Spoken English Grading: Machine Learning with Crowd Intelligence

Vinay Shashidhar, Nishant Pandey and Varun Aggarwal
Aspiring Minds
vinay.shashidhar@aspiringminds.com, nishant.pandey@aspiringminds.com,
varun@aspiringminds.com

ABSTRACT

In this paper, we address the problem of grading spontaneous speech using a combination of machine learning and crowdsourcing. Traditional machine learning techniques solve the stated problem inadequately as automatic speaker-independent speech transcription is inaccurate. The features derived from it are also inaccurate and so is the machine learning model developed for speech evaluation. We propose a framework that combines machine learning with crowdsourcing. This entails identifying *human intelligence* tasks in the feature derivation step and using crowdsourcing to get them completed. We post the task of speech transcription to a large community of online workers (crowd). We also get spoken English grades from the crowd. We achieve 95% transcription accuracy by combining transcriptions from multiple crowd workers. Speech and prosody features are derived by force aligning the speech samples on these highly accurate transcriptions. Additionally, we derive surface and semantic level features directly from the transcription. We demonstrate the efficacy of our approach by predicting expert graded speech sample of 566 adult non-native speakers across two different countries - India and Philippines. Using the regression modeling technique, we are able achieve a Pearson correlation of 0.79 on the Philippines set and 0.74 on the Indian set with expert grades, an accuracy much higher than any previously reported machine learning approach. Our approach has an accuracy that rivals that of expert agreement. We show the value of the system through a case study in a real-world industrial deployment. This work is timely given the huge requirement of spoken English training and assessment.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech Recognition and synthesis*

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2788595 .

Keywords

Spontaneous speech grading; Spoken English evaluation; Crowdsourcing; Speech recognition; Machine Learning

1. INTRODUCTION

The grading of constructed (open) response items, popularly known as subjective evaluation, provides a more holistic and accurate assessment of a candidate's skills as compared to selected response items (multiple-choice questions) [1]. The primary limitation of a selected response item is that it asks the candidate to *choose* the right answer, providing implicit hints and the structure of the solution [2]. However they are traditionally preferred over constructed response as their evaluation can be automated fairly easily. With the recent interest in MOOCs (Massively Online Open Courseware) [3], scalable education/training and automated recruitment assessment [4, 5], the interest in reducing manual effort in the assessment of constructed responses has increased manifolds.

Machine learning has been used to automatically grade these responses. There are many examples of successfully using machine learning for constructed response grading [2, 6]. A framework and general principles for using machine learning to grade constructed responses are presented in [2].

However machine learning techniques fall short of providing accurate assessment for a number of problems [3, 7]. Secondly, these automated approaches have come under criticism since the test-takers can fake high-scoring responses [8]. For instance, automated assessment of spontaneous speech for spoken language evaluation remains largely an unsolved problem [6]. On the other hand, automatic essay grading algorithms can be tricked by inserting the right words in random order or writing long essays [8].

We have identified that one of the key limitations of the current automated techniques is their inability to automatically derive the right set of *features* with high precision for assessing the response. Interestingly, many of these features can be easily derived by (or with help of) *humans*, even if they are non-experts. For instance, word sense disambiguation [9] or gibberish identification for essay grading, text transcription from noisy handwriting [10] and speech transcription [11] for speech grading are easily doable tasks by humans through crowdsourcing [12].

With this insight, we propose an extended machine learning framework with a crowdsourcing layer for constructed response assessment. In this approach, one identifies *human intelligence* tasks in the feature derivation step and posts

them to be completed by a non-expert crowd.[1] The responses from the human intelligence tasks are then used to create relevant features for machine learning.

We illustrate our technique for evaluation of spontaneous English speech[2].

Evaluation of spoken English is in high demand for recruitment in the industry and admissions to the universities [17]. Universities teaching in English expect a level of English proficiency in incoming students whereas various job roles require different levels of spoken English capabilities. With the massive increase in the number of applicants applying for both university and companies, there is a great to scale this process by automating it. Having an automated/semi-automated way of evaluation for the purpose of giving feedback to candidates cannot be overstated.

The problem acquires far reaching significance given the evidence that better English leads to better employment outcomes, wages and promotions [18, 19].

Figure 1 shows the application of technique to spoken English grading. In the task of spontaneous speech evaluation, the text of the speech is not known a priori. We post the task of transcription of the sample speech to the crowd who transcribe it fairly accurately [11]. We force-align [18, 20] the speech of the candidate on this text to derive various features which go into a machine learning engine. We also collect non-expert grades of the speech from the crowd, which can be optionally used as additional features. With these accurately identified features with the help of the crowd, machine learning (specifically the modeling step) becomes a powerful technique for constructed response grading.

