Speech- and Text-driven Features for Automated Scoring of English Speaking Tasks

Anastassia Loukina **Nitin Madnani** **Aoife Cahill**

Educational Testing Service
Princeton, NJ, 08541 USA
{aloukina,nmadnani,acahill}@ets.org

Abstract

We consider the automatic scoring of a task for which both the content of the response as well the pronunciation and fluency are important. We combine features from a text-only content scoring system originally designed for written responses with several categories of acoustic features. Although adding any single category of acoustic features to the text-only system on its own does not significantly improve performance, adding *all* acoustic features together does yield a small but significant improvement. These results are consistent for responses to open-ended questions and to questions focused on some given source material.

1 Introduction

English language proficiency assessments designed to evaluate speaking ability often include tasks that require the test takers to speak for one or two minutes on a particular topic. These responses are then evaluated by a human rater in terms of how well the test takers addressed the question as well as the general proficiency of their speech. Therefore, a system designed to automatically score such responses should combine NLP components aimed at evaluating the content of the response as well as text-based aspects of speaking proficiency such as vocabulary and grammar, and speech-processing components aimed at evaluating fluency and pronunciation. In this paper, we investigate the automatic scoring of such spoken responses collected as part of a large-scale assessment of English speaking ability.

Our corpus contains responses to two types of questions — both administered as part of the same speaking ability task — that we will refer to as “source-based” and “general”. For source-based questions, test-takers are expected to use the provided materials (e.g., a reading passage) as a basis for their response and, therefore, good responses are likely to have similar content. In contrast, general questions are more open-ended such as “What is your favorite food and why?” and, therefore, the content of such responses can vary greatly across test takers. In total, our corpus contains over 150,000 spoken responses to 147 different questions, both source-based and general.

We focus our system on two dimensions of proficiency: content, that is how well the test-taker addressed the task, and delivery (pronunciation and fluency). To evaluate the content of a spoken response, we use features from an existing content-scoring NLP system developed for written responses that uses the textual characteristics of the response to produce a score. We apply this system to the 1-best ASR (automatic speech recognition) hypotheses for the spoken responses.

To evaluate the fluency and pronunciation of the speech in the response, we use features from an existing speech-scoring system that capture information relevant to spoken language proficiency and cannot be obtained just from the ASR hypothesis. We compare the contributions of several types of features: speech rate, pausing patterns, pronunciation measures based on acoustic model scores and ASR confidence scores as well as more complex features that capture timing patterns and other prosodic properties of the response.

We combine the two types of features (text-driven and speech-driven) and compare the performance of this model to two baseline models, each using only one type of features. All models are evaluated by comparing the scores obtained from that model to the scores assigned by human raters to the same responses. We hypothesize that:

Proceedings of the First Workshop on Speech-Centric Natural Language Processing, pages 67–77
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

- Given the characteristics of the two types of questions, the model with *only* text-driven features will exhibit better performance for source-based questions as opposed to general ones.

- Since human raters reward how well the response addresses the question as well as higher spoken proficiency, the combined model that uses *both* text-driven features (for content) & speech-driven features (for proficiency) will perform better than the individual text-only and speech-only models.

We find that our results generally meet our expectations but interestingly the improvement in performance by combining text-driven & speech-driven features — while significant — is not as large as we had expected, i.e., the combination does not add much over the text-driven features. We conclude by discussing possible reasons for this observation.

2 Related work

Most systems for scoring proficiency of spoken responses rely on ASR to obtain a transcription of the responses. Since work on automated scoring predates the availability of accurate ASR, the majority of earlier automated scoring systems focused on tasks that elicited restricted speech such as read-aloud or repeat-aloud. Such systems either did not consider the content of the response at all or relied on relatively simple string-matching (see Eskenazi (2009) and Zechner et al. (2009) for a detailed review). Even when the task required answering open-ended questions, e.g. in the PhonePass test (Townshend et al., 1998; Bernstein et al., 2000), fluency was considered more important than content.

