Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning

Stefano Banò¹,²,³, Kate M. Knill²,⁴, Marco Matassoni², Vyas Raina³, Mark J.F. Gales³,⁴

¹University of Trento, Italy
²Fondazione Bruno Kessler, Italy
³ALTA Institute, Cambridge University, UK
⁴Enhanced Speech Technology Ltd., UK

{sbanno, matasso}@fbk.eu, {kmk1001, vr313}@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

A standard pipeline for automated spoken language assessment is to start with an automatic speech recognition (ASR) system and derive features that exploit transcriptions and audio. Although efficient, these approaches require ASR systems that can be used for second language (L2) speakers and preferably tuned to the specific form of test being deployed. Recently, a self-supervised speech representation-based scheme requiring no ASR was proposed. This work extends the initial analysis to a large-scale proficiency test, Linguaskill. The performance of a self-supervised, wav2vec 2.0, system is compared to a high-performance hand-crafted assessment system and a BERT-based system, both of which use ASR transcriptions. Though the wav2vec 2.0-based system is found to be sensitive to the nature of the response, it can be configured to yield comparable performance to systems requiring transcriptions and shows significant gains when appropriately combined with standard approaches.

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

1. Introduction

In recent years, the growing number of learners of English as a second language (L2) on a global scale has created an increasing demand for automated spoken language assessment systems for applications in Computer-Assisted Language Learning (CALL) in private practice and classroom situations and to validate language examinations. One of the compelling reasons for automatic assessment is the need to evaluate and provide feedback to increasing numbers of learners and return results in a timely manner. Secondly, compared to human graders, not only can automatic systems ensure greater speed, but they can do it at a lower cost, since the recruitment and training of new human experts is expensive and sometimes offers only a small increase in performance. Finally, the use of automatic assessment methods can enhance consistency, reliability, and objectivity of scoring, since machines are not susceptible to rater effects and - more simply-to tiredness [1].

Automated systems for assessing L2 speaking proficiency typically receive sequential data from a learner as input to predict their grade or level with respect to overall proficiency or specific facets of proficiency. This input data may consist of phones, recognised words, acoustic features, or other information derived directly from audio or from automatic speech recognition (ASR) transcriptions. In some cases, sets of hand-crafted features related to specific aspects of proficiency, such as fluency [2], pronunciation [3], prosody [4] and text complexity [5], are extracted and fed into graders to predict analytic grades related to those specific aspects. An alternative but similar approach consists of concatenating multiple hand-crafted features related to multiple aspects of proficiency in order to obtain overall feature sets, which are subsequently projected through graders to predict holistic grades, as shown in [6, 7, 8, 9]. The efficacy of hand-crafted features for scoring either overall proficiency or individual aspects heavily relies on their specific underlying assumptions, and salient information about proficiency may risk being discarded. For holistic grading, this limitation has been addressed by substituting hand-crafted features with automatically derived features for holistic grading prediction, either through end-to-end systems [10] or in multiple stages [11, 12]. Other studies have investigated the use of graders that are trained on holistic scores but are defined with both their inputs and structure adapted to target particular aspects of proficiency, such as pronunciation [13], rhythm [14] and text [15, 16]. In these cases, a possible limitation might be the absence of information concerning facets of proficiency that are not present in the input data fed to the grader, although it has been shown that it is possible to combine multiple graders targeting different aspects of proficiency [17]. Such a limitation is particularly evident in systems using ASR transcriptions only, since a) they will contain errors and so will not provide faithful verbatim transcriptions of a learner’s performance, thus not rendering its content appropriately; b) transcriptions do not contain any information relating to the message realisation, such as fluency, pronunciation, intonation, rhythm, or prosody. Instead, they represent a precious resource for highly specific tasks in CALL applications, e.g., for spoken grammatical error correction and feedback, as shown in [18].

In this paper, in order to tackle these issues and address these limitations, we propose an approach based on self-supervised speech representations using wav2vec 2.0 [19]. Recent studies have shown that self-supervised learning (SSL) is an effective approach in various downstream tasks of speech processing applications, such as ASR, keyword spotting, emotion recognition, intent classification, speaker identification, and speaker diarisation [19, 20]. In these studies, contextual representations were applied by means of pre-trained models. In particular, it has been demonstrated that these models are able to capture a vast range of speech-related features and linguistic information, such as audio, fluency, suprasegmental pronunciation, and even semantic and syntactic text-based features for L1, L2, read and spontaneous speech [21]. In the field of CALL, SSL has been explored for mispronunciation detection and diagnosis [22, 23] and automatic pronunciation assessment of read speech [24, 25], and recently for fluency assessment of...
both read-aloud and open responses [26].