In summary, our approach is to identify human-intelligence tasks in the feature derivation step and let the crowd complete these in an accurate and reliable way. This enables machine learning to grade spoken English accurately. Our approach is an interesting addition to the extant machine learning only approaches and peer/crowd grading approaches. The approach could help in other grading problems such as that of essay's by outsourcing word sense disambiguation [9], detecting insensible sentences or grammar checks to the crowd. It could also be used to grade handwritten assignments of students [10], where the crowd digitizes the responses. It needs to be seen how one creatively gleans human intelligence tasks for other grading tasks and what is the incremental predictive value added by our technique.

In this paper, we show how we can solve a hitherto unsolved problem [6] of English spontaneous speech evaluation using our approach.

The paper makes the following contributions:

- We introduce a general framework to add a crowdsourcing step for feature derivation in a machine learning based constructed response grading systems. This adds a new alternative to the set of techniques used for automated and semi-automated grading.
- We find that the features derived from crowdsourced transcription do better in predicting expert grades than the machine learning only approach. These features when combined with crowd grades predict spoken English scores with accuracy comparable to expert agreement.
- We provide a scalable and an accurate way to do spoken English grading, a huge requirement in the industry and elsewhere.

The paper is organized as follows– Section 2 describes previous work, compares our technique with current incumbent techniques; Section 3 describes the procedure and aim of the speech assessment task; Section 4 describes the features classes used in the prediction algorithm; Section 5 describes the crowdsourcing framework which is used as input to machine learning methods; Section 6 shows how this framework was used with machine learning techniques to predict a spoken English composite score and a case study of a real world implementations and results; Section 7 discusses the future work and concludes the paper.

2. PREVIOUS WORK

Machine learning and human intelligence have been used in tandem since a long time. The classic application is to use the crowd to label samples and then use supervised learning to predict the labels. Within grading of constructed responses, recently [21] learnt a grade using ML and used its confidence score to determine the number of expert evaluations required for the task. We believe that our work is the first proposal and implementation of using crowdsourcing specifically for the feature derivation step of ML in the context of grading. To our best knowledge, this has not been attempted for other problems either. Some loosely related works include [22, 23, 24].

Our technique, when compared to a ML only approach, has a promise of higher accuracy, but makes some trade-offs. Firstly, there is a cost for every assessment done and the scalability depends on the number of non-expert workers available. These drawbacks are strictly speaking, valid, but they didn't lead to practical challenges in our case. We paid 24 cents for each completely graded sample, a very affordable cost (per assessment) which is significantly lower than that of any expert grading efforts. Recently we have had crowd rate hundreds of samples in a day without any challenge or the need of altering the payment significantly.

Secondly, machine learning based assessments generally provide evaluations in real-time, whereas there is a time lag when using crowdsourcing. This works fine in many scenarios, but doesn't cater well to provide real-time feedback. Real time crowdsourcing has been an active area of research [25, 26] and a focus area for our group as well.

In comparison to peer grading methods [13] or crowdsourcing assessments [14, 15, 16], our technique works differently. In these techniques, the crowd/peers directly evaluates/grades the response on a rubric and a combination of their grades mimics the grades given by experts. We, on the other hand, use the crowd to help complete a human intelligence task, which helps in better feature derivation. One may additionally combine crowd grades too, if needed. This offers a new trade-off point among techniques with interesting possibilities. Crowd grades do not work for evaluating

[1]Our approach is different from peer grading [13] or crowd grading [14, 15, 16] approaches. These approaches directly ask the crowd to grade the response. The primary feature of our technique is using the crowd in the feature extraction step of machine learning. We provide a detailed comparison in Section 2.

[2]Spontaneous speech refers to text–independent speech samples.

