Automatically Grading Learners' English Using a Gaussian Process

Rogier C. van Dalen, Kate M. Knill, Mark J. F. Gales

ALTA Institute / Department of Engineering, University of Cambridge, United Kingdom

{rcv25,kmk1001,mjfg}@eng.cam.ac.uk

Abstract

There is a high demand around the world for the learning of English as a second language. Correspondingly, there is a need to assess the proficiency level of learners both during their studies and for formal qualifications. A number of automatic methods have been proposed to help meet this demand with varying degrees of success. This paper considers the automatic assessment of spoken English proficiency, which is still a challenging problem. In this scenario, the grader should be able to accurately assess the learner's ability level from spontaneous, prompted, speech, independent of L1 language and the quality of the audio recording. Automatic graders are potentially more consistent than humans. However, the validity of the predicted grade varies. This paper proposes an automatic grader based on a Gaussian process. The advantage of using a Gaussian process is that as well as predicting a grade, it provides a measure of the uncertainty of its prediction. The uncertainty measure is sufficiently accurate to decide which automatic grades should be re-graded by humans. It can also be used to determine which candidates are hard to grade for humans and therefore need expert grading. Performance of the automatic grader is shown to be close to human graders on real candidate entries. Interpolation of human and GP grades further boosts performance.

Index Terms: spoken language assessment, Bayesian methods, Gaussian process

1. Introduction

English is the modern-day *lingua franca*, and many non-native speakers around the world are learning it. Currently, language tests are often graded by human graders, for example, the tests from Cambridge English, one of the largest providers of assessment of spoken English. To meet demand from learners the introduction of automatic approaches to testing would be beneficial, especially for practice situations. This could be fully automatic or combined with a human grader to boost the reliability.

Assessing spoken English is a challenging problem for automatic systems. In addition to the issues seen in English text based assessments, such as grammatical errors, depending on the proficiency level of the learner, the speech will contain the accent of the L1 language and pronunciations may be incorrect, affected by the L1. To get a proper indication of ability the speech should be spontaneous, and not simply be readings of a known text. This introduces further challenges since spontaneous speech typically contains disfluencies such as hesitations and false starts. Also whilst the question text is known e.g. "describe what is happening in the picture", the vocabulary used in the answer is likely to be unknown unless significant recordings are made of typical answers as in [1]. Finally, there is likely to be a large variation in the quality of the audio recordings in terms of levels of background noise and volume levels. Despite these issues, a number of methods have been proposed to assess different aspects of a learner's spoken language abilities [2, 3, 4, 5, 1].

Automated graders are potentially more consistent than human graders. However, the validity of the grade may suffer: not all aspects of learners' speech can be captured by current automated grading systems. As long as a candidate is similar enough to speakers in the training data, the quality of the automatic grading may be sufficient. For speakers that are unlike those seen by the automated system, however, the grade predicted can be poor. In those cases, ideally the system would know to back off to human graders. Previous work in this area [5] used a filter, essentially a separate classifier, to decide whether or not a recording is gradeable. This paper introduces a method - based on a Gaussian process - of computing a grade and a measure of the uncertainty at the same time and from the same data.

Gaussian processes [6] give a mathematically consistent method for approximating an unknown function that also provides a measure of the uncertainty around this estimate. In this case, the function maps a feature vector representing a candidate's spoken English to a grade. By relating a new candidate to the training data, a distribution over the result of the function for the new candidate can be produced. The variance of this distribution will be used for rejecting the grades given. Combination of this automatic method with human grades is also considered.

This paper is organised as follows. Section 2 will introduce Gaussian processes. Section 3 will describe the automated grader; section 4 will then present experimental results.

2. Gaussian processes

A Gaussian process (see e.g. [6]; for applications to speech processing, see [7, 8]) is a model that can be used to perform regression. One way of viewing it is as modelling a distribution over functions. The Gaussian process is a nonparametric model, which means that the functions themselves are not parameterised. However, the covariance between any two inputs x and x' is given by a function $k(x, x')$. All training data points are stored (though sparsification methods exist). When a prediction is required for a new test point (a new candidate, in this case), the covariance between it and each training point is computed. The prediction, in the form of a Gaussian, can be computed from that.

