[Automatic Proficiency Assessment in L2 English Learners](https://arxiv.org/html/2505.02615v1)

**Objective**:  
This paper explores deep learning techniques for comprehensive L2 proficiency assessment, addressing both the speech signal and its correspondent transcription.

Current systems primarily focus on structured speech tasks, limiting their applicability to unscripted communication. Deep learning models, while effective in structured speech classification, require further investigation in spontaneous settings. Similarly, adapting text-based models like BERT for spoken language presents unique challenges due to differences in discourse and tokenization.

Dialogic speech assessment remains a significant hurdle. Evaluating interactional competence and real-time responsiveness requires models beyond traditional metrics

A screenshot of a computer program

AI-generated content may be incorrect.

**Architectures**:

* We analyze spoken proficiency classification prediction using diverse architectures, including 2D CNN, frequency-based CNN, ResNet, and a pretrained wav2vec 2.0 model.
* Text-based proficiency assessment by fine-tuning a BERT language model within resource constraints.

**Related Work:**

Previous models showed the potential of AI models in automating the assessment process. However, early models had limitations due to manual feature engineering and ASR challenges. Deep learning advancements enabled end-to-end learning from raw data.

In text-based assessment, BERT-based frameworks have been explored to enrich contextual embeddings, achieving high accuracy on large datasets [[27](https://arxiv.org/html/2505.02615v1#bib.bib27), [28](https://arxiv.org/html/2505.02615v1#bib.bib28)]. Multimodal approaches, like combining wav2vec 2.0 and Longformer, have further improved dialogic assessment [[6](https://arxiv.org/html/2505.02615v1#bib.bib6)].

While several studies have explored wav2vec 2.0 for L2 assessment, many rely on speaker-dependent setups.

**Method:**

1. ***Preprocessing and Segmentation***

Models like 2D CNNs and ResNet require the pre-processing of the speech signal. The preprocessing pipeline consisted of multiple stages, including audio downsampling to 16 kHz, segmentation into 8-sec excerpts for models requiring fixed length inputs, feature extraction using FBanks with 40 Mel-frequency bins, and input normalization using z-score.

1. ***Dialogue-Specific Preprocessing***

Two datasets were created: one with full dialogues and another containing only learner speech.

The recordings lacked transcripts, they used Whisper to generate automatic speech transcriptions and evaluated multiple variants to identify the most accurate for our data. Despite its strong performance, Whisper occasionally produced errors due to overlapping speech and real-world noise conditions. A subset of the data was manually reviewed to verify segmentation accuracy.

**Datasets:**

ANGLISH 🡪 three proficiency levels:

native English speakers (NES), the highest proficiency group, including 13 females and 10 males with an average age of 31 years from various regions in England;

French speakers subdivided into 11 females and 11 males with an average age of 37.5 years without specialized phonetics training (FR1)

university students, 11 females and 11 males, aged between 19 to 22 years with formal phonetics training (FR2).

**Speech tasks:**

* **Reading passages:** 4 texts per speaker, ~1.5 hours, 1,260 sentences.
* **Spontaneous monologues:** 63 speakers, 4-min talks on pre-suggested topics (e.g., holidays).
*  **Recording conditions:** All in **anechoic chamber** (high-quality, noiseless audio).
*  **Annotation:** Transcribed and segmented at **phoneme level** using PRAAT; includes pauses, hesitations, and truncated words.

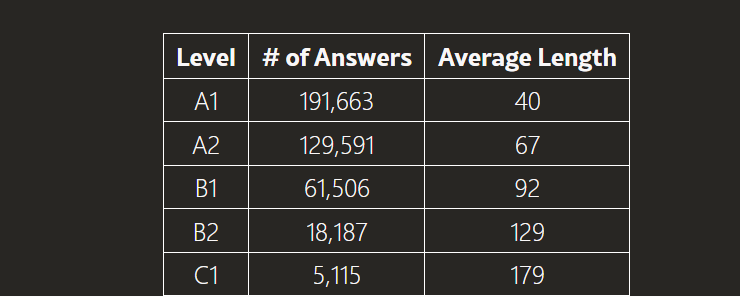
**During Training:**  
10-fold cross-validation (CV) using 61 speakers. Stratification ensured balanced gender and proficiency representation across folds. test set composed of 6 speakers (one male and one female from each proficiency level), never seen during training or validation, to assess the generalization of the models.

**EFCamDat:**

The EF-Cambridge open language dataset comprises 1,180,310 essays written by learners (172 nationalities) and evaluated by professional language instructors. The dataset represents 16 different proficiency levels mapped to five common European framework of reference for languages (A1-C1). Each essay is accompanied by rich metadata, including the learner’s native language, age, gender, and time spent learning English.

**During Training:**

carefully curated subset with 2,000 samples for the training set, 200 samples for the validation set, and 200 samples for the test set. To ensure representativeness and fairness, they stratified the subset according to the 16 proficiency levels present in the full dataset.



**Private:**

A black screen with white text

AI-generated content may be incorrect.

A black and white table with numbers and text

AI-generated content may be incorrect.

**Training**:

To ensure a fair and robust evaluation, we partitioned the data into training, validation, and test sets. For each level, we randomly selected 12 speakers for the validation and test sets, ensuring that their total audio duration matched 10% of the total duration in L5, the least represented level. The remaining speakers were used for training. Crucially, we enforced strict speaker separation across partitions, ensuring that no speaker appeared in both training and evaluation sets. This approach guarantees that model performance is assessed on entirely unseen speakers, providing a more reliable measure of generalizability.

**Results:**

A black and white chart with white text

AI-generated content may be incorrect.

A screen shot of a black screen

AI-generated content may be incorrect. A black and white table with numbers and text

AI-generated content may be incorrect.

A black and white table with white text

AI-generated content may be incorrect.

**Recommendation:**

Future research could explore multimodal fusion strategies, integrating speech and text-based features to capitalize on their complementary strengths.