

AUTOMATED ASSESSMENT OF SPOKEN ENGLISH AT CONSUMER SCALE: AN ASR + LLM PIPELINE EVALUATED AGAINST HUMAN EXAMINERS

<https://doi.org/10.5421/zenodo.22759634>

Bobirjon Mardonov

CEO of Humaid.tech & ex-CTO of Speaklish.uz

Abstract. The rapid growth of online English-learning platforms has created demand for spoken-language assessment that is instant, low-cost, and consistent, yet human examining remains the accepted standard. This study evaluates an automated pipeline that couples automatic speech recognition (ASR) with a large language model (LLM) scorer against certified human examiners. A private corpus of 1,240 monologue responses from 310 learners, collected on the basis of the Speaklish.uz platform, was scored on a CEFR-aligned analytic rubric by two trained raters and by the pipeline. Agreement between the automated overall score and the human consensus reached a Pearson correlation of $r = 0.87$ and a quadratic weighted kappa of 0.79, approaching the inter-rater ceiling of $r = 0.91$. The mean absolute error was 0.42 band on a nine-band scale, and 84.6 % of automated scores fell within half a band of the human consensus. Word error rate of the ASR stage was 11.3 % overall but rose to 19.8 % for CEFR A2 speakers, and this degradation propagated into pronunciation scoring. The findings indicate that an ASR + LLM pipeline can deliver examiner-comparable reliability for formative, consumer-scale assessment, provided that low-proficiency speech and high-stakes decisions retain human oversight.

Keywords: automated speech assessment, automatic speech recognition, large language models, spoken English, CEFR, human examiners, inter-rater reliability, computer-assisted language testing, proprietary corpus, Speaklish.uz

INTRODUCTION AND RELEVANCE

The market for online language learning has expanded to a scale that manual assessment can no longer serve on its own. Industry estimates place the global digital language-learning sector above 25 billion US dollars in 2024, with English accounting for the largest share of enrolments across every major provider [1]. A single popular application may register tens of millions of active learners in a

month, and a meaningful fraction of them expect feedback not only on vocabulary and grammar, which are straightforward to check automatically, but also on speaking, the skill learners most often name as their goal and the one hardest to measure without a trained human listener.

Speaking assessment has traditionally depended on certified examiners applying a standardised rubric, as in the IELTS Speaking test or the CEFR-aligned oral interviews used by ministries of education. This model produces defensible, high-quality judgements, but it does not scale to consumer volumes: an examiner rates a candidate in real time, one at a time, and the marginal cost of each additional test is essentially the cost of expert labour. When a platform must return a speaking score within seconds to millions of users, human examining becomes economically and logistically impossible. The gap between demand and examiner capacity is the practical problem this study addresses.

Two technologies have matured to the point where automation is now plausible. Automatic speech recognition (ASR) has improved sharply since the introduction of large self-supervised acoustic models, with reported word error rates on clear read speech falling below five per cent for major English varieties [2]. Large language models (LLMs) have, in parallel, demonstrated an ability to apply detailed textual criteria such as coherence, lexical range and grammatical accuracy to a transcript and to produce a justified score rather than a mere classification [3]. Coupling the two suggests a pipeline in which ASR converts speech to text and timing features, and an LLM applies the rubric to that output.

The relevance of the problem for Uzbekistan is direct. The Presidential Decree on measures to raise the quality of foreign-language teaching (PQ-4406, 2019) and subsequent programmes have set national targets for English proficiency [7], and a growing number of domestic ed-tech products now compete to prepare learners for international examinations. A reliable, low-cost speaking assessment engine would allow these products to give formative speaking feedback at a scale that no examiner workforce in the country could match, while reserving scarce human expertise for certification and appeals.

Despite the promise, the literature leaves an important question only partially answered. Prior work on automated speaking assessment has largely evaluated purpose-built scoring models on constrained, read-aloud tasks or has reported agreement with a single rubric dimension [4; 8]. Far less is known about how a general-purpose ASR + LLM pipeline performs on spontaneous monologue at consumer scale, how its errors are distributed across proficiency levels, and

whether its reliability approaches the agreement that two human examiners reach with each other. The present study is designed to close that gap.

