Automated L2 Speech Assessment with Audio-Based Self-Supervised Neural Networks

Eli Stafford[1,2*], **Sylvain Coulange**[2,3,4]

[1]ALMAnaCH, INRIA Paris, Paris, France

[2]Service des Langues, Université Grenoble Alpes, Grenoble, France

[3] Laboratory of Linguistics and Didactics of Foreign and Mother Tongues (LIDILEM), Université Grenoble Alpes, France

[4] CNRS, Institute of Engineering, Grenoble Computer Science Laboratory (LIG), Université Grenoble Alpes, France

eli.stafford@inria.fr, sylvain.coulange@univ-grenoble-alpes.fr

Abstract

Automated L2 speech assessments traditionally use vectors of acoustic, lexical, and syntactic metrics as representations of speech for statistical learning. This work explores the use of audio-based self-supervised neural networks to supplement existing scoring strategies with robust computational representations. We explore architectural considerations of neural L2 scoring, probe their effects in the context of a reliable and equitable examination, and measure their linguistic coverage compared to heuristic-based and neural text-based systems. We find that their use in automated speech scoring corresponds to a measurable improvement in performance, while revealing evidence of major drawbacks in terms of explainability and negligible increases in robustness compared to traditional systems, especially in low-resource contexts.

1 Introduction

L2 automatic speaking proficiency assessment is an important advancement in second language education that enhances the practicality and deliverability of speech assessments facing ever-growing demand (Van Moere and Downey, 2016). Private testing bodies that employ automated speech assessments, such as Pearson (with the Test of English Academic), Cambridge University Press & Assessment (with the Cambridge English Testing System), or Duolingo, are now widely used. While publicly available descriptions of their assessment methods cite the usage of statistical learning with vectors of a variety of computed heuristics on large, professionally graded student corpora (Saito et al., 2023; Xu et al., 2021; Park et al., 2025), they do not mention the usage of neural methods in text or speech.

In the academic literature, machine learning models show a strong correlation with human judgments on this task (Qian et al., 2025a; Zechner and Evanini, 2019). The majority of competitive systems utilize cascaded architectures, with transcriptions of student speech (produced by an automatic speech recognition model like Whisper (Radford et al., 2023)) fed into text-based classifiers built on pre-trained text encoders or decoder-only Large Language Models (Qian et al., 2025a; Bannò et al., 2025; Wang et al., 2021). Although there have been explorations of multimodal models (Ma et al., 2025) and correlations of transcription model outputs with pronunciation accuracy (Ballier and Méli, 2024; Liu et al., 2025), the use of large pre-trained audio encoders that consider only speech without a transcription, such as Wav2Vec2.0 (Baevski et al., 2020) and WavLM (Chen et al., 2022), remains understudied, with few experiments investigating their application (Bannò and Matassoni, 2023).

These pre-trained audio encoders are encoder-only deep neural networks built on the transformer architecture (Vaswani et al., 2017), trained on unlabeled audio inputs to produce high-dimensional representations of speech. These self-supervised learning models (SSLs), also called foundation models, can be applied to a variety of tasks by adding specialized classification heads to their base architecture. They achieve high performance across a wide variety of spoken language processing benchmarks (Yang et al., 2024), including speaker identification, emotion recognition, speaker diarization, and low-resource automatic speech recognition (Yi et al., 2021). L2 speaker recognition, as an utterance-level classification task, is a prime use case for audio SSLs and has the advantage of targeting the speech quality of

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

*This work was carried out during an internship at Service des Langues.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 334–345, November 19–20, 2026, University of Tartu Library

an L2 language learner, which is unavailable in the text-only setting. Compared to state-of-the-art speech-based heuristics (Coulange, 2025), which are calculated directly from audio to augment text-only systems and are based on linguistic intuitions of speech quality, audio SSLs can learn the most pertinent features of speech automatically.

However, the usage of any neural model comes with a host of ethical concerns regarding validity, practicality, and transparency (De Jong et al., 2025; Vajjala et al., 2026). Audio SSLs suffer both from classic questions of validity and dataset overfitting due to their black-box nature (Tsai et al., 2025), as well as questions of speaker group bias (Herron et al., 2026, 2025). Utilizing these systems for reliable and equitable examinations requires close examination of these factors.

Further, while other tasks have well-defined architectures to leverage pertinent encoded features, such as graph-based mechanisms for synthetic voice detection (Jung et al., 2022) and temporal average pooling for speaker identification (Wang et al., 2024), there is no agreed-upon architecture for L2 spoken language assessment when using audio encoder models.

In this work, we explore a wide range of possible architectures for audio SSLs based on their application to other tasks, with the goal of optimizing the aggregation of information over time (phoneme level to speaker level) and the abstractness of representation at each model layer (low-level acoustic information to abstract speech representation). We then compare the performance of these models to heuristic-based and text-only neural baselines, and combine their usage to see if audio SSLs are a promising supplement to existing pipelines by targeting synergistic aspects of speech. Additionally, we explore the information encoded in our best-trained model via layerwise probing to determine possible avenues of exploitation of non-linguistic information, as well as the model's encoding of linguistic features like pragmatics, pronunciation, and grammar.

