Automatic Assessment of the Speech of Young English Learners

Jian Cheng[1], **Yuan Zhao D'Antilio**[1], **Xin Chen**[1], **Jared Bernstein**[2]

[1]Knowledge Technologies, Pearson, Menlo Park, California, USA
[2]Tasso Partners LLC, Palo Alto, California, USA

jian.cheng@pearson.com

Abstract

This paper introduces some of the research behind automatic scoring of the speaking part of the *Arizona English Language Learner Assessment*, a large-scale test now operational for students in Arizona. Approximately 70% of the students tested are in the range 4-11 years old. We cover the methods used to assess spoken responses automatically, considering both what the student says and the way in which the student speaks. We also provide evidence for the validity of machine scores. The assessments include 10 open-ended item types. For 9 of the 10 open item types, machine scoring performed at a similar level or better than human scoring at the item-type level. At the participant level, correlation coefficients between machine overall scores and average human overall scores were: Kindergarten: 0.88; Grades 1-2: 0.90; Grades 3-5: 0.94; Grades 6-8: 0.95; Grades 9-12: 0.93. The average correlation coefficient was 0.92. We include a note on implementing a detector to catch problematic test performances.

1 Introduction

Arizona English Language Learner Assessment (AZELLA) (Arizona Department of Education, 2014) is a test administered in the state of Arizona to all students from kindergarten up to grade 12 (K-12) who had been previously identified as English learners (ELs). AZELLA is used to place EL students into an appropriate level of instructional and to reassess EL students on an annual basis to monitor their progress. AZELLA was originally a fully human-delivered paper-pencil test covering four domains: listening, speaking, reading and writing. The Arizona Department of Education chose to automate the delivery and scoring of the speaking parts of the test, and further decided that test delivery via speakerphone would be the most efficient and universally accessible mode of administration. During the first field test (Nov. 7 - Dec. 2, 2011) over 31,000 tests were administered to 1st to 12th graders on speakerphones in Arizona schools. A second field test in April 2012 delivered over 13,000 AZELLA tests to kindergarten students. This paper reports research results based on analysis of data sets from the 44,000 students tested in these two administrations.

2 AZELLA speaking tests

AZELLA speaking tests are published in five *stages* (Table 1), one for each of five grade ranges or student levels. Each stage has four fixed test forms. Table 1 presents the total number of field tests delivered for each stage, or level.

Table 1: Stages, grades, and number of field tests

Stage	I	II	III	IV	V
Grade	K	1-2	3-5	6-8	9-12
N	13184	10646	9369	6439	5231

Fourteen different speaking exercises (item-types) were included in the various level-specific forms of the test. Some item-types were accompanied by images; some only had audio prompts. Note, however, that before the change to automatic administration and scoring, test forms had only included speaking item-types from a set of thirteen different types, of which ten were not designed to constrain the spoken responses. On the contrary, these ten item-types were designed to elicit relatively open-ended displays of speaking ability, and most test forms included one or two items of most types. A *Repeat Sentence* item type was added to the test designs (10 Repeat items per test form at every level), yielding test forms with around

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 12–21,
Baltimore, Maryland USA, June 26, 2014.

27 items total, including Repeats. Table 2 lists all the speaking item types that are presented in one AZELLA test form for Stage III (Grades 3-5). Some items such as *Questions on Image*, *Similarities & Differences*, *Ask Qs about a Statement*, and *Detailed Response to Topic* are presented as a sequence of two related questions and the two responses are human-rated together to produce one holistic score.

Table 2: Stage III (Grades 3-5) items.

Descriptions	items/test	Score-Points
Repeat Sentence	10	0-4
Read-by-Syllables	3	0-1
Read-Three-Words	3	0-1
Questions on Image	3	0-4
Similarities & Differences	2	0-4
Give Directions from Map	1	0-4
Ask Qs about a Statement	1	0-4
Give Instructions	1	0-4
Open Question on Topic	1	0-4
Detailed Response to Topic	1	0-4

Table 3: Item types used in AZELLA speaking field tests.

Description (restriction)	Score-Points
Naming (Stage I)	0-1
Short Response (Stage I)	0-2
Open Question (Stage I)	0-2
Read-by-Syllables	0-1
Read-Three-Words	0-1 or 0-3
Repeat Sentence	0-4
Questions on Image	0-4
Similarities & Differences (III)	0-4
Give Directions from Map	0-4
Ask Qs about a Thing (II)	0-2
Ask Qs about a Statement (III)	0-4
Give Instructions	0-4
Open Questions on Topic	0-4
Detailed Response to Topic	0-4

All the speaking item-types used at any level in the AZELLA field tests are listed in Table 3. Item-types used at only one stage (level) are noted. From Table 3 we can see that, except for *Naming*, *Repeat Sentence*, *Read-by-Syllables*, and *Read-Three-Words*, all the items are fairly unconstrained questions. Engineering considerations did not guide the design of these items to make them be more suitable for machine learning and automatic scoring, and they were, indeed, a challenge to score.

