effective string metric for measuring the amount of difference between two word sequences, including insertions, deletions and substitutions, we use it to capture the sequence distance between the test and reference responses.

Given a test response, we can separately calculate the edit distance by comparing it with training responses from each score level. Afterwards, the minimum edit distance from each score level can be extracted as *ed_1*, *ed_2* and *ed_3*, where *ed_3* is selected as feature for our experiments. Furthermore, by comparing these three edit distances, the score category with the minimum value is taken as another feature, *ed_score*.

CII. Word Error Rate (WER)

By dividing the edit distance by the length of the reference response, we obtain the word error rate (WER) metrics, commonly used in speech recognition, and two additional features, *wer_3* and *wer_score*, similarly as above, can be calculated.

Compared to the above category of *n*-gram related features, which capture the *n*-gram fragment

matching between the test and reference samples, the category of edit distance features try to find the most similar reference sample to the test sample at the whole-response level.

Finally, all the proposed features are implemented and then examined based on both the ideal human transcription and the realistic ASR output. The speech recognizer used with the small in-house data is the same as the ASR system described in Section 3.4, but its language model is adapted with the much smaller set of 789 training text responses. The WER of this system is 17.8%, evaluated on 444 spoken responses.

In addition, in order to increase the robustness of the extracted features, a preprocessing stage is introduced to remove all the disfluencies from the ASR output, such as filler words, recognized partial words and repeated words. Afterwards, each feature is evaluated on both the transcription and the ASR output of the 444 collected spoken responses, and its corresponding Pearson correlation coefficient with human scores is presented in Table 3.

Based on overall correlation, inter-correlation analyses, as well as on construct[4] considerations, seven content features from three categories are selected and will be evaluated on a larger scale on ETLA pilot data in the next section: *re_match (A1)*, *num_fragments (A2)*, *percent_sub_keywords (A3)*, *bleu_3 (B1)*, *ed_score (C1)*, *wer_3 (C2)* and *wer_score (C3)*.

5 Experiments and Results

This section first describes experiments related to the performance of the seven selected content features on a larger corpus from an ETLA pilot administration (described above in Section 3.2). Then, a similar analysis is conducted based on human rater analytic content scores on a subset of this data. Finally, the selected content features are combined with other features related to pronunciation, prosody and fluency to build a scoring model for the prediction of human scores.

[4] A construct is the set of knowledge, skills and abilities measured by a test. The term "construct considerations" in the context of feature selection refers to the process of ensuring that the selected feature set obtains a high coverage of all aspects of the relevant construct.

	Feature	Trans	ASR
A	re_match	0.789	***0.537***
	num_fragments	0.629	***0.523***
	num_keywords	0.269	0.254
	percent_keywords	0.419	0.375
	num_sub_keywords	0.249	0.239
	percent_sub_keywords	0.482	***0.417***
B	lm_3	0.482	0.461
	pos_3	0.270	0.270
	pos_score	0.315	0.339
	bleu_3	0.531	***0.458***
	bleu_score	0.144	0.194
C	ed_3	-0.362	-0.337
	ed_score	0.642	***0.614***
	wer_3	-0.573	***-0.513***
	wer_score	0.585	***0.557***

Table 3. Pearson correlation coefficients (*r*) of content features with human holistic scores.

5.1 Feature Evaluation on Pilot Data

In the following experiments, we use the asrTrain set to train the content features. Then these features are examined on the smTrain and smEval data sets. In order to extract the edit distance, WER- and BLEU-related features for each item, three text reference corpora according to different score levels, are needed. Duplicate reference responses with the same content are removed within each score level.

Furthermore, we improve two RegEx features using the reference responses from the highest score level 3 in the asrTrain set. (1) Since the previously obtained *re_match* feature based on the in-house data may not be able to match multiple content-correct responses in the pilot data, we need to augment the set of RegEx for this feature based on correct responses from score level 3 in the asrTrain set. (2) Since the maximum number of candidate fragments varies across different ETLA items, the *num_fragments* feature values are not comparable across items. Therefore, we redesign this feature by assigning a list of manually selected keywords for each fragment. During feature extraction, we count the number of distinct keywords associated with all the matched fragments and divide this number by the number of predefined keywords for each item (as in AII. Keyword Detection), which results in another feature: *perc_fragment_kw (A2)*.

Based on the ASR output of smTrain and smEval data sets, seven content features are extracted and their Pearson correlation coefficients with the holistic human scores are calculated and shown in Table 4.