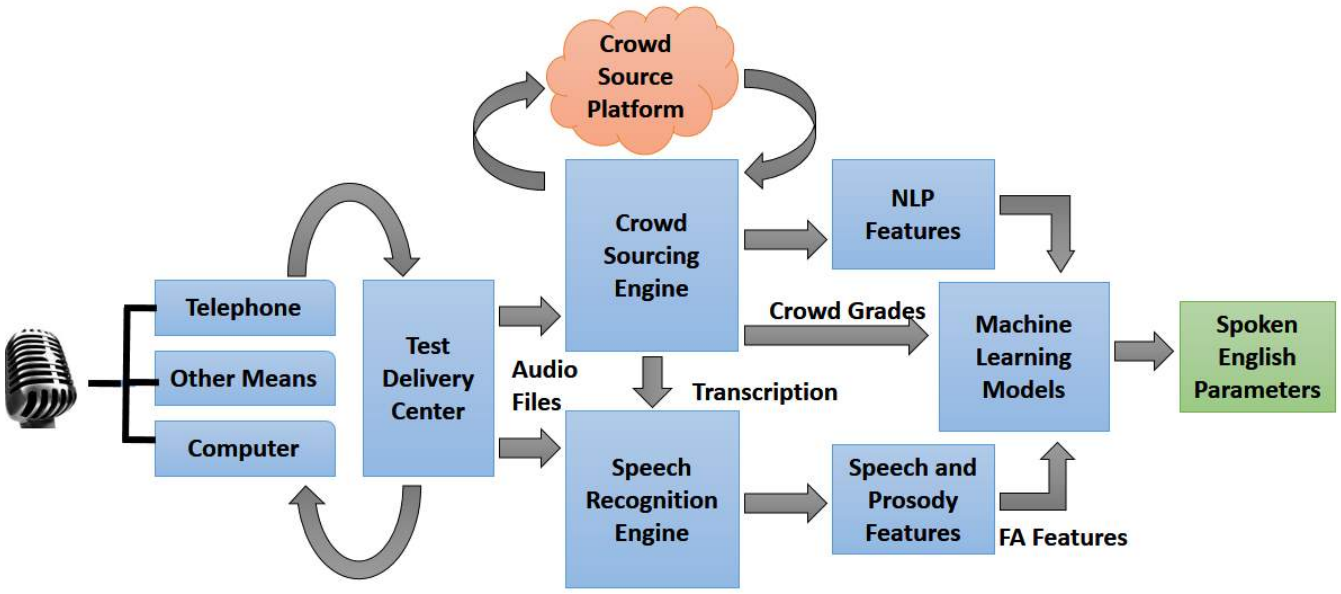

Figure 1: System Design

expert tasks, say a computer program or an advanced electronics question and such evaluations need peers with some exposure to the subject. Our technique has the possibility to make these amenable with cheaper and more scalable crowd, non-expert raters. It may also help improve the accuracy of peer/crowd grading techniques or reduce workers, given that we use a complementary set of information from the crowd. Futuristically, our technique could be more defensible for mid/high stake testing, given it uses the crowd for human intelligence tasks only and not direct evaluation.

In summary, our approach lies somewhere in between the machine learning only approaches and that of drawing grades from the consensus of crowd/experts.

3. GRADING TASK

We want to assess the quality of spoken English of candidates based on their spontaneous speech samples. The speech samples of the candidates were collected using Aspiring Minds' automated speech assessment tool– SVAR [5]. SVAR is conducted over phone as well as on a computer. The test has multiple sections where the candidate is required to: read sentences aloud, listen and repeat sentences, listen to a passage or conversation and answer multiple choice questions and finally spontaneously speak on a given topic. In the spontaneous speech section, the candidates[3] are provided with a topic and given 30 seconds to think, take notes and then speak on the topic for 45 seconds. The topic is repeated to ensure task clarity. The complete test takes 16-20 minutes to complete, depending on the test version.

Currently, SVAR evaluates speech samples from the read and repeat sections with high accuracy [5]. Our goal in this paper is to evaluate the spontaneous speech of the candidate and provide a composite score based on it.

[3]The subjects of our study use English as their second language and hail from various backgrounds, dialects and educational qualifications.

A 5 point rubric for the composite score, similar to CEFR [27], was prepared with the help of experts. This score is a function of the pronunciation, fluency, content organization and grammar quality of the speech sample. Broadly speaking, Pronunciation [28] refers to the correctness in the utterance of the phonemes of a word by the students as per neutral accent. Fluency [29] refers to a desired rate of speech along with the absence of hesitations, false starts and stops etc. Content organization [30] measures the candidate's ability to structure the information disposition and present it coherently. Grammar [31] measures how well the syntax of the language was followed by the candidate.

In the next section we discuss the features which are used in the prediction algorithm.

4. FEATURES

We use three classes of features– Crowd Grades (CG), Force Alignment features (FA) and Natural Language Processing features (NLP). The spoken English samples are posted to the crowd to get the transcription and spoken English grades (Figure 1). Each task was completed by three workers. The crowd grades become one set of features. A second set, i.e., FA features, are derived by aligning [32] the speech sample on the crowdsourced transcriptions. A third set, i.e., NLP features, are also derived from the crowdsourced text. These are explained in the succeeding paragraphs.

- *Crowd Grades:* The crowd transcribes the speech in addition to providing scores on each of the following– pronunciation, fluency, content organization and grammar. These grades are combined to form a composite score per worker per candidate. These are further averaged across workers to give a final score.[4]

[4]Advanced Expectation-Maximization techniques [33] may also be used for an aggregation strategy, once the number of tasks done by every individual worker increases. In our current experiments, this number wasn't very high.

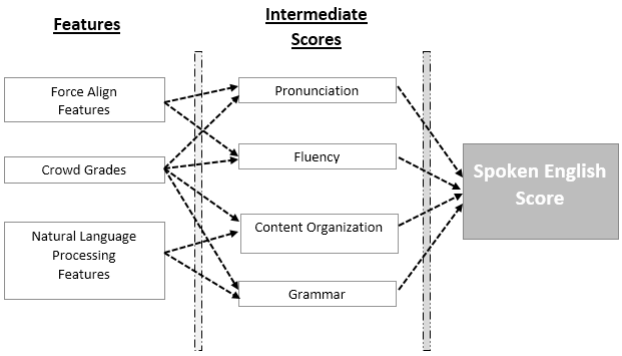

Figure 2: Our intuition of how different features predict the holistic score.

