View-Specific Assessment of L2 Spoken English

S. Bannò[1,2,3], B. Balusu[3], M.J.F. Gales[3], K.M. Knill[3], K. Kyriakopolous[3]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Trento, Trento, Italy
[3]Cambridge University Engineering Department, Cambridge, UK.
sbanno@fbk.eu, {bb562,mjfg100,kmk1001,kk492}@cam.ac.uk

Abstract

The growing demand for learning English as a second language has increased interest in automatic approaches for assessing and improving spoken language proficiency. A significant challenge in this field is to provide interpretable scores and informative feedback to learners through individual viewpoints of learners' proficiency, as opposed to holistic scores. Thus far, holistic scoring remains commonly applied in large-scale commercial tests. As a result, an issue with more detailed evaluation is that human graders are generally trained to provide holistic scores. This paper investigates whether view-specific systems can be trained when only holistic scores are available. To enable this process, view-specific networks are defined where both their inputs and structure are adapted to focus on specific facets of proficiency. It is shown that it is possible to train such systems on holistic scores, such that they provide view-specific scores at evaluation time. View-specific networks are designed in this way for pronunciation, rhythm, text, use of parts of speech and grammatical accuracy. The relationships between the predictions of each system are investigated on the spoken part of the Linguaskill proficiency test. It is shown that the view-specific predictions are complementary in nature and capture different information about proficiency.

Index Terms: automatic assessment of spoken language proficiency, computer assisted language learning

1. Introduction

Automatic scoring of language proficiency is becoming a point of growing interest and importance in the field of second language assessment because the number of English-as-a-second-language (ESL) learners has been steadily increasing worldwide [1]. Internationally recognized language tests, such as International English Language Testing System (IELTS) and Test of English as a Foreign Language (TOEFL), are often composed of sections specifically dedicated to the assessment of speaking, listening, writing and reading skills, that are evaluated by human experts according to certain sets of criteria. The introduction of automatic graders for spoken language assessment would be beneficial for examiners and learners for formal settings and also for practice situations in Computer Assisted Language Learning (CALL). In fact, compared to human graders, automatic graders can ensure greater consistency and speed at a lower cost, since the recruitment and training of new human experts is expensive and can offer only a small increase in performance [2].

Despite some limitations [3], established standards such as the Common European Framework of Reference for languages (CEFR) [4] are recognized throughout the world as effective measures for grading the proficiency of second language (L2) speakers. The CEFR scales are structured according to 'can-do' descriptors of language proficiency outcomes, especially in relation to communicative competence. Therefore, these guidelines expect graders to grade proficiency by means of holistic assessments rather than separated facets. Nonetheless, it has been demonstrated that such holistic evaluations do have a modularizable structure, which can be divisible into single facets of proficiency, such as phonetic pronunciation, intonation, speed of speech, vocabulary and grammar, each of which is assigned a score that strongly correlates with the holistic grade [5]. In light of this, automatic grading can be a valuable resource a fortiori, as it has been suggested that it might be used to make consistent assessments of specific linguistic phenomena, whereas human grading focuses on more global aspects of performance, as shown in [6] for written and in [7] for spoken proficiency. Furthermore, although the CEFR descriptors mainly target communicative competence, it has been proven that even learner errors can be connected to CEFR proficiency levels [8] and can be considered as criterial features for each level, together with other linguistic features, as illustrated in [9].

CALL applications also distinguish between different facets of proficiency during teaching, with different systems used to separately teach specific linguistic skills, such as pronunciation [10], prosody [11], and vocabulary [12]. As a result, the ability to analytically assess a learner's progress according to each of these facets should be useful for feedback and in order to inform further teaching in an adaptive fashion.

