Automatic Assessment of Conversational Speaking Tests

S.W. McKnight, A. Civelekoglu, M.J.F. Gales, S. Bannò, A. Liu and K.M. Knill

ALTA Institute, Cambridge University Engineering Department, Trumpington Street, UK
{swm35,mjfg100,kmk1001}@cam.ac.uk

Abstract
Many speaking tests are conversational, dialogic, in form with an interlocutor talking to one or more candidates. This paper investigates how to automatically assess such a test. State-of-the-art approaches are used in a multi-stage pipeline: diarization and speaker assignment, to detect who is speaking and when; automatic speech recognition (ASR), to produce a transcript; and finally assessment. Each presents challenges which are investigated in the paper. Advanced foundation model-based auto-markers are examined: an ensemble of Longformer-based models that operates on the ASR output text; and a wav2vec2-based system that works directly on the audio. The two are combined to yield the final score. This fully automated system is evaluated in terms of ASR performance, and related impact of candidate assignment, as well as prediction of the candidate mark on data from the Occupational English Test. This is a conversational speaking test for L2 English healthcare professionals.

Index Terms: dialogic assessment, candidate detection

1. Introduction
Auto-marking of free speaking tests, whether for practice or examination, has increased in the last few years due to improvements in technology combined with a need to support learners worldwide. More advanced speaking auto-markers, such as those deployed for TOEFL iBT [1], Linguaskill Speaking [2] and Duolingo English Test [3], are applied to prompt-response type scenarios. In this monologic spoken language assessment (SLA), a single speaker, the candidate, speaks freely in response to a prompted question. Whilst this can achieve a good measure of a learner’s ability to speak fluently in a language, these semi-direct speaking assessments are limited in assessing a candidate’s ability to interact with an interlocutor. Most communication-oriented oral proficiency tests tend to be mediated by a human assessor who also forms the interlocutor role. This dialogic assessment is typically 1-1 or 1-2. There is significant interest in dialogic assessment for L2 English speakers generally [4, 5, 6, 7] and developing auto-markers to evaluate recordings and reliably assess dialogic L2 English speaking performance [8, 9, 10]. Some research has been done on simulated human-computer dialogues [11], but no systems that automatically mark complex or human-human dialogues currently appear to exist. This paper investigates advanced SLA for a human-to-human L2 English dialogic test, and the challenges to address.

The Occupational English Test (OET) is a set of four English tests covering listening, reading, writing and speaking [12]. It focuses on English language proficiency required for healthcare professionals, in 12 specific industries such as medicine and pharmacy. It is commonly used by L2 English speakers seeking healthcare jobs in English speaking countries. The speaking component test, OET Speaking, is the one of interest. It comprises two role-plays between the candidate (in their healthcare role, such as doctor, nurse or pharmacist) and an interlocutor (e.g. as a patient). The role-plays are face-to-face either in-person or over video. The audio recordings are subsequently scored by assessors. This paper considers the typical in-person case where the test is recorded on a single-channel microphone set on the table between candidate and interlocutor.

Recently foundation model-based auto-markers have been shown to work well for SLA of monologic tests (e.g. [11]), particularly BERT-based [13] text neural grader models [14, 15, 16] and wav2vec2-based [17] speech neural graders [15, 16]. As well as performing similarly to systems based on hand-crafted features, these have the advantage that they do not require the best features (e.g. [1, 18]) for conversational SLA to be determined so are adopted here. Prior to assessment, speaker diarization must be applied to the audio to separate it into homogeneous speaker segments. Each segment is assigned to candidate or interlocutor. The segments may overlap when both speakers speak at the same time. To handle the much longer text sequences, a Longformer [19] model is used for a text neural grader. A Whisper [20, 21] ASR model from OpenAI is used to generate the input transcriptions of the diarized speech segments. As audio is not directly input, the text grader cannot fully cover all salient components of speech e.g. pronunciation. These missing aspects (and others) are scored through combination with predictions from a wav2vec2-based speech neural grader which applies attention over input audio segments.