- *FA features:* The speech sample is forced aligned [18, 20] on the crowdsourced transcription using the HTK speech recognizer [34]. We used an acoustic model based on TIMIT [35] for our experiments. TIMIT is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects.

  A number of speech quality features are derived, which include– rate of speech, position and length of pauses, log likelihood of recognition, posterior probability, hesitations and repetitions, etc. [36, 6, 37]. These features are predictive of the pronunciation and fluency of the candidate.

- *NLP features:* These features predict the content quality and grammar of the spoken content[5]. They were derived using standard NLP packages [38, 39] on the crowdsourced transcription. The package calculates surface level features such as the number of words complexity or difficulty of words and the number of common words used. It also calculates semantic features like the coherency in text, context of the words spoken, sentiment of the text and grammar correctness. These features are predictive of the grammar and content organization of the sample.

All the features described above were obtained for the spontaneous speech sample. We also derived features similar to FA features for the candidate's read and repeat speech samples collected during his/her SVAR test. The speech and prosody features are calculated by force aligning the speech on the known text. One of the models (RS/LR) in our experiments is based on these features and has been included for comparison. These features do not have any bearing on our final model for spontaneous speech evaluation.

5. CROWDSOURCING

The spoken English sample was given to the crowd to transcribe and provide grades. The task was posted on a popular crowdsourcing platform– Amazon Mechanical Turk (AMT) [40]. AMT is a popular crowdsourcing marketplace. It is inspired by the famous 18^{th} century automated chess playing machine, running on the intelligence of a hidden human operator. It has more than $500,000$ online workers from 190 countries [41]. One can post tasks on the platform online and offer fixed remuneration for their completion.

A clean and simple interface was provided to the worker with standard features needed for transcription. Additionally, an advanced audio player was embedded with the ability to play the speech sample in repeat mode, rewind and forward, apart from standard play/pause functionality to help the worker. The different transcriptions were combined using the ROVER algorithm [42]. ROVER is a sophisticated voting algorithm to combine multiple transcriptions with errors, to obtain the best estimate of the correct transcription. It is reported to lead to an error reduction of 20-25%. ROVER proceeds in two stages: first the outputs are aligned and a single word transcription network (WTN) is built. The second stage consists of selecting the best scoring word (with the highest number of votes) at each node.

Several methods have been used in the past for increasing the reliability of the grades given by the crowd by identifying and correcting any biases and removing non-serious/low quality workers [43]. One of the key techniques for this involves inserting gold standard tasks with known answers to get an estimate of the worker's ability [44]. The gold standard tasks are similar to real tasks and the workers have no way to distinguish between the two. Our tasks took workers a reasonable amount of time (8-10 minutes). It wasn't hence feasible to insert a gold standard task, as done typically, with every task to be completed.

To overcome this problem, we propose an innovative approach where a risk is assigned to a worker based on his/her performance on the gold standard tasks. We conceptualized this system as a state machine that determines the risk level of a worker and proposes actions based on it (Refer to Figure 3). All workers started with an initial risk level of 0.2. Gold standard tasks were probabilistically inserted among real tasks based on the worker's risk level. Workers with a higher risk level saw more gold standard tasks. Also, the risk level of the worker was updated based on his/her performance on the gold standard tasks. Workers who consistently performed poorly on gold standard tasks were allocated a higher risk level and a notification was sent to them with a corrective course of action. Beyond a certain level, the worker was barred from attempting future work. We did not do any retrospective correction of the barred worker's completed tasks and simply stopped him/her from attempting newer tasks. This approach allowed us to control for the quality of workers, provide feedback, remove unsuitable workers and also adaptively control the balance between real and gold standard tasks.[6]

We describe the experimental setup and the results in the next section.

6. EXPERIMENTS

We conducted the experiments to answer the following questions:

- Can read/repeat features predict spontaneous speech grades accurately?

[5]We were looking at prompt independent features only, at this point.

[6]Specific details of the implementation are beyond the scope of the paper.

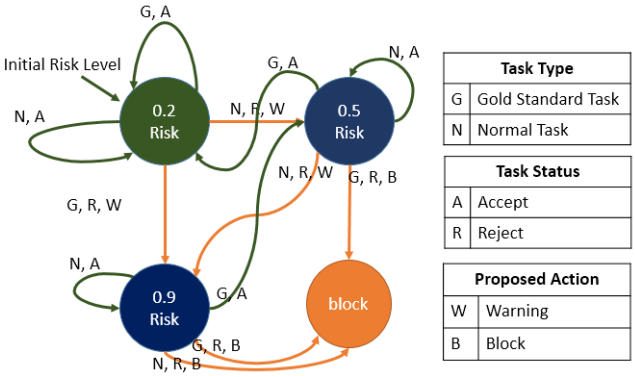

Figure 3: Risk Level State Diagram: In the above figure, each node corresponds to a risk level associated with a worker. The values range between 0 (min) - 1 (max). The worker is either assigned a gold standard task (G) or a normal task (N) on the basis of his/her present risk level. The risk level changes every time a task is Accepted (A) or Rejected (R). Additionally worker may be warned (W) or blocked (B) in case of rejection.