In automatic assessment of L2 spoken language proficiency, input sequential data from a learner is used to predict a holistic grade and/or a grade representing proficiency with respect to a particular facet (single-view grading). The input may consist, as needed, of acoustic features, recognised words, phones and/or time-alignment information, or other information, such as fundamental frequency, extracted directly from the audio or from automatic speech recognition (ASR) output. Most approaches in the literature extract sets of hand-crafted features to capture views including fluency [13], pronunciation [14], prosody [15] and text complexity [16], which are then fed into graders, trained with human-annotated single-view scores, to predict single-view scores. Since CEFR descriptors do not provide accurate information regarding the use of analytic scores, annotated data containing such human-annotated single-view scores are hard to obtain and are likely to suffer from inconsistency between and within human raters. This approach can be used for holistic grading by concatenating multi-

This paper reports on research partially supported by Cambridge University Press and Assessment. The authors would like to thank Dr. Linlin Wang for providing the ASR transcription, Vyas Raina for the text model, Yiting Lu for the GEC model, and the ALTA Speech Technology Project Team for general discussions and contributions to the evaluation infrastructure.

ple view-specific hand-crafted features targeting more than one facet in order to produce holistic feature sets, which are then passed through graders, to predict holistic grades, as shown in [17, 18, 19, 20], with the grader trained on human assigned holistic scores. The efficacy of hand-crafted features for either view-specific or holistic grading relies heavily on their particular underlying assumptions and they risk discarding potentially salient information about proficiency. This issue for holistic grading has been addressed by replacing hand-crafted features with automatically derived features for holistic grading prediction, either through an end-to-end system [21] or in multiple stages [22, 23]. However, neither can be used for multi-view assessment.

This paper investigates whether view-specific systems can be trained when only holistic scores for a test-taker are available. Section 2 presents a framework for multi-view assessment using graders trained on only holistic grades. Section 3 describes the single-view graders used in this work, while Section 4 illustrates the experimental setup and data. Finally, the experimental results are analysed and discussed in Section 5.

2. View-Specific Training

As previously discussed, for most spoken language assessment training data sets only overall holistic scores are available. Thus, the training data set comprises $\mathcal{D} = \{\boldsymbol{x}^{(i)}, y^{(i)}\}$ where $\boldsymbol{x}^{(i)}$ is the set of features, or sequence of features, extracted from the audio and ASR system, and $y^{(i)}$ the associate reference score. This section motivates how this training data can be used to train view-specific models.

The assessment process can be split into two distinct stages. where initially the features $\boldsymbol{x}$ are mapped to view-specific features $\boldsymbol{v}$, and then fed into the score-prediction network. Thus, for a particular view

$$\hat{y}_{\text{v}}^{(i)} = \mathcal{F}_{\text{v}}(\boldsymbol{x}^{(i)}) = f_{\text{v}}(\mathbf{g}_{\text{v}}(\boldsymbol{x}^{(i)})) = f_{\text{v}}(\boldsymbol{v}^{(i)}) \tag{1}$$

where the desired training data comprises $\mathcal{D}_{\text{v}} = \{\boldsymbol{x}^{(i)}, y_{\text{v}}^{(i)}\}$. Unfortunately, there are no view-specific reference grades, $y_{\text{v}}^{(i)}$, associated with each of the training observations, $\boldsymbol{x}^{(i)}$, just overall holistic grades, $y^{(i)}$. To address this problem, the form of the feature extractor $\mathbf{g}_{\text{v}}(\boldsymbol{x}^{(i)})$ is constrained so that only information about a specific view is contained within $\boldsymbol{v}^{(i)}$ (see Figure 1).

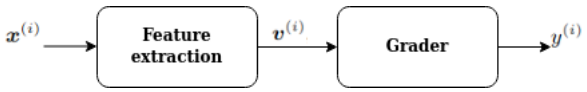

Figure 1: *View-specific training.*

For example, if only information about the text spoken is in $\boldsymbol{v}^{(i)}$, irrespective of the pronunciation of the words[1], then the same feature vector $\boldsymbol{v}$ can be obtained from the different values of $\boldsymbol{x}$.

