Automatic Spontaneous Speech Grading: A Novel Feature Derivation Technique using the Crowd

Vinay Shashidhar
Aspiring Minds
vinay.shashidhar@aspiringminds.com

Nishant Pandey
Aspiring Minds
nishant.pandey@aspiringminds.com

Varun Aggarwal
Aspiring Minds
varun@aspiringminds.com

Abstract

In this paper, we address the problem of evaluating spontaneous speech using a combination of machine learning and crowdsourcing. Machine learning techniques inadequately solve the stated problem because 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. To address this, 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. To demonstrate the efficacy of our approach we performed experiments on an expert–graded speech sample of 319 adult non–native speakers. Using these features in a regression model, we are able achieve a Pearson correlation of 0.76 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. This work is timely given the huge requirement of spoken English training and assessment.

1 Introduction

Automatic evaluation of spoken English has been of keen interest for more than two decades (Zechner et al., 2007; Neumeyer et al., 1996; Franco et al., 2000; Cucchiarini et al., 1997). It can help learners get feedback in a scalable manner, help build better English training software and also help companies and institutions filter and select prospective employees more effectively. The problem acquires significance given the evidence that better English leads to better employment outcome, wages and promotions (Guven and Islam, 2013).

There has been a considerable success in automatically scoring spoken English, when the spoken text is known a priori (Cucchiarini et al., 2000; Franco et al., 2000). In these cases, the candidate is asked to either read a given text or listen to some speech and repeat it. For these tasks, the scores generated by an automatic system on parameters such as pronunciation and fluency closely mimic those given by human experts. The primary approach behind a majority of these systems is to force align the speech sample on the known text using an HMM–based acoustic model. Features such as likelihood, posterior probability and fluency related features are derived from the aligned speech and a machine learning model is used to predict expert grades (Neumeyer et al., 1996; Franco et al., 2000; Cucchiarini et al., 1997). Some approaches additionally use prosody and energy related features (Dong et al., 2004). More recently, this research has moved towards the assessment of higher granularity metrics like the mispronunciation of particular phonemes (Li et al., 2009; Ito et al., 2006; Koniaris and Engwall, 2011).

In spontaneous speech evaluation, the candidate is asked to speak on a topic or answer a question and what he/she speaks isn't known priori. Evaluation of spontaneous speech is the ultimate test of a candidate's proficiency in speaking a language (Hagley, 2010; Halleck, 1995). While scores from the evaluation of read/repeat speech do correlate with spontaneous speech evaluation, there remains

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1085–1094,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

an unexplained variance in the spontaneous speech scores (see Section 5). Generally, candidates who score high on spontaneous speech also score high on read speech and not vice versa.

Given the primacy of spontaneous speech evaluation in judging a person's language capability, there is considerable interest in doing it automatically (Cucchiarini et al., 1997; Dong et al., 2004). Automated approaches for the same have not worked well (Powers et al., 2002; Cucchiarini et al., 2000) primarily because speaker-independent speech recognition is a tough computer science problem. This is exacerbated when the speakers are not proficient in the language or are non-natives (Powers et al., 2002). Given that speech to text conversion for such candidates has a low accuracy, force alignment of the speech on this inaccurate text makes the features and the model inaccurate.

We present a semi-automated approach to grade short duration (45 seconds) spontaneous speech. We accurately predict a holistic score which is based on the pronunciation, fluency, content characteristics and grammar of the speech sample, as determined by experts. Multiple previous studies in language acquisition and second language research conclusively show that proficiency in a second language can be characterized by these factors (Bhat et al., 2014). Being able to provide a holistic score is of high interest in both educational testing (Zechner et al., 2009) and job related testing (Streeter et al., 2011). Institutions and firms look for a holistic score, say based on CEFR, a standard to describe spoken English assessment (Little, 2006; Little, 2007), to make an accept or reject decision on candidates. Currently, an expert based assessment is used for these purposes.

Our method involves combining machine learning with a crowdsourcing layer. Crowdsourcing (Estellés-Arolas and González-Ladrón-de Guevara, 2012) is the process of getting *human intelligence* tasks performed by a large community of online workers (crowd) as opposed to traditional employees.[1] The responses from the human intelligence tasks are then used to create relevant features for machine learning. Human intelligence tasks are defined as those which most humans find easy, but are hard for machines. For instance, a classic example is the task of finding a particular object in an image. There is a large research community that uses crowdsourcing and has demonstrated that it can help perform tasks inexpensively, in large volumes and within reasonable time (Howe, 2006; Whitla, 2009).