By tradition and by design, human scoring of AZELLA responses is limited to a single holistic score, guided by sets of Score-Point rubrics defining scores at 2, 3, 4, or 5 levels. The column *Score-Points* specifies the number of categories used in holistic scoring. One set of five abbreviated holistic rubrics for assigning points by human rating is presented below in Table 4. For the *Repeat Sentence* items only, separate human ratings were collected under a pronunciation rubric and a fluency rubric.

Table 4: Example AZELLA abbreviated holistic rubric (5 Score-Points).

Points	Descriptors
4	*Correct understandable English using two or more sentences.* 1. Complete declarative or interrogative sentences. 2. Grammar (or syntax) errors are not evident and do not impede communication. 3. Clear and correct pronunciation. 4. Correct syntax.
3	*Understandable English using two or more sentences.* 1. Complete declarative or interrogative sentences. 2. Minor grammatical (or syntax) errors. 3. Clear and correct pronunciation.
2	*An intelligible English response.* 1. Less than two complete declarative or interrogative sentences. 2. Errors in grammar (or syntax). 3. Attempt to respond with clear and correct pronunciation.
1	*Erroneous responses.* 1. Not complete declarative or interrogative sentences. 2. Significant errors in grammar (or syntax). 3. Not clear and correct pronunciation.
0	Non-English or silence.

3 Development and validation data

From the data in the first field test (Stages II, III, IV, V), for each AZELLA Stage, we randomly sampled 300 tests (75 tests/form x 4 forms) as a validation set and $1,200$ tests as a development set. For the data in the second field test (Stage I), we randomly sampled 167 tests from the four forms as the validation set and $1,200$ tests as the

development set. No validation data was used for model training.

3.1 Human transcriptions and scoring

In the development sets, we needed from 100 to 300 responses per item to be transcribed, depending on the complexity of the item type. In the validation sets, all responses were fully transcribed. Depending on the item type, we got single or double transcriptions, as necessary.

All responses from the tests were scored by trained professional raters according to predefined rubrics (Arizona Department of Education, 2012), such as those in Table 4. Departing from usual practice in production settings, we used the average score from different raters as the final score during machine learning. The responses in each validation set were double rated (producing two final scores) for use in validation. Note that five of the 1,367 tests in the validation sets had no human transcriptions and ratings, and so were excluded from the final validation results.

4 Machine scoring methods

Previous research on automatic assessment of spoken responses can be found in Bernstein et al. (2000; 2010), Cheng (2011) and Higgins et al. (2011). Past work on automatic assessment of children's oral reading fluency has been reported at the passage-level (Cheng and Shen, 2010; Downey et al., 2011) and at the word-level (Tepperman et al., 2007). A comprehensive review of spoken language technologies for education can be found in Eskanazi (2009). The following subsections summarize the methods we have used for scoring AZELLA tests. Those methods with citations have been previously discussed in research papers. Other methods described are novel modifications or extensions of known methods.

Both the linguistic content and the manner of speaking are scored. Our machine scoring methods include a combination of automatic speech recognition (ASR), speech processing, statistical modeling, linguistics, word vectors, and machine learning. The speech processing technology was built to handle the different rhythms and varied pronunciations used by a range of natives and learners. In addition to recognizing the words spoken, the system also aligns the speech signal, i.e., it locates the part of the signal containing relevant segments, syllables, and words, allowing the system to assign independent scores based on the content of what is spoken and the manner in which it is said. Thus, we derive scores based on the words used, as well as the pace, fluency, and pronunciation of those words in phrases and sentences. For each response, base measures are then derived from the linguistic units (segments, syllables, words), with reference to statistical models built from the spoken performances of natives and learners. Except for the Repeat items, the system produces only one holistic score per item from a combination of base measures.

4.1 Acoustic models

We tried various sets of recorded responses to train GMM-HMM acoustic models as implemented in HTK (Young et al., 2000). Performance improved by training acoustic models on larger sets of recordings, including material from students out of the age range being tested. For example, training acoustic models using only the Stage II transcriptions to recognize other Stage II responses was significantly improved by using more data from outside the Stage II data set, such as other AZELLA field data. We observed that the more child speech data, the better the automatic scoring. The final acoustic models used for recognition were trained on all transcribed AZELLA field data, except the data in the validation sets, plus data from an unrelated set of children's oral reading of passages (Cheng and Shen, 2010), and the data collected during the construction of the Versant Junior English tests for use by young children in Asia (Bernstein and Cheng, 2007). Thus, the acoustic models were built using any and all relevant data available: totaling about 380 hours of data (or around $176,000$ responses). The word error rate (WER) over all the validation sets using the final acoustic models is around 35%.

