Using Machine Learning for Speech and Writing Ability Assessment: an investigation

Daniel Karp, February 2024

# Introduction

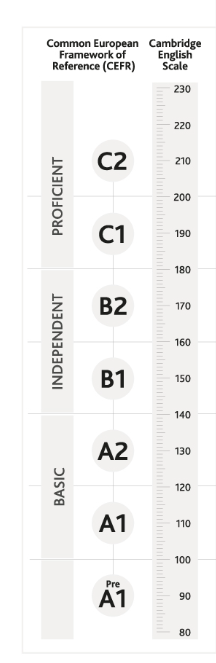
The Common European Framework of Reference for Languages (CEFR) provides a comprehensive framework for assessing language proficiency across speaking and writing. With the increasing demand for accurate and efficient language assessment tools, the integration of modelling techniques offers a promising approach to enhance the evaluation process.

Figure 1

Of course there are other methods of language assessments such as the Cambridge English Scale, which is broadly in-line with the CEFR. See Figure 1 to the left which shows the comparison (derived from *https://www.cambridgeenglish.org/exams-and-tests/cefr/*).

This report explores the application of modelling techniques in assessing speech and writing proficiency, aligned with the CEFR standards. By leveraging computational models, linguistic analysis, and machine learning algorithms, educators and language experts can gain valuable insights into learners' language capabilities, leading to more precise and reliable assessments.

The CEFR offers a standardized framework that divides language proficiency into six levels, ranging from A1 (beginner) to C2 (proficient) (see Appendix A: *Global scale - Table 1 (CEFR 3.3): Common Reference levels*). Each level encompasses specific descriptors related to speaking and writing skills, providing a detailed roadmap for evaluating language competency. However, traditional assessment methods often rely on subjective judgments and limited sample sizes, which may not accurately capture an individual's true proficiency level.

Modelling techniques offer a solution to these challenges by providing objective, data-driven approaches to language assessment. By analyzing linguistic features, such as vocabulary usage, grammatical structures, and discourse coherence, models can generate quantitative metrics that align with CEFR proficiency levels.

These metrics are essentially a breakdown of the components of speech/writing as can be see in the table in Appendix B: *Qualitative aspects of spoken language use - Table 3 (CEFR 3.3): Common Reference levels*.

Furthermore, machine learning algorithms can process large datasets of speech and writing samples, allowing for more comprehensive evaluations across diverse language contexts and learner populations.

The current report will examine a number of different modelling approaches used in assessing speech and writing proficiency, by predicting the CEFR from a range of derived features.

Ultimately, the use of modelling techniques for language assessment holds great promise for enhancing the accuracy, efficiency, and fairness of evaluating speech and writing proficiency according to CEFR standards. It would allow assessment tools to be integrated into online services that could be utilized across the board, standardizing the approach more widely.

# A Short Literature Review

According to van Dalen et al (2015), There is a need to assess the proficiency level of learners both during their studies and for formal qualifications. In the almost decade since they published, this need has not gone away. They were also one of the first to propose an ‘automatic grader’ in order to accurately assess the learner’s ability level from spontaneous, prompted speech, independent of the quality of the audio recording and their primary language. They proposed a Gaussian-based grader, that would be used as a type of assistant to determine which assessments required further grading by a human.

In fact Talia Isaacs (2018) underscored the need for machine learning being complemented by human ratings, as the ability to transcribe text accurately enough from machine recordings at that time was not as sophisticated as we now find ourselves in 2024.

Gaillat et al (2021) outlined various methods of automating the assessment process. However they only looked at CEFR grading for writing assessments. They concluded that a “…combination of lexical, syntactic, cohesive and accuracy features yields the most efficient classification across several corpora”. With a balance accuracy of around 59%. They utilized logistic regression and the elastic net models.

