



Linguistics
GRADUATE SCHOOL



LINGUISTIQUE
Sociétés et Humanités
Université Paris Cité

UNIVERSITÉ PARIS CITÉ

Faculté Sociétés et Humanités

UFR LINGUISTIQUE

Thesis for the diploma Master Science du Langage: Linguistique Informatique

**Predictive Modeling for Oral Proficiency of L2
Speaker of English with Self-Supervised Neural
Networks**

By

Eli STAFFORD

Under the supervision of Mr. the Doctor Sylvain COULANGE

In coordination with

Université Grenoble-Alpes
Laboratoire Informatique de Grenoble
Service des Langues



Abstract

Automated L2 speech assessment traditionally utilizes vectors of acoustic, lexical, and syntactic metrics as representations of speech for statistical learning. This work explores the use of Self Supervised Neural Networks, in both the audio and text domain, to supplement existing scoring strategies with robust computational representations. We experiment with architectural considerations for their application to this task, considerations for their use in reliable and equitable speech scoring, and measure their linguistic coverage compared to traditional systems. We find their use in automated speech scoring corresponds with a measurable improvement in performance, while finding evidence that their use comes with major drawbacks in terms of explainability and robustness compared to traditional systems.

Abstract	2
Table of content	4
1 Context of this Internship	1
1.1 SELF	1
1.2 Context of this work	2
2 Related Work	3
2.1 Use in Industry	3
2.2 Neural Approaches	3
2.3 Related Datasets	4
3 Methodology	5
3.1 Architectural Choices	5
3.1.1 Layerwise Aggregation Approaches	6
3.1.2 Pooling Approaches	7
3.2 Evaluation	8
3.2.1 Dataset	8
3.2.2 Evaluation Methodology	8
3.2.3 Baseline System	9
3.3 Data Augmentation	9
3.4 Experiments	10
3.4.1 Architecture and Hyperparameters	10
3.4.2 Data Augmentation	11
3.5 Probing	11
3.5.1 Motivation	11
3.5.2 Probing	12
4 Results	14
4.1 SSL Experiments	14
4.1.1 Audio SSL Architecture	14
4.1.2 BERT Data Augmentation	15
4.2 Overall Performance	16
4.2.1 Combination of Models - Recording Level	16
4.2.2 Combination of Models - Speaker Level	17
4.2.3 CEFR Disparity	17
4.3 Speech Encoder Probing Results	18
5 Discussion	21
6 Conclusion	23
Bibliography	23
List of figures	27
Additional Figures	27

Context of this Internship

This internship was carried out in the context of the SELF Production orale project, headed by Monica Masperi, with supervision by Sylvain Coulangue, 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.

1.1 SELF

SELF, or "Système d'Évaluation en Langues à visée Formative" is an online language assessment designed to automatically evaluate student proficiency in 7 languages, used as a placement test by over 40 universities around the world ¹. Developed at Université Grenoble Alpes as part of the IDEFI Innovalangues project (ANR-11-IDFI-0024/IDEX 2012-2020 and Formation projets Emergants 2022), carried out by the LIDILEM (Laboratoire de linguistique et didactique des langues étrangères et maternelles), SELF aims to address a national demand for formative assessment by providing a research-based, multilingual test that:

- Meets real institutional needs and logistical constraints for test administration
- Offers cross-linguistic coherence within a multilingual training framework
- Supports pedagogy through level-based skill-specific grouping

Over a 1h examination period, the test evaluates students across reading comprehension, listening comprehension, and short written expression, with questions designed by specialists in language teaching, formally verified by statisticians and psychometricians. Items are designed to test competencies outlined in the Common European Framework of Reference for Languages, or an adapted rubric for non-European languages. After evaluation, students are ranked with a corresponding CEFR level, from A1 to B2, C1, or C1+ depending on

¹<https://self.univ-grenoble-alpes.fr/english/>

language. These placements are used by partner universities to place students in their language programs, with over 40,000 tests carried out per year.

SELF Production Orale (SELF-PO) is a continuation of the of existing SELF examination focusing on the development of a module for the evaluation of oral production, a hitherto unaddressed fundamental competency. LIDILEM, in collaboration with the Laboratoire d'Informatique de Grenoble (LIG), seeks to mobilize advancements in natural language processing to develop a systematic, standardized, and large-scale module for automated L2 speech assessment.

1.2 Context of this work

The development of the Production Orale module faces unique challenges compared to those present in existing modules. Comprehension competencies are currently measured with multiple choice or short answer questions about question content, and written production with fill-in-the-blank style vocabulary and grammar questions. These questions target specific CEFR descriptors, generally specific "I can" descriptions of language use, and are validated through a lengthy psychometric validation.

Creating a CEFR aligned oral productive competency test requires a completely different approach. The official descriptions of oral productive competency use descriptors like "fluently", "spontaneously", and "effortlessly", creating challenges for previous methods of validation. To create such an automated scoring module, representational learning of these properties in student speech is necessary, and extensive work has already been carried out to create a dataset representing the target population of SELF-PO, and to create a baseline model utilizing traditional acoustic and lexical scoring for statistical learning of a classification model.

This work centers on deep neural representations of student speech, rather than traditional acoustic/lexical metrics, and its use in an automated L2 speech assessment, with the collected dataset as a benchmark. We seek to leverage Self-Supervised Learning models (SSLs) to aid the creation of a performant predictive scoring system for the SELF-PO module, systematically experiment with various proposed architectures, and explore trade-offs of deep neural methods in this context.

Related Work

2.1 Use in Industry

Preliminary explorations of L2 automatic speech assessment leads us to private assessments which utilize automatic methods for oral production evaluation. Prolific testings bodies, like Pearson (with the Test of English Academic), Cambridge Assessment English (with the Cambridge English Testing System), or DuoLingo, do not generally publish full descriptions of their proprietary systems, but often publish rubrics and certain important methodological choices, some of which were taken into account when designing the baseline model.

Partial outlines of examinations systems (Saito et al. 2023; Xu et al. 2021; Park et al. 2025), used as reference for the prior development in SELF-PO, show that the best performing tests of spontaneous spoken proficiency are based on statistical learning of professionally graded student speech, collected into large corpora, in the same vein as our baseline system. They utilize a range of linguistic scores, which are fed into a regression or machine learning model as vectors representations of student speech. There is generally no mention of the use of pre-trained encoder or Large Language models, however they are mentioned by some related studies.

2.2 Neural Approaches

In the academic literature, there is more evidence of the efficacy of neural systems in spoken language assessment (De Jong et al. 2025). Neural approaches in this domain have historically used cascaded systems with a transcription component and a text based language model, sometimes augmented with acoustic metrics. For example, baseline and competitive systems described in the Speak and Improve Challenge 2025 sandi-results utilized this type of architecture. Recent work has also experimented with text based LLMs (Wang et al. 2021), and multi-modal approaches (Fu et al. 2024).

Most of these approaches, being in the text-domain, only take into account acoustic information via secondary surface analysis, essentially combining a traditional metric-based scorer with a text-based deep neural model. In this work,

we seek to explore the efficacy of both text-based neural models and pretrained audio encoders for this task. We find evidence of this in [Bannò and Matassoni \(2022\)](#), which includes a direct utilization of Wav2Vec2.0 as a speech grader, and compares it to a text based BERT grader. The neural approaches of this work are largely based off this study, though we explore more models and architectures, and do direct comparisons to traditional methods.

