Toward Automated Content Feedback Generation for Non-native Spontaneous Speech

Su-Youn Yoon, Ching-Ni Hsieh, Klaus Zechner,
Matthew Mulholland, Yuan Wang, and Nitin Madnani
Educational Testing Service
660 Rosedale road, Princeton, USA
{syoon, chsieh, kzechner, mmulholland, ywang, nmadnani}@ets.org

Abstract

In this study, we developed an automated algorithm to provide feedback about the specific content of non-native English speakers' spoken responses. The responses were spontaneous speech, elicited using integrated tasks where the language learners listened to and/or read passages and integrated the core content in their spoken responses. Our models detected the absence of key points considered to be important in a spoken response to a particular test question, based on two different models: (a) a model using word-embedding based content features and (b) a state-of-the art short response scoring engine using traditional n-gram based features. Both models achieved a substantially improved performance over the majority baseline, and the combination of the two models achieved a significant further improvement. In particular, the models were robust to automated speech recognition (ASR) errors, and performance based on the ASR word hypotheses was comparable to that based on manual transcriptions. The accuracy and F-score of the best model for the questions included in the train set were 0.80 and 0.68, respectively. Finally, we discussed possible approaches to generating targeted feedback about the content of a language learner's response, based on automatically detected missing key points.

1 Introduction

In this study, we propose an automated algorithm which provides feedback about the specific content of non-native English speakers' spoken responses. It is designed to help language learners preparing for a speaking test that is part of an assessment of English proficiency for academic purposes. The speaking test includes questions eliciting spontaneous speech. In particular, the items require language learners to read and/or listen to stimulus materials and then integrate and reproduce the key content from the source materials into their speaking performances (hereafter, integrated tasks). Research in integrated task performance (Brown et al., 2005; Cotos, 2011; Frost et al., 2012; Xi, 2010) has shown that human raters pay substantial attention to test-takers' speech content. A speaker's performance is evaluated by the content completeness and accuracy of the reproduced information, in addition to linguistic criteria including fluency, pronunciation, grammar, and vocabulary.

The current study investigated automated feedback through the dimension of content completeness. This content-aspect of speech performance refers to the degree to which an individual can process, select, integrate, and reproduce key source information into a subsequent oral response. The ability to reproduce complete content represents a critical aspect of integrated speaking task performance and is evaluated by the number of key points reproduced from the input materials (Frost et al., 2012). Key points are brief descriptions of content elements that test developers determine to be important in responses to a particular test question.

Providing feedback on content aspects of speech can help language learners discern the quality of their speech performance beyond linguistic dimensions such as fluency or grammar. This type of feedback is particularly relevant and crucial when we consider integrated task performance, because the ability to accurately and adequately recreate the source materials is an essential language skill required in real-world academic or workplace contexts.

Despite the importance of content as a component of speech, few studies have explored automated content feedback. To address this gap, we aim to develop a content feedback algorithm. In

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 306–315
Florence, Italy, August 2, 2019.

this study, we trained automated models to detect the absence of key points that are the core content expected in correct answers. Next, we discussed possible ways to generate content feedback based on the output of the automated models.

2 Previous studies

In the past two decades, feedback has become a central issue for second language education research, and language teachers and researchers have continued to identify guidelines and best practices for providing learners with effective feedback (Lyster et al., 2013). Advances in technology have led to increased research efforts in developing automated feedback systems that can support language learners (Xi, 2010). Automated feedback systems can provide practical benefits, such as making teaching and learning more individualized, efficient, and cost effective. However, research on automated feedback is still scarce and primarily focused on aspects of learners' writing performance, rather than speech (Cotos, 2011). In automated feedback for spoken responses, previous studies focused on pronunciation (Franco et al., 2010) and prosody (Eskenazi et al., 2007) from restricted speech.