Figure 1 illustrates a Gaussian process trained on five data points (the dots). The horizontal axis represents the input (1-dimensional for illustration), and the vertical axis represents the

Thanks to Cambridge English, University of Cambridge for supporting this research and providing access to the data. Thanks to Nahal Khabbazbashi for useful comments on an earlier draft. Thanks to Mohammad Rashid for information about the neural network grader.

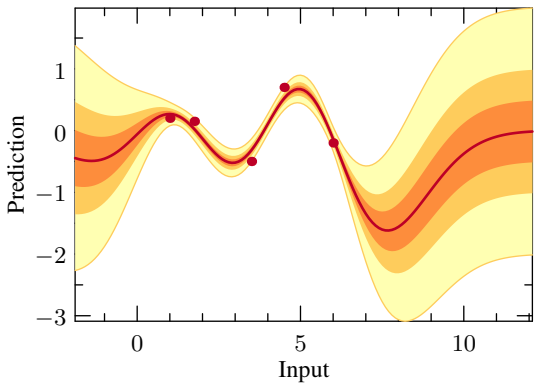

Figure 1: *A Gaussian process trained on a few data points. The mean and variance contours are indicated. When the test point is further away from the training data, the predicted mean and variance revert to the prior.*

target values. The bands show the predicted Gaussian distribution for any input point. The middle line indicates the mean, and the coloured bands the variance contours at ½, 1, and 2 times the variance around the means. Close to the data points, the predictions have low variance, and the mean interpolates, and to some degree extrapolates, between the points. The data is assumed to be observed with noise, so the mean does not quite go through the training points.

The key aspect for this work is that when the prediction is requested for points further away from the training data points, the predicted distribution increases in variance. For these points the distribution reverts to the prior probability. This corresponds to the intuition that when no training data points are in the vicinity of the test point, there is little to base a prediction on, and the uncertainty is great.

2.1. Mathematical description

In more detail, the Gaussian process works as follows. Functions are viewed as a mapping from an infinite number of inputs to corresponding output values. They are assumed to be Gaussian-distributed, that is, the joint distribution of the infinite number of output values is Gaussian. It is impossible to deal with an infinite-dimensional vector, so a property of Gaussians must be exploited.

This property is that if variables (here $\mathbf{y}$ and $\mathbf{y}'$) are jointly Gaussian (here with zero mean),

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}' \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\mathsf{T} & \mathbf{B} \end{bmatrix} \right), \tag{1a}$$

then any subset of the variables with the other variables marginalised out are also Gaussian distributed, and the parameters are trivially found, in this case:

$$\left[\mathbf{y}' \right] \sim \mathcal{N}(\mathbf{0}, \mathbf{B}). \tag{1b}$$

Thus, even if the distribution of functions is theoretically characterised by an infinite number of values, it is possible to consider only a finite number of them by marginalising the rest out. Once this joint Gaussian has been set up, the conditional distribution of the test data point given the training data becomes necessary. If $\mathbf{y}$ and $\mathbf{y}'$ are again jointly Gaussian distributed as in (1a), then the conditional distribution of $\mathbf{y}'$ given $\mathbf{y}$ is also Gaussian (see e.g. [9]):

$$\mathbf{y}' \mid \mathbf{y} \sim \mathcal{N}\big(\mathbf{C}^\mathsf{T}\mathbf{A}^{-1}\mathbf{y}, \mathbf{B} - \mathbf{C}^\mathsf{T}\mathbf{A}^{-1}\mathbf{C}\big). \tag{1c}$$

For a Gaussian process, the values of interest are normally the training data points and the test data points. The training data consists of input points $\mathbf{x} = [x_1 \dots x_N]^\mathsf{T}$ and corresponding outputs $\mathbf{y} = [y_1 \dots y_N]^\mathsf{T}$. The observed outputs are assumed to be Gaussian-distributed around the real function values $f(x)$: The real function values $f(x_n)$ are observed as y_n, with additive observation noise $\mathcal{N}\big(0, \sigma_o^2\big)$:

$$y_n \sim \mathcal{N}\big(f(x_n), \sigma_o^2\big) \tag{2}$$

Assume that the value of the function f is to be predicted at test point x_*. The joint distribution of the observed outputs $\mathbf{y}$ and the output $f(x_*)$ to be predicted is