We separate these trained neural approaches from *pseudo-neural* approaches, which are already included in our baseline system. These pseudo-neural approaches utilize properties of pre-trained deep neural networks, such as the Whisper transcription model ([Radford et al. 2022](#)) or pre-trained sentence embeddings ([Reimers and Gurevych 2019](#)), to add information to our baseline representation vectors. These approaches utilize statistical correlations between model outputs, like Whisper confidence scores ([Ballier and Méli 2024](#)) or cosine similarity of sentence embeddings, to our reference scores, without intermediary training on our data. The same could be said about the use of LLMs in this context ([Ma et al. 2025](#)), which are difficult to fine-tune on a specific task but are rather used in a zero-shot setting.

This work focuses specifically on the fine-tuning of pretrained models with a classification head, which can be used independently or in tandem with classical linguistic scores.

2.3 Related Datasets

Neural models in general require large quantities of annotated speech data for effective performance. Compared to similar datasets ([Ishikawa 2019](#); [Gilquin et al. 2010](#); [Knill et al. 2024](#)), the data collected for the SELF-PO module is orders of magnitude smaller (10 hours of speech), with a more ambitious scoring range. Datasets like ICNALE (*The International Corpus Network of Asian Learners of English*, a benchmark for previous systems) includes 75 hours of student speech, and SandI (*The Speak and Improve Corpus*) around 315 hours of speech. Additionally, our dataset includes the entire range of CEFR levels, grouped from A1-C1 + , whereas these datasets lack data covering students at the A1 level, and in the case of ICNALE even the C1 and C2 level.

This is an important observation, as systems with good performance on these benchmarks may not perform well on our dataset due to dataset size. Conversely, it may be pertinent to utilize these datasets for data augmentation, which we will explore in our methodology.

Methodology

3.1 Architectural Choices

We explore neural approaches to L2 automatic speech assessment in two main modalities: audio and text. These two approaches are much more similar than they are different, architecturally speaking. They consist of a pre-trained base model (SSL), and an untrained classifier or regression head which is then fine-tuned for our task. We consider encoder-only architectures, namely BERT-like architectures for text, and Wav2Vec2.0-like architectures for speech.

The majority of our experiments center on a major architectural question for L2 automatic speech assessment. These models, given an utterance, produce 3-dimensional representations including model layer, time/token position, and embedding space, which we can utilize to train a scoring system. At the embedding dimension, we have a fixed length abstract embedding representing a specific token or audio slice, with its fixed length being determined by the model. At the time/token dimension, we have some number of these fixed length embeddings, either determined by the feature encoder of an audio model (audio slices 10-40 milliseconds), or based on the number of tokens as determined by the tokenizer for a text model. Finally, we can consider the output of each layer (*hidden layers*) as another dimension of important information, as it has been shown that the layers of a pre-trained SSL contains different information about each timestep/token, from low-level acoustic/semantic information in the first layers, to higher level contextual information in later layers (Rogers et al. 2020).

Different tasks require different methods of aggregating this information (Wen Yang et al. 2024). Sequence level classification tasks, i.e the classification of a series of tokens or time slices in an input, do not aggregate along the time dimension, and generally benefit from full finetuning (leaving parameters open to training in the base model). Utterance level classification tasks like emotion recognition often utilize abstract, higher level layers of these models via learnable weighted mean pooling, while tasks like synthetic voice detection use low level convolutions to identify artifacts. Classification of student recordings can be theoretically formulated as a utterance-level classification task, but evaluations of student speech often center on specific disfluencies. It is for this reason that we

experiment with different methods for the aggregation of the layer and time dimensions.

3.1.1 Layerwise Aggregation Approaches

Layerwise aggregation strategies seek to combine disparate representational information in each layer into a single output. We experiment with three major aggregation approaches, which can be applied independent of pooling strategies along the time dimension, and therefore aggregate information across layers at a given token or time slice into a single embedding.

- **Layerwise:** The simplest of our approaches trains a classifier head on each layer of our audio model to output a global score, and trains a linear regression classifier on these scores. A model with 25 hidden layers therefore requires 25 classifier heads, producing a feature vector of 25 scores. The advantage of this approach is that we can control which information we give our final model, and that we consider the full breadth of information encoded in each layer. Conversely, this approach only weakly combines information from each layer, as whatever linguistic information was contained in each layer is collapsed into a single continuous score, and combined linearly, without regard to contextual combinations of scores.
- **Weighted Mean:** This approach, standard for frozen SSLs (wen Yang et al. 2024), trains a vector with the same length as 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. Similarly to our layerwise approach, this allows us to identify the most important layers for our task automatically, and does not combine information in a non-linear way. This approach is different from layerwise classification in that the aggregation strategy and prediction head are trained jointly, so each layer can contribute to the final prediction with more nuance than a linear combination of the best predictions by each layer, but may erode encoded information in a layer that is important for a specific example in lieu of more generalizable layers.
- **FT/PEFT:** In this approach we feed only the final layer of our SSL into a classification head, and leave all layers previous layers unfrozen for training, and in a generic sense can be called *full-finetuning* or FT. This approach updates the gradients of the entire model during training, rather than just those of the classifier head. This allows the model to take advantage of its deep architecture for our particular task, and update its internal representations to better take advantage of information present in the signal and lower layers, and propagate that information directly to the more abstract final layers. In contrast to our other two approaches, this approach aggregates layer information through the model architecture, combining their representation in a non-linear fashion. As SSLs are not normally

pretrained on non-native utterances, updating model weights our dataset should lead to more robust modelling of student speech. Conventional full fine-tuning of audio SSLs is prone to catastrophic forgetting, especially in low-resource settings, so we opt for a Parameter-Efficient-Finetuning method, namely Low-Rank Approximation on the query and key matrices of each transformer layer. *Low-Rank Approximation* (LORA) replaces direct gradient updates to model weights with two low-rank matrices, which are combined at prediction time and added to base model weights. Our approach, based off of [Hu et al. \(2021\)](#), freezes all model weights save for the query and key matrices of the model’s transformer modules, which reduces necessary parameters to 1% of model size, and has been shown to reduce sensitivity during fine-tuning ([He et al. 2022](#)).

3.1.2 Pooling Approaches

Pooling approaches aggregate sequences of vectors, embeddings of token or time slices at specific positions into a single embedding representing the entire utterance. We experiment with three pooling approaches, applied after layerwise aggregation (or directly on layers, with the multilayer strategy), collapsing the outputs of each model into a single vector.

- Mean Pooling: the simple average of each embedding across the time or token dimension.
- Attention Pooling: each vector is given an attention value by a linear feedforward layer, and we perform the weighted average based on the softmax of these values. This allows the model to choose pertinent time slices or tokens based on their learned relevance.
- Fourhead: 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. 2024](#)). This approach considers the entire context of the phrase when determining the weight of a timeslice or token, and does so with four distinct heads. The resulting weighted means are concatenated, and a nonlinear regression head utilizes this vector to predict student level. Theoretically this allows us to have multiple specialized heads which consider different nonlinear relationships between different parts of the utterance.

3.2 Evaluation

3.2.1 Dataset

Models were evaluated on a novel dataset collected for the SELF-PO project. It consists 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 levels¹. 14 raters, comprised of CLES² certified language teachers and university professors of English as a second language, were recruited to create gold ratings of these recordings. At 150 ratings per rater, this resulted in 2,100 ratings (an average of 2.4 ratings per recording), then normalized and averaged to produce continuous, reliable, gold ratings. Prior to this work an automatic statistical scoring model (Linear Mixed Model) was created utilizing acoustic, lexical, and semantic metrics as representational vectors. It is against this non-neural baseline that we compare all of our results.

In paradigm called *holistic scoring*, students were also rated with three *sub-scores* in addition to their overall score. These sub-scores measure more specific axes of linguistic competency, including Pragmatics, Vocabulary, and Pronunciation. These scores allow us to discover the weight raters give to certain aspects of productive competency via their correlation with overall score, and to perform the same analysis on our models.

