Automated speech scoring for non-native middle school students with multiple task types

Keelan Evanini, Xinhao Wang

Educational Testing Service
Princeton, NJ, USA

kevanini@ets.org, xwang002@ets.org

Abstract

This study presents the results of applying automated speech scoring technology to English spoken responses provided by non-native children in the context of an English proficiency assessment for middle school students. The assessment contains three diverse task types designed to measure a student's English communication skills, and an automated scoring system was used to extract features and build scoring models for each task. The results show that the automated scores have a correlation of $r = 0.70$ with human scores for the Read Aloud task, which matches the human-human agreement level. For the two tasks involving spontaneous speech, the automated scores obtain correlations of $r = 0.62$ and $r = 0.63$ with human scores, which represents a drop of 0.08 - 0.09 from the human-human agreement level. When all 5 scores from the assessment for a given student are aggregated, the automated speaker-level scores show a correlation of $r = 0.78$ with human scores, compared to a human-human correlation of $r = 0.90$. The challenges of using automated spoken language assessment for children are discussed, and directions for future improvements are proposed.

Index Terms: automated speech scoring, children's speech, non-native speech

1. Introduction

The continued spread of English as a global language has resulted in an increase in the number of children in many countries who are exposed to English as a Foreign Language while they are middle school students. Depending on the situation, there are many variables that can have an effect on the quality of instruction for these children, and, thus, a corresponding effect on their learning outcomes. Some of these factors include: the English proficiency of the instructor (e.g., native vs. non-native speaker), the venue of the instruction (e.g., public school class vs. private English academy), pedagogical style (e.g., grammar-translation vs. communicative approach), etc. Since there is such widespread variation in the English proficiency levels of these children who are exposed to English instruction, it would be desirable to have an objective and reliable means of assessing their English proficiency for a variety of purposes, including placement, monitoring, and advancement. A recently released global English assessment for middle school students, the TOEFL Junior Test from ETS, was designed to serve this purpose. This study reports on an investigation of the applicability of automated scoring technology for the spoken responses from the version of this test that includes a Speaking section, the TOEFL Junior Comprehensive.

While automated spoken language assessments for adult speech have been studied widely, e.g. [1, 2, 3], little work has been done in the domain of automated speaking proficiency assessments for children. Most of the applications that have been developed for children focus specifically on the tasks of oral reading assessment and oral reading tutoring [4, 5, 6]. None of these systems, however, provide an assessment of the speaker's English proficiency in terms of communicative ability. In contrast, the TOEFL Junior Comprehensive Test includes task types that elicit spontaneous speech from the children, since the assessment was designed to evaluate a student's ability to produce meaningful English utterances in response to specific tasks.

2. Previous Research

One of the major challenges for automated assessment of children's speech is the difficulty of building accurate Automatic Speech Recognition (ASR) systems for children's speech. Due to the differences in vocal tract length between children and adults, acoustic models trained on adult speech will produce worse results on children's speech. In addition, children may have different speech patterns in linguistic areas such as pronunciation, prosody, lexical choice, and syntax. To overcome these problems, several corpora containing only children's speech have been collected [7, 8, 9, 10, 6], and have been used to train or adapt ASR systems so that they will perform better on children's speech.

By far the most common application of automated language assessment for children involves oral reading assessment and tutoring. Several systems have been developed which present written material to the child and use ASR technology to process the child's oral reading proficiency, and, in some cases, also provide feedback. For example, the Reading Tutor from CMU's Project LISTEN presents reading passages one sentence at a time and has implemented sophisticated dialogue strategies based on metrics such as response latency, reading accuracy, etc., to determine the appropriate type of feedback or assistance to provide to the child [11]. The system then evaluates the child's proficiency based on metrics associated with oral reading fluency and prosody. Another system, IBM's Reading Companion, also produces feedback to children based on the accuracy of their reading, and provides several types of scaffolding assistance, depending on whether the child is struggling or not. The system furthermore generates reports for teachers that contain statistics about the overall accuracy of the child's oral reading as well as frequencies of the specific types of reading errors that the child made [6]. Another system, developed for the TBALL project, contains a wider range of task types that are designed to assess whether a child possesses certain basic literacy skills that are building blocks to proficient oral reading, such as reading single words out loud, combining syllables to form words, naming letters, producing the sounds represented by letters, and reading comprehension [5]. In addition to these systems, several additional ones have been developed for the specific purpose of assessing chilren's oral reading fluency, e.g. [12], [13], and others.