Accordingly, the aim of this study is to evaluate the reliability and error structure of an ASR + LLM pipeline for spoken-English assessment by benchmarking it directly against certified human examiners on a CEFR-aligned analytic rubric. The specific objectives are: to quantify agreement between the automated overall score and the human consensus; to compare that agreement with the inter-rater agreement of the examiners themselves; to analyse how ASR accuracy and scoring error vary across proficiency levels; and to identify the conditions under which the pipeline is, and is not, a safe substitute for human judgement.

MATERIALS AND METHODS

Corpus and its provenance

The study used a private corpus of 1,240 spoken monologue responses produced by 310 adult learners of English ($n = 310$). The corpus was collected on the basis of Speaklish.uz, an Uzbek online platform for practising spoken English, over a four-month collection period. It is proprietary to the platform, was assembled for internal model development, and is not distributed with this article; accordingly, only aggregate statistics are reported here and no individual recording or transcript is reproduced. The same team has separately released an open dataset on the Hugging Face Hub which illustrates its collection and annotation practice and may be consulted by readers who wish to see the format of the underlying records [16].

Each learner completed four speaking tasks drawn from a bank modelled on the IELTS Part 2 long-turn format, in which the speaker talks for up to two minutes on a familiar prompt after one minute of preparation [15]. Learners spanned CEFR levels A2 to C1 [5], with self-reported levels confirmed by a placement test; the distribution was deliberately weighted toward the A2 to B2 range that dominates the consumer population.

Reference scoring

Two reference scores were obtained for every response. Two certified examiners, each with more than five years of experience rating IELTS-style speaking, independently scored all responses using an analytic rubric with four dimensions, namely fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation, each expressed on a nine-band CEFR-aligned scale [15]. The examiners were blind to each other's scores and to the automated output. Where their overall band differed by no more than one band, the mean was

taken as the human consensus; the 6.1 % of cases exceeding that threshold were adjudicated by a third senior examiner.

The automated pipeline

The automated pipeline comprised two stages. In the first, a self-supervised ASR model transcribed each recording and produced word-level timing, from which fluency features such as speech rate, pause frequency and duration, and filled-pause ratio were derived [2; 9]. In the second, a large language model received the transcript, the timing-derived features, and the full text of the analytic rubric, and was instructed to return a band score for each of the four dimensions together with a short justification [11; 12]. The overall automated band was computed as the mean of the four dimension scores, matching the procedure used by the human examiners.

Analysis

Analysis proceeded on three fronts. Agreement between the automated overall score and the human consensus was quantified using the Pearson correlation coefficient, the quadratic weighted kappa, the mean absolute error (MAE) in bands, and the proportion of automated scores falling within half a band of consensus. The same statistics were computed between the two human examiners to establish an inter-rater ceiling against which the pipeline could be judged [8]. Finally, ASR word error rate (WER) and scoring error were disaggregated by CEFR level to expose any systematic degradation. Statistical processing was carried out in Python; a significance threshold of $p < 0.05$ was applied to all comparisons, and 95 % confidence intervals were obtained by bootstrap resampling over learners to respect the clustering of responses within speakers.

RESULTS

Across the full corpus, the automated overall band agreed closely with the human consensus. The Pearson correlation was $r = 0.87$ (95 % CI 0.85 to 0.89) and the quadratic weighted kappa was 0.79, both significant at $p < 0.001$. The mean absolute error was 0.42 band on the nine-band scale, and 84.6 % of automated scores fell within half a band of the human consensus, while 97.1 % fell within a full band. As Table 1 shows, this level of agreement approached but did not reach the inter-rater ceiling set by the two examiners, who agreed with each other at $r = 0.91$ and a weighted kappa of 0.85.

Table 1. Agreement of the automated pipeline and of human examiners with the reference

Comparison	Pears on r	Weigh ted κ	MAE (bands)	Within 0.5 band, %
Pipeline vs. human consensus	0.87	0.79	0.42	84.6
Examiner A vs. Examiner B	0.91	0.85	0.31	90.2
Pipeline vs. Examiner A	0.85	0.77	0.45	82.8
Pipeline vs. Examiner B	0.86	0.78	0.44	83.5

Source: author's calculations on the 1,240-response corpus; all correlations significant at $p < 0.001$.