Our system design for evaluation of spontaneous speech is illustrated in Figure 1. We post the task[2] of speech transcription to the crowd. We get a final accurate transcription by combining the transcriptions from more than one crowd worker for the same speech sample. Once we have this accurate transcription, we force-align (Erling and Seargeant, 2013; Sjölander, 2003) the speech of the candidate on this text to derive various features which go into a machine learning engine. We also collect spoken English grades of the speech from the crowd (Lejk and Wyvill, 2001), which are used as additional features. With these accurately identified features and crowd grades, machine learning is able to grade spontaneous speech with high accuracy. We found that this approach does much better than a pure machine learning approach.

Crowdsourcing has been used for almost a decade in various problems in speech analysis, grading and language learning (Kunath and Weinberger, 2010; Peabody, 2011; Wang et al., 2014). Within assessment of speech, currently all such approaches use the crowd to directly grade certain parts of the speech (Wang and Meng, 2012). Our work is uniquely positioned where we use the crowd to do accurate transcription, a human intelligence task, and use it in a machine learning based algorithm.[3] We show that such a system provides an accuracy rivaling that of experts.

In this paper, we solve a hitherto unsolved problem of spontaneous speech evaluation (Zechner et al., 2009). The paper makes the following contributions:

- We show that spoken English can be graded with accuracy by combining machine learning and crowdsourcing higher than a pure machine learning approach.

[1] Our approach is different from peer grading (Lejk and Wyvill, 2001) or crowd grading (Van Houdnos, 2011; Tetreault et al., 2010; Madnani et al., 2011) 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.

[2] Even though speaker-independent speech recognition is a hard problem for machines, it is fairly easy for a native speaker or anyone with reasonable command over the language.

[3] Again, speech transcription has been done previously using crowdsourcing (Zaidan and Callison-Burch, 2011), but not used for a grading purpose or combined with machine learning.

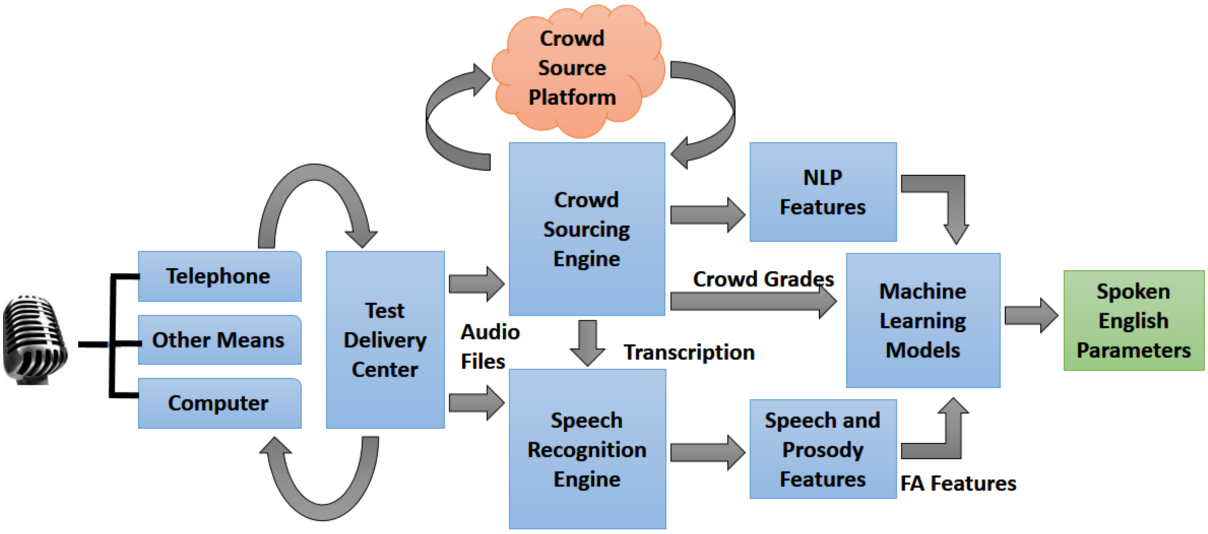

Figure 1: System Design

- We show that the features derived from crowdsourced transcriptions perform as well as crowd grades in predicting expert grades. However, crowd grades add additional predictive value.
- We propose a scalable and accurate way to perform evaluation of spontaneous speech, a huge requirement in the industry and elsewhere.