2 Architectures

While audio SSLs can sometimes be used end-to-end by feeding the final layer as input into a classifier head, many tasks require preliminary aggregation of model representations. As SSLs are deep neural networks given audio as input (a sequential medium), each input produces an embedding vector for each time slice (usually ∼20 ms of audio) at each hidden layer. A model like Wav2Vec2.0, with 13 layers, given an input of 4,000 time slices (∼80 seconds of audio), produces 52,000 available embeddings, each representing a specific slice of audio at a specific hidden layer (level of abstraction (Rogers et al., 2020)). We separate possible aggregation approaches into *Layerwise Aggregation* approaches, which aggregate each model layer into a single dimension, and time-wise *Pooling* approaches, which aggregate a sequential audio input into a single representation. Together, these approaches produce a single embedding vector, which is fed to a feed-forward regression head to predict student level.

Layerwise Aggregation strategies seek to combine disparate representational information across layers into a single output. Embeddings in lower layers of an SSL generally contain raw acoustic information, while higher layers contain higher-level learned abstractions like word classes or intonation. We experiment with three major aggregation strategies, which can be applied independently of pooling strategies along the time dimension:

- Layerwise: The simplest of our approaches trains a classifier head on each layer of our frozen SSL to output a global score for that layer, and then trains a linear regression classifier on these scores.

- Weighted Mean: This approach, standard for frozen SSLs in utterance classification (Yang et al., 2024), trains a vector with length equal to the number of layers in our SSL during fine-tuning, and performs weighted mean averaging based on these weights. The weights for each layer are learned at training time.

- FT/PEFT: In this approach, we feed only the final layer of our SSL into a classification head and leave all previous layers unfrozen for training; in a generic sense, this can be called *full fine-tuning* (FT). Conventional full fine-tuning of audio SSLs is prone to catastrophic forgetting, especially in low-resource settings, so we opt for a Parameter-Efficient Fine-Tuning (PEFT) method, namely Low-Rank Adaptation (LoRA) (Hu et al., 2022) on the query and key matrices of each transformer layer.

Pooling approaches aggregate sequences of vectors (along the time dimension) into a single embedding representing the entire utterance. We experiment with three pooling approaches, applied after layerwise aggregation (or directly on layers in the case of the multilayer strategy):

- Mean Pooling: The simple average of each embedding across the time or token dimension. This erodes information specific to a time slice (such as particular mispronunciations or accent features) in favor of a generalized representation.
- Attention Pooling: Each vector is assigned an attention value by a linear feed-forward layer, and we compute a weighted average based on the softmax of these values. This allows the model to select pertinent time slices or tokens based on their learned relevance.
- Four-Head Pooling: The third approach is a specific variant of attention pooling utilizing a more complex attention mechanism with four heads instead of one, based on the baseline approach of the Speak and Improve Challenge (Qian et al., 2025b). This approach considers the entire context of the phrase when determining the weight of a time slice, doing so with four distinct heads. The resulting weighted means are concatenated, and a non-linear regression head utilizes this vector to predict student level. This allows us to have multiple specialized heads that consider different non-linear relationships across parts of the utterance.

Probing models (probes) are secondary, auxiliary models trained on top of a pre-trained primary model to test whether specific information is encoded within the primary model's internal representations. Probes can identify the presence of specific information in model outputs by training on secondary classification objectives, where high performance suggests the presence of the information necessary to complete that task. Previous explorations of audio encoder layers have found that these models generally do not encode non-acoustic linguistic information like word meaning (Pasad et al., 2021), therefore any spoken language assessment system solely comprised of these models will inherit this acoustic bias. To determine which linguistic features of speech our best system encodes, we probe model layers for phonetic, grammatical, and pragmatic fluency according to holistic scoring data provided in our dataset. Furthermore, to quantify speaker group bias in our best-performing model, we train probes for each model layer on a speaker identification task, where each input is classified with the student's speaker ID. Finally, we probe for task identification, which identifies the specific task being performed by the student since our dataset includes recordings of different tasks. From a pedagogical standpoint, student speaker features and the task being performed have a non-causal correlation with the proficiency level of an utterance and should not be prominently present in our model's outputs.

3 Experiments

3.1 Dataset

Models were evaluated on a novel dataset collected for the SELF Production Orale project[*]. It consists of 873 recordings of spontaneous speech from 339 L2 learners of English responding to 10 diverse tasks designed for this project. Each recording is rated on a continuous scale between 1.0 and 7.0, with integer values representing discrete CEFR levels[†]. Fourteen raters, comprising CLES[‡] certified raters and teachers of English as a second language, were recruited to create gold ratings for these recordings. At 150 ratings per rater, this resulted in 2,100 ratings (an average of 2.4 ratings per recording), which were then normalized and averaged to produce continuous, reliable gold ratings.