2. OET speaking test
Each OET Speaking test takes place between the candidate and an interlocutor. The assessors are not present. It consists of three key stages: Introductory; Role-play One; Role-play Two. In the unmarked Introductory stage, the candidate’s identity is verified, followed by questions from the interlocutor designed to put the candidate at ease. At the end of this stage the interlocutor presents the candidate with a written description of the first role-play scenario. This describes: the setting, e.g. a medical clinic; the person they are meeting, e.g. a 50 year old diabetic with blood-sugar level problems; and a number of tasks they should carry out, e.g. confirm the reason for the appointment. After 3 minutes for preparation, Role-play One begins and lasts up to 5 minutes. At the end of Role-play One, the
interlocutor presents the candidate with the Role-play Two information and the same process is repeated.

Since the role-plays are a (fairly) natural conversation, they contain spontaneous speech effects, such as disfluencies, and overlapping speech. The latter can occur when one speaker speaks before the other has finished or can take the form of back-channel markers [22]. Directly before and after each role-play there is some chat where the interlocutor checks whether the candidate is ready, notes the end of the role-play, and sets up for the next one. In general, there is no sound or other marker to indicate when a role-play has started/ended. Between role-plays there is mostly silence, but some non-speech sounds may be recorded or the candidate may ask for clarification.

The recordings are marked independently by assessors evaluating the candidate’s linguistic and clinical communication ability [23, 24]. A raw score of up to 39 is made up of scores for linguistic and clinical communication criteria. The combined assessor scores are mapped to a score out of 500.

3. Dialogic assessment auto-marking

Figure 1 shows the overall dialogic assessment process. Unlike monologic automatic assessment, the system needs to know which speech is pertinent to the test and, of that, which corresponds to the candidate and which to the interlocutor. The input audio is therefore first passed through a speaker diarization system. This is used to establish the role-play sections of the recording and from these identify who spoke when. Following [16], two auto-markers are used to predict the holistic test score: a Longformer-based neural text grader; and a wav2vec2-based neural speech grader. For the former, ASR is used on the diarized speaker segments to produce the text transcript of each participant in the test conversation. The auto-marker predictions are combined to yield the final score.

![Figure 1: Overall assessment process (⊕ denotes ensemble).](image)

As noted in Section 1, BERT [13] models have been shown to work well for monologic speaking auto-markers based on ASR generated text [14, 15, 16]. However, BERT models, and improved models such as RoBERTa [25], are limited to 512 tokens in their inputs. As the role-plays are each around 5 minutes long and generate significantly more than 512 tokens, a Longformer-based model [19] is preferred here. This is a transformer model based on RoBERTa but uses a sliding window approach to increase the maximum number of tokens to 4,096, sufficient to cover each role-play in full. The model is a sequence-to-one regression model where the encoder-only part of the pre-trained Longformer model is used with an additional final fully connected layer for regression. This operates on the first vector of the final Longformer layer to produce a single score [13, 26].

The Longformer input is the relevant tokenized ASR text for a particular role-play, and the text can be candidate-only, interlocutor-only or both. The token "<<" is automatically inserted at the start. "]]></s>" with no spaces is inserted between each utterance and at the end, e.g.

<s>Hello, I’m Dr ...</s><s>Hi, I’m ...</s><s>How can I help ...</s>

Where text from both speakers is used, models are also built using candidate and interlocutor identifier tokens, "<<c>" and "<<i>" respectively, at the start of each utterance. These tokens were added to the tokenizer dictionary.

In [15, 16] it was shown that an auto-marker based on wav2vec2 self-supervised speech representations [17] can work well for responses of 10-20s. The model structure from [15, 16] is used (convolutional neural network to encode audio, masking applied to latent representations spans, then transformer) with the mean pooling layer at the output of the transformer replaced by four intra-segment attention heads [27]. Unlike [15], multiple audio segments from a test are input to wav2vec2 models. The outputs of each model are fed into four cross-segment attention heads before the predicted score is output.

The auto-marking models are trained and tested on a role-play. In training, the same target label, the candidate’s average raw human-assessed test score, is used for both role-plays. The auto-marked score for a test is the average of the two role-play score predictions. For the Longformer model, an ensemble of M models is comprised of models with the same architecture but different initialisation seeds for the fully connected layer and some different hyperparameters. The final ensemble prediction is the average of the individual auto-marking model predictions.