$$\begin{bmatrix} \mathbf{y} \\ f(x_*) \end{bmatrix} \triangleq \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{x},\mathbf{x}) + \sigma_o^2\mathbf{I} & \mathbf{k}(x_*,\mathbf{x}) \\ \mathbf{k}(x_*,\mathbf{x})^\mathsf{T} & k(x_*,x_*) \end{bmatrix}\right), \tag{3a}$$

where $\mathbf{I}$ is the identity matrix, and the functions $\mathbf{k}(x_*,\mathbf{x})$ and $\mathbf{K}(\mathbf{x},\mathbf{x})$ are convenience functions that apply the covariance function $k(\cdot,\cdot)$ to each combination of elements of their arguments:

$$\mathbf{k}(x_*,\mathbf{x}) \triangleq \begin{bmatrix} k(x_*,x_1) \\ \vdots \\ k(x_*,x_N) \end{bmatrix}; \tag{3b}$$

$$\mathbf{K}(\mathbf{x},\mathbf{x}) \triangleq \begin{bmatrix} k(x_1,x_1) & \dots & k(x_N,x_1) \\ \vdots & \ddots & \vdots \\ k(x_1,x_N) & \dots & k(x_N,x_N) \end{bmatrix}. \tag{3c}$$

Having set up the model, the posterior distribution (i.e. after seeing the training data) over the output value of $f(x_*)$ involves the training outputs $\mathbf{y}$, and all the covariances. Substituting (3a) into (1c),

$$\begin{aligned} f(x_*) \mid \mathbf{y} \sim \mathcal{N}\Big(&\mathbf{k}(x_*,\mathbf{x})^\mathsf{T}(\mathbf{K}(\mathbf{x},\mathbf{x}) + \sigma_o^2\mathbf{I})^{-1}\mathbf{y}, \\ &k(x_*,x_*) - \mathbf{k}(x_*,\mathbf{x})^\mathsf{T}(\mathbf{K}(\mathbf{x},\mathbf{x}) + \sigma_o^2\mathbf{I})^{-1}\mathbf{k}(x_*,\mathbf{x})\Big). \end{aligned} \tag{4}$$

Thus, the prediction for the grade is a Gaussian with parameters derived from the training outputs $\mathbf{y}$ and the covariance between the training input $\mathbf{x}$ and the new input x_*.

The next section will discuss the form of the covariance function $k(x, x')$.

2.2. Covariance function

The input, whether training or test data, only enters the Gaussian process model, in (3), as arguments to the covariance function $k(x, x')$ (this is also often true for support vector machines). This allows something known as the "kernel trick". This allows the input data point to be any type of object, as long as they can be used with a suitable kernel $k(\cdot,\cdot)$. In this work, the input data will consist of feature vectors extracted from the audio (see section 3).

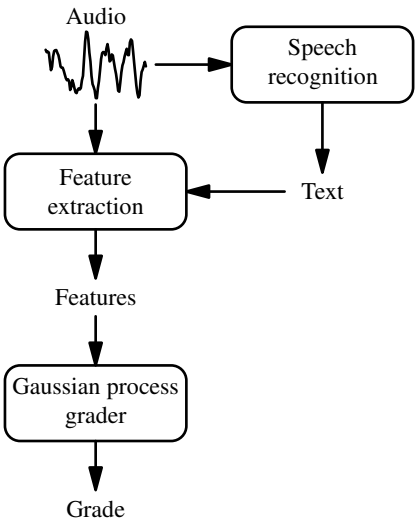

Figure 2: *The grader system, schematically.*

An often-used covariance function, used in this work, is the *radial basis function*. Defined on vectors $\mathbf{x}$ and $\mathbf{x}'$:

$$k(\mathbf{x}, \mathbf{x}') \triangleq \sigma_y^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right), \quad (5)$$

where ℓ is the length scale, which governs how much the covariance tails off as the points are further away from each other, and σ_y^2 is the output variance, which affects the variance of the output.

Though this work does not exploit this, it is also possible for Gaussian processes to take other types of objects than scalars or vectors, if a suitable covariance function is defined.