It is important to note that this data suffers from class imbalance at extreme ranges, with $\sim 10\%$ of all data at the A1 and C1+ levels (Appendix A). Determining if audio SSLs can deal with this imbalance is a goal of this work, as statistical learning of heuristics do not handle this imbalance sufficiently.

In a *holistic scoring* paradigm, students were also rated on three *sub-scores* in addition to their overall score. These sub-scores, on a A1–C2 categorical scale, measure more specific axes of linguistic competency, including Pragmatics, Vocabulary, and Pronunciation.

[*] https://self.univ-grenoble-alpes.fr/self-production-orale-1664940.kjsp

[†] $1.0 \leq x < 2.0$ for A1, $2.0 \leq x < 3.0$ for A2, up to $5.0 \leq x$ for C1+

[‡] https://www.certification-cles.fr/

3.2 Architecture and Hyperparameters

We choose model and architecture combinations for evaluation on this dataset based on best-performing previous explorations of their application to automated L2 assessment and, where literature is absent, their performance on related tasks.

We test the proposed architectures on Wav2Vec2.0, the only previously used SSL for this task, and WavLM-Large, a much larger and more performant pre-trained audio encoder. We explore all possible combinations of the aggregation strategies previously mentioned on Wav2Vec2.0. WavLM-Large is prohibitively expensive for full fine-tuning, even with PEFT methods, so we experiment only with the multilayer and weighted-mean layer aggregation methods, while testing all pooling methods.

After aggregation and pooling, audio models were passed to a two-layer regression head with a 256-dimensional intermediate layer and a dropout rate of 0.1. Dropout, learning rate, optimizer choice, and intermediate layer size were determined via layerwise prediction results on the target task using mean pooling, comparing results for the best-performing layer under each configuration.

3.3 Baseline System

Our heuristic baseline system represents student speech with a vector of 51 linguistic metrics, input into a Linear Mixed Model predicting student level. These 51 metrics come from a wide range of literature on automatic L2 spoken language assessment and include surface-level heuristics such as pause frequency and position, word count and word per second, measurements of vowel stress contrast, intelligibility scores derived from ASR system confidences, syntactic complexity, and the proportion of vocabulary present in the A1 Oxford Learner's Dictionary. This wide range of scores, covering pronunciation, intelligibility, grammar, and lexical context, serves as a strong baseline against which to compare our audio-based neural system, specifically its acoustic bias.

Additionally, we provide a BERT-based baseline trained on our data using the architecture from Qian et al. (2025a). This system is text-only, using BERT-base (Devlin et al., 2019) as a backbone model and the same four-head classifier head as in our audio model experiment.

3.4 Evaluation Methodology

Models are evaluated on continuous outputs using the standard spoken language assessment metrics of RMSE (Root Mean Squared Error) and PCC (Pearson Correlation Coefficient) against continuous gold predictions. We additionally evaluate accuracy and macro F1 after binning continuous outputs into discrete CEFR classes. While continuous metrics are standard and better represent overall model performance, our binning metrics highlight specific weaknesses across CEFR levels, specifically performance on outlier classes with macro F1 score. We consider the linguistic heuristic system to be our primary baseline with respect to these metrics.

Due to our small dataset, we evaluate each of our models using a stratified 10-fold cross-validation strategy, where ten 10% data samples are withheld during training for evaluation. Each of these 10 samples is balanced in terms of CEFR level (binned rating). All metrics proposed are the average of performance on each fold. As student level and recording level are highly correlated, we ensure that recordings featuring the same student are never split across training and test sets.

3.5 Probing

All probing tasks were carried out with a two-layer linear probe featuring an intermediate layer size of 256. Each task was formulated as a classification task, with target classes being sub-score CEFR levels, task IDs, or speaker IDs. Sub-score (Pragmatics, Pronunciation, and Grammar) and task ID probes were trained on a single train/test split (fold) taken from other experiments, whereas speaker ID probes were trained on a modified split ensuring each student present in the training set had at least one recording in the test set.

3.6 Combination

To determine whether pre-trained SSLs predict based on the same information as the baselines—and are therefore alternative rather than synergistic approaches—we experiment with combinations of scores from the heuristic-based baseline model, the neural text-based model, and the best audio SSL model. The combinations of neural text-based baseline + audio SSL, our heuristic baseline + neural text-based baseline, and all 3 scores combined, we train a linear regression model with the combined scores as vector

Model	Aggregation	Pooling	RMSE	PCC	Acc	Macro F1
wav2vec2.0	lora	meanpool	.626	.760	51.3	.425
		attnpool	.632	.757	51.5	.409
		fourhead	.651	.739	51.6	.413
	wmean	meanpool	.695	.691	50.3	.394
		attnpool	.704	.683	49.0	.371
		fourhead	.640	.748	53.6	.466
	multilayer	meanpool	.606	.778	55.4	.471
		attnpool	.586	.794	54.2	.482
		fourhead	.611	.774	55.0	.497
wavlm	**wmean**	meanpool	**.564**	**.814**	**57.2**	**.499**
		attnpool	.633	.756	50.1	.460
	multilayer	meanpool	.643	.745	51.3	.424
		attnpool	.693	.694	47.8	.357

Table 1: Audio SSL Architecture Comparison

input (either 2 or 3 scores) and the gold continuous scores as targets. Each regression model is fit on the same training data as our main models and evaluated on the unseen test set.