4. Experimental set-up

4.1. OET data

The OET data for these experiments was collected in a number of test centres worldwide. The primary data set (OET1) was divided into non-overlapping training (train), development (dev) and evaluation (eval) sets in a 80%/10%/10% split. Table 1 gives statistics for the data sets. The candidates for the test are mostly around B2/C1 CEFR level.

<table>
<thead>
<tr>
<th># tests</th>
<th>train</th>
<th>dev</th>
<th>eval</th>
<th>OET1</th>
</tr>
</thead>
<tbody>
<tr>
<td>hrs. of speech</td>
<td>1,847</td>
<td>232</td>
<td>233</td>
<td>2,312</td>
</tr>
<tr>
<td>308</td>
<td>39</td>
<td>39</td>
<td>385</td>
<td></td>
</tr>
</tbody>
</table>

A smaller set of 59 manually transcribed tests (10 hours of speech), H-man, was used to assess diarization and ASR performance. These transcriptions only cover the within role-play speech. To assess the diarization and ASR performance, the start and end points of each role-play were manually marked and a ground truth RTTM file automatically created. Unknown words were first mapped to the closest acoustic matching word from a set of ~3,500 words using constrained recognition [28]. HTK forced alignment [29] was then run on all the utterances to yield word and utterance level timings. Audio from the Introductory stage had been removed previously from all the data.

4.2. Speaker diarization

Speaker diarization distinguishes different speakers. It is used for two main purposes here: role-play detection; and candidate/interlocutor assignment. The pre-trained speaker diarization system, version 2.0.1 [30, 31, 32], was selected as it has an in-built voice activity detection (VAD) and performs well compared to the state-of-the-art [33].

Role-play detection. This splits the recording into the two role-plays for assessment. All diarized speech segments within 10s of one another are first joined, regardless of who spoke. The two longest resulting segments are then treated as the two role-
plays. Strictly this is the role-plays plus immediate pre and post role-play conversation between the candidate and interlocutor. The role-play detection fails if all the speech is merged into a single segment, usually caused by the diarization incorrectly assigning a third speaker to the noise (background chatter and other sounds) between role-plays. The automated procedure resulted in role-play detection failing in 0.4% of test recordings. In a live system these would be given to a human to assess, so they were excluded from this research and Table 1.

Candidate/interlocutor assignment. The diarized speech segments have to be assigned to either the candidate or the interlocutor for assessment. Various methods to do this are evaluated in Section 5.1. Occasionally the diarization attributes a third speaker to some segments. Those segments are ignored in the candidate/interlocutor assignment. They are not passed to the ASR, so are excluded from training/testing the auto-marker.

In some recordings, the speaker diarization assigned all or most of the speech segments to one speaker. As this is unlikely to be correct, files where over 85% of the speech was assigned to a single speaker were counted as diarization failures. This occurred 28 times on OET1, ∼1% of the data, with 2 failures each on dev and eval, and 1 on H-man. These diarization failures were primarily caused by the candidate and interlocutor being very similar, from the same region and of the same gender, with similar pitch and other vocal features. The remainder came where the recording equipment set-up meant one speaker was much louder in the recording with the other barely audible. These are removed prior to the experiments in Section 5.

Metrics. Standard voice activity detection (VAD) and diarization error rate (DER) metrics were calculated for H-man. The VAD accuracy was 92.0% (precision 97.7%, accuracy 93.9%, F1 score 95.7%). The binary VAD’s DER was 8.3% and therefore a significant contributor to the actual DER. The DER metrics are shown in Table 2, without forgiveness collars. Although pyannote.audio identifies overlapping speech, the ASR used (Section 4.3) cannot handle it so DERs are reported excluding these regions. As can be seen missed speech is the largest error component.

<table>
<thead>
<tr>
<th>Stat.</th>
<th>Miss</th>
<th>False alarm</th>
<th>Spkr error</th>
<th>DER</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>6.3</td>
<td>4.4</td>
<td>2.1</td>
<td>12.8</td>
</tr>
<tr>
<td>stdev</td>
<td>4.2</td>
<td>2.7</td>
<td>1.6</td>
<td>5.4</td>
</tr>
</tbody>
</table>

There is considerable overlapping speech in H-man, averaging 3.6% per test with standard deviation 1.6% and a high of 7.2%, with similar levels in OET1. If overlapping speech is included the mean DER is 15.7%.