Settles, LaFlair and Hagiwara (2020) claim to have been the first to develop a machine learning/natural language processing methodology to align with proficiency scales and then use linguistic models to estimate difficulty directly for human adaptive testing. They developed the online Duolingo English Test, demonstrating that its scores “align significantly with other high-stakes English assessments”. They utilized a six-way multinomial regression classifier to predict CEFR level as well as a linear regression model for comparison. Their linear regression models appeared to be the most accurate. They also used a variety of features, including character length, corpus frequency, log-likelihood and Fisher score of each word.

# Data

Data was derived from both mp3 audio files and text. The audio files were provided as five short responses for each respondent, and the average metrics of pronunciation, vocab, fluency, cohesion, grammar and CEFR score were calculated across them per respondent. Associated with 445 individual responses, were 328 transcripts. As such, this study was required to generate its own transcripts to fill in the missing data. This will be discussed further in the method section.

# Method

## Data Engineering

The mp3 audio files were loaded and file properties extracted. Utilising both the AudioSegment and Librosa Python libraries, several features were extracted from the audio files. These included sampling rate, number of channels, duration, formants, pitch, intensity, speech rate, duration of pauses, frequency of pauses, mel frequency ceptral coefficients, and spectral centroid statistics.

The audio files were subsequently transcribed utilizing the ChatGPT whisper-1 model. This was used as it was found to be the most accurate at transcribing the sound files. Other models used included deepspeech v0.3.9, huggingface, vosk en-us-0.22-lgraph and en-us-0.42-gigaspeech models.

Once transcribed, additional features were derived from them. These included Brunets Index, average sentence length, text cohesion (the GloVe 6B (100d) embeddings were used here), the Flesch-Kincaid Grade Level and Coleman-Liau Index to assess readability, the Automated Readability Index, the number of detected repetitions and finally, filler words.

## Models

Two separate CEFR scoring models were required – a speech-based model, to score audio recordings, and a writing-based model, to score written text.

While the literature review uncovered different approaches to modelling, including variations on regression, there was no consistent approach. In fact, most of the literature dealing with automated scoring dealt only with written texts, and not audio files. It appeared that the regression approaches were the most successful, so this investigation decided to utilize three different types of regression models – Scikit-Learn: RandomForestRegressor, Scikit-Learn GradientBoostingRegressor, and XGBoost. It also decided to implement a classification model using a version of the BertTokenizer Transformer model to attempt to classify a particular text into one of the 6 CEFR classifications. Once classified, the scores were converted into numbers and evaluated.

These models were coded within Anaconda Jupyter notebooks using Python 3.9. Note: coding started with version 3.11, though deepspeech required version 3.9.

Prior to ingestion into the models, all data was scaled.

The models were optimized using parameter Grid Search, and the results compared to find the best of the models.

The best speech model was then converted into a microservice so that it could be made available for easy testing.