Feature	smTrain (*r*)		smEval (*r*)	
	Trans	ASR	Trans	ASR
A1	0.53	0.415	0.534	0.441
A2	0.576	0.458	0.583	0.48
A3	0.42	0.286	0.419	0.297
B1	0.597	0.478	0.564	0.452
C1	0.535	0.412	0.52	0.39
C2	-0.588	-0.469	-0.564	-0.446
C3	0.554	0.433	0.51	0.428

Table 4. Pearson correlation coefficients between content features and human holistic scores, based on both the transcription and the ASR output of smTrain and smEval.[5] Features include A1 (*re_match*), A2 (*perc_fragment_kw*), A3 (*percent_sub_keywords*), B1 (*bleu_3*), C1 (*ed_score*), C2 (*wer_3*) and C3 (*wer_score*)

5.2 Evaluations Using Human Rater Analytic Content Scores

In addition to the human rating of all spoken responses of the ETLA pilot data set with holistic scores that take into account both the dimensions of "delivery" (fluency, pronunciation, prosody) and "content," a subset of the data was further scored by human expert raters in these two dimensions separately, resulting in so-called analytic scores for delivery and content. The inter-correlation for content analytic scores was 0.79.

1,410 responses from the smTrain set and 1,402 responses from the smEval set received such analytic content scores. On this subset, table 5 shows the Pearson correlation coefficients between the content features and the analytic content scores, as well as the holistic scores, for comparison.

5.3 Scoring Model Comparison

We further examine these content features by introducing them in a scoring model to predict human rater holistic proficiency scores, using smTrain for training of the models and smEval for their evaluation. The baseline system employs 14 features related to the construct dimension of delivery, such as pronunciation, prosody and fluency.

[5] The evaluation is conducted on recognition output with more than one word. In addition, due to technical problems, such as high background noise, some responses are non-scorable for human raters, and these responses are removed from the evaluation sets. Finally, there are 3176 responses included in smTrain, and 3084 responses in smEval.

Feature	smTrain (*r*)			
	Holistic		Content	
	Trans	ASR	Trans	ASR
A1	0.529	0.415	0.563	0.434
A2	0.564	0.46	0.646	0.525
A3	0.422	0.283	0.452	0.277
B1	0.6	0.499	0.654	0.504
C1	0.527	0.43	0.555	0.46
C2	-0.588	-0.473	-0.627	-0.488
C3	0.542	0.434	0.563	0.462
Feature	**smEval (*r*)**			
	Holistic		**Content**	
	Trans	**ASR**	**Trans**	**ASR**
A1	0.525	0.424	0.538	0.436
A2	0.579	0.472	0.621	0.512
A3	0.423	0.308	0.454	0.321
B1	0.563	0.442	0.606	0.471
C1	0.521	0.4	0.539	0.422
C2	-0.543	-0.42	-0.584	-0.457
C3	0.514	0.417	0.529	0.439

Table 5. Pearson correlation coefficients between content features and human analytic content scores as well as human holistic scores.

Furthermore, an extended scoring model is built by adding the selected seven content features to the model. Table 6 provides the comparison between these two scoring models, reporting both quadratic weighted kappa and Pearson correlation coefficients between automatically predicted scores and human holistic scores on the smEval data set.

Scoring Model	Kappa	*r*
Baseline (Delivery only)	0.30	0.33
Extended (Delivery+Content)	0.53	0.56

Table 6. Scoring model comparison: quadratic weighted kappa and Pearson correlation coefficients between predicted scores (unrounded) and human holistic scores.

6 Discussion

The goal of this paper was to conceptualize, implement and evaluate features that can determine the content correctness of spoken item responses in an English language test for teachers of English who are not native speakers of English.

Based on observations from a small in-house data collection, where human test developers and content experts created example responses to 22 test items for three different score levels, we decided to implement a range of features that can capture the content correctness of test takers' responses in varying degree of precision. Our fea-

tures belong to three classes: features related to fixed expressions, with potential small variations, such as regular expressions or keywords; features based on *n*-grams of words or POS tags, including the BLEU metrics frequently used for evaluations of machine translation output; and features related to measures of string edit distance, including the WER metrics commonly used in speech recognition evaluations.

It should be noted that we use the term "content" in a fairly broad way in this paper, namely, everything in a spoken response that is not related to lower-level aspects of speech production such as fluency or pronunciation. Since the scoring rubrics for ETLA place a high emphasis both on the grammatical accuracy, as well as on the correct content (in a more narrow sense), this situation is reflected by our choice of features that focus both on elements traditionally associated with content (such as matching of keywords), as well as on elements more related to correct grammatical expressions (e.g., sequences of POS tags).

Our initial evaluations on the small in-house data collection showed that most of these features correlate well with human expert scores, both when using transcribed speech as well as when using ASR output. The absolute correlations for human transcriptions of speech range from r = 0.144 (*bleu_score*) to r = 0.789 (*re_match*), and for ASR output from r = 0.194 (*bleu_score*) to r = 0.614 (*ed_score*). The relative drop in correlation between these two conditions varies across features, but is generally around 5%-15%, with *re_match* having a much larger performance drop from r = 0.789 for transcribed speech to r = 0.537 for ASR output (32% relative decrease in performance). [6]

From this initial set of 15 features, we selected seven features based on feature performance, inter-correlation analyses (i.e., avoiding features that have a high inter-correlation and measure a similar aspect of content), and considerations of construct, i.e., which features are representing content in a way that is consistent with what human experts would consider important in determining the content correctness of a response. This subset of seven features includes three features each from the classes of flexible string matching and string edit distance, and one feature (*bleu_3*) from the *n*-gram class.