Zechner et al. (2009) were one of the first to attempt automatically scoring tasks that not only elicited open-ended responses but where content knowledge was also an integral part of the task. They did not use any explicit features to measure content because of the high ASR word error rates (around 50%). Instead, they focused on fluency-related features on which ASR errors had little impact. They reported a correlation of 0.62 between the system and human scores.

More recent studies have explored different approaches to evaluating the content of spoken responses. Xie et al. (2012) explored content measures based on the lexical similarity between the response and a set of reference responses. A content-scoring component based on word vectors was also part of the automated scoring engine described by Cheng et al. (2014). In both these studies, content features were developed to supplement other features measuring various aspects of speaking proficiency. Neither study reported the relative contributions of content and speech features to the system performance.

Although it may seem obvious that, given the nature of the task, a model using both speech-based and content-based features should outperform models using only one of them, it may not turn out that way. Multiple studies that have developed new features measuring vocabulary, grammar or content for spoken responses have reported only limited improvements when these features were combined with features based on fluency and pronunciation (Bhat and Yoon, 2015; Yoon et al., 2012; Somasundaran et al., 2015). Crossley and McNamara (2013) used a large set of text-based measures including Coh-Metrix (Graesser et al., 2004) to obtain fairly accurate predictions of proficiency scores for spoken responses to general questions similar to the ones used in this study based on transcription only, without using any information based on acoustic analysis of speech. It is not possible to establish from published results how their system would compare to the one that also evaluates pronunciation and fluency. They did not compute any such features and their results based on text are not directly comparable to the other papers discussed in this section since some of their features required a minimum length of 100 words and, therefore, required them to combine several responses to meet this text length requirement.

Most recently, Loukina and Cahill (2016) compared the performance of several text- and speech-based scoring systems and found that even though each system individually achieved reasonable accuracy in predicting proficiency scores, there was no improvement in performance from combining the systems. They argued that the majority of speakers who perform well along one dimension of language proficiency are also likely to perform well along other dimensions (cf. also Xi (2007) who reports similar results for human analytic scores). Consequently, the gain in performance from combining different systems is small or nonexistent. Their work focused on general language proficiency features and did not consider the content of the responses.

This study has several significant differences from previous work. We consider content-scoring features that go well beyond word vectors and instead build a textual profile of the response. Furthermore, we conduct more fine-grained analyses and report the *types* of speech-driven features that add the most information to content-scoring features. We also examine how the interactions between content and speech features vary by types of questions. Finally, we conduct our analyses on a very large corpus of spoken responses which, to our knowledge, is the largest used so far in studies on automated scoring of spoken responses. The size of the data allows us to identify patterns that persist across responses to multiple questions and are more reliable.

3 Methodology

3.1 Data

The data used in this study comes from a large-scale English proficiency assessment for non-native speakers administered in multiple countries. Each test-taker answers up to 6 questions: two general and four source-based. For source-based questions, test-takers are provided with spoken and/or written materials and asked to respond to a question based on these materials while general questions have no such materials. Test-takers are given 45 seconds to answer general questions and one minute to answer source-based questions.

Each response was scored by a professional human rater on a scale of 1–4. When assigning scores, raters evaluated both how well the test taker addressed the task in terms of content as well as the overall intelligibility of the speech. A response scored as a "1" would be limited in content and/or largely intelligible due to consistent pronunciation difficulties and limited use of vocabulary and grammar. On the other hand, a response scored as a "4" would fulfill the demands of the task and be highly intelligible with clear speech and effective use of grammar and vocabulary. The raters are provided with the description of typical responses at each score level and are asked to provide a holistic score without prioritizing any particular aspect.

For this study, we used responses to 147 questions (48 general questions and 99 source-based questions) from different administrations of the assessment. We excluded responses where the ASR hypothesis contained fewer than 10 words (0.2% of the original sample). The final corpus used for model training and evaluation included 153,461 responses from 33,503 test takers.[1] As shown in Table 1, the number of responses for a question was consistent for the two question types.