Training the model parameters, $\boldsymbol{\theta}$, on the holistic training data, $\mathcal{D}$, aims to minimise the loss $\mathcal{L}(\boldsymbol{\theta})$

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathcal{L}(y^{(i)}, \mathcal{F}_{\text{v}}(\boldsymbol{x}^{(i)})) = \sum_{j=1}^{\tilde{N}} \sum_{i \in \mathcal{S}^{(j)}} \mathcal{L}(y^{(i)}, f_{\text{v}}(\boldsymbol{v}^{(j)})) \tag{2}$$

[1]There will be some influence of pronunciation on the performance of the ASR system and associated confidence scores.

where $\mathcal{S}^{(j)}$ is the set of samples such that $\boldsymbol{g}_{\text{v}}(\boldsymbol{x}^{(i)}) \approx \boldsymbol{v}^{(j)}$ and $\tilde{N}$ is the number of distinct values of $\boldsymbol{v}$. In this work, a least-squares cost function, $\mathcal{L}(y, \hat{y}_{\text{v}})$, is used. When training the model it is not necessary for the loss function to be "correct" provided the gradients for training the model parameters are suitable. Thus

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &\propto \sum_{j=1}^{\tilde{N}} \sum_{i \in \mathcal{S}^{(j)}} \left(f_{\text{v}}(\boldsymbol{v}^{(j)}) - y^{(i)} \right) \frac{\partial f_{\text{v}}(\boldsymbol{v}^{(j)})}{\partial \boldsymbol{\theta}} \\ &= \sum_{j=1}^{\tilde{N}} |\mathcal{S}^{(j)}| \left(f_{\text{v}}(\boldsymbol{v}^{(j)}) - \frac{\sum_{i \in \mathcal{S}^{(j)}} y^{(i)}}{|\mathcal{S}^{(j)}|} \right) \frac{\partial f_{\text{v}}(\boldsymbol{v}^{(j)})}{\partial \boldsymbol{\theta}} \end{aligned} \tag{3}$$

Thus, the gradient, and associated minima, will be consistent with training against view specific training data $\mathcal{D}_{\text{v}}$ provided

$$y_{\text{v}}^{(j)} \approx \frac{\sum_{i \in \mathcal{S}^{(j)}} y^{(i)}}{|\mathcal{S}^{(j)}|} \tag{4}$$

Here it is assumed that the view score contributes to the overall holistic score. By averaging over samples with similar view-specific features, $\boldsymbol{v}$, the resulting scores should be biased to the view-specific grades even if (4) is not exactly satisfied.

In this analysis the precise concept of how the set $\mathcal{S}^{(j)}$ is derived has not been strictly specified. Assuming that there is sufficient data and the $\mathbf{g}_{\text{v}}()$ is a smooth function, the standard training, the LHS expression in (2), can be run. The model implicitly smooths the view-specific predictions.

3. Single-view graders

In the present study, we implement 5 grading models for as many views of proficiency, namely pronunciation, rhythm, text, GEC edit sequences and POS tag sequences. For all graders, an ensemble of 10 models was trained.[2]

Pronunciation: the pronunciation model is described in detail in [24]. Sequences of acoustic observations corresponding to phone instances are projected to fixed-length phone instance representations, with those corresponding to a specific phone label attended over to obtain an overall representation for that phone. Euclidean distances between phone representations are then passed through a feed-forward layer to predict the score. The objective is for information from the observation vectors to only be preserved insofar as it characterises the way the speaker pronounced each phone in relation to the pronunciation of the other phones.

Rhythm: we implement the rhythm grader as described in [25]. In this case, the grader is constrained in a way that the input only consists of durations of phones and silences, grouped into consonant and inter-consonant intervals, such that the grader can only exploit duration patterns for scoring.

Text: we use the text grader presented in [26], which consists of an LSTM with attention over its hidden representation. The inputs are word embeddings obtained by passing the words of each utterance through a trained BERT language model [27].

