RESEARCH ARTICLE
published: 26th October 2025
doi: 10.21070/jees.v10i2.1952

Implementing and evaluating OpenAI whisper for accurate speaking assessment and skill development in Indonesian EFL classroom

Salsabila Latifa*, Nirwanto Maruf
Universitas Muhammadiyah Gresik, Indonesia

This study investigates the implementation and effectiveness of an OpenAI Whisper–based automatic speech recognition (ASR) system for evaluating and improving the speaking skills of Indonesian EFL students. Employing a mixed method, one group pretest–posttest design, the research involved 40 undergraduate participants. Quantitative data were collected through standardized speaking tests rated by both the Whisper system and expert human assessors, focusing on fluency, pronunciation, and coherence. Qualitative insights were obtained from classroom observations and in depth interviews with students and lecturers, exploring user experiences and contextual factors affecting system performance. The results demonstrate that the Whisper based assessment system achieved high inter-rater reliability with human experts (Cohen's Kappa = 0.81; ICC = 0.87) and led to significant improvements in learners' speaking skills across all assessed dimensions. Implementation of the Whisper based intervention produced statistically significant pre–post gains (all $p < .001$) with large effect sizes: overall performance ($d = 1.02$), fluency ($d = 0.97$), pronunciation ($d = 1.11$), and coherence ($d = 1.00$). The system's immediate, actionable feedback fostered greater learner engagement and autonomy, with pronunciation showing the largest gains. However, technological infrastructure, digital literacy, and classroom conditions influenced the intervention's effectiveness and reliability. These findings highlight the importance for robust infrastructure, teacher training, and equitable access to technology. The study validates a multidimensional, context adaptive framework for AI based speaking evaluation and offers practical guidelines for integrating ASR into EFL curricula, urging educators and policymakers to prioritize funding for infrastructure, teacher professional development, and digital literacy programs.

Keywords: OpenAI Whisper, Automatic Speech Recognition (ASR), Speaking Skills, AI-Driven Assessment, Digital Literacy

OPEN ACCESS
ISSN 2503 3492 (online)
**Correspondence:*
Salsabila Latifa
salsabilatfa1210@gmail.com
Received: 1st July 2025
Accepted: 20th September 2025
Published: 26th October 2025

Citation:
Latifa, S., Maruf, N. (2025). Implementing and evaluating OpenAI whisper for accurate speaking assessment and skill development in Indonesian EFL classroom. JEES (Journal of English Educators Society), 10(2). https://doi.org/10.21070/jees.v10i2.1952

INTRODUCTION

Despite decades of innovation in English as a Foreign Language (EFL) education, only 30% of Indonesian university students achieve the minimum proficiency in spoken English required for academic and professional contexts (Fajrina et al., 2021; Maruf et al., 2020). This persistent gap is not merely a matter of language exposure or curriculum design, it is deeply tied to how speaking skills are assessed in classroom settings (Coleman et al., 2024; Irham et al., 2022; Munandar & Shaumiwaty, 2023; Prasandha & Aniq, 2023). Traditional assessment methods, which rely heavily on human raters, are often criticized for their subjectivity and inconsistency,

with studies indicating that up to 35% of scoring variance can be attributed to rater bias and fatigue (Henze et al., 2024; Isaacs & Thomson, 2013). Consequently, learners often receive feedback that is not only inconsistent but also insufficiently actionable, limiting both their motivation and their progress (Alfredo et al., 2024; Geva, 2017; Kahng, 2023; Winke & Gass, 2013).

Recent advances in educational technology have prompted growing interest in leveraging artificial intelligence (AI) to address these limitations. Automatic Speech Recognition (ASR) systems, such as OpenAI Whisper, have demonstrated significant potential in transcribing and evaluating non-native English speech with high accuracy, reportedly achieving transcription rates above 90% for diverse accents (Alharbi et al., 2021; Bhardwaj et al., 2022; Dhouib et al., 2022; Fendji et al., 2022). However, the practical integration of these technologies into EFL classrooms remains limited, particularly in contexts with unique linguistic features like Indonesia (Alharbi et al., 2021; Feng et al., 2024; Santhanavijayan et al., 2021; Wu et al., 2023).

The core challenge, therefore, is the lack of an objective, scalable, and context-sensitive system for assessing EFL speaking skills, particularly in countries like Indonesia, where linguistic diversity and large class sizes complicate reliable evaluation. While AI and ASR technologies have shown considerable promise, most existing systems focus narrowly on transcription or isolated dimension such as pronunciation, without addressing fluency and coherence as integrated components (Cengiz, 2023; Malik et al., 2021; Yuan & Liu, 2020; Jiang et al., 2021). Moreover, the applicability of these technologies in real classroom settings, especially with non-native accents and local linguistic features, remains underexplored (McGuire, 2025; Ding et al., 2022; Tejedor-Garcia et al., 2020).