# Results

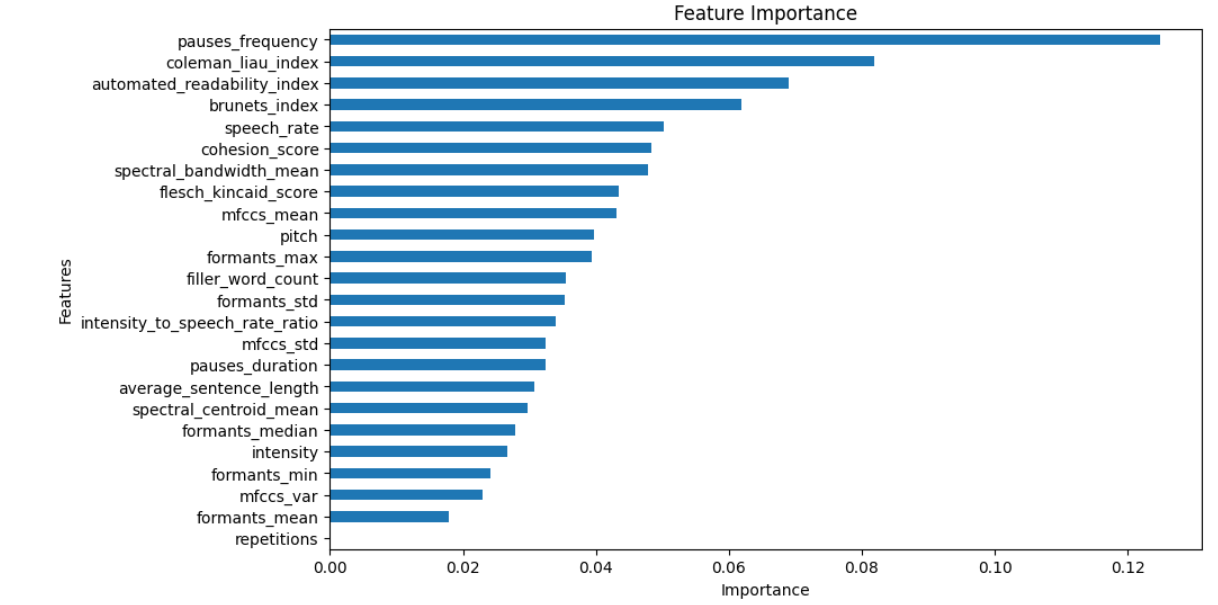
The speech model returned results almost double as accurate as the writing model.

## Speech Model

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **R²** |
| RandomForestRegressor | 0.6390628421636154 | 0.5205851270299446 |
| GradientBoostingRegressor | 0.6164068350131547 | 0.5539749388098199 |
| XGBoost | 0.6015915152550213 | 0.5751576770379657 |
| BertForSequenceClassification |  | -0.6687352326689657 |

## It appears that the XGBoost regression model was the most accurate, with the features being able to explain almost 60% of the variance between scores.

The BertForSequenceClassification model appeared to be the worst performing model, with results being worse than chance.