Agreement varied by rubric dimension. The pipeline matched the examiners most closely on lexical resource and grammatical range, where scoring depends on features recoverable from an accurate transcript, and least closely on pronunciation, where it depended on the acoustic stage and inherited that stage's errors [13]. Fluency and coherence occupied an intermediate position, benefiting from the timing features but suffering where disfluent speech degraded the transcript [10].

The disaggregation by proficiency level, reported in Table 2, revealed the pipeline's principal weakness. ASR word error rate rose from 6.9 % for C1 speakers to 19.8 % for A2 speakers, and scoring error tracked this degradation: the mean absolute error nearly doubled between the highest and lowest levels. In other words, the learners who most need reliable formative feedback are precisely those for whom the current pipeline is least accurate, because low-proficiency speech is harder for the acoustic model to transcribe and the resulting transcript errors propagate into every downstream dimension.

Table 2. ASR accuracy and scoring error by CEFR level

CEFR level	Responses, n	ASR WER, %	Scoring (bands)	MAE	Within band, %
A2	356	19.8	0.61		74.2
B1	402	12.4	0.44		83.1
B2	331	8.7	0.34		89.4
C1	151	6.9	0.29		91.4

Source: author's calculations. WER denotes word error rate; MAE denotes mean absolute error.

DISCUSSION

The central result is that an ASR + LLM pipeline can reach a level of agreement with human examiners that is high in absolute terms and close to the examiners' agreement with each other. A Pearson correlation of 0.87 against a human consensus, with 84.6 % of scores within half a band, means that for the majority of consumer-scale interactions the automated score is practically indistinguishable from what a trained examiner would award. Because the inter-rater ceiling itself was 0.91 rather than a perfect 1.0, part of the remaining gap reflects the irreducible subjectivity of rubric-based scoring rather than a deficiency unique to the machine [8].

These findings are consistent with earlier reports that LLM-based scorers align well with human judgement on text-derived dimensions such as lexical and grammatical quality [3; 4; 12], and they extend that evidence from constrained tasks to spontaneous monologue at scale. They also align with the long-standing observation in the speech-technology literature that recognition accuracy is the binding constraint on downstream spoken-language tasks [2; 6]. Where the present study diverges from more optimistic accounts is in its explicit measurement of how that constraint bites unevenly: the pronunciation dimension and the A2 level are not merely slightly weaker but weak enough to matter for decisions.

The proficiency-dependent degradation is the most consequential result for practice. A pipeline that is most accurate for advanced speakers and least accurate for beginners inverts the pedagogical priority, since beginners depend most on trustworthy formative feedback. The mechanism is clear, in that higher word error rates for A2 speech corrupt the transcript on which every LLM judgement rests, and it implies that improvements to the acoustic stage for low-proficiency, accented speech would yield larger gains than further refinement of the scoring model [9; 14].

The practical implication is a division of labour rather than wholesale replacement. For formative, low-stakes feedback delivered at consumer scale, which is the setting that motivated the study, the pipeline is accurate enough to be useful immediately, provided that its confidence is communicated honestly and that pronunciation feedback for low levels is treated with caution [13]. For certification, admissions, or any high-stakes decision, the error rates observed here are too large to justify removing the human examiner, who should remain the authority while the pipeline serves as a triage and consistency check.

Several limitations qualify these conclusions. The corpus, though substantial, was drawn from a single platform and a single monologue task type, so

generalisation to interactive or interview formats is not established. Because the corpus is proprietary, the results cannot at present be reproduced independently; this is a known limitation of platform-derived research, and it is one reason the team has released an open dataset illustrating the same collection format [16]. The human consensus, while carefully constructed, still rests on only two primary raters, and the inter-rater ceiling would itself shift with a larger examiner panel. The ASR and LLM components are specific versions whose behaviour will change as models are updated, so absolute figures should be read as a snapshot rather than a fixed property of the approach. Future work should extend the evaluation to dialogic tasks, test robustness across first-language backgrounds, and investigate whether targeted acoustic adaptation can close the A2 accuracy gap that this study identifies as the pipeline's chief weakness.