The paper is organized as follows– Section 2 describes the procedure and aim of the speech assessment task; Section 3 describes the feature classes used in the prediction algorithm; Section 4 describes the crowdsourcing framework which is used as an input to machine learning methods; Section 5 demonstrates how this framework is used with machine learning techniques to predict a composite spoken English score; Section 6 discusses the future work and concludes the paper.

2 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 (SVAR, 2014). 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[4] are provided with a topic and given 30 seconds[5] 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 (SVAR, 2014). Our goal in this paper is to evaluate the spontaneous speech of the candidate and provide a composite score based on it.

A 5 point rubric for the composite score, similar to CEFR (Examinations, 2011), 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 (Dobson, 1957) refers to the correctness in the utterance of the phonemes of a word by the students as per neutral accent. Fluency (Brumfit and Brumfit, 1984) refers to a desired rate of speech along with the absence of hesitations, false starts and stops etc. Content organization (Stalnaker, 1999) measures the candidate's ability to structure the information disposition and present it coherently. Grammar (Brazil, 1995) measures how well the syntax of the language was followed by the candidate.

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

[5] This is as per global standards of spoken English assessment. High stake tests such as TOEFL provide the candidate 15-30 seconds to think before responding to a spontaneous speech task.

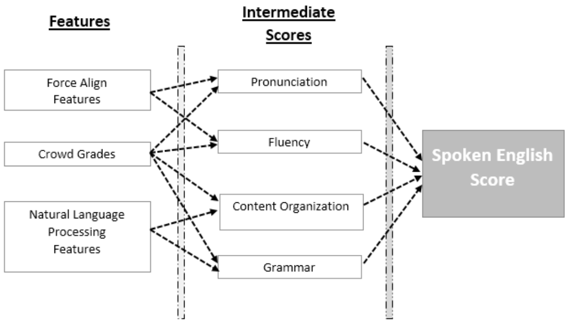

Figure 2: Our intuition of how different features predict the holistic score.

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

3 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 (Erling and Seargeant, 2013; Sjölander, 2003) 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.[6]

- *FA features:* The speech sample is forced aligned (Erling and Seargeant, 2013; Sjölander, 2003) on the crowdsourced transcription using the HTK speech recognizer (Young et al., 2006). We used an acoustic model based on TIMIT (Garofolo et al., 1993) 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. These features are well known in literature and may be referred from (Neumeyer et al., 1996; Zechner et al., 2009; Cucchiarini et al., 2000). 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[7]. They were derived using standard NLP packages (LightSide, 2013; AfterTheDeadline, 2014) 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. In the current system, we do not use any prompt specific features such as occurrence of specific words or phrases. 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.

4 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) (Paolacci et al.,

[6] Advanced Expectation-Maximization techniques (Hosseini et al., 2012) 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.

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

2010). AMT is a popular crowdsourcing marketplace. It is inspired by the famous 18th 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 (Turk, 2014). 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 (Fiscus, 1997). 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 (Aker et al., 2012). One of the key techniques for this involves inserting gold standard tasks with known answers to get an estimate of the worker's ability (Nguyen et al., 2013). 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.[8]

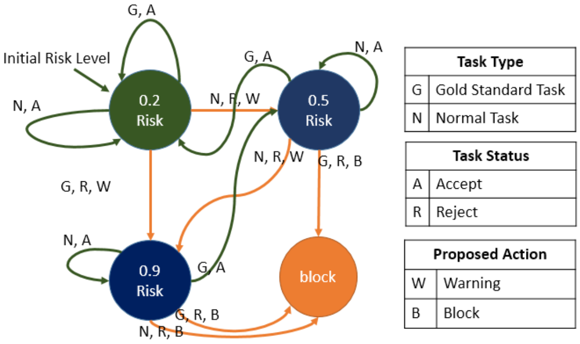

Figure 3: Risk Level State Diagram: In the above figure, each node corresponds to a risk level associated with a worker. The values range between 0 (min) - 1 (max). The worker is either assigned a gold standard task (G) or a normal task (N) on the basis of his/her present risk level. The risk level changes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

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

5 Experiments

We conducted the experiments to answer the following questions:

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

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

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 319 spontaneous speech samples which were graded by expert assessors. To answer the questions stated above, we used different sets of features to develop 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 with different features selection algorithms. The data set used in the experiments is discussed in the next section.