Type	**general**	**source-based**
N questions	48	99
N responses	50,811	102,650
Average responses	1058.6	1036.9
Median responses	902.5	936.0
Min responses	255	250
Max responses	2030	2,174
Average N words	90.8	120.3

Table 1: Total number of responses for each question type; the average, median, min and max number of responses per question; the average number of words in responses to each question computed based on ASR hypotheses.

Test-takers from each administration were randomly split between training and evaluation partition with about 70% of responses to each question allocated to the training set and 30% allocated to the evaluation set. We ensured that, across all 147 questions, responses from the same test taker were *always* allocated to the same partition and that test takers in training and evaluation sets had similar demographic characteristics.

3.2 Automatic Speech Recognizer

All responses were processed using an automatic speech-recognition system based on the Kaldi toolkit (Povey et al., 2011) using the approach described by Tao et al. (2016). The language model was based on tri-grams. The acoustic models were based on 5-layer DNN and 13 MFCC-based features. Tao et al. (2016) give further detail about the model training procedure.

The ASR system was trained on a proprietary corpus consisting of 800 hours of non-native speech from 8,700 speakers of more than 100 native languages. The speech in the ASR training

[1] Our sampling was done by question and some questions were repeated across administrations in combination with other questions not included in this study. The number of speakers who answered each question varied between 250 and 2,174, with an average of 1,043 responses to each question. For 68% of test takers, we had responses to all 6 questions.

corpus was elicited using questions similar to the ones considered in this study. There was no overlap of speakers or questions between the ASR training corpus and the corpus used in this paper. We did not additionally adapt the ASR to the speakers or responses in this study.

While no transcriptions are available to compute the WER of the ASR system on this corpus, the WER for this system on a similar corpus is around 30%.

3.3 Text-driven features

Scoring responses for writing quality requires measuring whether the student can organize and develop an argument and write fluently with no grammatical errors or misspellings. In contrast, scoring for content deals with responses to open-ended questions designed to test what the student knows, has learned, or can do in a specific subject area (such as Computer Science, Math, or Biology) (Sukkarieh and Stoyanchev, 2009; Sukkarieh, 2011; Mohler et al., 2011; Dzikovska et al., 2013; Ramachandran et al., 2015; Sakaguchi et al., 2015; Zhu et al., 2016).[2]

In order to measure the content of the spoken responses in our data, we extract the following set of features from the 1-best ASR hypotheses for each response:

- lowercased word n-grams (n=1,2), including punctuation
- lowercased character n-grams (n=2,3,4,5)
- syntactic dependency triples computed using the ZPar parser (Zhang and Clark, 2011)
- length bins (specifically, whether the log of 1 plus the number of characters in the response, rounded down to the nearest integer, equals x, for all possible x from the training set). For example, consider a question for which transcriptions of the responses in the training data are between 50 and 200 characters long. For this question, we will have 3 length bins numbered from 5 ($\lfloor \log_2 51 \rfloor$) to 7 ($\lfloor \log_2 201 \rfloor$). For a new response of length 150 characters, length bin 7 ($\lfloor \log_2 151 \rfloor$) would be the binary feature that gets a value of 1 with the other two bins getting the value of 0.

We refer to these features as "text-driven" features in subsequent sections.

3.4 Speech-driven features

We used five types of features that capture information relevant to the fluency and pronunciation of a spoken response and are extracted based on the acoustic properties of the spoken responses. These are primarily related to spectral quality (*how* the words and sounds were pronounced) and timing (*when* they were pronounced). All features are summarized in Table 2. Each feature type is computed as a continuous value for the whole response and relies on the availability of *both* the speech signal as well as the 1-best ASR hypothesis.

The first set of features ("speech rate") computes the words spoken per minute with and without trailing and leading pauses. Speech rate has been consistently identified as one of the major covariates of language proficiency and the features in this group have some of the highest correlations with the overall human score.

Name	Description	N_{feat}	r
speech rate	Speech rate	3	.42
quality	Segmental quality	6	.41
pausing	Location and duration of pauses	9	.34
timing	Patterns of durations of individual segments	9	.36
prosody	Time intervals between stressed syllables	6	.30

Table 2: The five sets of speech features used in this study along with the number of features in each group and the average correlations with human score across all features and questions (Pearson's r).