Finally, we introduce two novel graders for the specific views of grammatical accuracy and use of parts of speech. For brevity, we refer to them as *es* (edit sequence) and *pos* (part-of-speech) respectively.

[2]Note that the text, es and pos graders consist in turn of multiple graders trained on the scores of the 5 parts that compose the exam, whereas the pronunciation and rhythm graders have been trained on the overall scores of the exam.

Grammatical accuracy: the es grader is a transformer-based [28] model that takes GEC edit sequences as inputs. Prior to the grader, a GEC model is run on the ASR output, after removing hesitations and partial words. Both corrected and original ASR texts are passed through ERRANT [29] to yield the GEC edit sequences.

Use of parts of speech:the pos grader has the same tranformer-based architecture as the es grader, but takes POS tag sequences as inputs. These sequences are generated with spaCy[3].

In addition to a comparative analysis of the single-view graders, we investigate a possible combination by means of an OLS (Ordinary Least Squares) multiple linear regression model using the 5 graders' predictions $\hat{y}_{\mathrm{v}}^{(i)}$as predictors and setting the reference holistic score $\hat{y}^{(i)}$ as target:

$$\hat{y}^{(i)} = \beta_0 + \beta_{\mathtt{p_r}}\hat{y}_{\mathtt{p_r}}^{(i)} + \beta_{\mathtt{r_y}}\hat{y}_{\mathtt{r_y}}^{(i)} + ... + \beta_{\mathtt{p_s}}\hat{y}_{\mathtt{p_s}}^{(i)} + \epsilon \quad (5)$$

where β_0 represents the intercept and β_{v} is the coefficient for a specific view prediction $\hat{y}_{\mathrm{v}}$ and ϵ is the model's residual (see Table 1 for notation). The linear model is trained on the development set. The performance of the single-view graders is compared against a baseline assessment system. This is is a Deep Density Network (DDN) trained on a set of hand-crafted features across different views [30]. These features include: grade dependent language model and word level statistics; statistics of phone duration; statistics to capture rhythm; fluency metrics; and fundamental frequency statistics are used to represent intonation. As for the other graders, the baseline predictions are the result of an ensemble of 10 models. Further information about the features employed and about the ensemble approach can be found in [19, 31].

4. Data and experimental setup

The data used in our experiments are obtained from candidate responses to the spoken components of the Linguaskill examinations for L2 learners of English, provided by Cambridge English Language Assessment [32]. Each speaker is graded on a scale of 1-6 based on the CEFR holistic criteria, i.e. A1, A2, B1, B2, C1 and C2. Non-overlapping datasets of 31475 and 1033 speakers are used respectively as the training and development/calibration set. For evaluation, we consider two test sets, LinGen and LinBus, of 1049 and 712 speakers respectively. LinGen contains learners' answers to questions on General English, whereas LinBus includes answers to questions on Business English. Each test set is balanced for gender and proficiency and features around 30 L1s.

The first step before passing the data through each automarking system is recognising the text being spoken and, for the pronunciation and rhythm systems, aligning the audio to a sequence of phones. Both of these tasks are performed using a Kaldi-based ASR system, specifically the TDNN-F acoustic model and Kneser-Ney language model described in [33]. The average word error rate (WER) is 19.5%.

For the es grader, a transformer-based GEC system [28] is trained on the CLC [34] and BEA-19 GEC shared task [35] corpora. It is a base-sized model [28] with 512D hidden states, 6 encoder, and 6 decoder layers. The vocabulary is derived from CLC and Switchboard [36]. Model parameters are averaged over 5 best checkpoints, and greedy decoding is used. Training uses the Adam optimiser [37] with batch size 256, dropout 0.2, and learning rate 1e-3. The GEC edit sequences are derived from ERRANT run on the original and automatically corrected ASR hypotheses. These sequences are fed into our es model which consists of an embedding layer with size 128, a transformer-block with hidden layer size 128 and 8 heads, a dense layer of 128 nodes, and finally the output layer. Training uses the Adam optimiser with batch size set at 32 and learning rate at 2e-6. The pos grader model has the same structure.