To the best of our knowledge, [27] was the first to propose SSL for holistic and analytic proficiency assessment of spontaneous speech. Nevertheless, this study had two relevant limitations related to a) quantity and quality of the data and b) experimental setup. First, the datasets employed in the study were relatively small and were elicited in low-stakes settings from learners assigned with only certain proficiency levels from a limited range of L1 backgrounds. Secondly, the study compared the SSL-based grading systems to BERT-based baselines only, which, although fed with manual transcriptions containing disfluencies, did not consider purely acoustic features, thus potentially missing strictly speech-related aspects of proficiency.

To address these limitations, in the present contribution, we conduct our experiments using a significantly larger amount of L2 learner data obtained from a computer-based oral English test in which a wide variety of L1 backgrounds and all proficiency levels are represented and comparing the performance of wav2vec 2.0 to two types of grader: a BERT-based and a standard grader fed with hand-crafted features across different facets of proficiency (see Figure 1). Additionally, we test the effectiveness of wav2vec 2.0 on a multi-part examination predicting both the overall and individual grades of each part of the exam. Furthermore, we investigate various combinations between the wav2vec2-based, the BERT-based, and the standard graders. These elements of novelty should provide the scientific community with a measure of how current and future, neural grader models can complement existing systems in real-world applications.

2. Data

The data used in our study are obtained from candidate responses to the spoken parts of the LinguaSkill examinations for L2 learners of English, provided by Cambridge University Press & Assessment [28]. The speaking section is divided into five parts. In Part 1, the candidates should answer eight questions about themselves, of which the first two are not marked. The answers last about 10 or 20 seconds. Part 2 consists of a reading-aloud activity including eight sentences of 10 seconds each. Part 3 and Part 4 test the candidates’ ability to deliver a long turn, speaking for up to one minute. In the first, the candidates should talk about a given topic, whereas in the latter they should talk about one or more graphics, such as charts, diagrams, or information sheets. Finally, in Part 5, test-takers should give their opinions in form of responses of about 20 seconds to five questions related to a given topic. Since each part contributes 20% to the speaking exam, the overall grade is calculated as the average of the grades assigned to the five parts. Each test submission is scored on a scale of 0-6. Here, speakers were selected from the A1 to C1 and above range of the Common European Framework of Reference (CEFR) for languages [29] proficiency scale. Non-overlapping datasets of 31475 and 1033 speakers are used as the training and development/calibration set, respectively. For evaluation, we consider two test sets, LinGen and LinBus, of 1049 and 712 speakers, respectively. LinGen includes learners’ responses to questions on General English, while LinBus contains answers to questions on Business English. Each test set features around 30 L1s and is balanced for gender and proficiency level. The text transcriptions for training and test used by the standard and BERT-based graders are generated using a non-native English ASR system. A Kaldi-based system with a TDNN-F acoustic model and a trigram language model is used similar to that in [30] with average WER of ~20%.

3. Model architectures

Wav2vec2-based graders: in wav2vec 2.0, speech audio is encoded through a multilayer convolutional neural network (CNN). After encoding, masking is applied to spans of the resulting latent representations. Subsequently, these are fed into a transformer to build contextualised representations. Gumbel softmax is used to calculate the contrastive loss on which the model is trained, and speech representations are learned from this training.

For our experiments, we initialised the model configuration and processor from a version provided by HuggingFace. After the learners’ answers are fed into the model, wav2vec 2.0 provides contextualised representations. To handle representations of various audio lengths, we used a mean pooling method to concatenate 3D representations into 2D representations, which are finally passed through a regression head. Since we trained a grader for each part of the exam, after trying different architectures, for Parts 1 and 5 we used a regression head composed of a Dropout layer, a layer of 768 units, a Dropout layer, and the output layer, whereas, for Parts 2, 3 and 4, we used a deeper architecture, i.e., a Dropout layer, a stack of three layers of 768 units, a Dropout layer, a layer of 128 units, and, finally, the output layer. The graders use mean squared error (MSE) as the loss function. Training uses the AdamW optimiser [31] and hyperparameters vary depending on each part. For Part 1, we used batch size 16, gradient accumulation step 2, dropout rate 0.1, and learning rate 5e-5, and we trained the grader for 2 epochs. For Part 2, we used batch size 16, gradient accumulation step 2, dropout rate 0.5, and learning rate 1e-6, and we trained the grader for 3 epochs. For Part 3 and Part 4, the hyperparameters are the same: we set batch size to 8, gradient accumulation step to 4, dropout rate to 0.5, and learning rate to 1e-5, and we trained the grader for 2 epochs. Finally, the grader for Part 5 has batch size set to 8, gradient accumulation step to 2, dropout rate to 0.1, and learning rate to 5e-5, and we trained it for 1 epoch.