- How accurate is a pure machine learning approach (without crowdsourced transcription) in predicting grades as compared to grades given by human experts?
- How much better is the ML-CS approach in predicting grades as compared to a pure ML approach and to using Crowd Grades only?
- Do Crowd Grades add additional value in predicting grades over and above the features derived from the crowdsourced transcription?

We conducted the experiments on 566 spontaneous speech samples which were graded by expert assessors. Out of these, 319 were of native Indians (IN set), while the remaining 247 were of native Philippines (PH set). To answer the questions stated above, we developed models for the IN and PH sets separately.[7]

In the model building phase, we used different set of features to develop the models and compared their accuracy. The models were built against expert grades using supervised learning techniques. We experimented with three machine learning techniques– Ridge Regression, SVMs and Neural Networks. The data set used in the experiments is discussed in the next section.

6.1 Data Set

Our data set contains a total of 566 spontaneous speech responses comprising of 319 samples from India and 247 samples from Philippines. The speech samples in both sets were from seniors (non–native English speakers in final year of undergraduate education) pursuing bachelor's degree in their respective countries. The candidates were asked to describe one of the following six scenes: *a hospital, flood, a crowded market, an airport, favorite holiday destination* and *a school playground*. The candidates were given 30 seconds to think and take notes and were then asked to speak for the next 45 seconds. The responses were collected on the phone during the SVAR test [5]. Apart from the spontaneous speech response, each candidate was asked to read 12 given sentences and repeat 9 given sentences immediately after listening to each of them. Empty or very noisy responses (not humanly discernible) were not included in the final 566 sample set.

[7]We tried to build a single model over both samples, but it did not get good results. This could be because a good speech in one accent may not be equivalent to the other quantitatively. It may need further techniques for normalization.

Table 1: Inter-rater correlations

Sample Set	Inter correlation for read/repeat scores	Inter correlation for spontaneous speech score
IN	0.83	0.86
PH	0.80	0.82

The responses in both sets were rated by one common rater with more than thirty years' experience in grading spoken English and one specific rater each who had more than fifteen years' experience. There were two set of scores. The first was a holistic score on the spontaneous speech samples based on its pronunciation, fluency, content characteristics and grammar. The second was a score on the pronunciation and fluency quality of the read/repeat sentences. The correlation between grades given by the two experts is shown below in Table 1. For each of the two scores, the average of the scores by the two expert grades was used for further purposes.

The correlation between the expert scores on spontaneous speech and automated read/repeat speech for IN set was 0.54 and for PH set was 0.56. This shows that there is a considerable unexplained variance (approx. 70%) in the spontaneous speech score, not addressed by the read/repeat scores. This could be due to a difference in the pronunciation quality and fluency of the candidates in reading/repeating text vs. speaking spontaneously and also due to the additional parameters of grammar and content characteristics in the spontaneous speech score. Thus, an automatic score mimicking the read/repeat expert grades, which is a solved problem, is inadequate for our task.

The first score is used for all subsequent discussion and development of models.

6.2 Crowdsourced Tasks

The 566 speech sample assessment task was posted on Amazon Mechanical Turk (AMT). Each task was completed by three workers. In total, 312 unique workers completed the tasks. The majority of workers (90%) belonged to USA and India.

The task took on an average 8–9 minutes to complete and a worker was paid between 6–10 cents per task including a bonus which was paid on completion of every 4 tasks. The average transcription accuracy for a worker was 79.7%[8]. This significantly improved to 94.6% when the transcriptions of the three workers were combined using the ROVER algorithm. In comparison, the accuracy of automatic transcription of a speech recognition engine was 51.6%.

[8]PHP similar_text function was used as similarity metric.

Table 2: Regression Results

Technique	Model Code	Feature Type	IN Set		PH Set	
			Train r	Validation r	Train r	Validation r
Ridge Regression	RR-1	RS/LR	0.42	0.51	0.47	0.44
	RR-2	Pure ML	0.46	0.48	0.60	0.54
	RR-3	Crowd Grades	0.61	0.67	0.61	0.71
	RR-4	ML-CS	0.64	0.70	0.77	0.60
	RR-5	All	0.74	0.74	0.76	0.79
SVM	SVM-1	RS/LR	0.42	0.51	0.49	0.43
	SVM-2	Pure ML	0.44	0.46	0.60	0.54
	SVM-3	Crowd Grades	0.62	0.57	0.60	0.70
	SVM-4	ML-CS	0.60	0.61	0.76	0.61
	SVM-5	All	0.75	0.74	0.75	0.78
Neural Networks	NN-1	RS/LR	0.47	0.51	0.47	0.47
	NN-2	Pure ML	0.60	0.44	0.55	0.49
	NN-3	Crowd Grades	0.61	0.57	0.58	0.63
	NN-4	ML-CS	0.68	0.58	0.62	0.61
	NN-5	All	0.76	0.75	0.76	0.78

6.3 Regression Modeling

Regression Modeling and the steps described herein were performed separately for IN and PH set. Each set was split into two sets: train and validation. The train set had 67% of the sample points whereas the validation set had 33%. The split was done randomly making sure that the grade distribution in both the sets was similar. While learning the model, a 3-fold cross validation was performed on the train sample.