The performance of each grading system is evaluated using root-mean-square error (RMSE), whilst further comparisons also include Pearson's correlation coefficient (PCC), Spearman's rank correlation coefficient (SRC), and the percentage of the predicted scores that are equal to or lie within 0.5 (i.e. within half a grade) of the actual score ($\% \leq 0.5$).

5. Experimental results and analysis

Table 1 shows the performance of the 5 single-view graders and the baseline in terms of RMSE, considering both the individual models and the ensembles. As can be seen, the ensemble approach gives a significant improvement on all the graders, including the baseline.

Model	LinGen		LinBus	
	Indiv.	Ens.	Indiv.	Ens.
baseline	$0.578_{\pm 0.011}$	0.412	$0.522_{\pm 0.009}$	0.406
pron ($\mathtt{p_r}$)	$0.455_{\pm 0.004}$	0.452	$0.454_{\pm 0.003}$	0.451
rhythm ($\mathtt{r_y}$)	$0.571_{\pm 0.036}$	0.508	$0.551_{\pm 0.037}$	0.490
text ($\mathtt{t_x}$)	$0.402_{\pm 0.005}$	0.400	$0.409_{\pm 0.007}$	0.409
es ($\mathtt{e_s}$)	$0.547_{\pm 0.001}$	0.547	$0.497_{\pm 0.001}$	0.495
pos ($\mathtt{p_s}$)	$0.550_{\pm 0.001}$	0.550	$0.499_{\pm 0.003}$	0.497

Table 1: *Performance of the single-view graders and baseline in terms of RMSE. Individual models VS ensembles.*

In order to examine the differences between and the complementarity of each single-view grader, we only consider LinBus. In Table 2, we report the performance of various combinations of the single-view graders through the OLS multiple linear regression model introduced in Section 3. For each component we report the respective β coefficient. It is observed that the combination of all the graders improves on the performance of their component graders, and this is consistent with the single-view graders extracting complementary information to each other. Specifically, among the 5 graders, the text grader affects the linear model most, as can be inferred from its high β coefficient and from the drop in performance of the combination that excludes it. Based on the β coefficients, the pronunciation and rhythm graders always contribute equally to the linear model, but the presence of the first appears to have a more positive impact on the overall performance. The es grader seems to have a relatively smaller impact, except when the combinations exclude the pos or the text graders. We continue our analysis

Combination	$\beta_{\mathtt{p_r}}$	$\beta_{\mathtt{r_y}}$	$\beta_{\mathtt{t_x}}$	$\beta_{\mathtt{e_s}}$	$\beta_{\mathtt{p_s}}$	RMSE
$\mathtt{p_r r_y t_x e_s p_s}$	0.14	0.14	1.30	-0.05	-0.39	0.386
$\mathtt{p_r r_y t_x e_s}$	0.14	0.14	1.30	-0.31	—	0.405
$\mathtt{p_r r_y t_x\ \ p_s}$	0.14	0.14	1.30	—	-0.47	0.384
$\mathtt{p_r r_y\ \ e_s p_s}$	0.45	0.45	—	0.28	-0.05	0.432
$\mathtt{p_r\ \ t_x e_s p_s}$	0.29	—	1.30	-0.05	-0.39	0.385
$\mathtt{r_y t_x e_s p_s}$	—	0.29	1.30	-0.05	-0.39	0.392

Table 2: *RMSE and β coefficients of linear regression model with different combinations.*

[3] spacy.io

focusing on each grader's performance across proficiency level. Figure 2 shows the RMSE variation of the 5 graders across the 5 proficiency levels. The es and pos graders follow very similar trends as expected, since ERRANT labels are based on POS tags. In particular, the case of es is consistent with what the authors of [9] call 'inverted U patterns' in written proficiency, i.e. errors increase after B1 and then decline again by C2. In this regard, it is also interesting to note that there is a correspondence between oral and written proficiency when it comes to grammatical accuracy. Compared to the other graders, the pronunciation grader has the lowest RMSE on the lowest grade (1), which gradually decreases until grade 4 and then rises again after grade 5. On the other hand, the rhythm grader shows its best performance for grade 5, and this is consistent with the findings shown in [38], in which English speech rhythm is described as one of the most difficult aspects for learners to acquire. Finally, the text grader shows the lowest RMSE, in both absolute and relative terms, for the middle grades (3-4).