Automated scoring of, or automated feedback generation about content in spontaneous speech is a challenging task for a variety of reasons. First, an automated speech recognition (ASR) system is used to generate an automated transcription of a spoken response as an input of the content feature generator. Errors at the ASR stage may negatively affect the content features such that they are noisy and distorted to some extent. Secondly, and more importantly, spontaneous speech, unlike read speech, is highly variable, and particular aspects of content can be expressed in many different ways by different speakers. Consequently, relatively few studies have explored content of spontaneous spoken responses. Xie et al. (2012) and Cheng et al. (2014) assessed content using similarity scores between test responses and highly proficient sample responses, based on content vector analysis (CVA). Loukina and Cahill (2016) used a content-scoring engine based on many sparse features, such as unigrams and bigrams, trained on a large corpus of existing responses. These studies were based on traditional character or word n-grams. Recently, significant improvement in ASR systems, semantic modeling technology based on more advanced deep-neural networks (DNN), and larger training data sets encouraged researchers in the automated scoring field to explore content-modeling for spoken responses. For instance, Chen et al. (2018) and Qian et al. (2018) developed automated oral proficiency scoring models using diverse neural models and achieved comparable or superior performance to sophisticated linguistic feature-based systems. In addition, Yoon et al. (2018) and Rei and Cummins (2016) used similarity scores between the prompt texts and test responses based on word embeddings. Compared to the traditional word-matching based method, they have the advantage of capturing topical relevance that is not based on specific, identical words. However, these studies have focused only on scoring, and based on our knowledge, no study has explored content feedback for spontaneous speech.

3 Overview of the approach

In order to address this gap, we developed an automated algorithm which provides feedback about content completeness for non-native speakers' spontaneous speech. Distinct from previous content scoring approaches that look at correctness of overall content by calculating similarity scores with high-scoring responses, our algorithm first determines absence of *individual* key points. The absence of a key point signals an issue in the content completeness of a spoken response. Next, we provide a list of missing key points with feedback about how to improve content completeness to the speakers. Our approach is able to provide much more fine-grained and targeted feedback about the content of a response, as compared to a traditional holistic approach.

In order to determine the absence of the key points, we calculated similarity scores between a spoken response and a key point using a short response scoring engine (Heilman and Madnani, 2013) and new word-embedding based features. The short response scoring engine generally requires a sizable amount of response data for each question to achieve a reliable performance. Collecting question-specific data is a difficult task. Thus, the word-embedding features, that do not require any sample responses for each question for the feature training, have a strong advantage for practical systems. We evaluated the algorithm in two different conditions (questions in the training data vs. questions not in the training data) and ex-

plored the impact of a question-specific training dataset.

4 Data

We used a collection of spoken responses from an English proficiency assessment. 395 non-native speakers with a wide range of proficiency levels[1] and from 52 different native language backgrounds produced a total of 1,185 responses. Each response consisted of around one minute of spontaneous speech. We used four forms[2], and each student responded to the questions on one form. We collected approximately 100 speakers' responses per form.

When producing the integrated speaking tasks that were used for the current study, expert assessment developers first generated a list of key points to guide the creation of the reading and listening passages. These key points were provided to and used by human raters to evaluate content completeness of the spoken responses. Six key points were generated for each speaking task (henceforth, Key Point 1 to Key Point 6).

Each key point generally consisted of one complete sentence. Key Point 1 and 2 were about the mentioning of the concepts introduced in the source materials or the general opinions voiced (i.e., agree or disagree with a situation/change/proposal). Depending on the nature of the task questions, Key Point 3, 4, 5, and 6 involved brief definitions of the concepts, reasons provided for the opinions voiced, or detailed examples that illustrated the topics or concepts discussed. Key Point 1 and 2 were relatively straightforward whereas Key Point 3 to 6 contained more elaborated content.

To give an idea of what the key points look like, we provide one sample in Figure 1. Originally, a question, a reading material, and a listening material were one set, and there were three Key Points for the reading material and three Key Points for the listening material. Due to the page limit, we provide only the question, the reading material, and three Key Points relevant to the reading material.

The human transcripts of the audio files were

Reading Material

University administrators announced yesterday that the sculpture program, a division of the art department, will be eliminated. The main reason is a lack of student interest, reported one administrator. Although the number of art students has increased, fewer and fewer art majors are taking sculpture classes. Furthermore, the department's only sculpture professor is retiring this year. Given the art department's limited budget, the administrator explained, it just doesn't make sense to hire a new full-time professor to teach sculpture for only a handful of students.

Question

The woman expresses her opinion of the university's plan. State her opinion and explain the reasons she gives for holding that opinion.