For machine scoring (after recognition and alignment), native acoustic models are used to compute native likelihoods of producing the observed base measures. Human listeners classified student recordings from Stage II (grades 1-2) as native or non-native. For example, in Stage II data, 287 subjects were identified as native and the recordings from these 287 subjects plus the native recordings from the Versant Junior English tests were used to build native acoustic models for grading. (approximately 66 hours of speech data, or $39,000$ responses).

4.2 Language models

Item-specific bigram language models were built using the human transcription of the development-set as described in Section 3.1.

4.3 Content modeling

"Content" refers to the linguistic material (words, phrases, and semantic elements) in the spoken response. Appropriate response content reflects the speaker's productive control of English vocabulary and also indicates how well the test-taker understood the prompt. Previous work on scoring linguistic content in the speech domain includes Bernstein et al. (2010) and Xie et al. (2012).

Except for the four relatively closed-response-form items (*Naming, Repeat*, *Read-by-Syllables* and *Read-Three-Words*), we produced a $word\_vector$ score for each response (Bernstein et al., 2010). The value of the $word\_vector$ score is calculated by scaling the weighted sum of the occurrence of a large set of expected words and word sequences available in an item-specific response scoring model. An automatic process assigned weights to the expected words and word sequences according to their semantic relation to known good responses using a method similar to latent semantic analysis (Landauer et al., 1998). The $word\_vector$ score is generally the most powerful feature used to predict the final human scores.

Note that a recent competition to develop accurate scoring algorithms for student-written short-answer responses (Kaggle, 2012) focused on a similar problem to the content scoring task for AZELLA open-ended responses. We assume that the methods used by the prize-winning teams, for example Tandalla (2012) and Zbontar (2012), should work well for the AZELLA open-ended material too, although we did not try these methods.

For the responses to *Naming*, *Read-by-Syllables*, and *Read-Three-Words* items, the machine scoring makes binary decisions based on the occurrence of a correct sequence of syllables or words ($keywords$). In Stage II forms, for first and second grade students, the responses to *Read-Three-Words* items were human-rated in four categories. For this stage, the machine counted the number of words read correctly.

For the responses to Repeat items, the recognized string is compared to the word string recited in the prompt, and the number of word errors ($word\_errors$) is calculated as the minimum number of substitutions, deletions, and/or insertions required to find a best string match in the response. This matching algorithm ignores hesitations and filled or unfilled pauses, as well as any leading or trailing material in the response (Bernstein et al., 2010). A verbatim repetition would have zero word errors. For Repeat responses, the percentage of words repeated correctly ($percent\_correct$) was used as an additional feature.

4.4 Duration modeling

Phone-level duration statistics contribute to machine scores of test-takers' pronunciation and fluency. Native-speakers segment duration statistics from Versant Junior English tests (Bernstein and Cheng, 2007) were used to compute the log-likelihood of phone durations produced by test-takers. No data from AZELLA tests contributed to the duration models. We calculated the phoneme duration log-likelihood: log_seg_prob and the inter-word silence duration log-likelihood: $iw\_log\_seg\_prob$ (Cheng, 2011).

Assume in a recognized response that the sequence of phonemes and their corresponding durations are p_i and D_i, $i = 1...N$, then the log likelihood segmental probability for phonemes (log_seg_prob) was computed as:

$$log\_seg\_prob = \frac{1}{N-2}\sum_{i=2}^{N-1}\log(\Pr(D_i)), \quad (1)$$

where $\Pr(D_i)$ was the probability that a native would produce phoneme p_i with the observed duration D_i in the context found. The first and last phonemes in the response were not used for the calculation of the log_seg_prob because durations of these phonemes as determined by the ASR were more likely to be incorrect. The log likelihood segmental probability for inter-word silence durations, $iw\_log\_seg\_prob$, was calculated the same way (Cheng, 2011).

4.5 Spectral modeling

To construct scoring models for pronunciation and fluency, we computed several spectral likelihood features with reference to native and learner segment-specific models applied to the recognition alignment, computing the phone-level posterior probabilities given the acoustic observation X

that is recognized as p_i:

$$P(p_i|X) = \frac{P(X|p_i)P(p_i)}{\sum_{k=1}^{m} P(X|p_k)P(p_k)} \quad (2)$$

where k runs over all the potential phonemes. In a real-world ASR system, it is extremely difficult to estimate $\sum_{k=1}^{m} P(X|p_k)P(p_k)$ precisely. So approximations are used, such as substituting a maximum for the summation, etc. Formula 2 is the general framework for pronunciation diagnosis (Witt and Young, 1997; Franco et al., 1999; Witt and Young, 2000) and pronunciation assessment (Witt and Young, 2000; Franco et al., 1997; Neumeyer et al., 1999; Bernstein et al., 2010). Various authors use different approximations to suit the particulars of their data and their applications.