5.1 Data Set

Our data set contains 319 spontaneous speech responses. The speech samples were from seniors (non–native English speakers in final year of undergraduate education) pursuing bachelor's degree in India. The candidates were asked to describe one of the following scenes: *a hospital, flood, a crowded market* 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 (SVAR, 2014). 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 319 sample set.

These responses were graded by two experts who had more than fifteen years of experience in grading spoken English responses. 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 was 0.86 and 0.83 respectively for the two cases. 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 read/repeat speech was 0.54. This shows that there is a considerable unexplained variance (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.

5.2 Crowdsourced Tasks

The 319 speech sample assessment task was posted on Amazon Mechanical Turk (AMT). Each task was completed by three workers. In total, 71 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. We also got the speech transcribed by experts to find the accuracy we could get from turks. The average transcription accuracy for a worker was 82.4%[9]. This significantly improved to 95.4% when the transcriptions of the three workers were combined using the ROVER algorithm. In comparison, the average automatic transcription of a speech recognition engine was 59.8%.

5.3 Regression Modeling

The data set was split into two sets: train and validation. The train-set had 75% of the sample points whereas the validation set had 25%. The split was done randomly making sure that the grade distribution in both the sets was similar. While learning the model, a 4-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.

Regression parameters: For linear regression with regularization, optimal ridge coefficient λ,

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

Table 1: Regression Results

Technique	Model Code	Feature Type	Train r	Validation r
Ridge Regression	RR-1	RS/LR	0.51	0.47
	RR-2	Pure ML	0.54	0.47
	RR-3	Crowd Grades	0.63	0.57
	RR-4	ML-CS	0.55	0.60
	RR-5	All	0.76	0.76
SVM	SVM-1	RS/LR	0.50	0.46
	SVM-2	Pure ML	0.53	0.46
	SVM-3	Crowd Grades	0.62	0.57
	SVM-4	ML-CS	0.60	0.61
	SVM-5	All	0.75	0.74
Neural Networks	NN-1	RS/LR	0.56	0.51
	NN-2	Pure ML	0.60	0.44
	NN-3	Crowd Grades	0.63	0.57
	NN-4	ML-CS	0.66	0.57
	NN-5	All	0.80	0.76

between 1 and 1000, was selected based on the 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 out on five sets of features:

- RS/LR: A set of features generated by force aligning read/repeated by candidates.
- Pure ML: Features generated by automatic speech transcription of spontaneous speech using a speech recognizer.
- Crowd Grades: A set of features pertaining to grades given by the crowd.
- ML–CS: NLP and FA features generated by force aligning free speech on crowdsourced transcription.
- All: NLP and FA features from crowdsourced transcription and Crowd Grades.

Here, the first set, RS/LR, helps us to know how well we can predict spontaneous speech grades by simply using the read/speak speech of the candidate and without using his/her spontaneous speech sample. This provides a comparison baseline. The second approach evaluates how well we can grade spontaneous speech of the candidate using machine learning approaches only. The third feature shows the efficacy of directly using grades given by crowd, while the fourth finds how well machine learning can do if it has a fairly accurate transcription of the speech by the crowd. The final fifth set tests what happens if we combine the third and fourth set of features, i.e. make use of both the crowdsourced transcription and the 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.

5.4 Observations

The results of the experiments are tabulated in Table 1. We report the Pearson coefficient of correlation (r) for the different models against the expert grades. 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 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 ($r = 0.47$). 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 RR-2) is also inadequate to grade spontaneous speech and does worse than approaches that uses features from crowdsourced transcription (Model 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 RR-4). However, when we combine all the features from crowdsourcing including the crowd grades, we find much better prediction accuracy ($r = 0.76$). 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 RR-5 is 0.76. We find that the expert agreement on the validation sample is 0.78. 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 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. 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 (Bernstein et al., 2011; Lasecki et al., 2013) and is an area for future work for us as well.