4.3. ASR
The diarized segments are passed to the open source OpenAI Whisper ASR [20, 21]. This is a transformer-based encoder-decoder model, also referred to as a sequence-to-sequence model. The models were trained on 680k hours of labelled speech data annotated using large-scale weak supervision. The transcripts include capitals and punctuation, which were retained in this research for auto-marker model inputs. The four Whisper English models (September 2022 pre-trained model release, Whisper code updated March 2023) were evaluated on H-man. The lowest WER was achieved with the medium.en model (26.8%) but the improvement over the much faster and lower memory small.en model (28.5%) was relatively small. A larger difference was observed between them and the base.en and tiny.en models (both 34.2% but different error profiles). The small.en English model was therefore adopted for all the other experiments reported in this paper. As observed by others, Whisper often hallucinated text when the diarized segments were short or contained mostly silence. Running Whisper on overlapping speech often led to the same words being attributed to both candidate and interlocutor.

4.4. Longformer-based auto-marker
A Longformer-based model is trained and applied at the role-play level. The encoder-only part of the pre-trained 12 layer Longformer model Hugging Face version 4.26.1 was used. The first vector of the final Longformer layer is passed to a fully connected regression layer of size 768 × 1 to give a single output score. The input text is tokenized and padded to the maximum length of 4,096. The training fine-tunes the pre-trained Longformer weights as well as fitting the fully connected regression layer weights. 5 models, with different regression layer initialisation weights, were trained to quantify the epistemic uncertainty and enable ensemble model testing (M = 5 in Section 3). Each model is trained on 3 trials of up to 50 epochs per trial, with learning rate reduction patience 2 and early stopping patience 4. The best model selected on each training run is the one that gives the lowest mean squared error (MSE) loss on the dev set. The test set is not used at all during the training/validation process. Each model training used a different learning rate selected at random in the range [10⁻⁶, 10⁻²], the initial weights for the final linear layer were randomized and the data was shuffled. Batches of 6 role-plays were used to avoid overloading the memory on two GPUs (Tesla V100S-PCIE-32GB), and a random gradient accumulation hyperparameter was selected from [1, 2, 4, 8]. The Adam optimization was used. Each model took 8-12 hours to train and has 107 million parameters.

4.5. wav2vec2-based auto-marker
The wav2vec2-based auto-marker applies at test level (audio segments from both role-plays). The pre-trained wav2vec2 speech model on Hugging Face is used and model attention heads added to give a model with 94 million parameters. A target segment length of 15s was selected, a bit shorter than in [16], to match OET1 lengths. The model was initialised by training for one epoch on 180k Linguaskill Part 1 learner utterances. 15s speech segments were extracted from the candidate utterances in the OET1 train set, and any residual segments over 5s were also used. For each role-play an average of 33 segments were used. The training batch size was 4 tests with gradient accumulation steps 8. The AdamW optimization [34] was used. Although it has fewer parameters than the Longformer-based auto-marker, it takes much longer to train at around 12 hours for the initial Linguaskill training, 2 hours for a single OET epoch and 80 hours in total for the final model used in Section 5.3.

5. Experimental results
5.1. Candidate/interlocutor assignment
The diarized speech segments have to be assigned to either the candidate or the interlocutor for assessment. Since this is a test, the candidate is expected to talk more than the interviewer. In
addition, the interlocutor is expected to introduce and initiate the role-play. The following approaches were therefore investigated, either per role-play or across the test, to assign them:
1. Assign candidate as speaker who speaks for the longest time.
2. Assign candidate as speaker who utters the most words.
3. Assign interlocutor as first person to speak.

The speaker attributed WER (SAWER) [35] can be used to measure the assignment accuracy. Table 3 shows that assigning the candidate as the person who spoke the most across the test (time) yields the closest SAWER to the reference. This concurs with manual inspection of the H-man, dev, and eval set assignments (1, 0 and 2 tests incorrect, respectively). The number of words also gave a reasonable approximation but simply picking the first speaker worked poorly.