The second set of features ("quality") captures how much the pronunciation of individual segments deviates from the pronunciation that would be expected from a proficient speaker. This includes the average confidence scores and acoustic model scores computed by the ASR system for the words in the 1-best ASR hypothesis. Since the ASR is trained on a wide range of proficiency levels, we also include features computed using the two-pass approach (Herron et al., 1999; Chen et al., 2009). In this approach, the acoustic model scores for words in the ASR hypothesis are recomputed using acoustic models trained on native

[2] See Table 3 in Burrows et al. (2015) for a detailed list.

speakers of English.

The third set of features captures pausing patterns in the response such as mean duration of pauses, mean number of words between two pauses, and the ratio of pauses to speech. For all features in this group the pauses were determined based on silences in the ASR output. Only silences longer than 0.145 seconds were included.

The fourth set of features ("prosody") measures patterns of variation in time intervals between stressed syllables as well as the number of syllables between adjacent stressed syllables (Zechner et al., 2011).

The final set of features ("timing") captures variation in the duration of vowels and consonants. This category includes features such as relative proportion of vocalic intervals or variability in adjacent consonantal intervals (Lai et al., 2013; Chen and Zechner, 2011) as well as features which compare vowel duration to reference models trained on native speakers (Chen et al., 2009).

We refer to these five feature sets as "speech-driven" features in subsequent sections.

3.5 Scoring models

We combined the text-driven features and speech-driven features into a single set of features and trained a support vector regressor (SVR) model with an RBF kernel for each of the 147 questions, using the human scores in the training partition as the labels. We used the *scikit-learn* (Pedregosa et al., 2011) implementation of SVRs and the SKLL toolkit.[3] The hyper-parameters of each SVR model (γ and C) were optimized using a cross-validated search over a grid with mean squared error (MSE) as the objective function.

In addition to the combined scoring models, we also built the following scoring models for each question:

- A model using only the text-driven features (1 model)
- A model using only the speech-driven features (1 model)
- Models using each of the individual speech-driven feature sets (5 models)
- Combinations of the text-driven model with each of the individual speech-driven feature sets (5 models)

In total, we built 1,911 scoring models (13 models for each of the 147 questions).

We evaluated each of our models on a held-out evaluation partition for each of the questions. We used the R^2 between the predicted and human scores computed on the evaluation set as a measure of model performance:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad (1)$$

where y_i are the observed values (human scores), $\hat{y}_i$ are the predicted values and $\bar{y}$ is the mean of observed scores.

As shown in Eq. 1, R^2 standardizes the MSE by the total variance of the observed values leading to a more interpretable metric that generally varies from 0 to 1, where 1 corresponds to perfect prediction and 0 indicates that the model is no more accurate than simply using mean value as the prediction.

4 Results

4.1 Model performance

Table 3 shows the mean R^2 for different types of questions and models across the 147 questions in our study.

Model	**general**	**source-based**
text + speech	.352	.442
text-only	.335	.431
speech-only	.325	.394
speech rate	.275	.341
pausing	.259	.312
quality	.303	.365
prosody	.256	.309
timing	.282	.329
text + speech rate	.339	.433
text + pausing	.340	.434
text + quality	.343	.436
text + prosody	.341	.434
text + timing	.342	.434

Table 3: Average R^2 achieved by different models on different types of questions (N=99 for general questions and N=48 for source-based questions).

We used linear mixed-effect models (cf. Snijders and Bosker (2012) for a comprehensive introduction and Searle et al. (1992) who give an extensive historical overview) to identify statistically significant differences among the various

[3] http://github.com/EducationalTestingService/skll

models. The mixed-effect models were fitted using the `statsmodels` Python package (Seabold and Perktold, 2010). We used model R^2 as a dependent variable, question as a random factor, and model and question type (general or source-based) as fixed effects. We include both the main effects of model and question type as well as their interaction and used the text-driven model as the reference category.