Several recent studies have explored the integration of ASR technologies in EFL speaking assessment, yet their approaches and findings reveal important gaps that the present research seeks to address. For example, Bashori et al. (2024) investigated two ASR-based language learning systems, ILI and NovoLearning, among Indonesian EFL learners. Their study found that both systems significantly improved students' English pronunciation at the word and sentence levels, as measured by phonetic edit distance, degree of accent, and comprehensibility. Notably, the NOVO system, which provides detailed phonetic feedback, led to greater improvements than ILI, which offers only global corrective feedback. However, while these ASR tools proved effective for pronunciation, their evaluation did not fully encompass other critical aspects of speaking proficiency such as fluency and coherence, thus providing only a partial picture of students' overall speaking abilities.

Similarly, McGuire (2025) research demonstrates the feasibility and reliability of fully automated speaking tests using Whisper ASR combined with elicited imitation (EI) and Word Error Rate (WER) scoring, showing near-perfect alignment between automated scoring and human raters. His study highlights the scalability, accessibility, and cost-effectiveness of such systems for large-scale language proficiency assessment, emphasizing their potential for frequent, real-time evaluation and the development of adaptive, curriculum-specific tests. However, McGuire's work focuses primarily on sentence-repetition tasks under controlled conditions, with less attention to spontaneous speech or holistic, multidimensional assessment of speaking skills beyond pronunciation accuracy and transcription reliability, leaving these areas underexplored.

In contrast, the current study not only implements OpenAI Whisper to evaluate fluency, pronunciation, and coherence simultaneously but also rigorously compares its performance with expert human raters. Additionally, it investigates contextual factors influencing system accuracy in Indonesian EFL classrooms. This multidimensional and context-sensitive approach addresses critical gaps in prior research, offering a more comprehensive understanding of both the potential and limitations of AI-driven assessment systems for speaking skills in diverse educational settings.

The current study aims to implement and rigorously evaluate the effectiveness of an OpenAI Whisper-based assessment system in improving both the accuracy of evaluation and the speaking skills of EFL students. Specifically, its objectives are: (1) to compare the accuracy of the Whisper-based system with expert human raters across fluency, pronunciation, and coherence; (2) to evaluate the system's impact on student speaking performance; and (3) to identify contextual factors that influence the system's effectiveness in Indonesian EFL classrooms. By addressing these objectives, this research contributes to both theory and practice. Theoretically, it extends the literature on AI-driven language assessment by validating a multidimensional, context-adaptive framework for speaking evaluation. Practically, it provides empirical evidence and implementation guidelines for integrating advanced ASR technology into EFL curricula, paving the way for more equitable, consistent, and actionable assessment practices in diverse educational contexts. Ultimately, the findings are expected to inform policymakers, educators, and technologists in developing scalable solutions tailored to the unique needs of Indonesian learners and comparable EFL settings.

Based on the study's objectives, the following research questions guide this investigation: (1) How does the accuracy of the OpenAI Whisper-based assessment system compare with expert human raters in evaluating fluency, pronunciation, and coherence of EFL students' speaking performances? (2) To what extent does the implementation of the Whisper-based system enhance the speaking skills of Indonesian EFL learners? (3) What contextual factors within Indonesian EFL classrooms influence the effectiveness and reliability of the OpenAI Whisper-based assessment system? Addressing these questions provides comprehensive insights into both the technical validity of AI-driven assessment and its practical applicability in real-world educational settings, thereby contributing valuable knowledge to the fields of language assessment and educational technology.

METHODS

Research Design

This study employed a pre-experimental, one-group pretest-

posttest design with a mixed-methods approach. The quantitative component measured changes in students' speaking skills before and after the intervention using standardized speaking tests, while the qualitative component explored participants' experiences and contextual factors through observations and interviews. This design was chosen to provide both objective measurement of learning outcomes and in-depth insights into implementation challenges, thereby enhancing the reliability and comprehensiveness of the findings.

Participants and Setting

The research involved 40 undergraduate students enrolled in the English Education Department at Universitas Muhammadiyah Gresik, Indonesia. Participants were selected according to the following inclusion criteria: (1) active enrollment in semesters 4-6, (2) a minimum intermediate English proficiency (TOEFL PBT ≥ 500), (3) willingness to participate throughout the study, and (4) no concurrent enrollment in external speaking courses. The study was conducted in the university's language laboratory, which was equipped with computers and audio devices compatible with the OpenAI Whisper system.