Table 3: % Speaker Attributed WER (SAWER) on H-man.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Sub</th>
<th>Del</th>
<th>Ins</th>
<th>SAWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>(No Spkr Att.)</td>
<td>(7.8)</td>
<td>(8.5)</td>
<td>(12.2)</td>
<td>(28.5)</td>
</tr>
<tr>
<td>Reference</td>
<td>10.5</td>
<td>9.5</td>
<td>12.1</td>
<td>32.2</td>
</tr>
<tr>
<td>Time</td>
<td>12.1</td>
<td>9.5</td>
<td>12.1</td>
<td>33.6</td>
</tr>
<tr>
<td># Words</td>
<td>13.4</td>
<td>9.4</td>
<td>12.1</td>
<td>34.9</td>
</tr>
<tr>
<td>First Speaker</td>
<td>23.1</td>
<td>9.1</td>
<td>11.7</td>
<td>43.9</td>
</tr>
</tbody>
</table>

5.2. Longformer-based auto-marker

Table 4 shows the Longformer auto-marker performance given both candidate (Can) and interlocutor (Int) transcripts or the transcripts from only the candidate, measured against human assessor scores by root mean square error (RMSE), Pearson correlation coefficient (PCC) and Spearman’s rank coefficient (SRC). The ensemble always achieves better performance than the mean individual model scores. The small variation of individual model metrics indicates low epistemic uncertainty, which is surprising given the model performances are far from perfect.

Table 4: Individual and ensemble (Ens) auto-marking metrics.

<table>
<thead>
<tr>
<th></th>
<th>RMSE ↓</th>
<th>PCC ↑</th>
<th>SRC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can + Int</td>
<td>0.930 ± 0.014</td>
<td>0.923 ± 0.009</td>
<td>1.063 ± 0.022</td>
</tr>
<tr>
<td>Can</td>
<td>0.917</td>
<td>0.910</td>
<td>1.051</td>
</tr>
<tr>
<td>Int</td>
<td>0.570 ± 0.023</td>
<td>0.593 ± 0.008</td>
<td>0.349 ± 0.004</td>
</tr>
<tr>
<td>Ens</td>
<td>0.581</td>
<td>0.600</td>
<td>0.356</td>
</tr>
</tbody>
</table>

Table 5 shows that its performance is similar to the Longformer ensemble – worse RMSE, slightly worse PCC, better SRC. An ensemble of the two auto-markers gives substantially better performance on all metrics. This suggests that (a) there are speech characteristics relevant for scoring (e.g. pronunciation) that are lost in the conversion to text and (b) text contains information relevant for scoring not in the self-supervised speech features.

Table 5: Auto-marker performance comparisons.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE ↓</th>
<th>PCC ↑</th>
<th>SRC ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longformer Ens</td>
<td>0.910</td>
<td>0.600</td>
<td>0.525</td>
</tr>
<tr>
<td>wav2vec2</td>
<td>0.923</td>
<td>0.588</td>
<td>0.571</td>
</tr>
<tr>
<td>Longformer Ens &amp; wav2vec2</td>
<td>0.860</td>
<td>0.644</td>
<td>0.594</td>
</tr>
</tbody>
</table>

6. Conclusions

Conversational, dialogic, speaking tests are a core approach to assessing candidates’ speaking and communicative skills. Automating assessment of such tests would yield considerable savings in time and money, supporting growth in L2 English testing, and making tests more accessible to candidates (i.e., easier to scale up the assessment) and faster score reporting. This paper proposes an approach to auto-mark such tests, with the speaking part of the Occupational English Test (OET) used for experiments.

The recorded conversation, here consisting of two role-plays, is first passed to a speaker diarization system which separates the audio into speaker segments. Candidate/interlocutor labels are automatically assigned to each segment. Two foundation model-based auto-marking systems are applied. The input to the Longformer-based text neural grader is provided by ASR, OpenAI’s Whisper, run on the diarized segments. It was found that using only the text of the candidate’s speech yielded more accurate score predictions. To cover speech characteristics such as pronunciation, a speech neural grader based on wav2vec2 was run directly on the candidate’s audio. Individually the two auto-markers perform similarly. Combining them in an ensemble led to improvements across all auto-marking metrics. The final PCC of 0.644 is a reasonable starting performance to build future auto-marker systems upon.
7. References