As has been said, the first part of wav2vec 2.0 consists of a stack of CNN layers that are used to obtain acoustically meaningful - but contextually independent - features from the raw speech audio signal. This part of the model has already been initialized from a version provided by HuggingFace.
sufficiently trained during the pre-training stage and does not need fine-tuning. Therefore, for our experiments we froze the parameters of the feature extractor.

BERT-based graders: for comparison, we use the text grader presented in [16], 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 [32].

Standard graders: we also compare our SSL approach to a standard grader, which is a Deep Density Network (DDN) trained on a set of hand-crafted features designed to cover all the different aspects of proficiency and is described in [33]. These features include: word level statistics; statistics of phone duration; statistics to capture rhythm; fluency metrics; and fundamental frequency statistics are used to represent intonation.

For all graders, their predictions are the result of ensembles. Further information about the ensemble approach can be found in [34]. Systems are calibrated and in a final set of experiments combined using linear combination:

$$y(n) = \beta_0 + \sum_{p \in \mathcal{P}} \beta_p y_p(n)$$  \hspace{1cm} (1)

where $\mathcal{P}$ is the set of parts to combine, which may come from multiple systems, and $\beta_p$ are the coefficients associated with the parts. For the baseline submission performance the values of $\beta_p, p > 0$ are all constrained to be the same, and equal to 0.2, to yield simple averaging consistent with the combination of operational examiner scores. Ordinary Least Squares (OLS) estimation using the development/calibration set is used to find the values of $\beta_p$ when unequal weighting is used.

For the evaluation of the grading systems at the part-level, we use 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 about half a grade point) ($\% \leq 0.5$), and within 1.0 (i.e., about one grade point) ($\% \leq 1.0$) of the actual score.

### 4. Experimental results and analysis

#### Part-Level Performance: we start our series of experiments with grading each of the five parts of the exam. For this part of our analysis we only consider LinGen. Table 1 reports the results of the three grading systems in terms of RMSE.

<table>
<thead>
<tr>
<th>LinGen</th>
<th>Model</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinGen</td>
<td>std ((a_s))</td>
<td>0.625</td>
<td>0.662</td>
<td>0.671</td>
<td>0.686</td>
<td>0.633</td>
</tr>
<tr>
<td></td>
<td>BERT ((b_t))</td>
<td>0.628</td>
<td>0.683</td>
<td>0.681</td>
<td>0.694</td>
<td>0.629</td>
</tr>
<tr>
<td></td>
<td>wav2vec2 ((w_s))</td>
<td>0.601</td>
<td>0.827</td>
<td>0.845</td>
<td>0.845</td>
<td>0.674</td>
</tr>
</tbody>
</table>

**Note:** The wav2vec 2.0 performance varies across the parts, with close or better RMSE to the other two graders for Parts 1 and 5 and lower performance on Parts 2, 3 and 4. This appears to be due to the nature of the responses required for different parts. Parts 1 and 5 consist of several short spontaneous answers, whereas Parts 3 and 4 are also composed of spontaneous speech but a single longer and more complex response in each case. The lower wav2vec 2.0 performance may be due to our use of a mean pooling method, which may be giving too compressed a representation for longer utterances.

**Part 2, by contrast, is similar in length to Part 1 but consists of read speech responses. This part mainly targets pronunciation and fluency skills at the expense of content-related aspects of proficiency so the wav2vec 2.0 system might have been expected to do well. Its higher RMSE might be due to the absence of information related to the reference text read by the test-takers — consisting in more than a hundred different prompts. It is noticeable that the standard grader, which covers all aspects of a candidate’s speech, performs the best, with the BERT grader which cannot measure pronunciation or prosody slightly behind. I don’t think they mention any operational autotmarker.

**Submission-Level Performance:** the second part of our experiments is focused on the overall grades of the Linguaskill exam, i.e., the average of the grades assigned to the five parts. Table 2 shows the results of the three grading systems both on LinGen and LinBus.