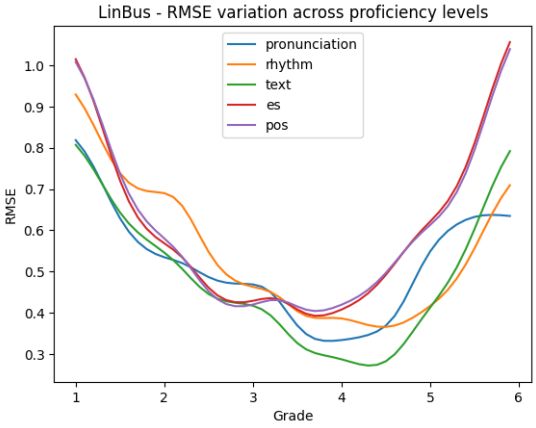

Figure 2: *RMSE variation across proficiency levels.*

Furthermore, we investigate the relationships between single-view graders through a repeated measures design. Arguably, the most well-known repeated measures design is repeated measures analysis of variance (rANOVA). However, since our data violate both the sphericity and normality assumptions required for rANOVA, we must opt for the Friedman test [39], which is considered the non-parametric equivalent of rANOVA and determines whether there are any statistically significant differences in ranks between the distributions of multiple paired groups. As we obtain a significant p-value, we find that there are significant differences among the graders.

In order to determine exactly which graders are significantly different, we perform post-hoc multiple comparisons using the Nemenyi test [40]. We report the test results in Figure 3. All paired comparisons, even those with the reference score, show significant differences (p-value<0.05), with the exception of the pairs es-pos and text-rhythm. As regards the first pair, we have already commented on the almost overlapping trends shown in Figure 2. As for the latter, we might argue that the non-significant p-value reflects the analogous trends of the RMSE variation curves followed by the text and rhythm grader, despite a remarkable gap between them.

Finally, in Table 3 we report a comparison of the baseline, our best performing single-view model i.e. the text grader, and the linear regression model considering the evaluation metrics mentioned in Section 4. The combination of the single-view

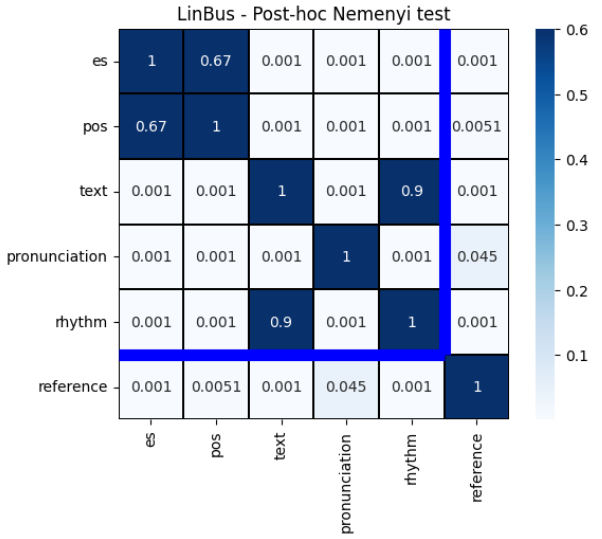

Figure 3: *Heatmap of the results of the post-hoc Nemenyi test.*

graders outperforms both the baseline and the text grader across all metrics.