3.2.2 Evaluation Methodology

Models are evaluated on continuous outputs using RMSE (Root Mean Squared Error), PCC (correlation with continuous gold predictions), and are evaluated on accuracy, and macro F1 after binning of continuous outputs to discrete CEFR classes. RMSE can be compared to the variance of our observed ratings, around 1.0, meaning any system significantly below this threshold models the data well.

Continuous scoring in this manner has many advantages:

- Continuous ratings hold important information relevant to classification, and allow us to subdivide CEFR levels if ever necessary (B1.1, B1.2).
- Benchmarks on prior datasets do not report classification metrics (Accuracy, F1 score), but rather regression metrics (RMSE, correlation). Regression scores allow us to calculate both types of scores, and compare our results to previous work.

¹ $1.0 \leq x < 2.0$ as A1, $2.0 \leq x < 3.0$ as A2, until $5.0 \leq x$ as C1 +

²<https://www.certification-cles.fr/>

- Regression scores both allow us to easily add these scores into our overall system, and calculate speaker level (the mean of ratings for all recordings by the same student) in a way which more accurately reflects the out of the system.

Discrete classes would allow us to do more sophisticated sampling in an attempt to correct for an unbalanced dataset, but our the benefits of continuous scores are more important for our eventual application.

Further, due to our small dataset, we evaluate each of our models with a stratified 10-fold validation strategy, where 10, 10% sections of data are withheld during training to be utilized during evaluation, and each of these 10 slices are balanced in terms of CEFR level (binned rating). We do this to better represent performance under deployment conditions, where the model will be trained on 100% of rated data. As student level and recording level are highly correlated, and our final system will be used on speakers the model has never seen, we ensure that recordings featuring the same student are never mixed between train and test splits.

3.2.3 Baseline System

Our baseline system, in development concurrently with this work, 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 metrics like number of words in an utterance, measurements of vowel stress contrast, intelligibility scores derived from ASR system confidences, average dependency length, and proportion of vocabulary present in the A1 Oxford Learner’s Dictionary. This wide range of scores, covering areas of pronunciation, intelligibility, grammar, and lexical context, is a strong baseline to compare our neural system.

We will perform the above validation first alone, comparing our best performing text model, audio model, and baseline model, and then combine each score. Scores will be combined using the above Linear Mixed Model, with the combinations of text-based + baseline metrics, audio-based + text-based, and text-based + audio-based + baseline metrics. The differences in performance will give us evidence of differences between model coverage, redundancy of scores, and standout areas of improvement.

3.3 Data Augmentation

Given the size of our dataset, it is pertinent to explore methods of leveraging other publicly available datasets for the same task. We find that the SandI dataset shares

certain key properties with our target task which means that, in a limited capacity, we can use it for additional training data (Ramanujan et al. 2023).

The important similarities are as follows

- SandI targets second language learners of English at the university level.
- SandI is majority comprised of responses to open-ended questions using spontaneous speech (with one section being comprised of read-speech).
- SandI utilizes the exact same rating scheme, namely a continuous scale of 1.0-7.0 with integer boundaries for level. This means there is no need to switch classification/regression architectures between "pre-training" and fine-tuning.

Importantly however, the exact distribution of students differs greatly from our target population. Our dataset is comprised of mostly French-native second language learners of English at a French university, while participants of SandI have more varied native language background at an English university. This type of data augmentation on related data usually gives rise to more robust representations, but due to differing accents and, we only experiment with the SandI corpus for the text-based approach. Our dataset also has a larger range of represented CEFR levels, from A1-C1+, compared to A2-B2 in the SandI corpus. Utilizing this corpus may weaken our model’s representation of extreme classes, but may strengthen our model’s ability to discriminate between the classes present.

3.4 Experiments

3.4.1 Architecture and Hyperparameters

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

For our audio models, the only direct application of pre-trained audio encoders to this task is in Bannò and Matassoni (2022), with a single architectural configuration. Their study, utilizing Wav2Vec2.0, does not include the experiments used to determine their proposed architecture. In this work, we explore all possible combinations of aggregation strategies previously mentioned on Wav2Vec2.0 (Baevski et al. 2020). As this model is the only attested model used on this task, we also experiment on a much larger and more performant pre-trained audio encoder, WavLM-Large (Chen et al. 2021), to measure the effect of model size on performance on our task. As determined by results on the SUPERB benchmark (Wen Yang et al. 2024), a large collection of benchmarks for audio models, WavLM-Large has by far the best generalized performance of pre-trained audio encoders, but is much larger and costlier to train. With 300M parameters, compared to 100M parameters of Wav2Vec2.0, WavLM-Large is

prohibitively expensive for full-finetuning, even with PEFT methods, so we only experiment multilayer and weighted-mean layer aggregation methods, but all other pooling methods.

For our text based approach, while we would have liked to perform experiments on all model configurations, time permitted only the evaluation of the baseline approach of the SandI Challenge. This configuration utilizes the fourhead pooling strategy. While their baseline uses a frozen version of *bert-base-uncased* (Devlin et al. 2019), we additionally experiment with the LORA method mentioned above.

Regression heads and hyperparameter choices were kept the same between experiments in the same modality.

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

3.4.2 Data Augmentation

BERT models, in line with the baseline system of SandI, have a three layer classification head after pooling, with two intermediate layers of size 600 and 20, and a dropout of 0.5 at each layer. Qian et al. (2024) cites the explainability and simplicity of this configuration to be their motivating factors. We experiment with this configuration under the conditions of frozen-finetuning on our task, PEFT fine-tuning on our task, and PEFT fine-tuning on the SandI transcriptions for the same scoring task, performed for one-third of real training epochs, followed by PEFT fine-tuning on our task. These results will reveal if there are necessary changes in this architecture to better adapt to a smaller dataset (namely PEFT vs frozen fine-tuning), and if we can naively leverage similar datasets in our pipeline without sophisticated transfer learning methods.

3.5 Probing

3.5.1 Motivation

Separately from our performance experiments, it is important that we verify, if only loosely, some of the results of our neural methods. Whereas our baseline metric-based model can easily report the effect of each of its metrics on the final outcome, deep neural models make predictions in a black box manner. As the SELF examination is based on the principles of "égalité" (equality) and "équité" (equity), it is important we analyze our models closely to ensure they adhere to these values, regardless of performance improvements.

Inspired by work in speaker bias for audio encoder models (van Niekerk et al. 2021), our probing analysis centers on linguistic and speaker group bias. It has been found that ASR models often contain large amounts of speaker bias, and perform differently on different speaker groups (Herron et al. 2025). Information which identifies a student (in our case mother tongue, gender, and more subtle cues) should be minimized in determining their overall level. It has also been found that audio encoders only weakly model semantic information (Gimeno-Gómez et al. 2025). Whereas we can prove our baseline model measures a wide variety of linguistic competencies during prediction, we would like to see if our audio encoders can similarly leverage different axes of speaking ability when providing a score, or if they only utilize surface features without strong pedagogical basis.