4 Results

4.1 Audio SSL Architecture

Here we present results for Wav2Vec2.0 and WavLM on our primary task across architecture combinations.

We find that while there are clearly optimal strategies for each model, there is no single best layerwise aggregation strategy across models, nor a best pooling strategy across layerwise aggregation strategies (Table 1). WavLM showed a 0.11 decrease in RMSE (and a corresponding improvement across other metrics) when using the weighted-mean strategy compared to the multilayer strategy, whereas Wav2Vec2.0 achieved a 0.1 decrease in RMSE when using the multilayer strategy over the weighted-mean strategy. These architectures are conceptually distinct: weighted mean targets utterance-level features, while multilayer selects specific time slices and layers for aggregation. This points to a fundamental difference in necessary architectural considerations across different model sizes.

WavLM-Large with the weighted-mean and mean-pooling architecture achieved the highest performance, reaching 0.499 macro F1, 0.564 RMSE, and 0.814 PCC, representing a 0.038 decrease in RMSE compared to Wav2Vec2.0's best combination. While the RMSE values for the mean pooling and attention pooling architectures were comparable (a 0.016 difference in RMSE), gains in macro F1 were more significant (0.05).

4.2 Inter-Approach Comparisons

Included Scores	RMSE	PCC	Acc %	Macro F1
Heuristic LMM	0.654	0.758	50.5	0.524
WavLM-Large	**0.564**	**0.814**	**56.6**	0.499
Neural text Baseline	0.694	0.704	51.6	0.400
WavLM-Large + Neural Text	0.617	0.784	50.5	0.514
LMM + Neural Text	0.637	0.768	51.8	0.535
All	0.590	0.808	56.0	**0.567**

Table 2: Score Comparison/Combination

Here we compare the results of our heuristic-based baseline model, the neural text-based model, and the best audio SSL model, as well as all possible combinations. All models were evaluated on a single kfold iteration. The analysis of consistency over multiple iterations, with different splits, can be found in Appendix B.

The performance of our audio SSL strategy, WavLM-Large with weighted mean layerwise aggregation and mean pooling, outperforms the baseline system (Table 2) by 0.9 RMSE and 0.046 PCC, a significant margin. This is expected when using such a large model, but the gains are less dramatic compared to our binning statistics— with an under 3-percentage-point increase in accuracy, and a 0.025 decrease in macro F1 score. While the decrease in RMSE is promising, the lack of significant improvement in F1 score (which indicates fairness across CEFR levels) suggests that this performance increase may stem from better modeling of the unbalanced dataset distribution, or other statistical artifacts, rather than improved representa-

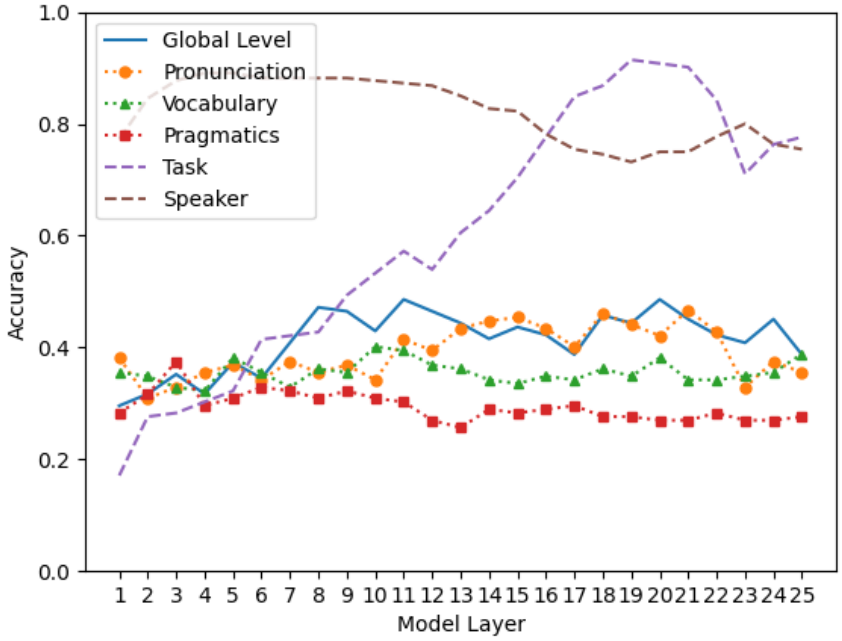


Figure 1: Layerwise holistic scoring, speaker, and task probing compared to base task: WavLM-Large

tional generalizability.