In the AZELLA spectral scoring, we approximated Formula 2 with the following procedure. After the learner acoustic models produce a recognition result, we force-align the utterance on the recognized word string, but using the native monophone acoustic models, producing acoustic log-likelihood, duration and time boundaries for every phone. For each such phone, again using the native monophone time alignment, we perform an all-phone recognition using the native monophone acoustic models. The recognizer calculates a log-likelihood for every phone and picks the best match from all possible phones over that time frame. For each phone-of-interest in a response, we calculated the average spectral score difference as:

$$spectral\_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{lp_i^{fa} - lp_i^{ap}}{d_i} \quad (3)$$

where the variables are:

- lp_i^{fa} is the log-likelihood corresponding to the i-th phoneme by using the forced alignment method;
- lp_i^{ap} is the log-likelihood by using the all-phone recognition method;
- d_i is its duration;
- N is the number of phonemes of interest in a response.

In calculating $spectral\_1$, all possible phonemes are included. We define another variable, $spectral\_2$, that only accumulates the log-likelihood for a target set of phonemes that learners often have difficulty with. We call the percentage of phones from the all-phone recognition that match the phones from the forced alignment the $percent\ phone\ match$, or ppm. We take Formula 3 as the average log of the approximate posterior probabilities that phones were produced by a native.

4.6 Confidence modeling

After finishing speech recognition, we can assign speech confidence scores to words and phonemes (Cheng and Shen, 2011). Then for every response, we can compute the average confidence, the percentage of words or phonemes whose confidences are lower than a threshold value as features to predict test-takers' performance.

4.7 Final models

AZELLA holistic score rubrics (Arizona Department of Education, 2012), such as those shown in Table 4, consider both the answer content and the manner of speaking used in the response. The automatic scoring should consider both too. Features $word\_vector$, $keywords$, $word\_errors$, $percent\_correct$ can represent content scores based on what is spoken. Features log_seg_prob, $iw\_log\_seg\_prob$, $spectral\_1$, $spectral\_2$, ppm can represent both the rhythmic and segmental aspects of the performance as native likelihoods of producing the observed base measures. By feeding these features to models, we can effectively predict human holistic scores, as well as human pronunciation and fluency ratings, although we did not model grammar errors in the way they are specifically described in the rubrics, e.g. in Table 4.

For each item, a specific combination of base scores was selected. So, on an item-by-item basis, we tried two methods of combination: (i) multiple linear regression and (ii) neural networks with one hidden layer trained by back propagation. Then we selected the one that was more accurate for that item. For almost all items, the neural network model worked better.

4.8 Unscorable test detection

Many factors can render a test unscorable: poor sound quality (recording noise, mouth too close to the microphone, too soft, etc.), gibberish (nonsense words, noise, or a foreign language), off-topic (off topic, but intelligible English), unintelligible English (e.g. a good-faith attempt to respond

in English, but is so unintelligible and/or disfluent that it cannot be understood confidently).

There have been several approaches to dealing with this issue (Cheng and Shen, 2011; Chen and Mostow, 2011; Yoon et al., 2011). Some unscorable tests can be identified easily by a human listener, and we reported research on a specified unscorable category (off-topic) before (Cheng and Shen, 2011). Dealing with a specified category could be significantly easier than dealing with wide-open items as in AZELLA. Also, because we did not collect human "unscorable" ratings for this data, we worked on predicting the absolute overall difference between human and machine scores; which is like predicting outliers. If the difference is expected to exceed a threshold, the test should be sent for human grading.

Many problems were due to low volume recordings made by shy kids, so we identified features to deal with low-volume tests. These included maximum energy, the number of frames with fundamental frequency, etc., using many features mentioned in Cheng and Shen (2011). The method used to detect off-topic responses did not work well here, but features based on lattice confidence seemed to work fairly well. If we define an unscorable test as one with an overall difference between human and machine scores greater than or equal to 3 (within the score range 0-14), our final unscorable test detector achieves an equal-error rate of 16.5% in validation sets; or when fixing the false rejection rate at 6%, the false acceptance rate is 44%. We are actively investigating better methods to achieve acceptable performance for use in real tests.

5 Experimental results

All results presented in this section used the validation data sets, while the recognition and scoring models were built from completely separate material. The participant-level speaking scores were designed not to consider the scores from *Read-by-Syllables* and *Read-Three-Words*. For each test, the system produced holistic scores for Repeat items and for non-Repeat items. For every Repeat item, the machine generated pronunciation, fluency and accuracy scores mapped into the 0 to 4 score-point range. Both human and machine holistic scores for a Repeat response are equal to: $50\% \cdot Accuracy + 25\% \cdot Pronunciation + 25\% \cdot Fluency$. Accuracy scores were scaled as $percent\_correct$ times four. Human accuracy scores were based on human transcriptions instead of ASR transcriptions. Holistic scores for Repeat items at the participant level were the simple average of the corresponding item-level scores.