Despite the fact that a wide variety of systems have been developed for the automated assessment of oral reading proficiency among children, no systems, to our knowledge, have been developed to assess spontaneous speech from non-native children. While automated spoken language assessment has been studied in recent years in the context of non-native adults, e.g. [2], this research has yet to consider the specific challenges that are related to the processing of children's speech. However, with the growth of English instruction programs for younger students across the globe and the larger numbers of non-native children with higher English proficiency, there is an increasing need for an automated assessment of English speaking proficiency for a younger population.

3. Design of the Assessment

As mentioned above, this study uses data from the TOEFL Junior Comprehensive assessment, which is a computer-based test containing four sections: Reading Comprehension, Listening Comprehension, Speaking, and Writing. It is intended for middle school students around the ages of 11 - 15, and is designed to assess a student's English communication skills through a variety of tasks. This study focuses on the Speaking section, which contains the following four task types eliciting spoken responses[1]:

- Read Aloud: the test taker reads a paragraph (containing approximately 90 - 100 words) presented on the screen out loud
- Picture Narration: the test taker is shown six images that depict a sequence of events and is asked to narrate the story in the pictures
- Listen Speak (Non-Academic): the test taker listens to an audio stimulus (approximately 2 minutes in duration) containing information about a school-related topic (for example, a homework assignment) and provides a spoken response containing information about specific facts in the stimulus
- Listen Speak (Academic): similar to the Listen-Speak (Non-Academic) item, except that the audio stimulus contains information about an academic topic relevant to middle school students (for example, the life cycle of frogs)

The responses to all task types are 60 seconds in duration, and they are scored on a scale of 1 - 4 by expert human raters. Responses containing anomolous test taker behavior (such as non-English responses or non-responses) and responses with severe technical difficulties (such as static or background noise) receive separate ratings and are excluded from this study.[2] For the purposes of this study, the Listen Speak (Non-Academic) and Listen Speak (Academic) were combined, since the responses to these two task types exhibited similar characteristics.

4. Data

The data used in this study were drawn from a pilot version of the TOEFL Junior Comprehensive assessment administered in late 2011 and includes a total of 16,925 spoken responses from 3,385 participants. Each pilot test form contained five questions (two Listen-Speak (Academic) tasks and one of each of the other task types), and six different test forms were used in the pilot. The average age of the participants was 13.1 years (std. dev. = 2.3), and there were 1847 females (54.6%) and 1538 males (45.4%). The following native language backgrounds are represented among the participants: Arabic, Chinese, French, German, Indonesian, Japanese, Javanese, Korean, Madurese, Polish, Portuguese, Spanish, Thai, and Vietnamese.

The speakers were divided into five different partitions for training and evaluating the speech recognizer and the scoring model. Table 1 presents the amount of data included in each partition as well as the score distributions of each partition.

All of the responses were provided with transcriptions using standard English orthography. Table 2 presents the means and standard deviations for the number of words contained in the responses by task and human score.

Task	1	2	3	4
RA	73.6 (28.0)	93.5 (11.4)	96.4 (8.0)	96.4 (6.0)
PN	45.9 (22.3)	65.3 (20.0)	78.5 (22.3)	96.4 (24.4)
LS	44.3 (23.1)	67.4 (23.3)	87.6 (24.6)	107.2 (25.4)

Table 2: *Mean (std. dev.) number of words contained in the responses by task and human score*

As expected, more proficient speakers produced longer responses for both the Picture Narration and Listen Speak tasks. For the Read Aloud task, all score points except 1 had mean values between 90 and 100 words (the target length of the stimulus passages) with small standard deviations; only Read Aloud responses receiving a score of 1 were typically much shorter (since the students with very low English proficiency were not able to complete the Read Aloud passage in the allotted time).