3.5.2 Probing

We achieve this through layerwise probing. Probing involves the training of a lightweight classifier on hidden representations of a neural network on standard inputs, usually on the same data as the primary task, but with different target properties. High performance of this classifier indicates that the information necessary to complete this task is present in the model, and can give insights into what information is available to the model for the completion of the primary task. In our case, we consider the primary task to be automated L2 speech scoring, and we perform probing along all layers of WavLM large for 5 classification tasks:

- As mentioned above, students were rated holistically, with three sub-scores (Pragmatics, Vocabulary, and Pronunciation) assigned by our raters in addition to their overall score. During probing, we consider each of these sub-scores to be a separate task, and consider the performance of our probes on the prediction of each sub-score to indicate which abstract facets of speaking proficiency our model could be using to predict student level.
- To explore the use of speaker identifying information in prediction, we probe each layer on a speaker identification task. We create a new train/test split of our dataset for speaker identification, and consider high performance on this task to indicate that the model could use speaker identify to complete predictions, rather than being speaker agnostic.
- To explore the generalizability of our model to new tasks, we perform probing for task classification, i.e. whether our model can predict the task a student is performing in their audio recording. High performance indicates that task type may influence final prediction.

All probing tasks were carried out with a two layer linear probe, with an intermediate layer of 256. Each task was formulated as a classification task, with classes being subscore CEFR levels, task ids, or speaker ids.

- Fine-tuning our model on the target task for a reduced number of epochs, to establish a direction of gradient descent.
- Fine-tuning our model on the target task, while training a secondary classifier on the intermediate representation produced by our classifier head to classify speaker. We negatively update the gradients of our model during this training, which has strong theoretical guarantees to reduce the amount of speaker information contained in this intermediate representation, while attempting to maximize performance.
- A short fine-tuning on the target task without the speaker classification adversary, to slightly increase performance.

Results

Here we present the results from our experiments in three main sections. The first details the results of our architecture and data augmentation experiments to determine the best models to compare to the baseline. The second compares these models to the baseline first in the recording prediction setting, then in the student level prediction setting, as well as detailing the difficulty all models had with extreme CEFR levels. The third section describes our probing results on WavLM-Large.

4.1 SSL Experiments

4.1.1 Audio SSL Architecture

Here we present results of Wav2Vec2.0 and WavLM on our primary task by architecture combination.

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 4.1: Architecture Comparison: Audio

We find that while there are clearly optimal strategies for each model, there is no

clear best layerwise aggregation strategy between models, and no best pooling strategy between layerwise aggregation strategies (Table 4.1). WavLM had a 0.11 decrease in RMSE (and similar decrease along other metrics) when using the weighted mean strategy compared to the multilayer strategy, while Wav2Vec2.0 had a 0.1 decrease in RMSE when using the multilayer strategy compared to the weighted mean strategy. These architectures are conceptually very different, with the wmean targeting utterance level features, and the multilayer selecting specific timeslices and layers for aggregation. This points to a fundamental difference between the necessary architectural considerations necessary for different model sizes.

WavLM-Large with the weighted-mean and meanpool architecture had the highest performance, with 0.546 macro F1, 0.548 RMSE and 0.822 PCC, with a 0.038 decrease in RMSE compared to Wav2Vec2.0’s best combination. While the RMSE’s for the mean pooling and attention pooling architectures were comparable (0.016 difference in RMSE), the gains in macro F1 were more significant (0.05). We will consider this strategy and model for future comparisons and combinations.

4.1.2 BERT Data Augmentation

Here we present results of bert-base-uncased with and without LORA training, and analyze the gains of pretraining on the SandI Corpus with LORA training (Table 4.2).

We find that a frozen BERT model alone is not enough to complete our task (Frozen SELF-PO), and performs at a random baseline, and that all BERT strategies perform worse than any of our audio model configurations. Disappointingly, we do not find a large difference in performance between a BERT model fine-tuned for transfer learning on a large corpus and then fine-tuned for our purposes (SandI + SELF-PO), and a model simply fine-tuned on our corpus (SELF-PO). RMSE only varies by .01, accuracy by less than 2 points, macro F1 by .03, with only PCC being somewhat significant, at .05 higher for the SandI fine-tuned model. These gains are also not directional, with accuracy and RMSE being better for SELF-PO, and PCC and Macro F1 being better for SandI + SELF-PO

Dataset(s) and Method	RMSE	PCC	Acc %	Macro F1
Frozen SELF-PO	0.980	N/A*	38.6	0.111
LORA SELF-PO	0.694	0.704	51.6	0.392
LORA SandI + SELF-PO	0.701	0.752	49.8	0.422

Table 4.2: BERT Models: Data comparison

As bert-base-uncased is a simple model compared to modern iterations of this architecture, our approach’s low performance compared to neural speech models is not surprising, and does not necessarily indicate a weakness of the text modality for L2 speech assessment. Further, our implementation of pretraining on a related

dataset was naive compared to existing literature (Zhuang et al. 2020; Jeddi et al. 2020), so it should not be ruled out as an area of further study. We will consider LORA SELF-PO (BERT, trained with LORA without data augmentation) for the following comparisons.

4.2 Overall Performance

4.2.1 Combination of Models - Recording Level

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

Table 4.3: Final Models: Recording Level

The performance of our best strategy, WavLM-Large alone, slightly beats the baseline system (Table 4.3). This is expected with usage of such a large model, but the gains are not as impressive as is usually the case when comparing neural and non-neural statistical systems, with an impressive 0.1 reduction in RMSE, but less than 3 point increase in accuracy, and a .025 decrease in macro F1 score. While the decrease in RMSE is promising, the lack of significant improvement along F1 score, which indicates fairness across CEFR levels, indicates that this performance could simply be better modeling the unbalanced distribution of our dataset, rather than better representational generalizability.

All approaches, save for the BERT model, are in the range of 0.56-0.65 RMSE, indicating an average error of just over 1/2 CEFR level. In the same manner, their accuracies range between 50% and 56%, and macro F1 score between 0.5 and 0.57. This small performance range either indicates that baseline and neural models are functionally equivalent (with a small range for improvement), or that our dataset has a ceiling on performance due to a low number of examples.

When combining BERT and the baseline approach, we can see that, despite poor performance on its own, the BERT model adds appreciable value (.017 decrease in RMSE).

WavLM-Large alone has the best RMSE at 0.564, while combining all scores leads to the highest performance in terms of absolute accuracy and Macro F1 score, and a 0.026 increase in RMSE compared to WavLM-Large. We believe this shows that none of the three tested approaches are completely redundant, with measure increases in performance, but we are in general forced to make CEFR coverage and RMSE trade-offs due to our unbalanced dataset.

When combining these approaches, we can observe the measurable effect of each score on prediction outcomes (predictor estimates), show in Figures 6.3 and 6.4. The scores produced by our deep neural models always have among the highest prediction estimates, often reducing the estimates of previously highly estimated metrics. This means that while other scores are important, our neural models may in some way take their measurements into account during prediction. This does not make other scores redundant, and we may see greater performance with a non-linear model.

4.2.2 Combination of Models - Speaker Level

As these models will eventually be used on classifying overall student level, rather than just recording level, it is pertinent to look at their speaker level performance. We define speaker level as the mean of scores for recordings made by the same student, both in calculating their gold level, and predicted level.

All models see an increase in performance on speaker level, around 0.1 RMSE, .05 PCC, 10 points or accuracy, and .1 macro F1, without significant differences in increase between strategies. The averaging mechanism of speaker level reduces the number of outlier classes, and perhaps makes those outlier classes more distinct as they stay high level across multiple recordings, which accounts for these increases if we assume our models already have good modeling of A2-B2 levels. All models have the same relative performance as in the recording level comparison (Table 4.4), but here the RMSE of WavLM-Large and our all-scores model is slightly closer, within 0.01 compared to .02 previously.