We observed that for both general and source-based questions:

1. The performance of the combined model (text + speech) using all five types of speech-driven features as well as the text-driven features was significantly better than both the text-only model as well as the speech-only model. The effect size of the improvement over the text-only model was small with the average R^2 increasing only slightly from .335 to .352 for source-based questions and from .431 to .442 for general questions ($p = 0.002$).
2. The performance of the text-only model was significantly better than the performance of each of the 5 models trained using only one group of speech-driven features ($p < 0.0001$).
3. There was no significant difference between the performance of the text-only model and the 5 models combining the text-driven features with each of the individual speech-driven feature sets.

In addition, as we predicted, there was a significant difference in model performance between general and source-based questions. Surprisingly, this difference was observed for *all* 13 models; all models achieved higher performance for source-based questions ($p < 0.0001$). We also observed a significant interaction between model type and question type: the difference between the speech-only model and the text-only model was higher for source-based questions than for general questions. Furthermore, while there was no statistically significant difference between the speech-only model and text-only model for general questions (.335 vs. .325, p=0.061), the difference between these two types of models *was* significant for source-based questions with the text-only model outperforming the speech-only model (R^2 = .431 vs. .394, $p < 0.0001$).

Finally, we compared the performance of our combined system to other published results on automated speech scoring reviewed earlier in this paper. Since most previous work reports their results using Pearson's correlation coefficients, we computed the same for our system for an easier comparison. Table 4 reports the correlations for our model as well as those reported in previous studies on automatically scoring responses to similar questions. It shows that our system performance is either comparable or better than previous results.

Model	**general**	**source-based**
text + speech	.60	.67
text-only	.59	.66
speech-only	.58	.63
Xie et al.	.40	.59
Loukina & Cahill	.64 (overall)	

Table 4: Average Pearson's r achieved by the three of the models in this study and the best performing models reported in the literature; Loukina and Cahill (2016) combine language proficiency features from speech and text and do not report performance by question type; Xie et al. (2012) use content features based on cosine similarity but no other language proficiency features. If a paper reports results based on both ASR hypothesis and human transcription, we only use the results based on ASR hypothesis.

4.2 Information overlap between text and speech: The role of disfluencies

A relatively minor improvement between the text-only model and the combined text + speech model suggests that text-driven features already incorporate some of the information captured by the speech-driven features or that the type of of information captured by two sets of features are highly correlated. We use disfluencies and pauses as a test case to explore this hypothesis further.

Our text-driven features computed on the ASR hypothesis included *all* information stored in that hypothesis including hesitation markers ("uh", "uhm" etc.) and silence markers. In other words, even though our text-driven features are designed to measure content for written responses, when applied to spoken responses they might also have captured some information related to fluency. In order to confirm this hypothesis, we removed hesitation markers and pauses from the 1-best ASR hypotheses and repeated our analysis with the primary models, i.e., text-only (with and without disfluencies), speech-only, and text + speech (with and without disfluencies) – a total of 5 models.

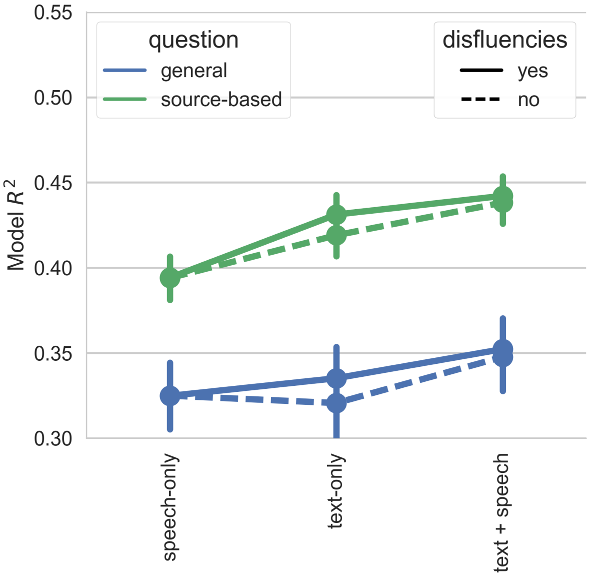

Figure 1: A plot showing the scoring performance across the two question types for two different conditions: including disfluencies and pauses in the 1-best ASR hypotheses and excluding them.