Linear ridge regression, Neural Networks and SVM regression with different kernels were used to build the models. The least cross-validation error was used to select the models. We used some simple techniques for feature selection including forward feature selection and the algorithm which removes all but the k highest correlating features.

In next paragraphs, we explain the steps performed in tuning regression parameters and feature set used to build regression model.

Regression parameters: For linear regression with regularization, optimal ridge coefficient λ, between 1 and 1000, was selected based on the least RMS error in cross-validation. For support vector machines we tested two kernels: linear and radial basis function. In order to select the optimal SVM model, we varied the penalty factor C, parameters γ and ϵ, the SVM kernel and the selected set of values that gave us the lowest RMS error in cross-validation. The Neural Networks model had one hidden layer and 5 to 10 neurons.

Feature sets used: The experiments were carried on five sets of features: first, features generated by automatic speech transcription of spontaneous speech using a speech recognizer (Pure ML approach); second, a set of features generated by force aligning read/repeated by candidates (another ML only approach referred to as RS/LR approach, see Section 4); third, a set of features pertaining to grades given by the crowd; fourth, NLP and FA features generated by force aligning spontaneous speech on crowdsourced transcription (ML-CS approach) and fifth, NLP, FA features from crowdsourced transcription and Crowd Grades.

In the following subsection, the features pertaining to ML-CS approach are referred to as ML-CS, those pertaining to natural language processing on crowdsourced transcription are referred to as NLP features while the one pertaining to crowd grades are referred to as Crowd Grades.

6.4 Observations

The results of the experiments are tabulated in Table 2. We report the Pearson coefficient of correlation (r) for the different models against the expert grades for IN and PH samples. These are the results for the models selected according to least cross-validation error. The best cross-validation error in case of SVMs was obtained for the linear kernel.

All the following observations are based on the validation error. All the three techniques perform similarly with Neural Networks doing slightly worse in some cases. The broad trends across feature–sets remain similar across different modeling techniques. We will be referring to the ridge regression results for further discussion.

Firstly, it is observed that the read/repeat features predict the spontaneous speech score with low accuracy ($r_{\text{PH}} = 0.47$, $r_{\text{IN}} = 0.47$) for both samples. This implies that read/repeat speech and derived features are inadequate to grade a person's spontaneous speech, the ultimate test of a person's spoken language skills. The second observation is that the ML-only approach using spontaneous speech features (Model IN_RR-2, PH_RR-2) is also inadequate to grade spontaneous speech and does worse than approaches that uses features from crowdsourced transcription (Model PH/IN_RR-4). This clearly shows the value of getting accurate transcription from workers towards better features and model.

Further, among the crowdsourcing approaches, we find that the crowd grades (Model RR-3) does equivalently well (and sometimes worse) than the model using features derived from the crowdsourced speech (Model IN/PH_RR-4). However, when we combine all the features from crowdsourcing including the crowd grades, we find much better prediction accuracy ($r_{\text{IN}} = 0.74$ and $r_{\text{PH}} = 0.79$). This shows that the crowd grades feature provides some orthogonal information as compared to the features from the crowdsourced transcription, towards predicting the grade given by experts.

The validation r for Model IN_RR-5 is 0.74 and Model PH_RR-5 is 0.79. We find that the expert agreement on the validation sample is 0.77 and 0.82 respectively for each model. Thus, our predicted score rivals the agreement of experts. This shows great promise for the technique to be used in a high stake test setting.

In summary, we show the following:

- Read/repeat speech features are inadequate to predict spontaneous speech scores.
- ML only approach based on spontaneous speech samples is also inadequate for the purpose.
- Features derived from crowdsourced transcription (or even crowd grades) do better than a ML only approach.
- When considering features from crowdsourced transcription and crowd grades together, we can predict spontaneous speech scores as well as those done by experts.

6.5 Case Study

We studied the deployment of our spontaneous speech scoring algorithm at a hiring event of a potential customer in Philippines. The company was using another automated speech evaluation product. The scoring of this product was based on the read and repeat speech of the candidate and could not do spontaneous evaluation. The customer was not satisfied with the product and experienced high type-1 error. The recruitment event had 500 applicants for the role of a customer support executive, who had to talk to native English speakers. The new scoring algorithm was tested on a subset of 150 students. All these candidates took the complete SVAR test along with the spontaneous speech section. The same set of students also took the existing speech grading product of the other vendor. These candidates were also evaluated by an expert designated by the company. The expert graded each candidate's speech as hirable or not-hireable based on the requirements for the position.