For every non-Repeat item, we generated one holistic score that considered pronunciation, fluency and content together. The non-Repeat holistic scores at the participant level were the simple average of the corresponding item level scores after normalizing them to the same scale. The final generated holistic scores for Repeats were scaled to a $0-4$ range and non-Repeat holistic scores were scaled to a $0-10$ range to satisfy an AZELLA design requirement that Repeat items count for 4 points and non-Repeats count for 10 points. The overall participant level scores are the sum of the Repeat holistic scores and the non-Repeat holistic scores (maximum 14). All machine-generated scores are continuous values. In the following tables, H-H r stands for the human-human correlation and M-H r stands for the correlation between machine-generated scores and average human scores.

Table 5: Human rating reliabilities and Machine-human correlations by item type. Third column gives mean and standard deviation of words per response.

S	Item types	Words/response $\mu \pm \sigma$	H-H r	M-H r
I	Naming	2.5 ± 2.5	0.83	0.67
I	Short Response	5.7 ± 3.8	0.71	0.73
I	Open Question	8.7 ± 7.9	0.70	0.76
I	Repeat Sentence	5.0 ± 2.5	0.91	0.83
II	Questions on Image	14.0 ± 10.8	0.87	0.86
II	Give Directions from Map	10.9 ± 9.7	0.82	0.84
II	Ask Qs about a Thing	6.8 ± 5.9	0.83	0.64
II	Open Question on Topic	11.6 ± 10.6	0.75	0.72
II	Give Instructions	11.5 ± 10.0	0.83	0.80
II	Repeat Sentence	6.1 ± 2.9	0.95	0.85
III	Questions on Image	14.5 ± 10.2	0.87	0.77
III	Similarities & Differences	19.5 ± 11.6	0.75	0.75
III	Give Directions from Map	16.3 ± 11.2	0.74	0.85
III	Ask Qs about a Statement	16.7 ± 13.4	0.79	0.82
III	Give Instructions	17.0 ± 12.8	0.77	0.81
III	Open Question on Topic	13.9 ± 11.1	0.85	0.85
III	Detailed Response to Topic	13.8 ± 10.5	0.81	0.80
III	Repeat Sentence	6.4 ± 3.2	0.97	0.88
IV	Questions on Image	13.9 ± 11.8	0.84	0.84
IV	Give Directions from Map	13.7 ± 13.3	0.84	0.90
IV	Open Question on Topic	17.2 ± 15.2	0.82	0.82
IV	Detailed Response to Topic	13.9 ± 11.4	0.85	0.87
IV	Give Instructions	16.5 ± 15.7	0.87	0.90
IV	Repeat Sentence	6.9 ± 3.2	0.96	0.89
V	Questions on Image	17.3 ± 12.0	0.80	0.76
V	Open Question on Topic	18.7 ± 14.9	0.84	0.82
V	Detailed Response to Topic	17.7 ± 15.2	0.88	0.87
V	Give Instructions	17.2 ± 16.6	0.90	0.90
V	Give Directions from Map	22.4 ± 16.8	0.86	0.85
V	Repeat Sentence	6.4 ± 3.5	0.95	0.89

We summarize the psychometric properties of different item types that contribute to the final scores in Table 5. For each item-type and each stage, the third column in Table 5 presents the mean and standard deviation of the words-per-response produced by students, showing that older students generally produce more spoken material. We found that the number of words spoken is a better measure than speech signal duration to represent the amount of material produced, because young English learners often emit long silences while speaking. The difference between the two measures in columns 4 and 5 is statistically significant (two-tailed, $p < 0.05$) for item types *Naming (Stage I)*, *Ask Qs about a Thing (Stage II)*, *Questions on Image (Stage III)*, and *Repeat Sentence (all Stages)*, in which machine scoring does not match human; and for item types *Give Directions from Map (Stage III, IV)*, in which machine is better than a single human score. For almost all open-ended items, machine scoring is similar to or better than human scoring. We noticed that machine scoring of one open-ended item type, *Ask Qs about a Thing* used in Stage II test forms, was significantly worse than human scoring, leading us to identify problems specific to the item type itself, both in the human rating rubric and in the machine grading approach. Arizona is not using this item type in operational tests.

Figures 1, 2, 3, 4, 5 present scatter plots of overall scores at the participant level comparing human and machine scores for test in each AZELLA stage. Figure 6 shows the averaged human holistic score distribution for participants in the validation set for Stage V. The human holistic score distributions for participants in other AZELLA stages are similar to those in Figure 6, except the means shift somewhat.