Included Scores	RMSE	PCC	Acc %	F1
Baseline - Reg	0.550	0.798	61.2	0.643
WavLM-Large	0.476	0.854	64.3	0.633
BERT	0.603	0.760	55.0	0.469
WavLM-Large + BERT	0.508	0.831	62.0	0.635
BERT + SCORES	0.540	0.805	62.6	0.656
All	0.486	0.845	66.1	0.684

Table 4.4: Final Models: Speaker Level

4.2.3 CEFR Disparity

In a global sense, all models have relatively poor performance across CEFR levels. Though we approach RMSE's observed on other benchmarks (around 0.4 for competitive SandI models, with a dataset variance of 0.6), all of our proposed models demonstrably struggle with A1 and C1 + CEFR levels. Our best performing model by macro F1 score (including all scores) has >50% classification rate for the A2, B1, and B2 classes, but only 30% for A1 and C1 + classes (Table 4.2). This

most likely stems from our unbalanced dataset, which contains a high and balanced number of recordings rated at A2-B2, but vastly fewer for both A1 and C1+, as shown in figure 4.1. The addition of deep neural models does not ameliorate this, with negligible improvements over the baseline model in terms of macro F1 (+0.043).

Despite this, when measured in terms of ± 1 level accuracy, i.e a prediction one level above or below the given rating, all of our models perform $> 90\%$, with our all-scores model approaching 98%. This suggests that there is not a major aspect linguistic competency being overlooked, but rather simply that our models do not have a sufficiently robust representation of these outlier classes.

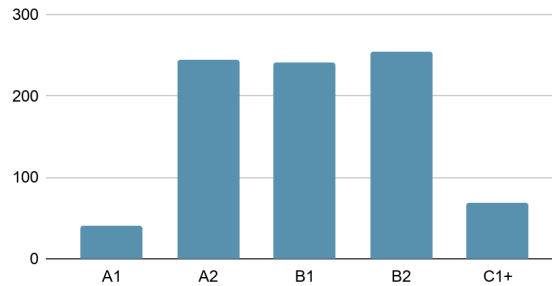


Figure 4.1: Distribution of CEFR levels in the final dataset by recording

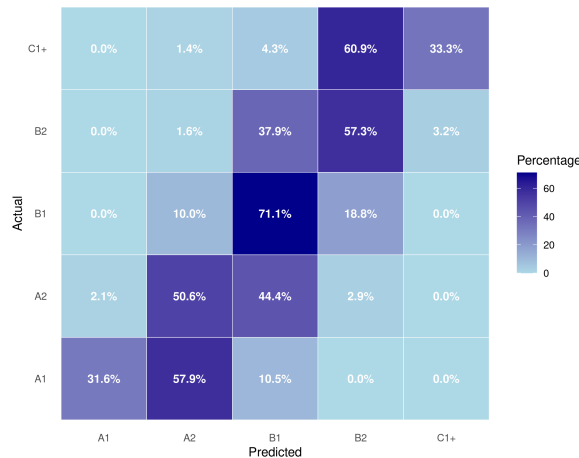


Figure 4.2: Confusion matrix of All Scores Model: by recording

4.3 Speech Encoder Probing Results

Figure 4.3 displays layerwise probing accuracy results on WavLM-large for our target CEFR predictions, holistic score classification, task identification, and speaker identification.

Given the performance of WavLM-Large, probing its representations is an important guide for our discussion. Subscore probing of WavLM-Large acts as we

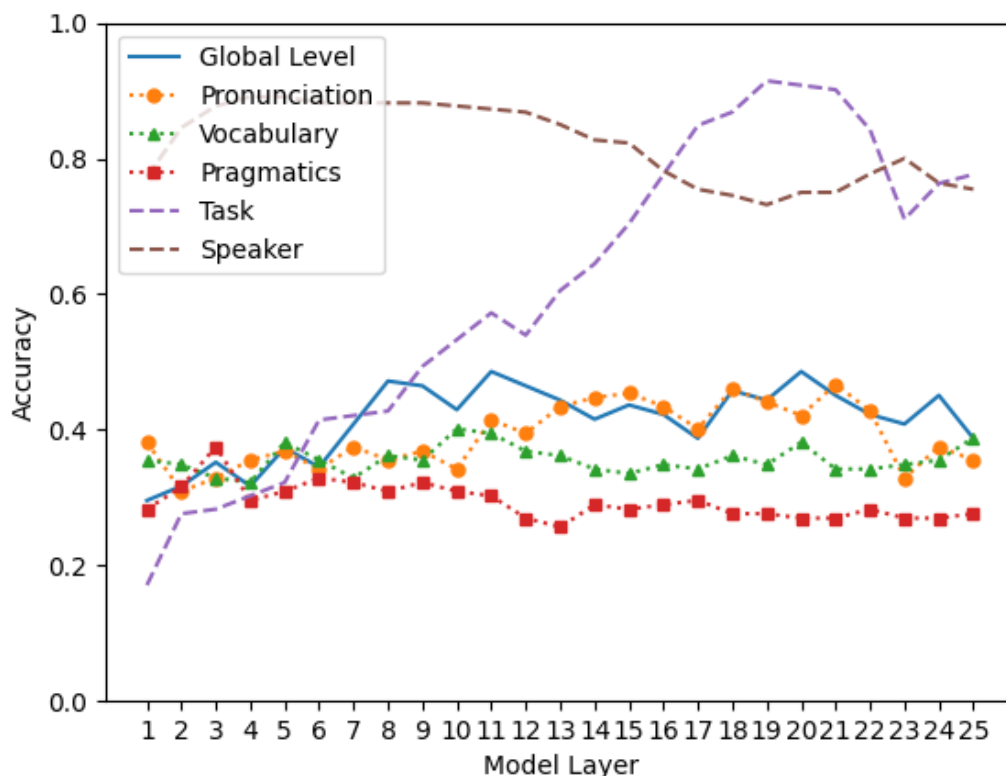


Figure 4.3: Layerwise holistic scoring, speaker, and task probing compared to base task: WavLM-Large

expect. Audio encoder models encode acoustic level information (Pronunciation) at levels near total predictive power (both Pronunciation accuracy and Global Level accuracy are around 50% at layer 18), while poorly encoding semantic or discourse level information (Vocabulary and Pragmatics) at $< 40\%$ accuracy even utilizing the best layers. Even with good performance, WavLM-Large does not take into account a wide variety of linguistic competencies, but rather focuses on acoustic properties of speech.

Our probe’s performance on speaker identification is also not surprising. On the CelebVOX speaker identification benchmark (Nagrani et al. 2017), WavLM-Large is a high performance model, reaching accuracies of over 90%. If there is any correlation between speaker and CEFR level, classifiers built on large models like WavLM-Large can exploit non-linguistic characteristics of speakers to predict CEFR score. This may lead to poor performance on new speakers, and is not desirable pedagogically.

Interestingly our probes were able to accurately predict the task (question id) a student was performing near the last layers of the model, with up to 90% accuracy. This is a surprising result, as even though WavLM-Large also performs very well on keyword identification, our mean-pooling approach would theoretically remove any identifying information about the task a student is performing. A possible prediction strategy of these models, in the worst case scenario, would be to identify the student

speaking and the task being performed, which acts as a unique identifier of recording rating.

These results confirm hypotheses about the black-box nature of these models, in that solid performance on a particular dataset in a particular setting is not guaranteed to generalize across groups of tasks or speakers, as it is possible they utilize information that would harm generalization. Additionally, speech-encoder models only robustly model the acoustic information in an utterance, and transcription based methods are necessary to capture relevant vocabulary and pragmatic information.