3. Grader

The automatic grader used in this work has a simple architecture, illustrated in figure 2. The input to the grader is a set of audio and fluency features. Audio features are extracted directly from the audio signal. The fluency features are extracted from a time-aligned hypothesis produced by a speech recogniser. A Gaussian process is trained to map these features to grades, and then used to predict a distribution over the grade, in the form of a mean and a variance.

Table 1 lists the features that the system uses. The features are similar to those used in other systems [2, 3, 10, 5]. Audio features are extracted without reference to the hypothesised transcriptions.

4. Experiments

The grader is tested on data from the BULATS (Business Language Testing Service) corpus of learners' speech made available by Cambridge English, the on-line version of which is described in [11]. The BULATS test has five sections, all with material appropriate to business scenarios. The first section contains questions about the candidate and their work (e.g. "How do you use English in your job?"). The second section is a read-aloud section. The last three sections have longer utterances of spontaneous speech elicited by prompts. In the third section the prompts are generic questions about business scenarios. In the fourth section, the candidate is asked to describe a visual such as a pie chart or bar chart. The prompt for the last section asks the candidate to imagine they are in a specific conversation and to respond to questions that may be asked in that situation (e.g. advice about planning a conference).

Table 1: *Grader input features, extracted from the audio.*

Item	**Statistics**
	Audio features
Fundamental frequency	mean mean-weighted: minimum, maximum, extent, mean absolute deviation
Energy	mean, standard deviation mean-weighted: minimum, maximum, extent, mean absolute deviation
	Fluency features
Long silence	number
Long silence duration	mean, standard deviation, median, mean absolute deviation
Silence duration	mean, standard deviation, median, mean absolute deviation
Disfluencies	number
Words	number, number per second, mean duration
Phones	mean, standard deviation, median, mean absolute deviation

Each section is graded between 0 and 6; the overall grade is therefore between 0 and 30. These can be binned into CEFR (Common European Framework of Reference) ability levels [12] A1, A2, B1, B2, C1, and C2. In this work, the audio from all sections will be used to predict the overall grade. The Pearson correlation with grades assigned by expert human graders will be used to measure performance. This correlation implicitly normalises the dynamic range.

A state-of-the-art tandem GMM-HMM recogniser is trained on data collected from BULATS candidates using HTK [13]. For this paper, the recogniser is trained on 58.5 hours of data from candidates with Gujarati as their first language. This choice reflects the initial deployment focus of BULATS which was in Gujarat. Transcriptions are obtained through crowd-sourcing. Two crowd-sourced transcriptions are combined using the algorithm in [14]. The input features for the tandem models are 26-dimensional bottleneck features and 52-dimensional PLP+Δ+Δ^2+Δ^3 features. Cepstral mean and variance normalisation are applied. A heteroscedastic linear discriminant analysis (HLDA) transform is applied to the PLP features and global semi-tied transform to the bottleneck features [15], reducing the dimensionality to 65. The bottleneck features are extracted from a 5-hidden layer neural network trained using QuickNet [16] on the AMI meeting corpus [17] with 1000 units per hidden layer and 6000 context-dependent output layer targets. The AMI database was selected for the DNN training instead of the BULATS data for robustness. It is a larger corpus of (mostly) non-native speakers of English and closely manually transcribed. Discriminately trained speaker independent and speaker adaptively trained (SAT) models are estimated using the minimum phone error (mpe) criterion [18]. Speaker-adaptive training is performed using constrained maximum likelihood linear regression (CMLLR) [19] followed by MPE. Each model set has approximately 4000

context-dependent states, with an average of 16 Gaussians per state. At decoding time, the speaker-independent model with a trigram language model is used to produce hypotheses for the estimation of CMLLR transforms for the tandem SAT models. The speech hypotheses for the fluency features are derived from decoding with these SAT models and a trigram language model. On a separate speech recogniser evaluation set, the recogniser achieves a 37.6 % word error rate, underscoring that this is a real-world, noisy data set with non-native speakers.