<table>
<thead>
<tr>
<th>LinGen</th>
<th>Model</th>
<th>PCC</th>
<th>SRC</th>
<th>RMSE</th>
<th>$% \leq 0.5$</th>
<th>$% \leq 1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinGen</td>
<td>std ((a_s))</td>
<td>0.932</td>
<td>0.937</td>
<td>0.383</td>
<td>81.5</td>
<td>98.6</td>
</tr>
<tr>
<td></td>
<td>BERT ((b_t))</td>
<td>0.929</td>
<td>0.934</td>
<td>0.395</td>
<td>80.3</td>
<td>98.5</td>
</tr>
<tr>
<td></td>
<td>wav2vec2 ((w_s))</td>
<td>0.908</td>
<td>0.931</td>
<td>0.455</td>
<td>73.3</td>
<td>97.3</td>
</tr>
<tr>
<td>LinBus</td>
<td>Model</td>
<td>PCC</td>
<td>SRC</td>
<td>RMSE</td>
<td>$% \leq 0.5$</td>
<td>$% \leq 1.0$</td>
</tr>
<tr>
<td>LinBus</td>
<td>std ((a_s))</td>
<td>0.911</td>
<td>0.918</td>
<td>0.416</td>
<td>76.5</td>
<td>98.3</td>
</tr>
<tr>
<td></td>
<td>BERT ((b_t))</td>
<td>0.920</td>
<td>0.925</td>
<td>0.398</td>
<td>80.1</td>
<td>99.2</td>
</tr>
<tr>
<td></td>
<td>wav2vec2 ((w_s))</td>
<td>0.893</td>
<td>0.911</td>
<td>0.446</td>
<td>72.1</td>
<td>97.9</td>
</tr>
</tbody>
</table>

Table 2: Submission-level performance on LinGen and LinBus.

They all achieve good results across all metrics, with the standard grader and the BERT-based grader performing moderately better than the wav2vec2-based grading system. As regards the BERT-based grader specifically, analogous trends can be observed in [16] and [17], and this fact is quite significant, since it highlights the importance of content-related aspects of speaking proficiency, which is far from being a mere surrogate of fluency and pronunciation. Furthermore, it appears that the standard grading system outperforms the other two graders on LinGen, but has a slightly worse performance than the BERT-based grader on LinBus. This might be ascribable to the different language models, since the first test set contains questions on General English, whereas the latter includes questions on Business English, which is typically more specific and complex.

Figure 2 shows that the wav2vec2-based graders can discriminate between lower levels of proficiency, but we can clearly see that it is not able to distinguish between the highest levels as its maximum prediction is 4.6, i.e., between grades B2 and C1 and above (the plots for LinBus show a similar trend). Our hypothesis is that higher-level assessment tends to be more dependent on message construction (what is said) rather than message realisation (how it is said), and wav2vec 2.0 should not have actual knowledge of words, although, as mentioned in the Introduction, it has been demonstrated that SSL speech representations are able to encode a great amount of high-level linguistic features, including aspects of semantics and syntax [21].

**Combinations:** we investigated the application of a multiple linear regression model using the per-part predictions as predic-
For each individual grading system and for four combinations of them. The results on LinGen and LinBus are reported in Table 4. The combinations show performances that are aligned to or better than the individual models across all metrics, with the combination including all three graders achieving the best results. In particular, the improvements in terms of RMSE, ≤0.5, and ≤1.0 are quite noticeable compared to the results of the individual graders. This combination also helps with the issue of the wav2vec2-based grader related to scoring higher levels as can be seen in the lowest plot in Figure 2. Table 3 reports the β coefficients of the individual models described in Table 4 and the combination of all three. In the standard and the BERT-based graders, Parts 1, 2 and 5 affect the linear model most, whereas in the wav2vec2-based grader Part 1 and 5 are the most influential. In their combination, the highest β coefficients are Parts 1 and 5 of the wav2vec2-based grader and Part 2 of the BERT-based and the standard graders. These values seem to confirm the RMSE results shown in Table 1.

5. Conclusions and future works

In this study, we have extended a recent novel approach to proficiency assessment using a wav2vec2-based grader, applying it to a large quantity of L2 learner data obtained from a computer-based oral English test. First, we compared its performance on the five parts of the Linguaskill test to a high-performing standard grader fed with hand-crafted features and a BERT-based grader. Our proposed approach shows a good performance on Parts 1 and 5, which consist of short spontaneous answers. While the lower results on Part 2 (i.e., the only reading-aloud activity) might be due to the absence of information about the text, mean pooling seems to negatively affect performance on Parts 3 and 4, which consist of longer answers. In future work, we plan to investigate other SSL approaches, i.e., recent models such as HuBERT [35] or WavLM [36], as well as other combination methods.
6. References