All approaches, except for the neural text-based baseline, fall in the range of 0.56–0.65 RMSE, indicating an average error of just over half a CEFR level. Similarly, accuracies range between 50% and 56%, and macro F1 scores between 0.50 and 0.57. This narrow performance range indicates either that baseline and neural models are functionally equivalent (leaving little room for improvement) or that dataset size imposes a performance ceiling.

When combining the heuristic and neural text-based baselines, we observe that despite its weak standalone performance, the BERT model adds appreciable value (a 0.017 decrease in RMSE).

WavLM-Large alone achieves the lowest RMSE at 0.564, whereas combining all scores leads to the highest absolute accuracy and macro F1 score, alongside a 0.026 increase in RMSE compared to WavLM-Large alone. We believe this demonstrates that none of the three evaluated approaches are redundant, offering measurable performance gains; however, tradeoffs between CEFR coverage and RMSE remain necessary due to dataset imbalance. Analysis of predictor estimates (relative importance of model scores/heuristics in the all scores model) can be found in Appendix C.

5 Speech Encoder Probing Results

Figure 1 displays layerwise probing accuracy results on WavLM-Large for our target CEFR predictions, holistic score classification, task identification, and speaker identification.

Holistic score probing yields results aligned with SSL intuitions. WavLM-Large probes predict Pronunciation accuracy with performance similar to overall level prediction (both Pronunciation accuracy and Global Level accuracy hover around 50% at layer 18), while poorly predicting semantic or discourse-level information (Vocabulary and Pragmatics remain below 40% accuracy). This suggests that the model does not account for a broad range of linguistic competencies when predicting recording levels, focusing instead primarily on acoustic properties of speech.

Our speaker identification probe aligns with previous findings. On the VoxCeleb speaker identification benchmark (Nagrani et al., 2017), WavLM-Large achieves accuracies exceeding 90%. This indicates that if any correlation exists between speaker identity and CEFR level, classifiers built on large models like WavLM-Large can exploit non-linguistic speaker characteristics to predict CEFR scores. This risks poor generalization to unseen speakers and is pedagogically undesirable.

Interestingly, our task probes (question ID) were able to accurately predict the task a stu-

dent was performing in the final layers of the model, achieving up to 90% accuracy. This is a notable result: although WavLM-Large excels at keyword identification (a key signal for task type) under other classifier configurations, the time-wise mean-pooling approach evaluated here would theoretically remove identifying task information. This points to a potential prediction strategy where the model identifies student level (approximated via speaker characteristics) alongside the task being performed—features that correlate highly with recording level but non-causally with the underlying linguistic proficiency.

6 Discussion

By continuous metrics, WavLM-Large is our best performing model by a significant margin with 0.9 RMSE below baseline, and 0.4 RMSE below the linear combination of all scores. However, its usage comes with a host of issues that make it difficult to justify in this context, including lack of improvements at extreme class ranges (macro F1 score), speaker bias, predictor coverage, and computational efficiency.

As far as our data suggests, there is no definitive optimal architecture for SSLs. Our results, only examining two models, give contradictory evidence for each, where Wav2Vec2.0's optimal architecture targets extremely specific production information (specific layers at specific timesteps), whereas WavLM-Large aggregates all utterance information into a specific vector. Without experiments on more models, we can only tenuously say that smaller models (Wav2Vec2.0) work better with attention-based architectures, and larger models (WavLM-Large), with hypothetically better or more consistent representations, are better with more severe aggregation.

Our probing results follow from previous explorations in audio SSL tendencies, and suggest that the information they consider at prediction time may not be entirely linguistic. These results, in addition to the lack of large macro F1 improvement indicate that our best audio SSL may be capitalizing on statistical artifacts related to task type and speaker distribution to find their improvements, rather than discovering more pertinent linguistic data that is not covered by our heuristic or text-based models.

We hoped that the introduction of deep neural networks would easily contend with extreme CEFR and levels and lead to a large increase in macro F1 score. It is clear now that SSLs suffer from the same dataset dependence as heuristic based models. Simply collecting more A1/C1+ levels, or perhaps native-level responses to the same tasks, would probably improve performance for all approaches at these levels, and at this time it is not clear if neural approaches would perform significantly better than traditional approaches if given more data.

7 Conclusion

In this work we have explored the possible uses of SSLs in the context of an automated assessment of L2 learners of English. We determined the best architectures for leveraging pre-trained audio encoder models for scoring spontaneous speech utterances, the fact that this decision changes between base models, and found slight improvements over traditional heuristic baseline and neural text-based approaches. Additionally, we find evidence that learned representations of text and speech can be used to improve heuristic-based systems, suggesting that their coverage of linguistic competencies differs in some way from traditional systems, that this coverage would be limited to acoustic phenomena, but that these models have the capacity for utilizing non-causal correlations between both the task being performed and speaker bias to determine recording level. Overall, audio SSLs represent a promising avenue of work in the creation of automated speech assessments, but their merely minor performance improvement still leave much room for debate about their nature as black-box systems, and their adherence to principles of language education.

8 Further Work

The use of zero-shot systems (text-based/multimodal LLMs) are a possible avenue of improvement with respect to extreme class levels, which could identify these outliers without more annotated A1/C1+ data, which is scarce in publicly available corpora.