The grader is then trained and tested using the system in figure 2: the speech recogniser is run and audio and fluency features are extracted. These features are used as the input, with associated grades as the targets, for the Gaussian process (GP). The noise variance σ_o^2 is set to 0.2, and the hyperparameters of the covariance function (which is the radial basis function, as in (5)) are trained with maximum-likelihood estimation. In initial tests, a neural network grader, like the system used in [1], was also trained, but yielded a lower raw performance than the Gaussian process grader. A separate training set is used to train the grader. The training data consists of 994 candidates distributed evenly over six languages (Polish, Vietnamese, Arabic, Dutch, French, Thai) and over CEFR ability levels A1, A2, B1, B2, and C (which combines C1 and C2 because of data scarcity). The grades provided by the original local graders are used as the GP targets. The evaluation set has 226 candidates, distributed similarly to the training set, but in addition to the original grades, candidates were re-graded by expert graders. On another data set, this group of expert graders had an inter-grader Pearson correlation around 0.96, so in this work their grades are used as the ground truth.

The availability of expert grades makes it possible to assess the performance of the original human graders as well as schemes that combine human and automated grades. In the following section, 4.1, an approach to interpolate human and automated grades is presented. Section 4.2 will then discuss a rejection scheme that automatically detects when expert grades should be used.

4.1. Interpolation

One method of using the grades produced by the automatic grader is to treat them as just another grader and to interpolate between grades of both graders. The assumption is that the automated grader is more consistent, but less sophisticated than the human graders. Combining the grades may exploit the strengths of both.

Figure 3 shows the performance as the interpolation weight changes. At the left-hand side of the graph, only the standard human grades are used; at the right-hand side, only the automated grades. In between, each grade is interpolated with the given weight. The optimal interpolation weight is 0.44. That the human graders receive a higher weight is not surprising, since their performance by themselves is better.

The interpolated grades will be used in the next section. In a completely realistic scenario, it would be best to have a representative development set to estimate the interpolation weight, and an entirely separate evaluation set. However, for the current data this is not available.

4.2. Rejection

This section will consider a situation where a number of candidates are re-graded by expert graders. If it is possible to detect lower-quality grades automatically, then it can save time (and money) and/or improve the overall quality of the grades.

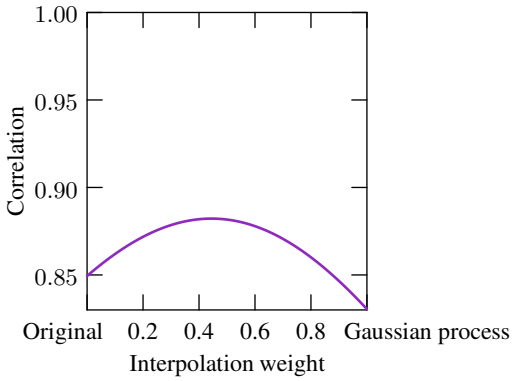

Figure 3: *Effect on Pearson correlation of interpolation between human (original) and automated (GP) grades.*

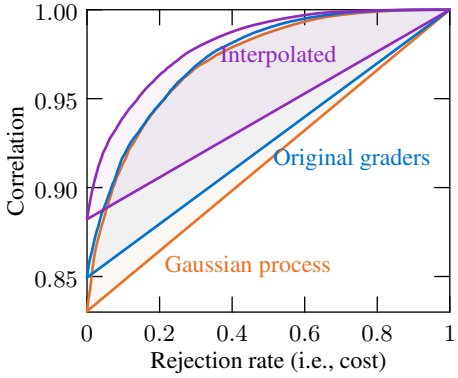

Figure 4: *Envelope of performance of rejection schemes.*

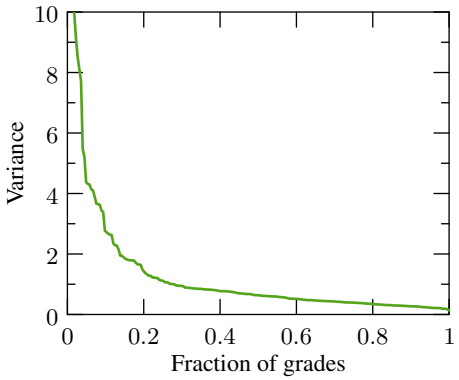

Figure 5: *Variances of the grades that the Gaussian process predicts, in descending order.*