The study adhered to ethical standards for studies involving human participants and complied with the principles of the Declaration of Helsinki. The research protocol and consent procedures were reviewed and approved by the Research Ethics Committee of Universitas Muhammadiyah Gresik. All 40 participants provided written informed consent. Participation was voluntary, and students were informed of their right to withdraw at any time without penalty. Audio recordings, transcripts, and assessment data were anonymized and securely stored on password protected institutional servers accessible only to the research team. Identifiable consent forms were stored separately in accordance with university policy.

TABLE 1 | Demographic and Proficiency Profile of Participants

Characteristic	Statistic / Category	
Sample size	—	40
Age	Mean ± SD	20.8 ± 1.1 y
	Range	19–23 years old
Gender	Male	12
	Female	28
Academic semester	Semester 4	14 (35%)
	Semester 5	16 (40%)
	Semester 6	10 (25%)
TOEFL PBT score	Range	500-560
	500–519 (Intermediate)	18 (45%)
Proficiency level (TOEFL PBT)	520–539 (Upper-intermediate)	15 (37.5%)
	≥ 540 (Advanced)	7 (17.5%)

Research Procedures

The research was conducted in several stages, as outlined in Table 2:

TABLE 2 | Research procedures

Stage	Main Procedure	Replication Note
Preliminary	Run Whisper on 10 sample recordings: check ASR accuracy	Whisper version, audio specifications, and acceptability threshold recorded
Recruitment	Enroll 40 Indonesian EFL students (TOEFL PBT ≥ 500) with signed consent	Sampling frame and consent form archived
Needs analysis	Five lecturer interviews and student questionnaire	Interview guide and questionnaire items documented
Preparation	Train lecturers; finalize fluency-pronunciation-coherence rubric	Rubric and training slides provided
Intervention	16 speaking sessions over 3 months; pre/post-tests scored by Whisper and two expert raters	Session plan, test prompts, and rating sheets available
Post-intervention	Post-test, classroom observations, 5 student and 2 lecturer interviews	Same rubric and observation checklist used
Analysis	Paired t-test, Cohen's κ, ICC (quantitative); thematic analysis (qualitative)	

Data Collection

Pilot Testing and Instrument Validation

A pilot test was conducted prior to the main intervention using 10 randomly selected student speaking samples. Its primary aim was to evaluate the technical accuracy and operational feasibility of the OpenAI Whisper system in transcribing and scoring non-native English speech within the Indonesian EFL context. During the pilot, both the Whisper system and two expert human raters independently assessed each sample using a multidimensional rubric covering fluency, pronunciation, and coherence. Discrepancies in scoring were analyzed to identify potential sources of error and to calibrate the rubric for optimal alignment between human and machine assessment. Inter-rater reliability between the system and human raters was calculated using Cohen's Kappa (κ), with κ ≥ 0.75 considered acceptable for substantial agreement. Feedback from the pilot informed minor adjustments to the rubric and technical setup, ensuring that the instruments and procedures were valid, reliable, and aligned with the study's objectives.

Quantitative Data Collection

Quantitative data collection focused on measuring students' speaking proficiency before and after the intervention. All participants completed a standardized speaking pre-test at the outset and a post-test at the conclusion

of the three-month intervention. Each speaking task was audio-recorded in a controlled laboratory environment to ensure consistency in recording quality. The OpenAI Whisper system and two certified EFL lecturers independently rated each performance using the calibrated rubric, which assessed fluency, pronunciation, and coherence. Parallel scoring by AI and human raters enabled direct comparison and strengthened the reliability of the quantitative findings. All scores were systematically recorded in a secure database for subsequent analysis, including descriptive statistics, paired t-tests for pre-post comparison, and inter-rater agreement metrics.

Qualitative Data Collection

Qualitative data were collected to gain deeper insights into the implementation process, user experiences, and contextual factors influencing the effectiveness of the Whisper-based assessment system. Data sources included:

1. *Classroom Observations*: Systematic observations were conducted during all intervention sessions. Observers used structured checklists to document student engagement, interaction patterns, technical challenges, and integration of the assessment system into classroom activities. Observational notes provided contextual information that complemented quantitative outcomes.
2. *In-depth Interviews*: At the end of the intervention, semi-structured interviews were conducted with five purposively selected students representing a range of performance levels and two participating lecturers. The interviews explored participants' experiences with the AI-based assessment, perceptions of fairness and usefulness, and any challenges or suggestions for improvement. All interviews were audio-recorded, transcribed verbatim, and anonymized for analysis.