5. Speech Recognizer

The children who participated in this study are all non-native speakers of English, so it is important to also use ASR training data from non-native speakers. Since the existing corpora of children's speech listed in Section 2 contain either exclusively native speech or non-native speech from speakers with different native languages than the students in this study, a new, in-domain corpus was collected based on the pilot data. As shown in Table 1, this ASR training corpus contained 137.2 hours of speech from 1625 children, with an average of 5 minutes of speech per child.

In order to obtain more reliable acoustic models for non-native speech, we first trained an HMM-based triphone ASR system using approximately 800 hours of non-native adult speech. Then, the Acoustic Model was subsequently adapted to children's speech by applying MAP adaptation to the responses contained in the ASR training partition. In addition, Language Model adaptation was applied using an interpolation weight of

[1] Sample test questions for each of the four spoken task types in the TOEFL Junior assessment are available at `http://toefljr.caltesting.org/sampletest/index.html`.

[2] Approximately 10% of the responses were excluded for this reason; since the data in this study was drawn from a pilot study, this percentage of non-scorable responses is substantially higher than in operational administrations.

Parition	Speakers	Responses	Duration (hrs)	1	2	3	4
ASR training	1625	7300	137.2	1434 (19.6)	3065 (42.0)	2088 (28.6)	713 (9.8)
ASR development	30	149	2.5	14 (9.4)	54 (36.2)	56 (37.6)	25 (16.8)
ASR evaluation	30	150	2.5	19 (12.7)	48 (32.0)	66 (44.0)	17 (11.3)
Model training	967	4338	81.7	798 (18.4)	1802 (41.5)	1277 (29.4)	461 (10.6)
Model evaluation	733	3297	62.0	664 (20.1)	1368 (41.5)	918 (27.8)	347 (10.5)

Table 1: *Data partitions used for the ASR system and the scoring model*

0.9 for the in-domain data.[3] Table 3 presents the recognition results on the ASR evaluation set.

Task	WER
Read Aloud	9.7
Picture Narration	26.5
Listen Speak	29.4

Table 3: *Performance of the ASR system on the three different task types*

6. Features

In order to train scoring models to predict the human scores, features were extracted from each response using the SpeechRater system for automated assessment of non-native speech [2]. A total of over 90 features were extracted from the speech signal and ASR hypotheses representing the following areas of speaking proficiency: fluency, pronunciation, prosody, language use, and content (for the Read Aloud task). Based on the Pearson correlations between the features and the human scores in the model training partition, inter-correlations with each other, and coverage of a range of proficiency areas, a subset of 10 features was selected to be used in the scoring models. These features are listed in Table 4, along with their correlations with human scores on the model training partition.[4] The features presented in Table 4 were used to train three separate linear regression models, one for each of the following three tasks: Read Aloud, Picture Narration, and Listen Speak (as mentioned in Section 3, the data for the Listen Speak (Non-Academic) and Listen Speak (Academic) task types were merged, since the responses for the two tasks types exhibit similar characteristics).

Feature	RA (N=882)	PN (N=874)	LS (N=2488)
Fluency			
rate of speech	0.587	0.582	0.577
number of words per chunk	-0.531	-0.473	-0.463
average number of pauses (silpwd)	0.532	0.460	0.450
average number of long pauses	0.463	0.509	0.481
Pronunciation			
normalized AM score	-0.587	-0.505	-0.573
average word confidence	0.522	0.310	0.365
average difference in phone duration from native speaker norms	0.420	0.449	0.301
Prosody			
mean duration between stressed syllables	0.549	-0.480	-0.507
Lexical choice / Grammar			
normalized LM score	0.593	0.411	0.437
Content			
reading accuracy	0.661	N/A	N/A

Table 4: *Correlations of the individual features with human scores on the model training partition*

7. Results

The linear regression scoring models were used to predict scores for each response. The Pearson correlations between these predicted scores and the human scores for the three different tasks are presented in Table 5, along with human-human correlations. The table also presents the speaker-level correlations, which were calculated by first summing up the scores for all 5 responses provided by each speaker. Speakers who did not have a complete set of 5 scorable responses were excluded from this analysis; thus, the score range for this analysis is 5 - 20.