Key Point 1: The university announced that they would eliminate the sculpture program.
Key point 2: The administrator explained that the main reason is because of a lack of student interest.
Key point 3: The second reason is that the sculpture professor is retiring and the department has limited budget to hire a new professor.

Figure 1: Question, reading material, and KPs

analyzed by three annotators who had backgrounds in linguistics and language education. In order to identify the Key Points that the students included or omitted in their responses, a binary scale, with 1 representing presence and 0 representing absence of each Key Point for the entire response[3], was used. The annotators paid attention to the ideas rather than the particular wording in Key Points and assigned a score of 1 (presence of Key Point) when students' conveyed the Key Points in semantically legitimate variations, not necessarily using identical expressions. The three annotators went through multiple rounds of training and calibration in order to establish inter-rater reliability. In the initial rounds of training, when there were disagreements in the annotation, the three annotators resolved the problematic cases through discussions until exact agreements were reached. After that, each annotator independently annotated roughly even numbers of responses. The inter-rater agreement was relatively strong, and Cohen's kappa based on the 22% of double-scored responses was 0.72. However, there were large variations across different Key Points and kappa ranged from 0.61 to 0.85. The number of responses and distribution of Key Point score

[1] We selected approximately 100 speakers per A2, B1, B2, and C1 levels based on Common European Framework of Reference for Languages (CEFR).

[2] A form is a set of three questions, and we used four forms. There were no question overlaps among different forms. Thus, we used a total of 12 questions.

[3] The annotators were not indicating the specific location of the Key Points in the responses.

are presented in Table 1.

5 Method

We used two different approaches to detect the absence of a Key Point in a spoken response. First, we trained classifiers using a set of features that calculate similarity scores between a student's response and a Key Point. Next, we trained automated models used for short response scoring.

5.1 Models based on word-embedding features

First, both Key Points and transcriptions of students' responses were normalized by removing stop words and disfluencies. After the normalization process, the length of the Key Points and responses were reduced into 60% and 40% of the original texts on average. After removing stop words, the average number of words in responses was 50.9 (based on the manual transcriptions) and 51.3 (based on the ASR hypotheses), respectively.

The number of words in the Key Point after the normalization was 3.85 on average. In particular, Key Point 1 and Key Point 2 were shorter than the other Key Points; the average number of words for Key Point 1 was 2.08, while it was 5.58 for Key Point 6. For each Key Point, we first created a word list containing all words (ALL) after the normalization. While some words (e.g., the topic or the concept name) appeared in multiple Key Points in the same question, some words were unique to a particular Key Point. Under the assumption that these unique words may be more important for detecting the absence of the specific Key Point, we created two additional word lists for each Key Point: a unique word list (Unique) [4] and a shared word list (Shared) that contained words not in the unique list.

The response was segmented into a sequence of word n-grams [5] with 5 words overlap between two consecutive n-grams. For each n-gram, the similarity with a particular Key Point was calculated using the following three word-embedding based metrics:

- Word Mover's Distance (WM-distance): This calculates a sum of the minimum distances between words in the two compared strings (a key-point and an n-gram of the response) where the distance between two words was the Euclidean distance between the two corresponding word vectors in the embedding space (Kusner et al., 2015).

- Weighted word embeddings: This calculates a cosine similarity score between a Key Point vector and a response n-gram vector. The Key Point vector was an average of the corresponding embedding vector with a tf-idf weight for each word in the Key Point. The n-gram vector was generated using the same process.

- Query-document Similarity (QD): Responses are generally much longer than Key Points and WM-distance may assign unfairly low similarity scores to responses with extra information. To address this issue, we use metrics designed for information retrieval (Kim et al., 2016). For each word in the Key Point, the algorithm finds the word with the maximum similarity from a response n-gram, where the similarity score is the cosine similarity between two corresponding word embeddings. Finally, this metric uses a sum of all maximum similarity scores normalized by the Key Point length.

Next, we generated response-level features by selecting the minimum and the maximum values among all n-grams in a response. From 9 n-gram-based features (3 Key Point word lists * 3 metrics), 18 values were selected for each response. We used the publicly available word embedding vectors trained on the Google News corpus by Mikolov et al. (2013) for all word-embedding based features, and WM-distance implementation in the gensim package (Rehurek and Sojka, 2010) for WM-distance calculation.