The results of this analysis are presented in Figure 1. As before, we used a linear mixed-effects model to evaluate which differences were statistically significant. Removing disfluencies and pauses from the hypotheses led to a significant decrease in the performance of the text-only model for both types of questions (R^2 = .335 vs. .321 for general question and .431 vs. .419 for source-based questions, $p = 0.001$).

We still observed no significant difference in performance between the text-only model without disfluencies and pauses and speech-only model for *general* questions. However, the difference between the text-only model and speech-only models for *source-based* questions remained significant even after removing the disfluencies and pauses from the ASR hypothesis (.394 vs. .419, $p < 0.001$).

Finally, for the combined text + speech model, there was no significant difference between including and excluding disfluencies and pauses from the ASR hypotheses.

4.3 Performance variation across questions

In Section 4.1, we presented general observations after we controlled for the individual question as a random effect. However, we also observed that all of the models showed substantial variation in performance across the 147 questions. The R^2 for the best performing model (text + speech) varied between .062 and .505 for the general questions and between .197 and .557 for the source-based questions. Given such a striking variation, we conducted further exploratory analyses into factors that may have affected model performance. We focused these analyses on the best performing model (text + speech).

First, we considered the sample size for each question. As shown in Table 1, the number of responses used to train and evaluate the models varied across questions and, therefore, we might expect lower performance for questions with fewer responses available for model training. A linear regression model with R^2 as the dependent variable and the sample size as the independent variable showed that the sample size accounted for 9.8% of variability in model performance for source-based questions ($p = 0.0016$) and 19.2% of variability in model performance for general questions ($p = 0.0018$). In other words, while there was a significant effect of the sample size, it was not the main factor.

Another possible source of variation in model performance may be the variation in ASR word error rate itself. Since no reference transcriptions are available for our corpus, we cannot test this hypothesis directly. As an indirect measurement, however, we consider the number of words in the ASR hypotheses across questions. If the ASR consistently failed to produce accurate hypotheses for some questions, this might manifest as consistently shorter ASR hypothesis for such questions, and, hence, discrepant scoring performance.

The average number of words varied between 83.6 and 100.2 for general questions and between 109.0 and 132.6 for source-based questions. While there was a statistically significant difference in number of words between the questions, we found that the average number of words in responses to a given question did not have a significant effect on the model performance (p = 0.09 for general questions and p = 0.03 for source-based questions[4]).

[4] Significance threshold was adjusted for multiple compar-

Of course, not all ASR failures necessarily result in shorter hypotheses and, therefore, further analysis based on the actual WER is necessary to reject or confirm any possible effect of ASR on model performance.

There are additional factors that might have contributed to the variation in model performance pertaining to both the properties of the question and the characteristics of test takers who answered each question. We plan to further explore the contribution of these factors in future work. Our results highlight the impact of the actual question in automated scoring studies and suggest that the results based on a small set of questions may be unreliable due to the large variation across questions.

5 Discussion

We considered a combination of text-driven and speech-driven features for automated scoring of spoken responses to general and source-based questions. We found that for both types of questions a combination of the two types of features outperforms models using only one of those two types of features. However, a significant improvement could only be achieved by combining several types of speech features. There was no improvement in model performance when text-driven features were combined with only one type of speech-driven features such as speech rate or pausing patterns.

Surprisingly we found that all models performed better for source-based questions than for general questions — a result we plan to explore further in future work. We also found that for general questions where the content of responses can vary greatly, the model that uses only speech-driven features achieves the same performance as the one only using text-driven features. We hypothesize that this is because in the absence of "pre-defined" content both systems measure various aspects of general linguistic proficiency and these tend to be closely related as we discussed in Section 2. At the same time, for source-based questions where the test-takers are expected to cover already provided content, the performance of the model using only text-driven features is significantly better than the performance of the model using only speech-driven features.

Although we do observe a significant improvement in scoring performance by combining text-driven features (to measure content) and speech-driven features (to measure fluency and pronunciation), the improvement is not as large as one might have expected. This may appear counter-intuitive considering the perceived role of fluency and pronunciation for this task. There are several possible reasons for this result.