Discussion

While macro F1 score and accuracy are important metrics for evaluating efficacy of our models for our final use case (ascribing discrete CEFR levels accurately to a diverse range of CEFR levels), given the small differences between models for these two metrics we consider RMSE and PCC with respect to continuous gold labels to the most important metrics for model evaluation in our scenario due to the "violent" nature of CEFR binning. A student rated 2.9 (A2) is much closer in proficiency to a student rated 3.1 (B1) than a student rated 2.0 (A2), and we would like to capture this fact in our quantitative evaluation. By this metric, 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. Its usage though comes with a host of issues that make it difficult to justify in this context, including lack of improvements at extreme class ranges, speaker bias, predictor coverage, and computational efficiency.

We hoped that the introduction of deep neural networks would assist in our difficulties with edge case CEFR levels, i.e a large increase in macro F1 score. It is clear now that this issue most likely requires a data-based solution, rather than an architecture-based one. Simply collecting more A1/C1+ levels, or perhaps native-level responses to the same tasks, would probably improve performance for all approaches, and at this time it is not clear if neural approaches would perform significantly better than traditional approaches if given more data.

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 the scores which have the most *reduction* in predictor estimates are surface level metrics like number of words and speech rate. 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 information during prediction.

Finally, undiscussed in our experimentation, WavLM-Large is an order of magnitude more computationally expensive than even our text-based approach. At prediction time, our BERT model takes 2 minutes to process 1,000 responses on CPU, and WavLM-Large takes more than 20 minutes for the same number of responses on GPU. The current SELF examination has a near instant feedback time,

which is appreciated by and sometimes necessary for educators utilizing the test, and adding such large models into its prediction pipeline while retaining this speed would come with engineering challenges. Future improvements to SELF will additionally take more extensive data collection (especially to cover non-French accents), and the time and computational resources necessary to train a WavLM-Large based model may become prohibitive.

A positive outcome of our experimentation is the slight performance gain of the addition of a BERT-like model to the baseline model. Though it is not a large performance improvement, we believe that it could be non-negligible, given that we only experimented with one architecture, and an older, less powerful model. It is likely that with more experimentation we can find an alternative architecture with more improvement gains, and much lower computational cost than for an audio encoder. As our baseline model is currently lacking in coverage over linguistic competencies like grammaticality and vocabulary choice, at least more than fluency-based competencies, a text-based language model would theoretically assist more than an audio-based one, and there is less concern over speaker identity bias in this modality. We have also found these coverage-based arguments are more convincing to the educators that will be affected by our models than the raw numbers proposed by audio encoders.

Overall, these models are fit for our use case with certain caveats. The extremely high ± 1 -level accuracy, approaching 98% for our all-scores model, and 0.55 RMSE are very encouraging. Only rarely do any of our models produce predictions which would be unsupportable in an educational environment, such as C1+ for and A2 level speaker.

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 utterances, the fact that this decision changes between base models, and found slight improvements over traditional metric based baselines. Methodologies used for utterance level classification tasks, like speaker identification and emotion detection, will most likely be applicable to automated speech assessment for large models like WavLM. Additionally, we find evidence that learned representations of text and speech can be used to improve linguistic metric-based systems, suggesting that their coverage of linguistic competencies differs in significant respects to traditional systems. There are however major drawbacks, especially in a low-resource setting, with evidence that these models have the capacity for speaker-overfitting with deep or large classifier heads, which is a major hurdle for their use in reliable and equitable systems. Adversarial training prevents this sort of overfitting to an extent, but removing speaker information seems to reduce performance by a non-negligible amount. Overall, SSLs represent a promising avenue of work in the creation of automated speech assessments, but their merely minor gains in performance still leave much room for debate about their nature as black-box, and adherence to principles of language education.

Further Work

The use of SSLs did not mitigate major problems stemming from dataset imbalance. The use of zero-shot systems (text-based/multimodal LLMs) are another possible avenue of improvement for L2 spoken proficiency assessments.

Though the literature suggests that pre-training on large corpora of a similar or related task yields improvements, our experiments with the SandI corpus did not yield any significant results. Success in this endeavor would aid in solving the classification outlier classes via more robust representations.

On the black-box nature of these systems, there is of course a large body of literature in interpretability which could be applied in this setting, especially input-output oriented approaches with existing disfluency annotated corpora.

Bibliography

- A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL <https://arxiv.org/abs/2006.11477>.
- N. Ballier and A. Méli. Investigating acoustic correlates of whisper scoring for L2 speech using forced alignment with the Italian component of the ISLE corpus. In T. Gaillat, C. Mallart, F. Moreau, J.-Y. Li, G. Drouet, D. Alfter, E. Volodina, and A. Jönsson, editors, *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 20–32, Rennes, France, Oct. 2024. LiU Electronic Press. URL <https://aclanthology.org/2024.nlp4call-1.2/>.
- S. Bannò and M. Matassoni. Proficiency assessment of l2 spoken english using wav2vec 2.0, 2022. URL <https://arxiv.org/abs/2210.13168>.
- S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *CoRR*, abs/2110.13900, 2021. URL <https://arxiv.org/abs/2110.13900>.
- N. De Jong, S. Raaijmakers, and D. Tigelaar. Developing high-quality, practical, and ethical automated l2 speaking assessments. *System*, 134:103796, 2025. ISSN 0346-251X. <https://doi.org/10.1016/j.system.2025.103796>. URL <https://www.sciencedirect.com/science/article/pii/S0346251X25002064>.
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL <https://arxiv.org/abs/1810.04805>.
- K. Fu, L. Peng, N. Yang, and S. Zhou. Pronunciation assessment with multi-modal large language models, 2024. URL <https://arxiv.org/abs/2407.09209>.
- G. Gilquin, S. De Cock, and S. Granger. *Louvain International Database of Spoken English Interlanguage*. 01 2010. ISBN 978-2-87463-245-7.

- D. Gimeno-Gómez, C. Botelho, A. Pompili, A. Abad, and C.-D. Martínez-Hinarejos. Unveiling interpretability in self-supervised speech representations for parkinson’s diagnosis. *IEEE Journal of Selected Topics in Signal Processing*, 19(5):717–730, 2025. ISSN 1941-0484. [10.1109/jstsp.2025.3539845](https://doi.org/10.1109/jstsp.2025.3539845). URL <http://dx.doi.org/10.1109/JSTSP.2025.3539845>.
- J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig. Towards a unified view of parameter-efficient transfer learning, 2022. URL <https://arxiv.org/abs/2110.04366>.
- F. Herron, S. Rossato, A. Allauzen, B. Favre, and F. Portet. *Speaker Group Encoding in Self-supervised Speech Recognition Models*, page 121–132. Springer Nature Switzerland, Aug. 2025. ISBN 9783032025487. [10.1007/978-3-032-02548-7_11](https://doi.org/10.1007/978-3-032-02548-7_11). URL http://dx.doi.org/10.1007/978-3-032-02548-7_11.
- E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021. URL <https://arxiv.org/abs/2106.09685>.
- S. Ishikawa. The icnale spoken dialogue: A new dataset for the study of asian learners’ performance in l2 english interviews. *English teaching*, 74:153–177, 2019. URL <https://api.semanticscholar.org/CorpusID:213608532>.
- A. Jeddi, M. J. Shafiee, and A. Wong. A simple fine-tuning is all you need: Towards robust deep learning via adversarial fine-tuning, 2020. URL <https://arxiv.org/abs/2012.13628>.
- K. Knill, D. Nicholls, M. J. F. Gales, M. Qian, and P. Strohinski. Speak and improve corpus 2025: an l2 english speech corpus for language assessment and feedback, 2024. URL <https://arxiv.org/abs/2412.11986>.
- R. Ma, M. Qian, S. Tang, S. Bannò, K. M. Knill, and M. J. F. Gales. Assessment of l2 oral proficiency using speech large language models, 2025. URL <https://arxiv.org/abs/2505.21148>.
- A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, interspeech₂₀₁₇, page 2616–2620. *ISCA*, Aug. 2017. [10.21437/interspeech.2017-950](https://doi.org/10.21437/interspeech.2017-950). URL <https://arxiv.org/abs/1708.04426>.
- Y. Park, R. Cardwell, S. Goodwin, B. Naismith, G. T. LaFlair, K. P. Yancey, and X. Zhang. Assessing speaking on the duolingo english test, 2025. URL <https://englishtest-static.duolingo.com/media/resources/media/resources/whitepapers/speaking-whitepaper.pdf?ref=blog.englishtest.duolingo.com>.
- M. Qian, K. Knill, S. Bannò, S. Tang, P. Karanasou, M. Gales, and D. Nicholls. Speak and improve challenge 2025: Tasks and baseline systems, 2024. URL <https://arxiv.org/abs/2412.11985>.