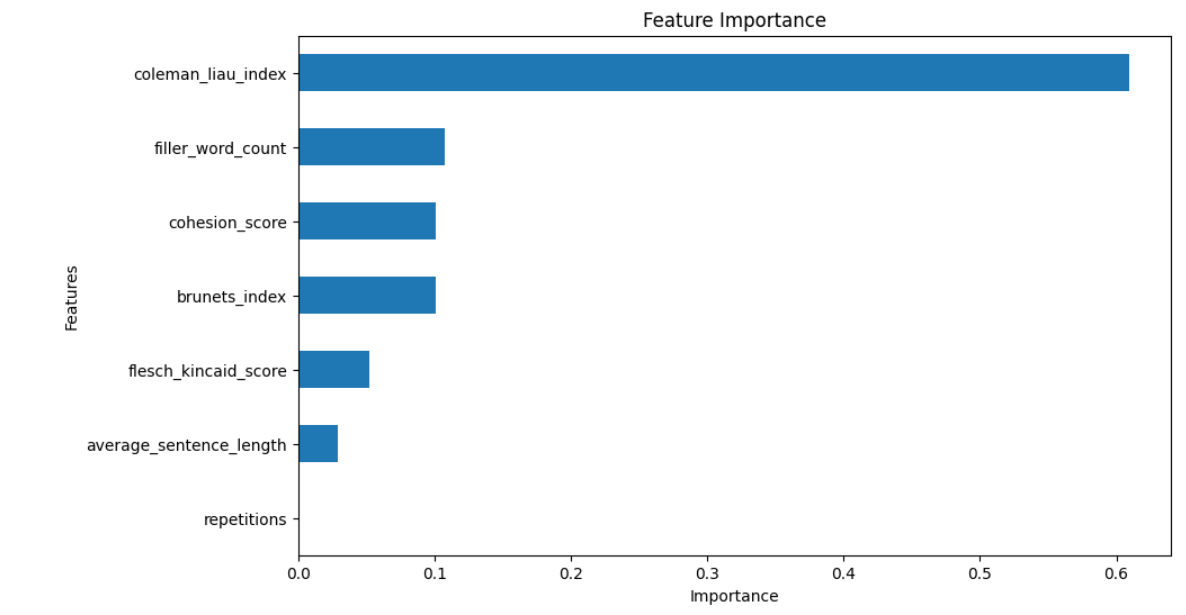


The most important features contributing to the speech model appear to be the frequency of pauses, then the Coleman-Liau Index, the automated readability index, the Brunets index and the speech rate.

## Writing Model

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **R²** |
| RandomForestRegressor | 0.8609932446424493 | 0.33890507689812954 |
| GradientBoostingRegressor | 0.9021420589314153 | 0.2742046269444769 |
| XGBoost | 0.9145461964224739 | 0.25410854865359633 |

The writing model appeared to perform much worse than the Speech model, with the RandomForestRegressor model working the best.



The most important features contributing to the writing model appear to be the Coleman-Liau Index, filler word count, cohesion score and the Brunets index and the speech rate.

# Discussion

The superior performance of the speech models can be attributed to the richness of acoustic features extracted from audio files, which capture nuances of speech delivery and pronunciation, providing richer information for CEFR assessment. In contrast, written text features may not encapsulate such nuances comprehensively, leading to lower predictive accuracy in the writing model. The poor performance of the BertForSequenceClassification model suggests that classifying text directly into CEFR levels may be challenging, requiring further exploration or feature engineering.

The observed differences in the importance of features between the speech and writing models likely stem from the distinct characteristics of spoken and written language, as well as the nature of the features extracted from each modality.

In spoken language, pauses play a crucial role in pacing, emphasis, and comprehension. A higher frequency of pauses may indicate hesitations, linguistic complexity, or deliberate pacing, all of which can provide insights into the speaker's fluency, coherence, and overall language proficiency. Additionally, speech rate, which measures the speed of speech delivery, is closely related to pause frequency. These features reflect the dynamic nature of spoken language and its temporal aspects, making them prominent indicators of language proficiency in the speech model.

Surprisingly, readability indices such as the Coleman-Liau Index and the Automated Readability Index, typically associated with written text, also emerged as important features in the speech model. This unexpected finding may suggest that speakers with higher language proficiency tend to exhibit characteristics of clearer articulation, better organization, and smoother delivery, akin to traits associated with readability in written texts. Therefore, these indices might capture aspects of clarity, coherence, and linguistic complexity in spoken language, contributing significantly to the predictive power of the speech model.

Brunet's Index, a measure of lexical diversity, appeared as an important feature in both the speech and writing models. This suggests that vocabulary richness and variation are essential markers of language proficiency across modalities. Speakers and writers with a broader vocabulary repertoire are likely to express themselves more effectively, demonstrating a higher level of linguistic competence and sophistication.

In contrast, features such as filler word count and cohesion score are more salient in the writing model. Filler words, including "um," "uh," and "like," are often indicative of hesitations, lack of fluency, or informal language use, which can detract from the clarity and coherence of written text. Cohesion, on the other hand, reflects the structural organization and flow of ideas within written text, capturing how well sentences and paragraphs are interconnected and logically linked. These features are particularly relevant in assessing the coherence and effectiveness of written communication, hence their prominence in the writing model.

It should be acknowledged, however, that although the ChatGPT whisper-1 model performed well, it was noted that it tended to excise many of the filler words from the transcripts. More experimentation with the hyperparameter settings of this model, especially Temperature, need to be conducted in order to determine if this can be rectified.