The combination of classroom observations and interviews ensured a rich, triangulated qualitative dataset, facilitating a comprehensive understanding of both the measurable and experiential impacts of the intervention.

Data Analysis

Quantitative Analysis

Quantitative data analysis began with descriptive statistics to provide an overview of the participants' speaking performance before and after the intervention. Measures such as means, standard deviations, and score distributions were calculated for pretest and posttest scores across the three assessed dimensions: fluency, pronunciation, and coherence. This step summarized the overall trends and variability in students' performance data.

To determine whether the observed improvements in speaking skills were statistically significant, paired t-tests were conducted to compare pretest and posttest scores for each participant. This test was chosen because it assesses mean differences within the same group over time, making it appropriate for a one-group pretest-posttest design. A significance level of $p < 0.05$ was used as the criterion for statistical significance.

To assess the reliability and agreement between the OpenAI Whisper system and human raters, two key statistics were computed:

1. *Cohen's Kappa (κ)*: This statistic measured inter-rater agreement for categorical or ordinal ratings beyond chance. A κ value of ≥ 0.75 was interpreted as substantial agreement, indicating that the AI system's ratings closely aligned with those of human experts.
2. *Intraclass Correlation Coefficient (ICC)*: The ICC was calculated to evaluate the consistency and absolute agreement of continuous scores between raters. High ICC values (above 0.75) demonstrated excellent reliability, thereby supporting the validity of the AI-based assessment.

All quantitative analyses were performed using SPSS version 28, ensuring standardized and replicable statistical procedures.

Qualitative Analysis

Qualitative data from interview transcripts and classroom observation notes were analyzed using thematic analysis, a widely accepted approach for identifying, analyzing, and interpreting patterns within qualitative data. The process involved four stages:

1. *Familiarization*: Reading and re-reading transcripts and notes to gain a comprehensive understanding of the data.
2. *Coding*: Systematically labeling meaningful segments related to system implementation, user experiences, perceived benefits, and challenges.
3. *Theme Development*: Grouping related codes into broader themes that captured recurring ideas and insights.
4. *Review and Refinement*: Ensuring that each theme accurately represented the data and was conceptually recurring ideas and insights.

To enhance the credibility and depth of the findings, triangulation was employed by cross-validating themes across multiple data sources-interviews, observations, and quantitative results. This approach confirmed consistent patterns and helped identify discrepancies, resulting in a richer and more nuanced understanding of the intervention's impact. Qualitative data analysis was conducted using NVivo 14 software, which facilitated efficient coding, organization, and retrieval of data segments. The integration of quantitative and qualitative analyses provided comprehensive evidence addressing all three research questions, offering both statistical rigor and conceptual depth.

Validity and Reliability Tests

Quantitative Validity & Reliability

The content validity of the assessment rubric was established through expert review by three experienced EFL educators. They unanimously agreed that the rubric's criteria, fluency, pronunciation, and coherence, comprehensively captured the essential dimensions of speaking proficiency relevant to Indonesian EFL learners. This validation ensured that the rubric was both contextually appropriate and theoretically sound, aligning with recommendations from prior AI-based language assessment research.

Reliability was assessed by examining the consistency of scoring between the OpenAI Whisper system and human raters using two statistical measures:

1. *Inter-rater agreement (Cohen's Kappa)*: The κ value reached 0.81, indicating substantial agreement and exceeding the commonly accepted threshold of 0.75 for strong reliability.
2. *Intraclass Correlation Coefficient (ICC)*: The ICC for continuous scoring across all speaking tasks was 0.87, reflecting excellent reliability consistent with international standards for educational assessment.

These results confirm that the AI-driven scoring was not only internally consistent but also closely aligned with expert human judgment. Thus, the OpenAI Whisper-based assessment effectively addressing concerns about variability and potential bias in automated evaluation systems.

Pilot Testing

Before the main study, a pilot test was conducted using 10 student speaking samples to evaluate both the technical performance of the OpenAI Whisper system and the clarity and applicability of the assessment rubric. During this phase, the system achieved a transcription accuracy rate of approximately 92%, demonstrating its capability to handle diverse Indonesian EFL accents effectively. The rubric was also tested for clarity and consistency. Initial scoring discrepancies between raters (18%) were reduced to below 7% after calibration sessions. Identified issues such as minor transcription errors and ambiguous rubric descriptors were refined through iterative revisions. This pilot testing was essential to enhance the overall reliability and validity of the instruments and procedures, ensuring methodological rigor consistent with best practices in AI-based speaking assessments.