While a major goal of this work was to see if audio SSLs could improve performance on an unbalanced dataset, our results regarding architecture choices must be verified on balanced data. For our conclusions to generalize across datasets, it will be important to attempt the same experiments in a more balanced setting, as the one utilized in the

Speak and Improve Challenge (Knill et al., 2025).

On the black-box nature of these systems, there is of course a large body of literature in interpretability which could apply in this setting, especially input-output oriented approaches with existing disfluency annotated corpora. Certain architectures covered (namely weighted mean layerwise pooling and timewise attention pooling) are ripe with data for close analysis on which layers and specific utterance contribute most to a recordings rating. This type of interpretability is important for justifying their use to educators, and can give insights to students without more complex error annotating models.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: Generative AI was utilized for grammar correction and spellchecking for the production of this document.

Acknowledgments

This work was carried out in the context of the SELF Production Orale project, headed by Monica Masperi, hosted between Service des Langues and Laboratoire d'Informatique de Grenoble of Université Grenoble Alpes. This project was funded by a FITInnovE grant provided by the Pôle Universitaire d'Innovation of Université Grenoble Alpes, with additional substantial funding from the Service des Langues of Université Grenoble Alpes.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Nicolas Ballier and Adrien Méli. 2024. Investigating acoustic correlates of whisper scoring for L2 speech using forced alignment with the Italian component of the ISLE corpus. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 20–32, Rennes, France. LiU Electronic Press.

Stefano Bannò, Rao Ma, Mengjie Qian, Siyuan Tang, Kate Knill, and Mark Gales. 2025. Natural language-based assessment of L2 oral proficiency using LLMs. Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE).

Stefano Bannò and Marco Matassoni. 2023. Proficiency assessment of L2 spoken English using wav2vec 2.0. In *2022 IEEE Spoken Language Technology Workshop (SLT)*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sylvain Coulange. 2025. *Évaluation automatique de la parole spontanée en anglais langue étrangère : le rôle des pauses et de l'accent lexical dans la compréhensibilité du locuteur*. Theses, Université Grenoble Alpes.

Nivja De Jong, Stephan Raaijmakers, and Dineke Tigelaar. 2025. Developing high-quality, practical, and ethical automated L2 speaking assessments. *System*, 134:103796.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Felix Herron, Solange Rossato, Alexandre Allauzen, Benoit Favre, and François Portet. 2025. *Speaker Group Encoding in Self-supervised Speech Recognition Models*, pages 121–132. Springer Nature Switzerland.

Felix Herron, Solange Rossato, Alexandre Allauzen, and François Portet. 2026. Identifying and typifying demographic unfairness in phoneme-level embeddings of self-supervised speech recognition models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33603–33620, San Diego, California, United States. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6367–6371.

Kate Knill, Diane Nicholls, Mark J. F. Gales, Mengjie Qian, and Pawel Stroinski. 2025. Speak and improve corpus 2025: an L2 English speech corpus for language assessment and feedback. Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE).

Yao Liu, Faizahani binti Ab Rahman, and Farah binti Mohamad Zain. 2025. A systematic literature review of research on automatic speech recognition in EFL pronunciation. *Cogent Education*, 12(1):2466288.

Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Assessment of L2 oral proficiency using speech large language models. In *Proc. Interspeech 2025*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In *Proc. Interspeech 2017*, pages 2616–2620. ISCA.

Yena Park, Ramsey Cardwell, Sarah Goodwin, Ben Naismith, Geoffrey T. LaFlair, Kevin P. Yancey, and Xiaowan Zhang. 2025. Assessing speaking on the Duolingo English Test. Whitepaper, Duolingo.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Mengjie Qian, Kate Knill, Stefano Bannò, Siyuan Tang, Penny Karanasou, Mark J. F. Gales, and Diane Nicholls. 2025a. Speak and improve challenge 2025. In *Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE 2025)*. ISCA.

Mengjie Qian, Kate Knill, Stefano Bannò, Siyuan Tang, Penny Karanasou, Mark J. F. Gales, and Diane Nicholls. 2025b. Speak and improve challenge 2025: Tasks and baseline systems. Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Kazuya Saito, Konstantinos Macmillan, Magdalena Kachlicka, Takuya Kunihara, and Nobuaki Minematsu. 2023. Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, 45(1):234–263.

Fong-Chun Tsai, Kuan-Tang Huang, Bi-Cheng Yan, Tien-Hong Lo, and Berlin Chen. 2025. Mitigating data imbalance in automated speaking assessment. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 688–693. IEEE.

Sowmya Vajjala, Bashar Alhafni, Stefano Bannò, Kaushal Maurya, and Ekaterina Kochmar. 2026. Opportunities and challenges of LLMs in education: An NLP perspective. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 368–404. Association for Computational Linguistics.