Despite the promising results, several limitations should be acknowledged. Firstly, the performance of the models may vary depending on the quality and diversity of the training data. It was felt that more training data was required, especially covering more diverse voices from different parts of the world. Also, not enough time was spend exploring other model architectures, especially deep learning variants which were mentioned occasionally in the literature. Future research could explore incorporating additional features, especially those associated with written text, as this model had much lower performance. Additionally, investigating interpretability techniques to understand the factors influencing CEFR predictions could enhance the practical utility of the developed models in educational and assessment contexts.

# Conclusion

In conclusion, this study demonstrated the feasibility of utilizing machine learning models for assessing language proficiency using both speech and written text data. The findings underscore the importance of feature selection and model choice in achieving accurate predictions, with speech-based models exhibiting superior performance over writing-based models. Moving forward, continued research in this domain could facilitate the development of robust and reliable language assessment tools with broader applicability and effectiveness.

# References

Dalen, R.C.v., Knill, K.M., Gales, M.J.F. (2015) Automatically grading learners’ English using a Gaussian process. Proc. Speech and Language Technology in Education (SLaTE 2015), 7-12, doi: 10.21437/SLaTE.2015-2

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), The Routledge handbook of contemporary English pronunciation (pp. 570–584). Abingdon, UK: Routledge

Gaillat, Thomas & Simpkin, Andrew & Ballier, Nicolas & Stearns, Bernardo & Sousa, Annanda & Bouyé, Manon & Zarrouk, Manel. (2021). Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. ReCALL. 34. 1-17. 10.1017/S095834402100029X.

Settles, B., T. LaFlair, G. and Hagiwara, M., 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8, pp.247-263.

# Appendices

## Appendix A: Global scale - Table 1 (CEFR 3.3): Common Reference levels

|  |  |  |
| --- | --- | --- |
| **PROFICIENT USER** | C2 | Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations. |
| C1 | Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices. |
| **INDEPENDENT USER** | B2 | Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options. |
| B1 | Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken.  Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans. |
| **BASIC USER** | A2 | Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters.  Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need. |
| A1 | Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help. |

*From the Council of Europe website:* [*https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale*](https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale)

## Appendix B: Qualitative aspects of spoken language use - Table 3 (CEFR 3.3): Common Reference levels

|  | **RANGE** | **ACCURACY** | **FLUENCY** | **INTERACTION** | **COHERENCE** |
| --- | --- | --- | --- | --- | --- |
| **C2** | Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms | Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions). | Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it. | Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc. | Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices. |
| **C1** | Has a good command of a broad range of language allowing him/her to select a formulation to express him/ herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say. | Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur. | Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language. | Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers. | Can produce clear, smoothly-flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices. |
| **B2** | Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much con­spicuous searching for words, using some complex sentence forms to do so. | Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes. | Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses. | Can initiate discourse, take his/her turn when appropriate and end conversation when he / she needs to, though he /she may not always do this elegantly.  Can help the discussion along on familiar ground confirming comprehen­sion, inviting others in, etc. | Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long con­tribution. |
| **B1** | Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circum-locutions on topics such as family, hobbies and interests, work, travel, and current events. | Uses reasonably accurately a repertoire of frequently used "routines" and patterns asso­ciated with more predictable situations. | Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production. | Can initiate, maintain and close simple face-to-face conversa­tion on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding. | Can link a series of shorter, discrete simple elements into a connected, linear sequence of points. |
| **A2** | Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to commu­nicate limited information in simple everyday situations. | Uses some simple structures correctly, but still systematically makes basic mistakes. | Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. | Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord. | Can link groups of words with simple connectors like "and, "but" and "because". |
| **A1** | Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations. | Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire. | Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication. | Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair. | Can link words or groups of words with very basic linear connectors like "and" or "then". |

[*https://www.coe.int/en/web/common-european-framework-reference-languages/table-3-cefr-3.3-common-reference-levels-qualitative-aspects-of-spoken-language-use*](https://www.coe.int/en/web/common-european-framework-reference-languages/table-3-cefr-3.3-common-reference-levels-qualitative-aspects-of-spoken-language-use)