Model	PCC	SRC	RMSE	%≤0.5
baseline	0.910	0.915	0.406	79.1
text ($\mathtt{t_x}$)	0.920	0.925	0.409	78.9
$\mathtt{p_r r_y t_x e_s p_s}$	0.920	0.926	0.386	80.5

Table 3: *Comparison of the performance of the baseline, text grader and linear regression model.*

6. Conclusions and future work

In order for CALL and automatic spoken language assessment systems to give learners interpretable scores and informative feedback on their speaking ability, specific aspects of their proficiency should be assessed, but for many real-world tasks single-view scores are not available or are often inconsistent for training automatic systems. This paper considers whether view specific systems can be trained when only holistic scores are available. Single-view graders are developed for views of pronunciation, rhythm, text, grammatical accuracy and use of parts of speech. The predictions made by these graders are seen to be complementary to all the others for the task of predicting holistic grades. Furthermore, we investigate a combination of the 5 graders by means of a multiple linear regression model and we find that it generally improves on the performance of each single-view grader. Since the single-view scores are also available, this multi-view system enables the holistic score to be significantly more interpretable enabling useful feedback to learners who need specific indications on how to improve their speaking skills. This is even more true for users of business English, as the ones represented in the LinBus test set. Further work should be undertaken in order to improve the performance of spoken grammatical error annotation, since current systems are generally designed for written texts and are not ideal for speech. Future work will also include other types of combinations, considering both shallow and deep fusion methods.

7. References

[1] P. Howson, *The English effect*. London: British Council, 2013.

[2] M. Zhang, "Contrasting automated and human scoring of essays," *R&D Connections*, no. 21, pp. 1–11, 2013.

[3] C.J Weir, "Limitations of the Common European Framework for developing comparable examinations and tests," *Language Testing*, vol. 3, no. 22, pp. 281–300, 2005.

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

[5] N.H. De Jong, M.P. Steinel, A.F. Florijn, R. Schoonen, and J.H. Hulstijn, "Facets of speaking proficiency," *Studies in Second Language Acquisition*, no. 34, pp. 5–34, 2012.

[6] M. Enright and T. Quinlan, "Complementing human judgment of essays written by English language learners with e-rater® scoring," *Language Testing*, vol. 3, no. 27, pp. 317–334, 2010.

[7] A. Loukina, M. Lopez, K. Evanini, D. Suendermann-Oeft, and K. Zechner, "Expert and crowdsourced annotation of pronunciation errors for automatic scoring systems," in *Proc. Interspeech 2015*, 2015, pp. 2809–2813.

[8] J. Thewissen, "Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus," *The Modern Language Journal*, vol. 97, no. 1, pp. 77–101, 2013.

[9] J.A. Hawkins and P. Buttery, "Criterial features in learner corpora: Theories and illustrations," *English Profile Journal*, vol. 1, no. 1, pp. 1–23, 2010.

[10] R. Thomson, "Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation." *Calico Journal*, vol. 3, no. 28, pp. 744–765, 2011.

[11] D. Hardison, "Generalization of computer assisted prosody training: Quantitative and qualitative findings." *Language Learning & Technology*, vol. 1, no. 8, pp. 34–52, 2004.

[12] P. Groot, "Computer assisted second language vocabulary acquisition," *Language Learning & Technology*, vol. 1, no. 4, pp. 56–76, 2000.

[13] H. Strik and C. Cucchiarini, "Automatic assessment of second language learners' fluency," in *Proc. International Congress of Phonetic Sciences (ICPhS) 1999*, 1999.

[14] L. Chen, K. Evanini, and X. Sun, "Assessment of non-native speech using vowel space characteristics," in *Proc. 2010 IEEE Spoken Language Technology Workshop*, 2010, pp. 139–144.

[15] E. Coutinho, F. Hönig, Y. Zhang, S. Hantke, A. Batliner, E. Nöth, and B. Schuller, "Assessing the prosody of non-native speakers of English: Measures and feature sets," in *Proc. 10th International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

[16] S. Bhat and S. Yoon, "Automatic assessment of syntactic complexity for spontaneous speech scoring," *Speech Communication*, vol. 67, pp. 42–57, 2015.