M. Qian, K. Knill, S. Bannò, S. Tang, P. Karanasou, M. Gales, and D. Nicholls. Speak and improve challenge 2025, 2025. URL https://www.isca-archive.org/slate_2025/qian25_slate.html#.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL <https://arxiv.org/abs/2212.04356>.

V. Ramanujan, T. Nguyen, S. Oh, L. Schmidt, and A. Farhadi. On the connection between pre-training data diversity and fine-tuning robustness, 2023. URL <https://arxiv.org/abs/2307.12532>.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL <http://arxiv.org/abs/1908.10084>.

A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in bertology: What we know about how bert works, 2020. URL <https://arxiv.org/abs/2002.12327>.

K. Saito, K. Macmillan, M. Kachlicka, T. Kuniyara, and N. Minematsu. Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, 45(1):234–263, 2023. [10.1017/S0272263122000080](https://doi.org/10.1017/S0272263122000080).

B. van Niekerk, L. Nortje, M. Baas, and H. Kamper. Analyzing speaker information in self-supervised models to improve zero-resource speech processing, 2021. URL <https://arxiv.org/abs/2108.00917>.

X. Wang, K. Evanini, Y. Qian, and M. Mulholland. Automated scoring of spontaneous speech from young learners of english using transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712, 2021. [10.1109/SLT48900.2021.9383553](https://doi.org/10.1109/SLT48900.2021.9383553).

S. wen Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang, T. hsun Feng, P.-H. Chi, Y. Y. Lin, Y.-S. Chuang, T.-H. Huang, W.-C. Tseng, K. Lakhota, S.-W. Li, A. Mohamed, S. Watanabe, and H. yi Lee. A large-scale evaluation of speech foundation models, 2024. URL <https://arxiv.org/abs/2404.09385>.

J. Xu, E. Jones, V. Laxton, and E. Galaczi. Assessing l2 english speaking using automated scoring technology: examining automarker reliability. *Assessment in Education: Principles, Policy Practice*, 28:1–26, 09 2021. [10.1080/0969594X.2021.1979467](https://doi.org/10.1080/0969594X.2021.1979467).

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning, 2020. URL <https://arxiv.org/abs/1911.02685>.

List of figures

4.1	Distribution of CEFR levels in the final dataset by recording	18
4.2	Confusion matrix of All Scores Model: by recording	18
4.3	Layerwise holistic scoring, speaker, and task probing compared to base task: WavLM-Large	19
6.1	Scoring Interface	28
6.2	Task "Tell a story". Directive: "Take a look at these pictures and tell the story." Open ended, designed to target competencies up to C2. . .	29
6.3	Predictor estimates of a model utilizing all baseline metrics, WavLM predictions, and BERT predictions	30
6.4	Predictor estimates of a model utilizing only baseline metrics	31

SELF PO Rating Interface
IGOR LOGOUT

Progress: 85 / 150 recordings

Task name: Describe_Jonathan
Task code: 10520

Context
0:00 / 0:02

Jonathan, 25

Question
0:00 / 0:04

Recording ID: 1762545566
Student ID: 133508

Recording to Rate
0:00 / 0:34
100%

Rating Form
☐ Non-scorable
☐ Off-topic

Task Completed
NO PARTIALLY YES

Pragmatics
A1 A2 B1 B2 C1 C2

Vocabulary and Grammar
A1 A2 B1 B2 C1 C2

Pronunciation / Prosody / Fluency
A1 A2 B1 B2 C1 C2

Recording's Global Level
According to you, what is the proficiency level of this recording?
A1 A2 B1 B2 C1 C2
Value: 4

Rating Confidence Score
LOW MEDIUM HIGH

Comments (optional)
Add any additional comments...

SUBMIT RATING AND LOAD NEXT

Figure 6.1: Scoring Interface

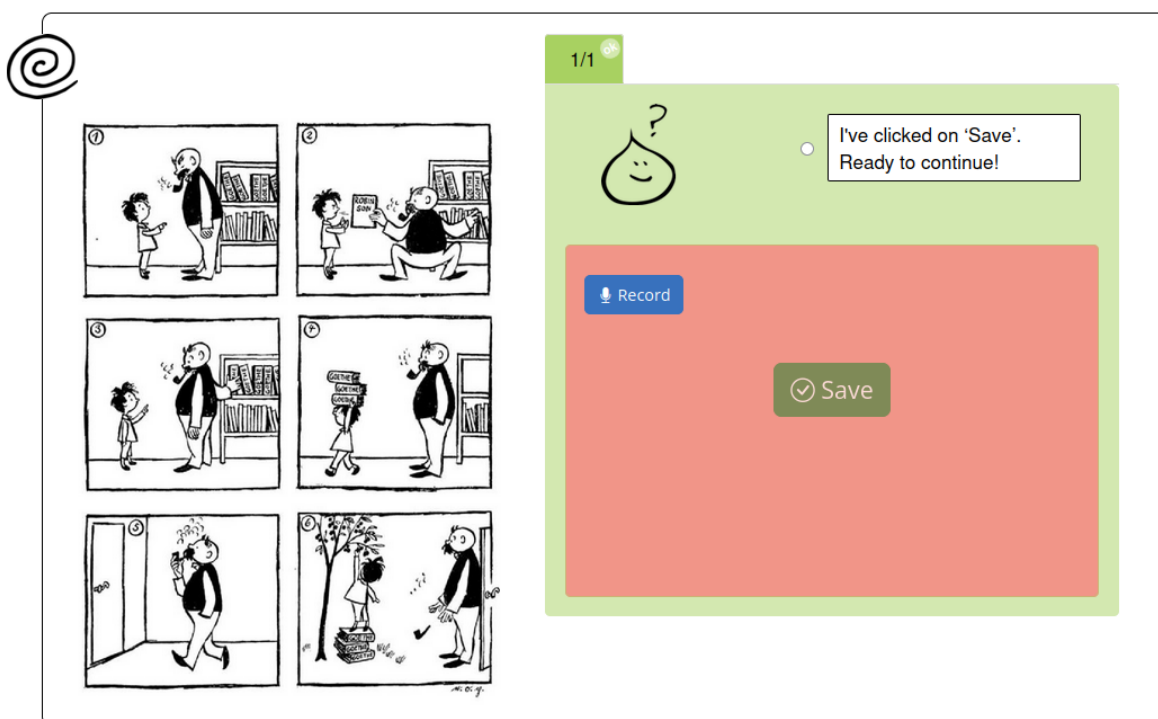


Figure 6.2: Task "Tell a story". Directive: "Take a look at these pictures and tell the story." Open ended, designed to target competencies up to C2.