First, it is possible that the speech-driven features in our study do not really capture the information present in the acoustic signal that is relevant to this task. However, this is unlikely given that the features we considered in this paper capture many aspects of spoken language proficiency and cover all major types of features used in other studies on automated evaluation of spoken proficiency. This is further illustrated by the fact that for general questions, the speech-only model performed as well as the text-only model. We also note that recent work by Yu et al. (2016) used neural networks to learn high-level abstractions from frame-to-frame acoustic properties of the signal and showed that these features provided a very limited gain over the features considered in this study.

Second, our results may be skewed because of poorly performing ASR. Although we cannot reject this hypothesis given the lack of human transcriptions for the responses, it is unlikely to hold because the same ASR system achieves a WER of 30% on another corpus of responses with similar demographic and response characteristics. Furthermore, previous studies compared the performance of speech and text features computed using manual transcriptions to those computed using ASR hypotheses (with a similar WER) and reported only a small drop in performance: $r = 0.67$ for transcriptions vs. $r = 0.64$ for ASR hypotheses (Loukina and Cahill, 2016).

Another possible reason may be the way in which the speech-driven and text-driven features are combined. For each response, we simply concatenate the small, dense vector of 33 continuous speech-driven features with the very large, sparse vector of tens of thousands of binary text-driven features. In such a scenario, the impact of speech-driven features may be mitigated due to the disproportionate number of sparse text-driven features. A better combination approach might be stacked generalization (Wolpert, 1992): building separate models for speech-driven features and text-driven

isons performed in this section to $\alpha = 0.0125$ using Bonferroni correction features and then combining their predictions in a third higher-level model. Sakaguchi et al. (2015) showed that stacking only improves over straightforward concatenation when there are a limited number of responses in the training data and we have a fairly large number of training responses available for each of our questions. However, the idea certainly merits further exploration.

A more likely explanation is that there is only a limited amount of information contained in the acoustic signal that is not already present in one way or another in the ASR hypothesis. We already discussed earlier in this paper that different aspects of language proficiency are highly correlated and thus one model can often achieve good empirical performance by measuring only one particular aspect. A related observation here is that many aspects of the spoken signal are already captured by ASR hypothesis. For example, while ASR hypothesis does not reflect the duration of pauses, it does contain information about the presence and location of pauses and whether they are accompanied by the hesitation markers. Similarly, the “choppiness” of speech would manifest itself in both prosody and syntax. This claim is supported by our results which show that removing disfluencies and pauses from the ASR hypotheses degrades the performance of the text-only system significantly but has no effect on the performance of the combined system since the same information is also captured by the speech-driven features.

In this study, we focused on content, fluency, and pronunciation and did not consider any features designed to measure other important elements of speaking proficiency such as grammar or choice of vocabulary. It is likely that some aspects of these are already indirectly captured by the content-scoring part of our system but future research will show whether system performance can be further improved by features that have been specifically designed to evaluate these aspects of spoken proficiency.

6 Conclusions

In this paper, we built automated scoring models for an English speaking task for which both content knowledge as well as an ability to produce fluent intelligible speech are required in order to obtain a high score. We applied an existing content-scoring NLP system (designed for written responses) to the 1-best ASR hypotheses of the spoken responses in order to extract text-driven features that measure content. To measure spoken fluency and pronunciation, we extracted a set of 33 features based on the acoustic signal for the response. Combining the two types of features results in a significant but smaller than expected improvement compared to using each type of features by itself. A deeper examination of the features yields that there is likely to be significant information overlap between the speech signal and the ASR 1-best hypothesis especially when the hypothesis includes pausing and silence markers. Based on these observations, we conclude that although our approach of extracting features from the speech signal and combining them with text-driven features extracted from the ASR hypothesis is certainly moderately effective, further research is warranted in order to determine whether a larger improvement can be obtained for this task.

Acknowledgments

We would like to thank Keelan Evanini, Su-Youn Yoon, Brian Riordan, and the anonymous SCNLP reviewers for their useful comments and suggestions. We also thank Matt Mulholland for help with processing the data.