As Table 5 shows, the automated system performs at a level comparable to humans for the Read Aloud task, with a degradation in correlation of only 0.01. The system's performance on the other two task types is somewhat lower than the human-human level, with degradations in performance of 0.08 and 0.09 for the Picture Narration and Listen Speak tasks, respectively. Finally, the sum of the automated speaker-level scores across all 5 responses correlates with the sum of the human scores at $r = 0.779$, which is 0.12 below the human-human agreement for the speaker-level score.

Figure 1 presents the distributions of the automated scores for responses receiving the four different human scores across all three task types. As expected based on the correlation results presented in Table 5, Figure 1 shows that the median value of the system score increases for each human score level. While there is some degree of overlap among the distributions, Figure 1 shows that the median value for each distribution is always greater than the 75th percentile of the distribution below it and smaller than the 25th percentile of the distribution above it. As can also be seen from Figure 1, the range of the system scores is compressed: the median system score for responses receiving a human score of 1 is 1.72 and the median system score for responses receiving a human score of 4 is 2.92. This indicates that the automated scoring system is not able to adequately model the extreme points of the score range (i.e., 1 and 4).

[3] The interpolation weight was tuned on the ASR development set.

[4] As mentioned in Section 3 this study excluded responses which received a human rating indicating a technical difficulty or other anomaly. Due to these exclusions, the counts in Tables 4 and 5 are lower than the speaker counts shown in Table 1.

Task	human-system	human-human
Read Aloud (N=684)	0.704	0.714
Picture Narration (N=668)	0.620	0.700
Listen Speak (N=1886)	0.629	0.720
Speaker-level (N=554)	0.779	0.896

Table 5: *Performance of the scoring model across the three task types*

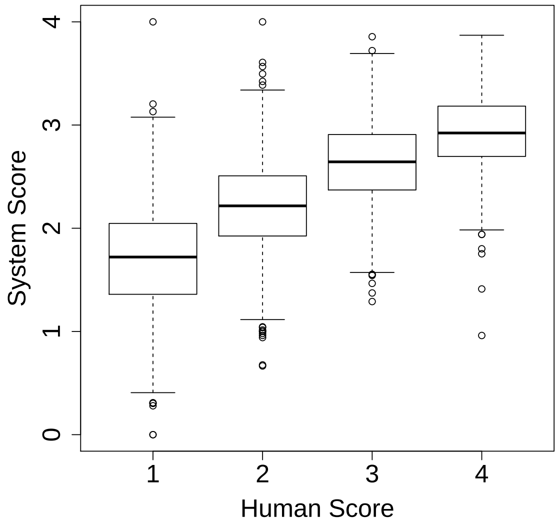

Figure 1: *Distribution of automated system scores for responses with the four human scores, all task types combined*

Table 6 presents the score distributions for the automated and human scores across the three task types. As the table shows, the mean value of the automated scores is nearly identical to the mean of the human scores across all task types. However, the standard deviation of the automated scores is always lower than for the human scores. This provides further evidence that the automated scores fall within a more restricted range than the human scores, and that the automated system is not able to predict the extreme ends of the score range well.

Task	system		human	
	Mean	Std. Dev.	Mean	Std. Dev.
Read Aloud	2.48	0.61	2.49	0.92
Picture Narration	2.28	0.54	2.25	0.87
Listen Speak	2.25	0.57	2.25	0.90
Speaker-level	11.93	2.14	11.97	3.39

Table 6: *Score distributions across the three task types*

8. Discussion

As shown in the previous section, the automated assessment system obtains a correlation with human scores for the Read Aloud task that is very close to the human-human agreement value. This indicates that the set of features used in the scoring model is sufficient to reliably model human behavior. For the other two task types, however, the performance of the automated system shows a degradation of 0.08 - 0.09 from the human-human agreement level. This result indicates that further features need to be investigated for the Picture Narration and Listen Speak tasks. This is not surprising, given the fact that the human raters are instructed to take the following three areas of language proficiency into account while providing the scores[5]: Delivery (fluency, pronunciation, prosody), Language Use (lexical choice, grammar), and Content (content appropriateness, coherence). The scoring models for these two tasks types, on the other hand, only contain a single feature representing the Language Use category (the Language Model score) and no features representing the Content category, as shown in Table 4. Thus, in order to obtain results closer to the human-human agreement level for the Picture Narration and Listen Speak tasks, the automated system would need to incorporate features addressing Content and additional features addressing Language Use. Studies have been conducted along these lines for the automated assessment of adult spontaneous speech using different task types [14, 15]; future research will investigate the utility of applying these features to non-native spontaneous speech produced by children. In addition, the performance of the speech recognizer is substantially worse on the Picture Narration and Listen Speak tasks, due to the fact that these tasks elicit spontaneous speech. Higher WER values on these types of responses result in less reliable Language Use and Content features. So, future research should also to address the task of improving the ASR performance on these types of responses in order to achieve more accurate scoring.