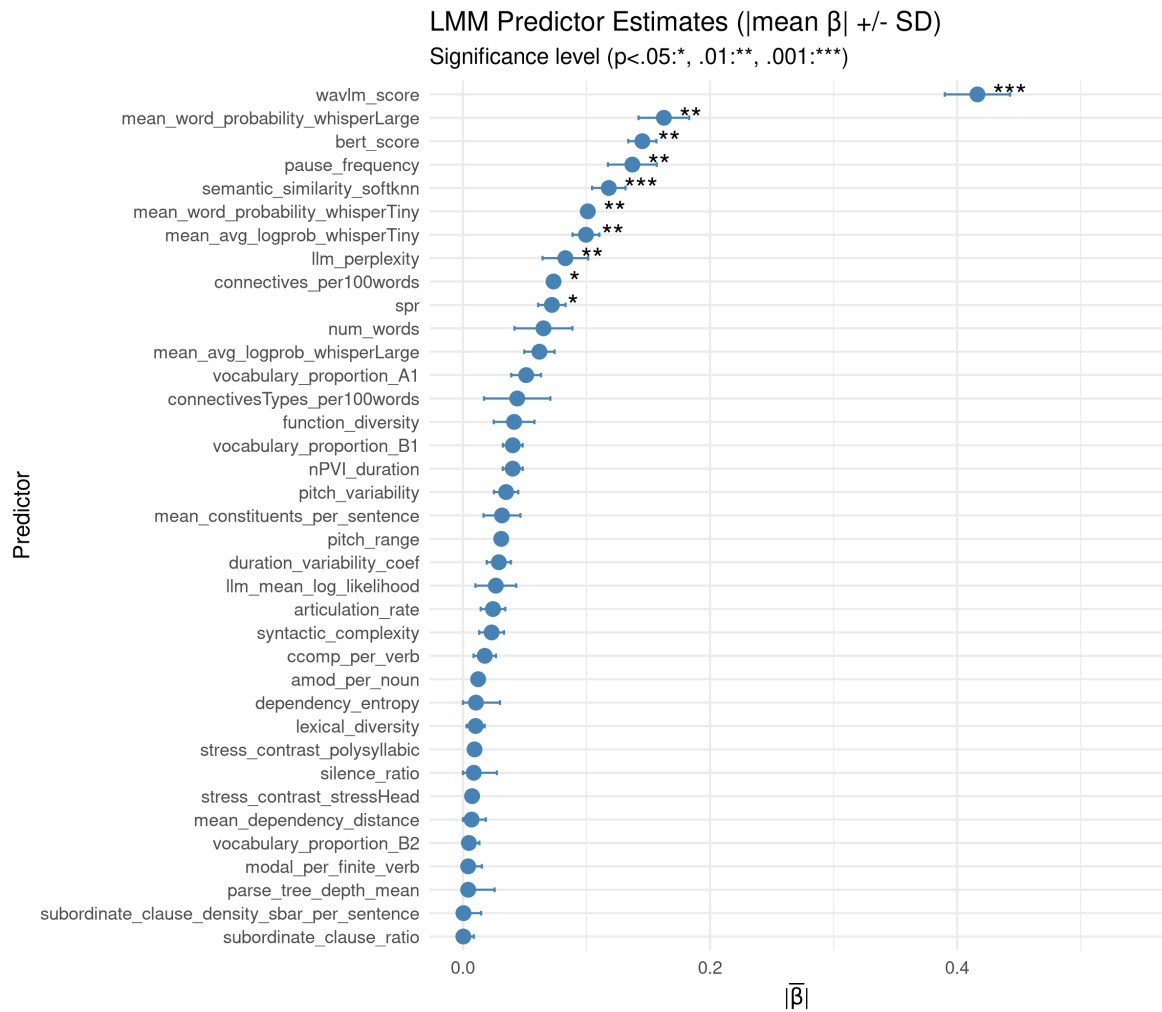


Figure 6.3: Predictor estimates of a model utilizing all baseline metrics, WavLM predictions, and BERT predictions

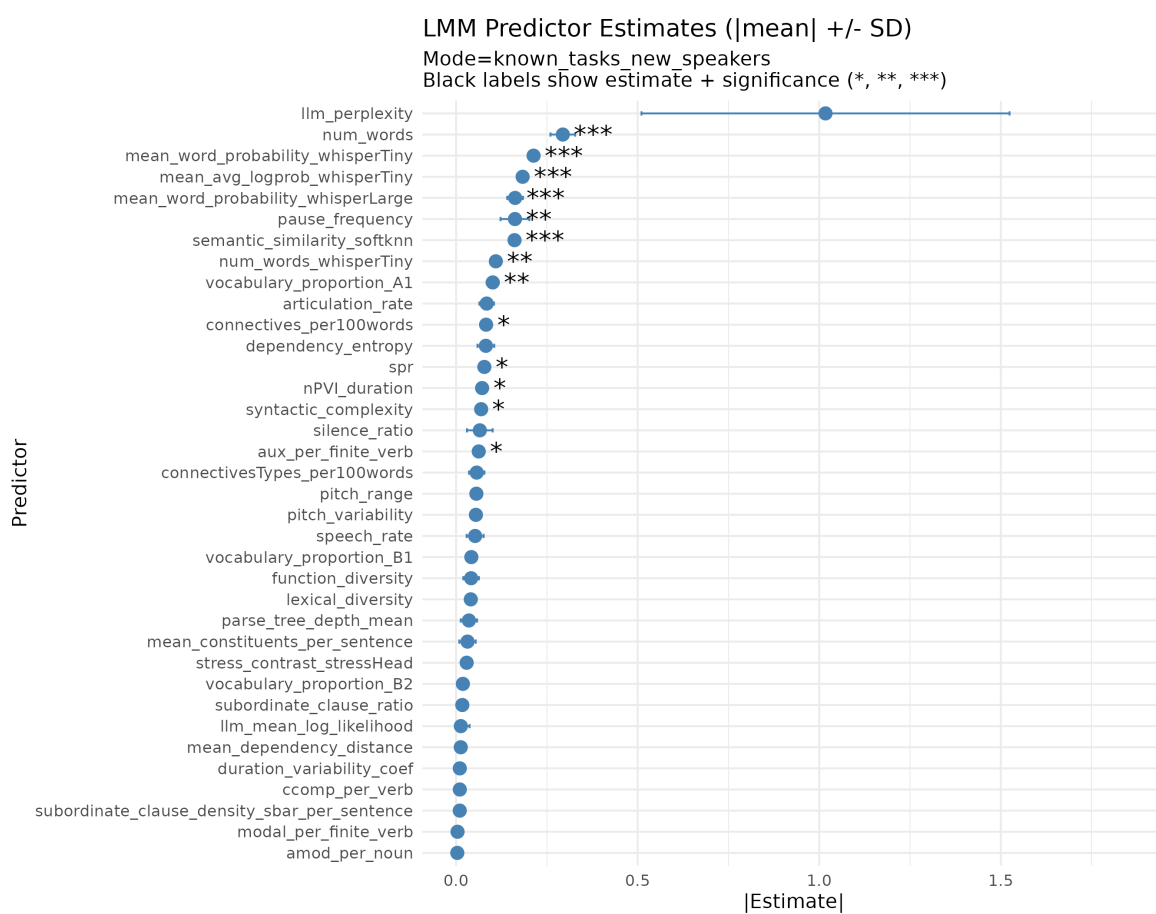


Figure 6.4: Predictor estimates of a model utilizing only baseline metrics

List of tables

4.1	Architecture Comparison: Audio	14
4.2	BERT Models: Data comparison	15
4.3	Final Models: Recording Level	16
4.4	Final Models: Speaker Level	17