We sweep thresholds across the scores from our spontaneous speech algorithm, read speech scores used by the company and SVAR read speech scores, to enable hirable/not-hireable decisions. Figure 4 shows the trade-off curve between rejecting good candidates (type-2 error) and selecting bad candidates (type-1 error). The spontaneous speech score ROC curve completely dominates the read speech by both SVAR and the existing product.

Based on one of the trade-off points on the curve, we are correctly able to identify 90% of the hirable candidates while allowing 21% of not-hireable candidates into the final hiring pool. The extant approach however could only identify 69% of the hirable candidates while allowing approximately the same amount of not-hireable candidates. Another trade-off point we consider is with a much lower type-1 error. By selecting an appropriate point, we were able to bring down the percentage of not-hireable candidates in the final set to 15%, while rejecting around 30% of the hirable candidates. On a similar selection percentage of not-hireable candidates, the other scoring algorithms were rejecting around 50% hirable candidates.

The above case study clearly shows the superior candidate selection capabilities of the spontaneous speech scoring algorithm over the score based on the read speech of the candidate. This preliminary case study demonstrates how our approach gives immediate benefit and saves time of the interviewers by massively reducing the number of good candidates being rejected while also reducing the number of bad candidates selected, thus improving the overall hiring quality.

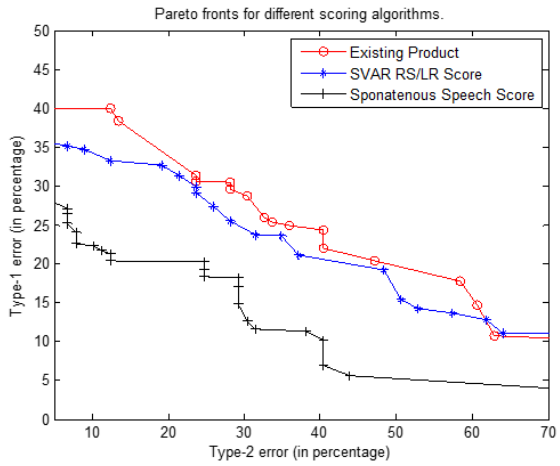

Figure 4: Classification errors of different scoring algorithms

7. CONCLUSIONS

We addressed the problem of evaluating spontaneous speech using a combination of machine learning and crowdsourcing. To achieve this, we post the task of speech transcription to the crowd. Additionally, we also get spoken English grades from the crowd. We are able to derive accurate features by force aligning the speech sample on the crowdsourced text. We experimented our technique on expert–graded speech samples of adult non–native speakers from India and Philippines. Using these features in a regression model, we are able to predict expert grades with much higher accuracy than a machine learning only approach. These features also predict equivalent or better than crowd grades and a combination of these two outperforms all other approaches. Our approach shows an accuracy that rivals that of expert agreement.

Our technique has a promise of higher accuracy but has some trade-offs compared to fully automated approaches. First, there is a cost for every assessment done and the scalability depends on the number of non-expert workers available. Though these drawbacks exist, we were able get tasks done inexpensively. We recently had the crowd rate a hundred samples in a day without any challenge. Second, our approach doesn't provide instant grades. This works fine in many scenarios, but doesn't cater well to providing real-time feedback. Real time crowdsourcing has been an active area of research [25, 26] and is an area for future work for us as well.

8. REFERENCES

[1] Menucha Birenbaum and Kikumi K Tatsuoka. Open-ended versus multiple-choice response formats - it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, 11(4):385–395, 1987.

[2] Varun Aggarwal, Shahank Srikant, and Vinay Shashidhar. Principles for using machine learning in the assessment of open response items: Programming assessment as a case study. *NIPS - Workshop on Data Driven Education*, 2013.

[3] P Mitros, Vikas Paruchuri, John Rogosic, and Diana Huang. An integrated framework for the grading of freeform responses. *MIT Learning International Networks Consortium*, 2013.

[4] Cliff E Beevers and Jane S Paterson. Automatic assessment of problem-solving skills in mathematics. *Active Learning in Higher Education*, 4(2):127–144, 2003.

[5] SVAR. 2014. http://www.aspiringminds.in/talent-evaluation/spoken-english-SVAR.html.

[6] Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895, 2009.

[7] Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. Powergrading: a clustering approach to amplify human effort for short answer grading. *TACL*, 1:391–402, 2013.

[8] Donald E Powers, Jill C Burstein, Martin Chodorow, Mary E Fowles, and Karen Kukich. Stumping *e-rater*: challenging the validity of automated essay scoring. *Computers in Human Behavior*, 18(2):103–134, 2002.

[9] Cem Akkaya, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 195–203. Association for Computational Linguistics, 2010.