References

AfterTheDeadline. 2014. www.afterthedeadline.com.

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

Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. 2011. 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.

Suma Bhat, Huichao Xue, and Su-Youn Yoon. 2014. Shallow analysis based assessment of syntactic complexity for automated speech scoring. In *Proceedings of the 52nd Annual Meeting of the Association* *for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1315. Association for Computational Linguistics.

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

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

Catia Cucchiarini, Helmer Strik, and Lou Boves. 1997. Automatic evaluation of dutch pronunciation by using speech recognition technology. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 622–629. IEEE.

Catia Cucchiarini, Helmer Strik, and Lou Boves. 2000. 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.

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

Bin Dong, Qingwei Zhao, Jianping Zhang, and Yonghong Yan. 2004. Automatic assessment of pronunciation quality. In *Chinese Spoken Language Processing, 2004 International Symposium on*, pages 137–140. IEEE.

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

Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200.

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

Jonathan G Fiscus. 1997. 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.

Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, John Butzberger, Romain Rossier, and Federico Cesari. 2000. The sri eduspeaktm system: Recognition and pronunciation scoring for language learning. *Proceedings of InSTILL 2000*, pages 123–128.

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

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

Eric Hagley. 2010. Creation of speaking tests for efl communication classes. *ł*, (8):33–41.

Gene B Halleck. 1995. Assessing oral proficiency: A comparison of holistic and objective measures. *The Modern Language Journal*, 79(2):223–234.

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

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Akinori Ito, Tadao Nagasawa, Hirokazu Ogasawara, Motoyuki Suzuki, and Shozo Makino. 2006. Automatic detection of english mispronunciation using speaker adaptation and automatic assessment of english intonation and rhythm. *Educational technology research*, 29(1):13–23.

Christos Koniaris and Olov Engwall. 2011. Perceptual differentiation modeling explains phoneme mispronunciation by non-native speakers. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5704–5707. IEEE.

Stephen A Kunath and Steven H Weinberger. 2010. The wisdom of the crowd's ear: speech accent rating and annotation with amazon mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 168–171. Association for Computational Linguistics.

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

Mark Lejk and Michael Wyvill. 2001. 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.

Hongyan Li, Shijin Wang, Jiaen Liang, Shen Huang, and Bo Xu. 2009. High performance automatic mispronunciation detection method based on neural network and trap features. In *INTERSPEECH*, pages 1911–1914.

LightSide. 2013. http://lightsidelabs.com/.

David Little. 2006. The common european framework of reference for languages: Content, purpose, origin, reception and impact. *Language Teaching*, 39:167–190, 7.

David Little. 2007. The common european framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4):645–655.

Nitin Madnani, Joel Tetreault, Martin Chodorow, and Alla Rozovskaya. 2011. 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.

Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. 1996. 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.

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

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

Mitchell Aaron Peabody. 2011. *Methods for pronunciation assessment in computer aided language learning*. Ph.D. thesis, Massachusetts Institute of Technology.

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

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

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

Lynn Streeter, Jared Bernstein, Peter Foltz, and Donald DeLand. 2011. Pearsons automated scoring of writing, speaking, and mathematics.

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

Joel R Tetreault, Elena Filatova, and Martin Chodorow. 2010. 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.

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

Nathan Van Houdnos. 2011. 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*.

Hao Wang and Helen Meng. 2012. Deriving perceptual gradation of l2 english mispronunciations using crowdsourcing and the workerrank algorithm. *Proc. of the 15th Oriental COCOSDA, Macau, China*, pages 9–12.

Hao Wang, Xiaojun Qian, and Helen Meng. 2014. Phonological modeling of mispronunciation gradations in l2 english speech of l1 chinese learners. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 7714–7718. IEEE.

Paul Whitla. 2009. Crowdsourcing and its application in marketing activities. *Contemporary Management Research*, 5(1).

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

Omar F Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1220–1229. Association for Computational Linguistics.

Klaus Zechner, Derrick Higgins, and Xiaoming Xi. 2007. Speechrater: A construct-driven approach to scoring spontaneous non-native speech. *Proc. SLaTE*.

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