Finally, we trained a binary classifier using response-level features with human Key Point scores as class labels. A total of 6 binary classifiers (one per Key Point) were trained using the random forest classifier algorithm[6] in scikit-learn (Pedregosa et al., 2011).

[4] words unique to the particular Key Point when comparing the 6 Key Points for a given question

[5] n = the number of words in a Key Point after the normalization

[6] During a pilot experiment, multiple machine learning algorithms such as decision tree, Support Vector Machine, AdaBoost were tested, and the random forest classifier was selected based on the consistently high performance.

CEFR	# speakers	# responses	# ratings	Percentage of Key Point absence (score = 0)						
				All	KP1	KP2	KP3	KP4	KP5	KP6
A2	95	285	1710	63	48	60	60	65	71	73
B1	100	300	1800	43	32	39	38	49	48	51
B2	100	300	1800	28	21	28	17	34	32	37
C1	100	300	1800	19	22	20	11	19	15	26
Total	395	1185	7110	31	37	31	42	41	47	38

Table 1: Data size and Key Point (KP) distribution by proficiency levels

5.2 Models based on the content scoring engine

We used an automated scoring system that achieved state-of-the-art performance in scoring content of short text responses (Heilman and Madnani, 2013) (hereafter, c-rater). This is also the same system used in Loukina and Cahill (2016).

The system first generated sparse lexicalized features including word and character n-gram features and syntactic dependency features. Unlike the word-embedding features, we used an entire spoken response as an input for the feature generator. Finally, we trained a Support Vector Regressor with a radial basis function kernel for each Key Point, resulting in a total of 6 regression models. Each model was a generic model that was trained on all 12 questions[7].

6 Experiment

The speakers were partitioned into two sets: train (49%), and test sets (51%). All responses from the same speaker belonged to one set, and thus the train and test sets did not share any speakers. The percentage of each form and speakers' proficiency levels were similar in each set. In order to investigate the impact of a question-specific training dataset, we conducted 4-fold cross-validation. As described in Section 4, the data was comprised of four forms (with three questions on each form). For each fold, three forms were used as the "seen form", and the remaining form was used as the "unseen form". The model was trained only on the seen form responses in the training partition. During evaluation, the model was evaluated on the seen form responses and the unseen form responses, separately. In the results section, we report the average of the four-folds.

We used two different transcription methods: manual transcriptions by professional transcribers and automated transcriptions by an ASR system trained on non-native speakers' speech. We used a gender-independent acoustic model (AM) trained on 800 hours of spoken responses covering over 100 native languages across 8,900 speakers using the Kaldi toolkit (Povey et al., 2011). A DNN-HMM model was adapted to test takers with fMLLR and i-vectors. The language model (LM) was a trigram model trained using the same dataset used for AM training. This ASR system achieved a Word Error Rate of 18.5% on 600 held-out responses. Detailed information about the ASR system is provided in Qian et al. (2016). In order to compare the performance of the content features with c-rater, we trained three models: EMB (model based on word-embedding features), c-rater (model based on the c-rater engine), and CMB (combination of two models). For CMB, we averaged the probabilities generated by EMB and c-rater with 0.5 as a decision boundary. Finally, for each transcription mode, we trained 18 binary classifiers.

7 Results

7.1 Performance on Seen form

Table 2 provides performance of the models on the seen forms where all questions in the test set appeared in the train set. The models were evaluated in terms of accuracy, F-score, and Cohen's kappa for detecting absence of the Key Points. We reported the average performance for 6 Key Points. In this study, the accuracy of the majority class baseline (classifying all responses as the Key Point presented) was 64% since the proportion of the responses without Key Point was 36% on av-

[7]We also trained a separate regression model for each question of each Key Point, resulting in $6 \times 12 = 72$ models (question-specific models). Because the overall performance of the question-specific models were not superior to the generic models, we reported only the generic model-based results. In a future study using a much larger numbers of questions, we will conduct more rigorous comparisons between the generic models and the question-specific models and select the final models.

erage.