[10] ASID Lang and Joshua Rio-Ross. Using amazon mechanical turk to transcribe historical handwritten documents. *The Code4Lib Journal*, 2011.

[11] Matthew Marge, Satanjeev Banerjee, and Alexander I Rudnicky. Using the amazon mechanical turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273. IEEE, 2010.

[12] Thierry Buecheler, Jan Henrik Sieg, Rudolf M Füchslin, and Rolf Pfeifer. Crowdsourcing, open innovation and collective intelligence in the scientific method-a research agenda and operational framework. In *ALIFE*, pages 679–686, 2010.

[13] Mark Lejk and Michael Wyvill. The effect of the inclusion of selfassessment with peer assessment of contributions to a group project: A quantitative study of secret and agreed assessments. *Assessment & Evaluation in Higher Education*, 26(6):551–561, 2001.

[14] Nathan Van Houdnos. Can the internet grade math? crowdsourcing a complex scoring task and picking the optimal crowd size. *Dietrich College of Humanities and Social Sciences at Research Showcase @ CMU*, 2011.

[15] Joel R Tetreault, Elena Filatova, and Martin Chodorow. Rethinking grammatical error annotation and evaluation with the amazon mechanical turk. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–48. Association for Computational Linguistics, 2010.

[16] Nitin Madnani, Joel Tetreault, Martin Chodorow, and Alla Rozovskaya. They can help: using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 508–513. Association for Computational Linguistics, 2011.

[17] Sherrie A Kossoudji. English language ability and the labor market opportunities of hispanic and east asian immigrant men. *Journal of Labor Economics*, pages 205–228, 1988.

[18] Elizabeth J Erling and Philip Seargeant. *English and development: Policy, pedagogy and globalization*, volume 17. Multilingual Matters, 2013.

[19] Cahit Guven and Asadul Islam. Age at migration, language proficiency and socio-economic outcomes: Evidence from australia. Technical report, 2013.

[20] Kåre Sjölander. An hmm-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik*, volume 2003, pages 93–96. Citeseer, 2003.

[21] Chinmay E Kulkarni, Richard Socher, Michael S Bernstein, and Scott R Klemmer. Scaling short-answer grading by combining peer assessment with algorithmic scoring. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 99–108. ACM, 2014.

[22] Marina Boia, Claudiu Cristian Musat, and Boi Faltings. Acquiring commonsense knowledge for sentiment analysis using human computation. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 225–226. International World Wide Web Conferences Steering Committee, 2014.

[23] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 467–474. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[24] Sivan Sabato and Adam Kalai. Feature multi-selection among subjective features. *arXiv preprint arXiv:1302.4297*, 2013.

[25] Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 33–42. ACM, 2011.

[26] Walter S Lasecki, Christopher D Miller, and Jeffrey P Bigham. Warping time for more effective real-time crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2033–2036. ACM, 2013.

[27] Cambridge EOCL Examinations. Using the *CEFR*: Principles of good practice. *at University of Cambridge*, 2011.

[28] Eric John Dobson. *English Pronunciation, 1500-1700: Phonology*, volume 2. Clarendon Press, 1957.

[29] Christopher Brumfit and Christopher J Brumfit. *Communicative methodology in language teaching: The roles of fluency and accuracy*, volume 129. Cambridge University Press Cambridge, 1984.

[30] Robert Stalnaker. The problem of logical omniscience, ii. context and content: Essays on intentionality in speech and thought (pp. 255–273), 1999.

[31] David Brazil. *A grammar of speech*. Oxford University Press, USA, 1995.

[32] Voxforge. 2014. http://www.voxforge.org/home/dev/autoaudioseg.

[33] Mehdi Hosseini, Ingemar J Cox, Nataša Milić-Frayling, Gabriella Kazai, and Vishwa Vinay. On aggregating labels from multiple crowd workers to infer relevance of documents. In *Advances in information retrieval*, pages 182–194. Springer, 2012.

[34] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book (for htk version 3.4). *Cambridge university engineering department*, 2(2):2–3, 2006.

[35] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403, 1993.

[36] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1457–1460. IEEE, 1996.

[37] Catia Cucchiarini, Helmer Strik, and Lou Boves. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, 107(2):989–999, 2000.

[38] LightSide. 2013. http://lightsidelabs.com/.

[39] AfterTheDeadline. 2014. http://www.afterthedeadline.com/.

[40] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[41] Amazon Mechanical Turk. 2014. https://requester.mturk.com/tour.

[42] Jonathan G Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 347–354. IEEE, 1997.

[43] Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour, and Udo Kruschwitz. Assessing crowdsourcing quality through objective tasks. In *LREC*, pages 1456–1461. Citeseer, 2012.

[44] Quoc Viet Hung Nguyen, Tam Nguyen Thanh, Tran Lam Ngoc, and Karl Aberer. An evaluation of aggregation techniques in crowdsourcing. In *The 14th International Conference on Web Information System Engineering (WISE), 2013*, number EPFL-CONF-187456, 2013.