Alistair Van Moere and Ryan Downey. 2016. Technology and artificial intelligence in language assessment. In Dina Tsagari and Jayanti Banerjee, editors, *Handbook of Second Language Assessment*, pages 341–358. De Gruyter Mouton.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Shuai Wang, Zhengyang Chen, Kong Aik Lee, Yanmin Qian, and Haizhou Li. 2024. Overview of speaker modeling and its applications: From the lens of deep speaker representation learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4971–4998.

Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew Mulholland. 2021. Automated scoring of spontaneous speech from young learners of english using transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712.

Jing Xu, Edmund Jones, Victoria Laxton, and Evelina Galaczi. 2021. Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education: Principles, Policy & Practice*, 28:411–436.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu-hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Shang-Wen Li, Abdelrahman Mohamed, Shinji Watanabe, and Hung-yi Lee. 2024.

A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2884–2899.

Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2021. Applying wav2vec2.0 to speech recognition in various low-resource languages. arXiv preprint arXiv:2012.12121.

Klaus Zechner and Keelan Evanini. 2019. Overview of automated speech scoring. In *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*, pages 3–20. Routledge.

Appendix A Dataset Imbalance

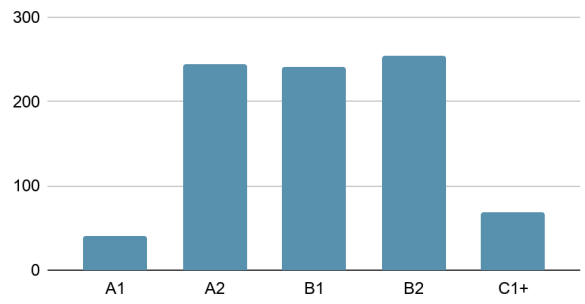


Figure 2: Distribution of CEFR levels in the final dataset by recording

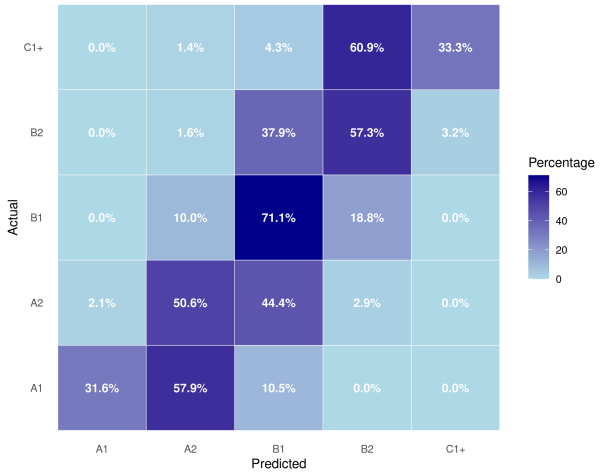


Figure 3: Confusion matrix of All Scores Model

Appendix B Effect of Train/Test Splits

Included Scores	RMSE	PCC	Acc %	Macro F1
Heuristic LMM	±0.003	±0.003	±0.3	±0.004
WavLM-Large	±0.020	±0.023	±1.9	±0.032
All	±0.034	±0.12	±3.6	±0.014

Table 3: Score Comparison/Combination

3 presents the variance of each approach by our four metrics. These metrics were collected from 5 different 10-fold iterations, where we combine all 10 test folds into a single set of predictions for comparison with our gold labels. Each value represents a 95% confidence interval. We find that while the the overall variance is not enough to affect our conclusions, there is a stark difference between the dataset dependence of our Heuristic LMM and neural speech approaches. The 10 times increase in variance indicates that neural speech models are much more susceptible to train/test splits than the simple regression model used by our heuristic LMM, at least in our low-resource setting.

Appendix C Predictor Estimates

Figures 4 and 5 show the measurable effect of each heuristic on prediction outcomes, firstly on the original heuristic baseline, and then when combining the heuristic baseline, the neural text-based model (`bert_score`) and the audio SSL model (`wavlm_score`). The scores produced by our deep neural models consistently receive the highest prediction estimates, often attenuating the weights of metrics previously estimated as highly significant. This suggests that while other scores remain important, our neural models implicitly capture their information during prediction. This does not make non-neural scores redundant, and performance could potentially improve further with a non-linear combination model.

While deep neural models exhibit very strong predictor estimations compared to most traditional scores, we believe it is not likely that their scores are strictly adding to coverage. When introducing SSL based scoring to the baseline model, we find that some of the scores which have the most *reduction* in predictor estimates are surface level metrics like number of words and articulation rate. These two metrics, the 2nd and 10th most significant scores in Figure 4, are reduced to the 11th and 24th most significant in Figure 5. While replacing these high estimators with more pedagogically pertinent metrics is always preferable, it seems that the scores produced by deep neural models rely on the same surface level information during prediction. It is true however, that some metrics retain high predictor estimates, like pause frequency.