	Model	accuracy	F-score	κ
Manual	EMB	0.77	0.65	0.47
	c-rater	0.76	0.65	0.43
	CMB	0.79	0.69	0.51
ASR	EMB	0.77	0.64	0.46
	c-rater	0.75	0.63	0.42
	CMB	0.80	0.68	0.51

Table 2: Average performance of six Key Points on seen form

For the experiment using the manual transcriptions, both the EMB and c-rater models achieved substantial improvement over the majority baseline. The performance of the EMB model was comparable to the c-rater model, and the combination of the two models resulted in further improvement. The accuracy and F-score of the CMB model were 0.79 and 0.69, respectively.

The results based on the ASR word hypotheses were comparable to those based on the manual transcriptions; the accuracy of the CMB model was 0.80 (0.79 for the manual transcription-based results) and F-score was 0.68 (0.69 for the manual transcription-based results). The EMB model achieved a slightly better performance than the c-rater model.

7.2 Performance on Unseen form

Table 3 provides the performance of the models on the unseen form where all questions in the test set did not appear in the train set.

	Model	accuracy	F-score	κ
Manual	EMB	0.71	0.56	0.35
	c-rater	0.61	0.56	0.23
	CMB	0.71	0.61	0.37
ASR	EMB	0.71	0.54	0.33
	c-rater	0.61	0.55	0.23
	CMB	0.71	0.60	0.36

Table 3: Average performance of six Key Points on unseen form

The performance of models for the unseen forms was substantially lower than that for the seen forms. For the manual transcription-based results, the accuracy and the F-score of the CMB model were 0.71 and 0.61, respectively, approximately $0.07 \sim 0.08$ lower than the results on the seen form. Notably, the performance drop of the c-rater model was much larger than that of the EMB model, and the accuracy of the c-rater model was lower than the majority baseline. The performance of the EMB model was relatively better than the c-rater model, but it was still substantially lower than the performance on the seen forms. Finally, the combination of the two models resulted in a slight improvement in the F-score, but not in accuracy. The results based on the ASR word hypotheses were comparable to those based on the manual transcriptions.

The low performance of the c-rater models for the unseen form was somewhat expected. The models learned characteristic n-grams of specific Key Points from the training data. The Key Points in this study were largely different by questions, and these characteristic n-grams for one question may not be useful for other questions. The EMB models, however, did not directly use the n-gram patterns in the training data. Instead, they calculated the similarity scores between Key Points and responses using the word-embeddings-based metrics and the train set was only used to determine the relationships between these features. This difference resulted in the performance difference between the two models on the unseen forms.

In summary, the models were relatively robust to the ASR errors, and performance based on the ASR hypotheses was comparable to the manual-transcription-based performance when using a high performing ASR system. Feedback that relies on manual transcription may be a critical challenge, or not even a feasible option, for automated feedback systems used for large-scale learning programs. Therefore, the robustness to the ASR errors is an important advantage of our method. In contrast, unseen questions had a strong negative impact on the models, and the performance of the best performing model (CMB model) decreased substantially when using the unseen questions. This may raise an important challenge to adding new questions in an operational learning program; in order to add new questions without lowering system performance, a sizable amount of responses may need to be annotated for each question.

8 Discussion

The proposed models achieved promising performance in detecting missing Key Points from responses to the questions included in the training

set. However, their performance was meaningfully lower than the performance of human raters; the κ between the algorithm and the human rater was 0.52, while the κ between two human raters was 0.72.

In this study, the raters did not penalize students who did not use the exact wordings in the Key Points; if a response contained a semantically comparable sentence to a Key Point, then the Key Point was considered to be present in the response. This approach may increase the difficulty of automated detection. In order to investigate how frequently students used expressions different from Key Points, we calculated a ratio of Key Point words that appeared in a response to all words in a particular Key Point (hereafter, Key Point ratio). For instance, if a Key Point is comprised of 5 words and only 2 words appear in a response, then the Key Point ratio is 0.4, and it roughly suggests that 3 words in the Key Point are realized in different expressions. If the Key Point ratios are generally low for the Key Point-present responses, then it suggests that students frequently use expressions other than those in the Key Point. We calculated Key Point ratio for each response using the manual transcription after the normalization process. Table 4 presents the average of the Key Point ratio.