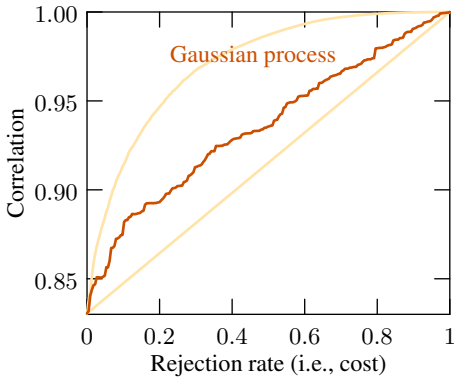

Figure 6: *Rejection of automated grades by Gaussian process variance. By rejecting a small number of grades with highest variance for re-grading by experts, performance increases far more than the expected improvement from random rejection.*

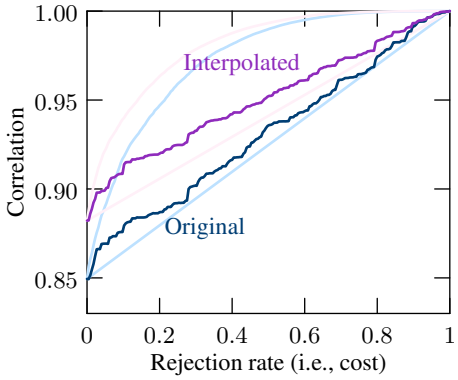

Figure 7: *Rejection of grades by Gaussian process variance. The grades are the original human grades, and the interpolated grades. That the Gaussian process variance is informative indicates that it identifies candidates that are hard to grade.*

Figure 4 shows baseline results for the original human graders and the automated grader. On the vertical axis is the Pearson correlation with the expert grades. On the left side of the graph, all candidates are graded by the original or automated grader, with a 0.85 and 0.83 correlation, respectively, to the expert graders. On the horizontal axis is the fraction of candidates whose original grades are rejected and replaced by the expert grades. By definition, the correlation at the right-hand side of the graph is 1: all grades have then been replaced by expert grades, and the Pearson correlation of these with themselves is 1. In between, the performance depends on the rejection scheme. Figure 4 shows the envelope of performance of any useful rejection scheme. The straight lines indicate the expected performance if candidates are chosen for re-grading randomly. The curves at the top indicate the upper bounds: grades that deviate most from the expert grades are replaced first. This is not a practical scheme, since it requires knowledge of the expert grades, but it indicates the best performance any rejection scheme can reach in theory.

A simple scheme for rejecting grades is to use the discrepancy between the original human grades and the automated grades as a measure of uncertainty. The grades for which the discrepancy (after normalising the mean and variance of both sets of grades) is greatest are rejected first. However, this yields no clear performance gain over using the human grades on the evaluation set used here, which will need to have been collected anyway. As an aside, on another evaluation set, with human grades that were suspected to be less reliable, this scheme did result in improvements. This suggests that this scheme may be a good way of finding outlier grades that deserve further investigation.

The rejection scheme that this paper proposes is to use the uncertainty measure that the grader itself provides. As discussed in section 2, a prediction from a Gaussian process is a distribution over the result of a function i.e. the prediction is a Gaussian distribution with a mean and a variance. The mean is used as the predicted grade; the variance is used to indicate confidence in the grade. Figure 5 shows the variances that the grader returns, sorted in descending order. This is the order in which the automatically predicted grades will be rejected and replaced by expert grades. Figure 6 shows performance as grades are rejected starting with the ones where the Gaussian process returns the highest variance, i.e. where the prediction is least certain. This produces a sizeable increase in performance: the variance turns out to be a good indicator of reliability of Gaussian process grades. From a practical perspective, a trade-off between cost and quality is a desideratum. By rejecting 10 % of the grades and having them re-graded by experts, overall performance can be improved from 0.83 to 0.88.