We identified several participants for whom the difference between human and machine scores is bigger than 4 in Figures 1, 2, 3, 4, 5. Listening to the recordings of these tests, we concluded that the most important factor was low Signal-to-Noise Ratio (SNR). Either the background noise was very high (in 6 of 1,362 tests in the validation set), or speech volume was low (in 3 of 1,362 tests in the validation set). Either condition can make recognition difficult. With very low voice amplitude and high background noise levels, the SNR of some outlier response recordings is so low that human raters refuse to affirm that they understand the

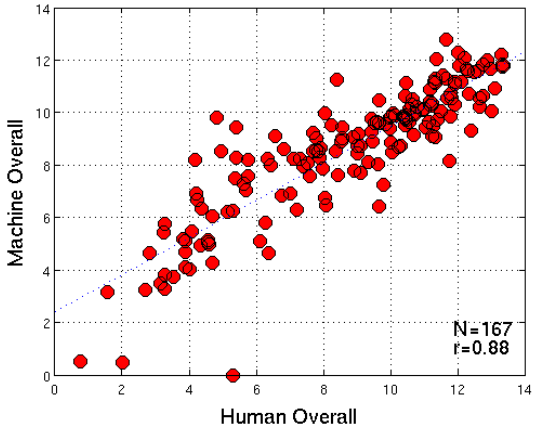

Figure 1: Overall human vs. machine scores at the participant level for Stage I (Grade K). Mean and standard deviation for human scores: $(8.74, 3.1)$.

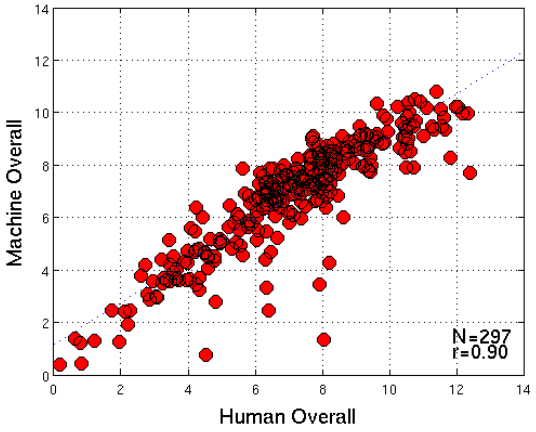

Figure 2: Overall human vs. machine scores at the participant level for Stage II (Grades 1-2). Mean and standard deviation for human scores: $(7.1, 2.5)$.

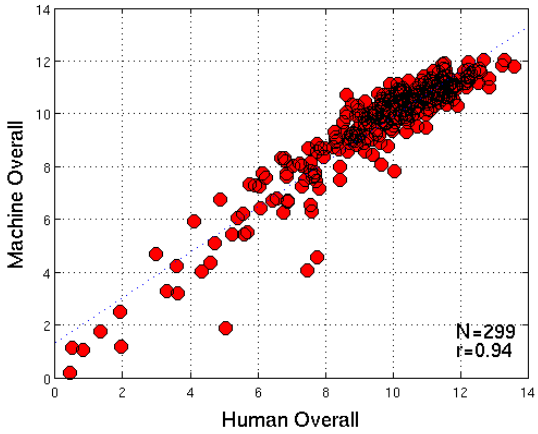

Figure 3: Overall human vs. machine scores at the participant level for Stage III (Grades 3-5). Mean and standard deviation for human scores: $(9.6, 2.3)$.

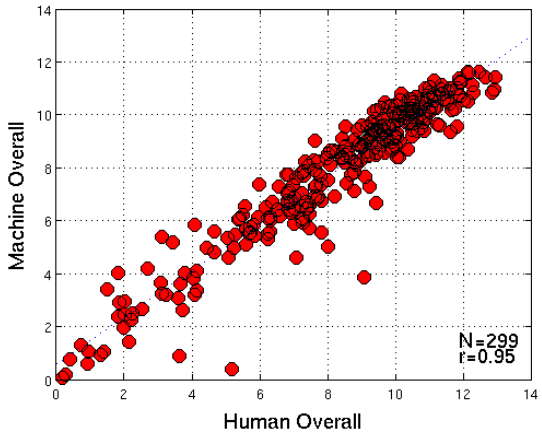

Figure 4: Overall human vs. machine scores at the participant level for Stage IV (Grades 6-8). Mean and standard deviation for human scores: $(8.3, 2.9)$.

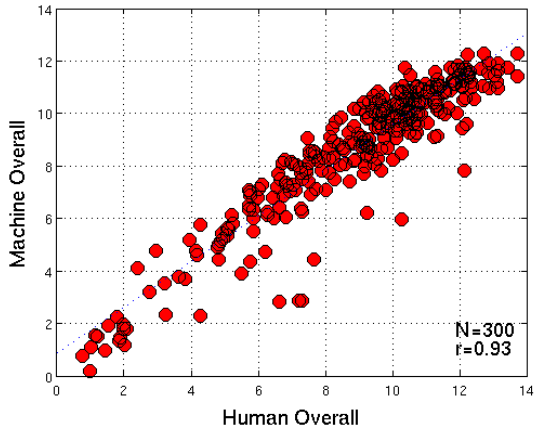

Figure 5: Overall human vs. machine scores at the participant level for Stage V (Grades 9-12). Mean and standard deviation for human scores: $(8.9, 2.9)$.