Key Point type	Proportion of Key Point words in responses
Key Point 1	0.69
Key Point 2	0.54
Key Point 3	0.60
Key Point 4	0.49
Key Point 5	0.41
Key Point 6	0.51

Table 4: Average of the Key Point ratios for the Key Point-present responses

The average of the ratios for Key Point-present responses was 0.54. It ranged from 0.41 to 0.69. This suggests that around half of the words in the Key Points were realized in the different wordings in these responses.

In order to understand the reason for the relatively low use of the exact wordings, we selected a subset of Key Point-present responses with low Key Point ratio and analyzed how the Key Points were expressed. Figure 2 shows one Key Point and two sample responses. For the responses, we provide only the segments that are relevant to the specific Key Point.

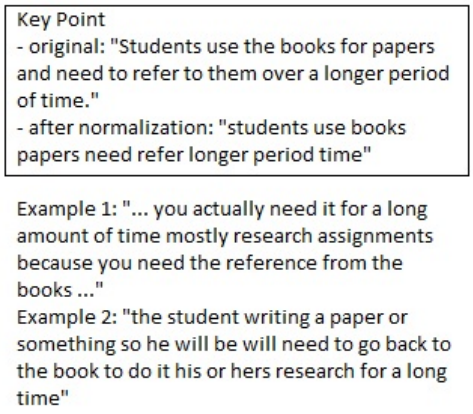

Figure 2: Sample Key Point and responses

Example 1 contained 3 Key Point words (“need”, “time”, “books”), and “reference”, “longer” were realized in their morphological variations (“refer”, “long”). “papers”, “students”, “use”, and “period” were replaced with contextually legitimate expressions (e.g., “research assignments” for “papers”) or omitted. In example 2, the Key Point was realized in very different wordings. For instance, the core concepts, “use books for papers” and “need the reference”, were expressed as “writing a paper” and “need to go back to the book”, respectively. In particular, spontaneous non-native speech includes frequent grammatical or vocabulary usage errors, and this results in even wider variations in the realization of Key Points in their responses. The Key Point in this study was generally short and 38 Key Point (53%) contained less than 3 content word types. The short Key Point length may increase the difficulty of automated detection further, since the impact of replacing one Key Point word with different wordings is large.

This analysis further motivates use of the word-embedding based features. In contrast to traditional lexical similarity features, which are limited to a reliance on exact word matching, the word embedding features have the advantage of capturing topically relevant words that are not identical. The students’ responses frequently included semantically legitimate expressions that were not same words with Key Points, and this has resulted in improvements over systems using only traditional lexical similarity features.

9 Targeted feedback based on the missing Key Points

In this section, we will discuss our future plan about how to generate targeted feedback based on the automated Key Point scores. There are several reasons that language learners may miss the key information from the source materials. When a student misses a key point, it may be an issue of reading and/or listening comprehension difficulty, or it could be an indication of lower speaking proficiency. When a language learner processes, selects, and synthesizes the key information from the source materials, the individual will need to recreate the key points using their linguistic knowledge to generate the speech content. If a speaker does not possess the required linguistic knowledge to produce a full response, a speaker may reproduce inaccurate or inadequate key points. In addition, previous research has suggested that reading and/or listening to source materials and reproducing them in an assessment context is a cognitively taxing task, especially for lower-proficiency students (Brown et al., 2005). This implies that some learners may not have the necessary linguistic working memory capacity to retain all the detailed information they read or heard that would enable them to reproduce the key information satisfactorily. Thus, providing feedback about missing key points can be helpful and revealing because it indicates the gaps in spoken summaries or responses.

To address this need, our preliminary feedback algorithm provides targeted feedback about the missing Key Points. Specifically, the feedback is comprised of four parts: (a) source materials, (b) a language learner's response, (c) actionable instructions, and (d) sample responses.

The first part (source materials) provides the listening passage and/or the reading passage of the question. The expert assessment developers annotate sentences relevant to each Key Point from the source materials, and the algorithm stores this information in advance. During feedback generation, the algorithm first automatically identifies Key Points missing from a response and displays the source materials relevant to the missing Key Points.

The second part (a language learner's response) provides a function for the language learner to replay their own responses. Listening to his or her own responses while paying attention to the missing Key Points provided in the first part may help the test taker to understand the gaps in the response better. Optionally, the algorithm provides the ASR-based transcriptions of the responses.