When evaluating these seven features on a larger data set, the smTrain and smEval sets of the 2012 ETLA pilot data, we find absolute correlations between features and human holistic scores ranging from r = 0.286 to r = 0.480 for ASR output, and from r = 0.419 to r = 0.597 for transcriptions. The relative decrease in correlation between transcriptions and ASR outputs ranges from 16% to 32% in these data sets (smTrain and smEval). The magnitude of content feature correlations observed in this study is similar to that of features related to fluency and pronunciation computed on spontaneous speech, as reported in Zechner et al. (2009). In fact, due to the brevity of the moderately predictable responses in ETLA, features related to fluency and pronunciation achieve correlations of less than 0.3 on this data set, making content features crucial for the assessment of speech here.

When comparing the six content features that are identical between the original feature set of 15 features (in-house data collection) and the final feature set, we observe a relative drop in feature correlation between the in-house data set and the smEval pilot data set between 1% (*blue_3*) and 36% (*ed_score*), with an average decrease of 20%. This performance decrease can be explained by (1) the more challenging data set of the pilot, as indicated, e.g., by a much higher word error rate of the ASR system (27% vs. 18%); and (2) the fact that the in-house data collection was much more constrained in terms of test taker response variation compared to the real-world pilot data.

Since a subset of the ETLA responses was also scored analytically by human raters, we could further compare the feature correlations between holistic vs. analytic content scores (Section 5.2). We find that on smEval, for all features, absolute correlations increase on human analytic content scores compared to human holistic scores. Although these differences are rather small (0.01 to 0.04), this is an indicator that our features are measuring what they are supposed to measure, since the holistic scores also take other dimensions of speech, such as fluency and pronunciation, into account.

[6] The correlation of one feature, *pos_3*, remained unchanged between the two conditions, and two features, *pos_score* and *bleu_score*, showed higher correlations for ASR output than for human transcriptions.

7 Conclusion and Future Work

This paper presented a study whose aim was to conceptualize, implement and evaluate features to measure the content correctness of test takers' responses in a new assessment for EFL teachers whose native language is not English.

We implemented and evaluated an initial set of 15 content features from three feature classes: flexible string matching, *n*-grams and string edit distance metrics. A subset of these features was then evaluated on a 2012 ETLA pilot administration, and we found correlations between features and human holistic scores in the range of $r = 0.29$ to $r = 0.48$ on ASR output. Correlations increased when comparing features with human analytic content scores.

Finally, we compared a baseline regression scoring model for prediction of human holistic scores without any content features to an extended model using seven content features and found that the model correlation substantially improved from $r = 0.33$ (baseline) to $r = 0.56$ (extended model).

Future work will include devising strategies on how to obtain RegEx features more quickly in a semi-automated way in order to reduce human labor. Further, we plan more in-depth analysis of the feature performance across different test items and item types which potentially could lead to further improvements and refinements of our content features.

References

Abeer Alwan, Yijian Bai, Matt Black, Larry Casey, Matteo Gerosa, Margaret Heritage, Markus Iseli, Barbara Jones, Abe Kazemzadeh, Sungbok Lee, Shrikanth Narayanan, Patti Price, Joseph Tepperman and Shizhen Wang. 2007. A system for technology based assessment of language and literacy in young children: the role of multiple information sources. *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, 26-30.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® V.2.0. *Journal of Technology, Learning, and Assessment*, 4(3): 159-174.

Miao Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. *Proceedings of ACL*, 722-731.

Peter W. Foltz, Darrell Laham and Thomas K. Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2).

Horacio Franco, Harry Bratt, Romain Rossier, Venkata Rao Gadde, Elizabeth Shriberg, Victor Abrash and Kristin Precoda. 2010. EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, 27(3): 401-418.

Dharmendra Kanejiya, Arun Kumary and Surendra Prasad. 2003. Automatic evaluation of students' answers using syntactically enhanced LSA. *Proceedings of Workshop on Building Educational Applications Using Natural Language Processing*, 53-60.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and Humanities*,37: 389–405.

Tom Mitchell, Terry Russell, Peter Broomhead and Nicola Aldridge. 2002. Towards robust computerised marking of free-text responses. *Proceedings of International Computer Assisted Assessment Conference*, 233-249.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. *Proceedings of EACL*, 567-575.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *Proceedings of ACL*, 311-318.

Shasha Xie, Keelan Evanini and Klaus Zechner. 2012. Exploring content features for automated speech scoring. *Proceedings of NAACL-HLT*, 103-111.

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