As was also demonstrated in Section 7, one problem with the automated assessment system developed for this study is that it is not able to model the extreme ends of the score range well. This is partly due to the fact that the distribution of scores is not balanced across the four score points, and that there is more data available for modeling score points 2 and 3. However, additional studies will investigate this further by exploring alternative modeling approaches that may result in a less compressed range of predicted scores.

9. Conclusions

This study described the development and evaluation of an automated scoring system for a spoken English proficiency assessment for non-native middle school students. The results show that the system matches human-human agreement for the Read Aloud task, but falls short of human-human agreement for the two tasks that elicit spontaneous speech. Nevertheless, this result represents the first attempt at developing an automated assessment system for spontaneous children's speech. The performance can likely be improved substantially by future research, since the results were obtained using a system that had been designed for adult speech in response to different task types, and no feature development was done for the specific task types contained in the assessment used in this study. The development of additional features related to Language Use and Content will likely lead to improved performance and an increase in the validity of the system.

[5] The scoring guides are available at `http://www.ets.org/s/toefl_junior/pdf/toefl_junior_comprehensive_speaking_scoring_guides.pdf`

10. References

[1] A. Chandel, A. Parate, M. Madathingal, H. Pant, N. Rajput, S. Ikbal, O. Deshmuck, and A. Verma, “Sensei: Spoken language assessment for call center agents,” in *Proceedings of ASRU*, 2007.

[2] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, “Automatic scoring of non-native spontaneous speech in tests of spoken English,” *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[3] J. Bernstein, A. V. Moere, and J. Cheng, “Validating automated speaking tests,” *Language Testing*, vol. 27, no. 3, pp. 355–377, 2010.

[4] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, “A prototype reading coach that listens,” in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 1994.

[5] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, “A system for technology based assessment of language and literacy in young children: The role of multiple information sources,” in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, 2007.

[6] A. Kantor, M. Cernak, J. Havelka, S. Huber, J. Kleindienst, and D. B. Gonzalez, “Reading Companion: The technical and social design of an automated reading tutor,” in *Proceedings of the Interspeech Workshop on Child, Computer, and Interaction*, 2012.

[7] LDC, “The CMU Kids Corpus,” http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC97S63, 1997.

[8] A. Hagen, B. Pellom, and R. Cole, “Children’s speech recognition with application to interactive books and tutors,” in *Proceedings of ASRU*, 2003.

[9] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, “Tball data collection: The making of a young children’s speech corpus,” in *Proceedings of Interspeech*, 2005.

[10] CSLU, “Kids speech corpus,” http://www.cslu.ogi.edu/corpora/kids, 2008.

[11] J. Mostow, “Why and how our automated reading tutor listens,” in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT)*, Stockholm, Sweden, 2002, pp. 43–52.

[12] D. Bolaños, R. A. Cole, W. Ward, E. Borts, and E. Svirsky, “FLORA: Fluent Oral Reading Assessment of children’s speech,” *ACM Transactions on Speech and Language Processing*, vol. 7, no. 4, pp. 1–19, 2011.

[13] A. Hagen, B. Pellom, and R. Cole, “Highly accurate children’s speech recognition for interactive reading tutors using subword units,” *Speech Communication*, vol. 49, pp. 861–873, 2007.

[14] S. Xie, K. Evanini, and K. Zechner, “Exploring content features for automated speech scoring,” in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 2012, pp. 103–111.

[15] M. Chen and K. Zechner, “Using an ontology for improved automated content scoring of spontaneous non-native speech,” in *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT*. Montréal, Canada: Association for Computational Linguistics, 2012.