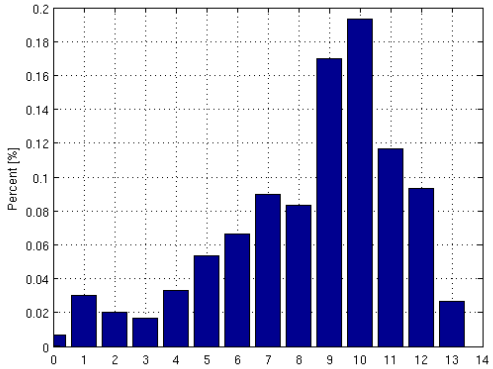

Figure 6: Distribution of average human holistic score for participants in the validation set for Stage V (Grades 9-12).

content of the response or rate its pronunciation. Since many young children in kindergarten and early elementary school speak softly, the youngest children's speech is substantially harder to recognize (Li and Russell, 2002; Lee et al., 1999). This probably contributes to the lower reliabilities in Stage I and II. When setting the total rejection rate at 6%, our unscorable test detector identifies only 7 of the 13 outlier tests.

Table 6: Reliability of human scores and Human-Machine correlations of overall test scores by stage.

Stage	H-H r	M-H r
I	0.91	0.88
II	0.96	0.90
III	0.97	0.94
IV	0.98	0.95
V	0.98	0.93
Average	0.96	0.92

Table 6 summarizes the reliabilities of the tests in different stages. At the participant level, the average inter-rater reliability coefficient across the five stages was 0.96, suggesting that the well-trained human raters agree with each other with high consistency when ratings are combined over all the material in all the responses in a whole test; the average correlation coefficient between machine-generated overall scores and average human overall scores was 0.92. This suggests that the machine grading may be sufficiently reliable for most purposes.

Table 7: Test reliability by stage, separating non-Repeat holistic scores and Repeat holistic scores.

Stage	H-H r NonRptH	M-H r NonRptH	H-H r RptH	M-H r RptH
I	0.85	0.83	0.99	0.94
II	0.93	0.89	0.99	0.90
III	0.95	0.92	0.99	0.92
IV	0.96	0.95	0.99	0.94
V	0.96	0.91	0.99	0.93
Average	0.93	0.90	0.99	0.93

Table 7 summarizes the reliabilities of test scores in the different stages considering the non-Repeat holistic scores and Repeat holistic scores separately to check the effect of adding the Repeat items. Repeat items improve the machine re-

liability in Stage I significantly, but not so much for other stages. This difference may relate to the difficulty in eliciting sufficient speech samples in non-Repeat items from the young EL students in Stage I. Eliciting spoken materials in Repeat items is more straightforward. Consideration of Table 7 suggests that using only open-ended item-types can also achieve sufficiently reliable results.

6 Discussion and future work

We believe that we can improve this system further by scoring Repeat items using a partial credit Rasch model (Masters, 1982) instead of the average of $percent\_correct$, which should improve the reliability of the Repeat item type. We may also be able to train a better native acoustic model by using a larger sample of native data from AZELLA, if we are given access to the test-taker demographic information.

The original item selection and assignment of items to forms was quite simple and had room for improvement. Currently in the AZELLA testing program, test forms go through a post-pilot revision, so that the operational tests only include good items in the final test forms. This post-pilot selection and arrangement of items into forms should improve human-machine correlations beyond the values reported here. If we effectively address the problem of shy-kids-talking-softly, the scoring performance will definitely improve even more. Getting young students to talk louder is probably something that can be best done at the testing site (by instruction or by example); and it may solve several problems. We are happy to report that the first operational AZELLA test with automatic speech scoring took place between January 14 and February 26, 2013, with approximately $140,700$ tests delivered.

Recent progress in machine learning has applied deep neural networks (DNNs) to many long-standing pattern recognition and classification problems. Many groups have now applied DNNs to the task of building better acoustic models for speech recognition (Hinton et al., 2012). DNNs have repeatedly been shown to work better than Gaussian mixture models (GMMs) for ASR acoustic modeling (Hinton et al., 2012; Dahl et al., 2012). We are actively exploring the use of DNNs for use in recognition of children's speech. We expect that DNN acoustic models can overcome some of the recognition difficulties mentioned in this paper (e.g. low SNR in responses and short response item types like *Naming*) and boost the final assessment accuracy significantly.

7 Conclusions

We have reported an evaluation of the automatic methods that are currently used to assess spoken responses to test tasks that occur in Arizona's AZELLA test for young English learners. The methods score both the content of the responses and the quality of the speech produced in the responses. Although most of the speaking item types in the AZELLA tests are unconstrained and open-ended, machine scoring accuracy is similar to or better than human scoring for most item types. We presented basic validity evidence for machine-generated scores, including an average correlation coefficient between machine-generated overall scores and human overall scores derived from subscores that are based on multiple human ratings. Further, we described the design, implementation and evaluation of a detector to catch problematic, unscorable tests. We believe that near-term re-optimization of some scoring process elements may further improve machine scoring accuracy.