CONCLUSION

This study set out to determine whether an ASR + LLM pipeline can assess spoken English with examiner-comparable reliability at consumer scale, and the evidence supports a qualified affirmative. The pipeline agreed with the human consensus at a level approaching the examiners' agreement with one another, while exhibiting a clear and explainable weakness at low proficiency levels. The specific conclusions are as follows.

1. The automated pipeline reached a Pearson correlation of $r = 0.87$ and a quadratic weighted kappa of 0.79 with the human consensus, against an inter-rater ceiling of $r = 0.91$, placing its reliability close to that of the examiners themselves.
2. Agreement was strong on transcript-derived dimensions and weakest on pronunciation, confirming that the acoustic stage is the limiting factor in the current design.
3. Accuracy degraded systematically with proficiency: ASR word error rate rose from 6.9 % at C1 to 19.8 % at A2, and scoring error nearly doubled over the same range, so the learners most in need of feedback are the least well served.
4. For formative, low-stakes, consumer-scale feedback the pipeline is already accurate enough to deploy, whereas high-stakes certification decisions should retain human examiners as the final authority.

In practical terms, the results argue for a hybrid deployment in which the pipeline handles the volume of formative speaking feedback and human examiners are reserved for certification and appeals, with acoustic-model improvement for low-proficiency, accented speech identified as the most valuable direction for further work.

REFERENCES:

1. HolonIQ. Global EdTech Market Sizing: Language Learning. 2024. URL: <https://www.holoniq.com/notes/global-language-learning-market>
2. Radford A., Kim J.W., Xu T. et al. Robust Speech Recognition via Large-Scale Weak Supervision // Proceedings of the 40th International Conference on Machine Learning (ICML). 2023. Vol. 202. P. 28492-28518.
3. Naismith B., Mulcaire P., Burstein J. Automated Evaluation of Written Discourse Coherence Using Large Language Models // Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications. 2023. P. 394-403.
4. Zechner K., Higgins D., Xi X., Williamson D.M. Automatic Scoring of Non-Native Spontaneous Speech in Tests of Spoken English // Speech Communication. 2009. Vol. 51, No. 10. P. 883-895.
5. Council of Europe. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume. Strasbourg: Council of Europe Publishing, 2020. 278 p.
6. Litman D., Strik H., Lim G.S. Speech Technologies and the Assessment of Second Language Speaking // Language Assessment Quarterly. 2018. Vol. 15, No. 3. P. 294-309.
7. O'zbekiston Respublikasi Prezidentining "Chet tillarni o'rganishni ommalashtirish tizimini tubdan takomillashtirish chora-tadbirlari to'g'risida"gi PQ-4406-son qarori. 2019-yil 19-may.
8. Xi X. Automated Scoring and Feedback Systems: Where Are We and Where Are We Heading? // Language Testing. 2010. Vol. 27, No. 3. P. 291-300.
9. Bannò S., Matassoni M. Proficiency Assessment of L2 Spoken English Using Wav2vec 2.0 // Proceedings of the IEEE Spoken Language Technology Workshop (SLT). 2022. P. 1088-1095.
10. Wang X., Evanini K., Zechner K., Mulholland M. Modeling Discourse Coherence for the Automated Scoring of Spontaneous Spoken Responses // Proceedings of the 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE). 2019. P. 132-136.
11. OpenAI. GPT-4 Technical Report. 2023. URL: <https://arxiv.org/abs/2303.08774>
12. Fu J., Ng S.-K., Liu P. GPTScore: Evaluate as You Desire // Proceedings of the 2024 Conference of the North American Chapter of the ACL. 2024. P. 6556-6576.

13. Isaacs T., Trofimovich P. Pronunciation and Second Language Assessment // Language Assessment Quarterly. 2017. Vol. 14, No. 2. P. 89-95.

14. Ramanarayanan V., Suendermann-Oeft D., Lange P. et al. Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System // Multimodal Interaction with W3C Standards. Cham: Springer, 2017. P. 295-310.

15. British Council, IDP: IELTS Australia, Cambridge Assessment English. IELTS Speaking Band Descriptors (Public Version). 2023. URL: <https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx>

16. Speaklish. ORPO-dataset. Hugging Face Datasets. URL: <https://huggingface.co/datasets/Speaklish/ORPO-dataset>