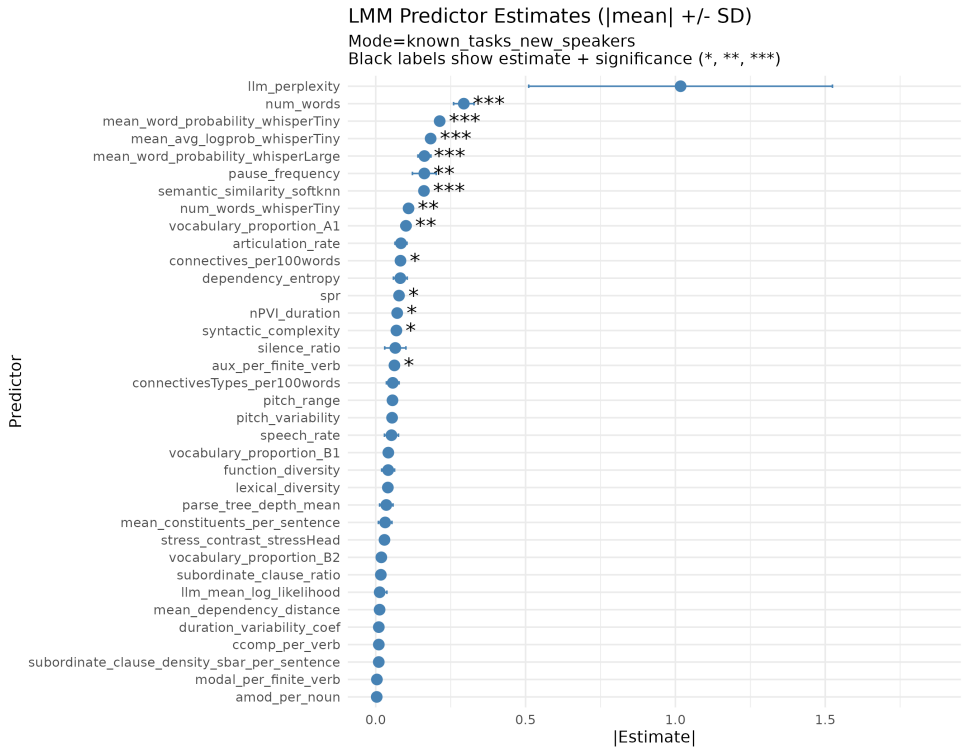


Figure 4: Predictor estimates of the heuristic based baseline

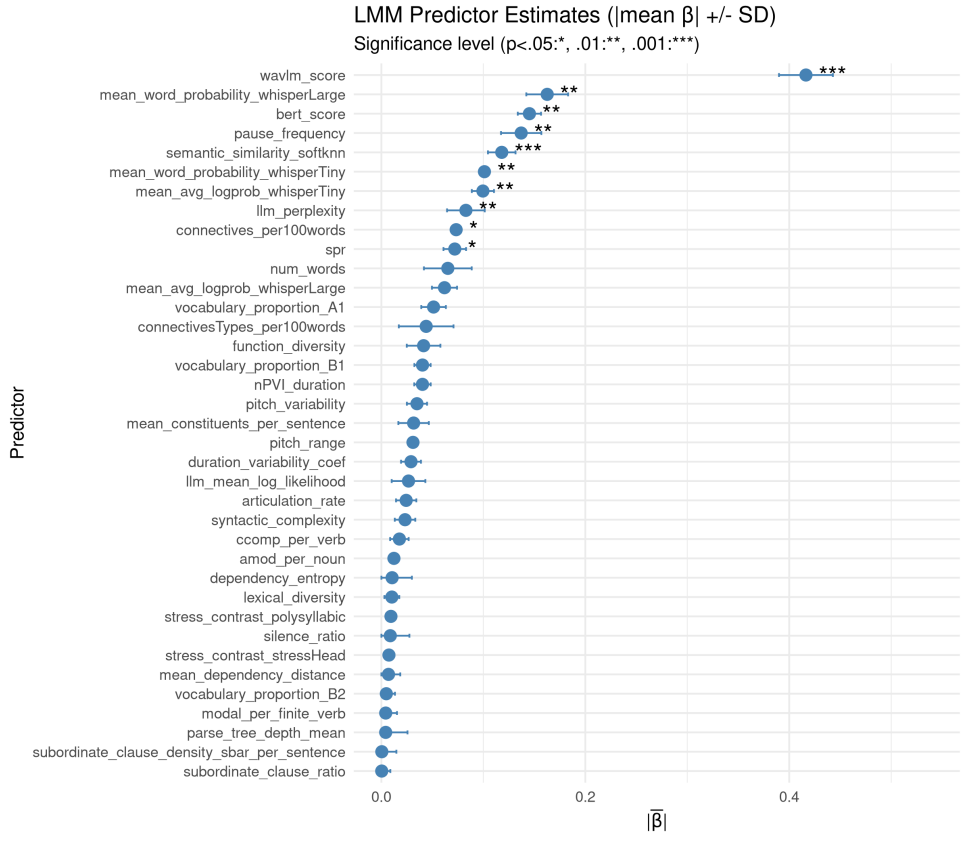


Figure 5: Predictor estimates of a model utilizing all baseline heuristics, WavLM predictions, and neural text-based predictions