[17] P. Müller, F. De Wet, C. Van Der Walt, and T. Niesler, "Automatically assessing the oral proficiency of proficient L2 speakers." in *Proc. Workshop on Speech and Language Technology for Education (SLaTE)*, 2009, pp. 29–32.

[18] S. Crossley and D. McNamara, "Applications of text analysis tools for spoken response grading," *Language Learning & Technology*, vol. 17, no. 2, pp. 171–192, 2013.

[19] Y. Wang, M.J.F. Gales, K.M. Knill, K. Kyriakopoulos, A. Malinin, R.C. van Dalen, and M. Rashid, "Towards automatic assessment of spontaneous spoken English," *Speech Communication*, vol. 104, pp. 47–56, 2018.

[20] Z. Liu, G. Xu, T. Liu, W. Fu, Y. Qi, W. Ding, Y. Song, C. Guo, C. Kong, S. Yang *et al.*, "Dolphin: a spoken language proficiency assessment system for elementary education," in *Proc. The Web Conference 2020*, 2020, pp. 2641–2647.

[21] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6234–6238.

[22] K. Takai, P. Heracleous, K. Yasuda, and A. Yoneyama, "Deep learning-based automatic pronunciation assessment for second language learners," in *Proc. International Conference on Human-Computer Interaction*. Springer, 2020, pp. 338–342.

[23] S. Cheng, Z. Liu, L. Li, Z. Tang, D. Wang, and T. F. Zheng, "ASR-Free pronunciation assessment," in *Proc. Interspeech 2020*, 2020, pp. 3047–3051.

[24] K. Kyriakopoulos, K.M. Knill, and M.J.F. Gales, "A deep learning approach to assessing non-native pronunciation of English using phone distances," in *Proc. Interspeech 2018*, 2018, pp. 1626–1630.

[25] ——, "A deep learning approach to automatic characterisation of rhythm in non-native English speech," in *Proc. Interspeech 2019*, 2019, pp. 1836–1840.

[26] V. Raina, M.J.F. Gales, and K.M. Knill, "Universal adversarial attacks on spoken language assessment systems," in *Proc. Interspeech 2020*, 2020, pp. 3855–3859.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[29] C. Bryant, M. Felice, and T. Briscoe, "Automatic annotation and evaluation of error types for grammatical error correction," in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 793–815.

[30] A. Malinin, A. Ragni, K.M. Knill, and M.J.F. Gales, "Incorporating uncertainty into deep learning for spoken language assessment," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 45–50.

[31] X. Wu, K.M. Knill, M.J.F. Gales, and A. Malinin, "Ensemble approaches for uncertainty in spoken language assessment," in *Proc. Interspeech 2020*, 2020, pp. 3860–3864.

[32] K. Ludlow, *Official Quick Guide to Linguaskill*. Cambridge University Press, 2020.

[33] Y. Lu, M.J.F. Gales, K. Knill, P. Manakul, L. Wang, and Y. Wang, "Impact of ASR performance on spoken grammatical error detection," in *Proc. Interspeech 2019*, 2019, pp. 1876–1880.

[34] D. Nicholls, "The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT," in *Proc. of the Corpus Linguistics 2003 conference; UCREL technical paper number 16*, 2003, pp. 572–581.

[35] C. Bryant, M. Felice, Ø.E. Andersen, and T. Briscoe, "The BEA-2019 shared task on grammatical error correction," in *Proc. of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, pp. 52–75.

[36] M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer, "Dysfluency annotation stylebook for the switchboard corpus," in *Tech. Rep. Linguistic Data Consortium*, 1995.

[37] D.P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *International Conference on Learning Representations*, 2014.

[38] D.S. Taylor, "Non-native speakers and the rhythm of english," *International Review of Applied Linguistics in Language Teaching*, vol. 19, no. 4, pp. 219–226, 1981.

[39] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[40] P.B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton University, 1963.