An interesting question is whether the grades that are rejected based on the Gaussian process variance are for candidates where the problem is the automated grader, or where the candidate was hard to grade. In the system in [5], the latter was what the separate filter aimed to detect. For insight into this, it is possible to take the original human grades and replace them with expert grades in descending order of the Gaussian process variance. In other words, the Gaussian process is used merely as a predictor of how hard the candidate is to grade. The blue squiggly line in figure 7 shows the result of this experiment. Interestingly, some of the gains that were made for the Gaussian process are mirrored here. The very first part of the curve even mimics the first part of the curve for the Gaussian process. This implies that in general the candidates with the greatest variance are those candidates which are hard to grade, rather than being purely those that the automatic grader has difficulty with.

The good results in rejecting and re-grading both automated and human grades based on the Gaussian process variances suggest that it should be possible to apply the same strategy to the

interpolated grades from section 4.1. Figure 7 also shows the curves for that experiment. The curve starts at the performance of the interpolated grades, at 0.88. Rejecting grades that the Gaussian process was less certain about again increases performance more than random rejection. By rejecting 10 % of grades and having them re-graded by experts, the Pearson correlation in this case can be improved from 0.88 to 0.91.

This means that two strategies to trade off cost and quality have been identified. In both cases a small fraction of the grades is classified as low-confidence by the automated grader, and re-graded by expert graders. In the first strategy, the remaining grades are produced by the automated grader itself, with performance improving from 0.83 to 0.88. In the second strategy, the standard human grades were at 0.85, an interpolation of them and automated grades at 0.88, and automated rejection increases this to 0.91.

5. Conclusion

Automatic assessment of spoken English proficiency of second language learners would be beneficial in helping to meet demand for testing, both for practice and in formal examinations. An automatic approach should be more consistent than human graders. However, there are a lot of challenges in assessing non-native spoken English and automatic approaches are less reliable than humans when a candidate's speech is not a good match to the data seen in training.

This paper has proposed a Gaussian process (GP) based automatic grader. The grades predicted by the GP grader are close to those of human graders, measured by Pearson correlation with expert graders. Its primary advantage though is that it gives a mathematically consistent framework for estimating not only grades, but also the uncertainty around them. The variance, the measure of uncertainty, is sufficiently accurate that it can be used to target candidates for which the automatic process has problems and so should be re-graded by humans. In addition, this measure is seen to be related to how hard a speaker is to grade for human graders. It can therefore be used to decide which candidates need to be assessed by expert graders. Interpolating between the automatic and human produced grades further boosts the overall grading performance.

The current automatic GP grader does not contain any features relating to content. This means candidates could potentially game this system if run in a fully automatic mode. Learners could also benefit from being provided with feedback as to why they were awarded a particular grade. Both of these issues will be investigated in future work.

6. References

[1] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children English language learners," in *Proceedings of Interspeech*, 2014.

[2] C. Cucchiarini, H. Strik, and L. Boves, "Automatic evaluation of Dutch pronunciation by using speech recognition technology," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, 1997, pp. 622–629.

[3] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The SRI Eduspeak system: Recognition and pronunciation scoring for language learning," in *Proceedings of InSTILL 2000*, 2000, pp. 123–128.

[4] J. Bernstein and J. Cheng, "Logic and validation of fully automatic spoken English test," *The path of speech technologies in computer assisted language learning: From research toward practice*, pp. 174–194, 2007.

[5] D. Higgins, X. Xi, K. Zechner, and D. Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech and Language*, vol. 25, no. 2, pp. 282–306, 2011.

[6] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press, 2006.

[7] H. Park and S. Yun, "Phoneme classification using constrained variational gaussian process dynamical system," in *Proceedings of the Conference on Neural Information Processing Systems*, 2011.

[8] G. E. Henter, M. R. Frean, and W. B. Kleijn, "Gaussian process dynamical models for nonparametric speech representation and synthesis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[9] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Nov 2008. [Online]. Available: http://matrixcookbook.com/

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

[11] L. Chambers and K. Ingham, "The BULATS online speaking test," *Research Notes*, vol. 43, pp. 21–25, 2011. [Online]. Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf

[12] *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4)," 2006. [Online]. Available: http://htk.eng.cam.ac.uk/docs/docs.shtml

[14] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Apr 2015.

[15] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE *Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[16] D. Johnson *et al.*, "Quicknet." [Online]. Available: http://www1.icsi.berkeley.edu/Speech/qn.html

[17] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction*. Springer, 2006, pp. 28–39.

[18] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[19] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.