For the third part, the algorithm first classifies a response into a sub-group based on the automated Key Point scores and provides feedback prepared for the particular group. The Key Points in this study were designed in a highly structured way, and each Key Point was tied to specific skill areas (e.g., listening and reading) or tasks (e.g., define a concept, express his/her position about the proposal). Thus, the combination of the Key Point scores for each response may reveal specific weaknesses of the language learner. For instance, a high proportion of missing Key Points related to the listening passage may indicate that the language learner has a weakness with regard to listening or integrating information from listening into speaking. The algorithm stores actionable instructions prepared based on these language learners' characteristics for each group. In addition, when applying the feedback algorithm with an automated proficiency scoring system, it further classifies a response into a sub-group based on an automatically detected proficiency level and provides feedback prepared for the particular proficiency level. The algorithm may provide different instructions for different proficiency levels, and this enables us to provide simpler and easier instructions for beginners while more complicated and sophisticated instructions are provided for the intermediate or advanced learners.

Finally, the fourth part (samples) provides representative samples from highly proficient language learners. The algorithm also provides explanations about how Key Points are expressed in their responses and what their strengths are. Optionally, the algorithm may provide some samples from low proficiency language learners with explanations about their weaknesses.

10 Conclusions

In this study, we aim to develop an automated content feedback algorithm for spontaneous speech from non-native English speakers. The algorithm was designed for integrated tasks where language learners listen to and/or read the passages and integrate the key content in their spoken responses. Focusing on the content completeness, the algorithm generated automated Key Point scores and

provided targeted feedback about the missing Key Points. It achieved promising performance for questions included in the training data and also was robust to ASR errors. In future work, we will conduct a user study and investigate whether our content feedback system could lead to improvement in learners integrated speaking task performance.

Acknowledgments

We thank Aoife Cahill, Keelan Evanini, Lin Gu, Beata Beigman Klebanov, and two anonymous reviewers for their comments and suggestions. We also thank Patrick Houghton for his careful edits.

References

Annie Brown, Noriko Iwashita, and Tim McNamara. 2005. An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, 2005(1):i–157.

Lei Chen, Jidong Tao, Shabnam Ghaffarzadegan, and Yao Qian. 2018. End-to-end neural network based automated speech scoring. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6234–6238. IEEE.

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.

Elena Cotos. 2011. Potential of automated writing evaluation feedback. *CALICO Journal*, 28(2):420–459.

Maxine Eskenazi, Angela Kennedy, Carlton Ketchum, Robert Olszewski, and Garrett Pelton. 2007. The $NativeAccent^{TM}$ pronunciation tutor: measuring success in the real world. In *Workshop on Speech and Language Technology in Education*.

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.

Kellie Frost, Catherine Elder, and Gillian Wigglesworth. 2012. Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers oral performances. *Language Testing*, 29(3):345–369.

Michael Heilman and Nitin Madnani. 2013. ETS: Domain adaptation and stacking for short answer scoring. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 275–279.

Sun Kim, W John Wilbur, and Zhiyong Lu. 2016. Bridging the gap: a semantic similarity measure between queries and documents. *arXiv preprint arXiv:1608.01972*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

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.

Roy Lyster, Kazuya Saito, and Masatoshi Sato. 2013. Oral corrective feedback in second language classrooms. *Language teaching*, 46(1):1–40.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, EPFL-CONF-192584. IEEE Signal Processing Society.

Yao Qian, Rutuja Ubale, Matthew Mulholland, Keelan Evanini, and Xinhao Wang. 2018. A prompt-aware neural network approach to content-based scoring of non-native spontaneous speech. In *Proceedings of the 2018 Workshop on Spoken Language Technology*.

Yao Qian, Xinhao Wang, Keelan Evanini, and David Suendermann-Oeft. 2016. Self-adaptive dnn for improving spoken language proficiency assessment. In *INTERSPEECH*, pages 3122–3126.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Marek Rei and Ronan Cummins. 2016. Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 283–288.

Xiaoming Xi. 2010. Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3):291–300.

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, Anastassia Loukina, Chong Min Lee, Matthew Mulholland, Xinhao Wang, and Ikkyu Choi. 2018. Word-embedding based content features for automated oral proficiency scoring. In *Proceedings of the Third Workshop on Semantic Deep Learning*, pages 12–22.