References

Jared Bernstein, John De Jong, David B. Pisoni, and Brent Townshend. 2000. Two experiments on automatic scoring of spoken language proficiency. In *Proceedings of InStil2000*. pages 57–61.

Suma Bhat and Su-Youn Yoon. 2015. Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication* 67:42–57.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education* 25(1):60–117.

Lei Chen and Klaus Zechner. 2011. Applying rhythm features to automatically assess non-native speech. In *Proceedings of Interspeech*. pages 1861–1864.

Lei Chen, Klaus Zechner, and Xiaoming Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of NAACL*. pages 442–449.

Jian Cheng, Yuan Zhao D’Antilio, Xin Chen, and Jared Bernstein. 2014. Automatic assessment of the speech of young English learners. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*. pages 12–21.

Scott Crossley and Danielle McNamara. 2013. Applications of text analysis tools for spoken response grading. *Language Learning & Technology* 17(2):171–192.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of SemEval*. pages 263–274.

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

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2):193–202.

Daniel Herron, Wolfgang Menzel, Eric Atwell, Roberto Bisiani, Fabio Daneluzzi, Rachel Morton, and Juergen a Schmidt. 1999. Automatic localization and diagnosis of pronunciation errors for second-language learners of English. In *Proceedings of EuroSpeech*. pages 855–858.

Catherine Lai, Keelan Evanini, and Klaus Zechner. 2013. Applying rhythm metrics to non-native spontaneous speech. In *Proceedings of SLaTE*. pages 159–163.

Anastassia Loukina and Aoife Cahill. 2016. Automated scoring across different modalities. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*. pages 130–135.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of ACL: HLT*. pages 752–762.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*. pages 97–106.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. In *Proceedings of NAACL*. pages 1049–1054.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the Python in Science Conference*. pages 57–61.

Shayle R. Searle, George Casella, and Charles E. McCulloch. 1992. *Variance Components*. Wiley-Interscience.

Tom A.B. Snijders and Roel J. Bosker. 2012. *Multilevel Analysis*. Sage, London, 2nd edition.

Swapna Somasundaran, Chong Min Lee, Martin Chodorow, and Xinhao Wang. 2015. Automated scoring of picture-based story narration. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*. pages 42–48.

Jana Z. Sukkarieh. 2011. Using a MaxEnt classifier for the automatic content scoring of free-text responses. In *Proceedings of the 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. AIP Press, pages 41–48.

Jana Z. Sukkarieh and Svetlana Stoyanchev. 2009. Automating model building in c-rater. In *Proceedings of the Workshop on Applied Textual Inference*. pages 61–69.

Jidong Tao, Shabnam Ghaffarzadegan, Lei Chen, and Klaus Zechner. 2016. Exploring deep learning architectures for automatically grading non-native spontaneous speech. In *Proceedings of ICASSP*. pages 6140–6144.

Brent Townshend, Brent Bernstein, Ognjen Todic, and Eryk Warren. 1998. Estimation of spoken language proficiency. In *Proceedings of the Workshop on Speech Technology in Language Learning (STiLL)*. pages 93–96.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks* 5:241–259.

Xiaoming Xi. 2007. Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing* 24(2):251–286.

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

Su-Youn Yoon, Suma Bhat, and Klaus Zechner. 2012. Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the Workshop on the innovative use of NLP for Building Educational Applications*. pages 180–189.

Zhou Yu, Vikram Ramanarayanan, David Suendermann-Oeft, Xinhao Wang, Klaus Zechner, Lei Chen, Jidong Tao, Aliaksei Ivanou, and Yao Qian. 2016. Using bidirectional LSTM recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. pages 338–345.

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 & Understanding*. pages 461–466.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational linguistics* 37(1):105–151.

Mengxiao Zhu, Ou Lydia Liu, Liyang Mao, and Amy Pallant. 2016. Use of automated scoring and feedback in online interactive Earth science tasks. In *Proceedings of the 2016 IEEE Integrated STEM Education Conference*.