References

Arizona Department of Education. 2012. AZELLA update. `http://www.azed.gov/standards-development-assessment/files/2012/12/12-12-12-update-v5.pdf`. [Accessed 19-March-2014].

Arizona Department of Education. 2014. Arizona English Language Learner Assessment (AZELLA). `http://www.azed.gov/standards-development-assessment/arizona-english-language-learner-assessment-azella`. [Accessed 19-March-2014].

J. Bernstein and J. Cheng. 2007. Logic and validation of a fully automatic spoken English test. In V. M. Holland and F. P. Fisher, editors, *The Path of Speech Technologies in Computer Assisted Language Learning*, pages 174–194. Routledge, New York.

J. Bernstein, J. De Jong, D. Pisoni, and B. Townshend. 2000. Two experiments on automatic scoring of spoken language proficiency. In *Proc. of STIL (Integrating Speech Technology in Learning)*, pages 57–61.

J. Bernstein, A. Van Moere, and J. Cheng. 2010. Validating automated speaking tests. *Language Testing*, 27(3):355–377.

W. Chen and J. Mostow. 2011. A tale of two tasks: Detecting children's off-task speech in a reading tutor. In *Interspeech 2011*, pages 1621–1624.

J. Cheng and J. Shen. 2010. Towards accurate recognition for children's oral reading fluency. In *IEEE-SLT 2010*, pages 91–96.

J. Cheng and J. Shen. 2011. Off-topic detection in automated speech assessment applications. In *Interspeech 2011*, pages 1597–1600.

J. Cheng. 2011. Automatic assessment of prosody in high-stakes English tests. In *Interspeech 2011*, pages 1589–1592.

G. Dahl, D. Yu, L. Deng, and A. Acero. 2012. Context-dependent pretrained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing, Special Issue on Deep Learning for Speech and Lang. Processing*, 20(1):30–42.

R. Downey, D. Rubin, J. Cheng, and J. Bernstein. 2011. Performance of automated scoring for children's oral reading. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 46–55.

M. Eskanazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51:832–844.

H. Franco, L. Neumeyer, Y. Kim, and O. Ronen. 1997. Automatic pronunciation scoring for language instruction. In *ICASSP 1997*, pages 1471–1474.

H. Franco, L. Neumeyer, M. Ramos, and H. Bratt. 1999. Automatic detection of phone-level mispronunciation for language learning. In *Eurospeech 1999*, pages 851–854.

D. Higgins, X. Xi, K. Zechner, and D. Williamson. 2011. A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, 25:282–306.

G. Hinton, L. Deng, Y. Dong, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Kaggle. 2012. The Hewlett Foundation: Short answer scoring. `http://www.kaggle.com/c/asap-sas;http://www.kaggle.com/c/asap-sas/details/winners`. [Accessed 20-April-2014].

T. K. Landauer, P. W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

S. Lee, A. Potamianos, and S. Narayanan. 1999. Acoustics of children's speech: developmental changes of temporal and spectral parameters. *Journal of Acoustics Society of American*, 105:1455–1468.

Q. Li and M. Russell. 2002. An analysis of the causes of increased error rates in children's speech recognition. In *ICSLP 2002*, pages 2337–2340.

G. N. Masters. 1982. A Rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.

L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. 1999. Automatic scoring of pronunciation quality. *Speech Communication*, 30:83–93.

L. Tandalla. 2012. ASAP Short Answer Scoring Competition System Description: Scoring short answer essays. `https://kaggle2.blob.core.windows.net/competitions/kaggle/2959/media/TechnicalMethodsPaper.pdf`. [Accessed 20-April-2014].

J. Tepperman, M. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. Narayanan. 2007. A Bayesian network classifier for word-level reading assessment. In *Interspeech 2007*, pages 2185–2188.

S. M. Witt and S. J. Young. 1997. Language learning based on non-native speech recognition. In *Eurospeech 1997*, pages 633–636.

S. M. Witt and S. J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108.

S. Xie, K. Evanini, and K. Zechner. 2012. 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*, pages 103–111.

S.-Y. Yoon, K. Evanini, and K. Zechner. 2011. Non-scorable response detection for automated speaking proficiency assessment. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 152–160.

S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. 2000. *The HTK Book Version 3.0*. Cambridge University, Cambridge, England.

J. Zbontar. 2012. ASAP Short Answer Scoring Competition System Description: Short answer scoring by stacking. `https://kaggle2.blob.core.windows.net/competitions/kaggle/2959/media/jzbontar.pdf`. [Accessed 20-April-2014].