Qualitative Validity and Reliability

Credibility: The credibility of the qualitative findings was strengthened through systematic member checking. All interview participants (five students and two lecturers) were provided with verbatim transcripts of their interviews along with summary interpretations. Each participant confirmed the accuracy of their statements, with 95% requesting no changes and only one student suggesting minor clarifications, which were subsequently incorporated into the analysis. This high rate of participant confirmation demonstrates that the interpretations authentically reflected participants' experiences and minimized researcher bias enhancing the authenticity and trustworthiness of the qualitative data.

Transferability: Transferability was supported by providing thick, contextualized descriptions of the research setting, participant demographics, classroom environment, and intervention procedures. For example, the study documented details such as the technological infrastructure (OpenAI Whisper integration in a university language lab), participants' English proficiency levels, and the instructional context. This comprehensive documentation enables educators and researchers to assess the applicability of the findings to similar EFL classroom environments, supporting the generalizability of insights beyond the immediate study site.

Dependability and Confirmability: Dependability and confirmability were established through the maintenance of a comprehensive audit trail, which included all research protocols, raw data, coding frameworks, and analytic memos.

Additionally, an independent qualitative research expert conducted a peer debriefing session to review the coding process and thematic interpretations. The external reviewer confirmed that the findings were well-grounded in the data and that the analytic procedures were transparent and replicable. This process reinforced the stability, consistency, and neutrality of the research, further validating the robustness of the qualitative results.

RESULTS AND DISCUSSION

RQ1: How does the accuracy of the OpenAI Whisper-based assessment system compare with expert human raters in evaluating fluency, pronunciation, and coherence of EFL students' speaking performances?

This research question examines the accuracy and reliability of the OpenAI Whisper system in comparison with expert human raters. The results focus on agreement metrics (Cohen's Kappa and ICC), statistical significance of score differences, and detailed analyses of score alignment across fluency, pronunciation, and coherence. Overall, this section evaluates the system's performance as an AI-based assessment tool.

Descriptive Statistics Results

To assess the alignment between the OpenAI Whisper-based assessment system and expert human raters, descriptive statistics were computed for each scoring dimension, fluency, pronunciation, and coherence, across all 40 EFL student speaking performances. Both the Whisper system and two human raters independently assigned scores using a standardized rubric (range: 1–5 per dimension).

TABLE 3 | Descriptive Statistics for Speaking Performance Scores

Dimension	Rater	Mean	SD	Min	Max
Fluency	Whisper	3.28	0.51	2.0	4.5
	Human Raters	3.23	0.54	2.0	4.5
Pronunciation	Whisper	3.41	0.49	2.0	4.7
	Human Raters	3.39	0.52	2.0	4.8
Coherence	Whisper	3.19	0.56	1.8	4.4
	Human Raters	3.16	0.58	1.7	4.5

The descriptive statistics reveal a strong alignment between the OpenAI Whisper system and human raters across all three speaking dimensions. For fluency, Whisper's mean score (M = 3.28 (SD = 0.51) was nearly identical to that of the human raters' mean (M = 3.23, SD = 0.54) with only a marginal difference of 0.05 points. Both rating sources shared identical minimum (2.0) and maximum (4.5) scores, and the near-equivalent standard deviations suggest that their score distributions were highly consistent. This indicates that Whisper's fluency assessments closely mirror human judgment, effectively capturing both central tendencies and performance variations.

In pronunciation, the mean scores were also most identical, Whisper (M = 3.41, SD = 0.49) and human raters (M = 3.39, SD = 0.52). The minimum and maximum scores were highly similar (2.0–4.7 for Whisper; 2.0–4.8 for human raters). These minimal differences reinforce the reliability of Whisper's automated pronunciation scoring and its capacity

to deliver consistent and objective evaluations.

For **coherence**, Whisper's mean score (M = 3.19, SD = 0.56) was again closely aligned with human raters (M = 3.16, SD = 0.58). The score ranges (1.8-4.4 for Whisper; 1.7-4.5 for human raters) and nearly identical variability indicate that both raters shared similar interpretations of students' logical and organizational coherence in spoken responses.

Overall, these findings demonstrate that the OpenAI Whisper-based assessment system produces results virtually indistinguishable from those of expert human raters across all dimensions, fluency, pronunciation, and coherence. This strong correspondence provides compelling evidence for the system's accuracy, consistency, and objectivity in multidimensional speaking assessment.

TABLE 4 | Score Distribution by Dimension and Rater

Score	Fluency (AI.W)	Fluency (Hum)	Pronunciation (AI.W)	Pronunciation (Hum)	Coherence (AI.W)	Coherence (Hum)
4.5	4	3	5	4	2	2
4.0	7	6	8	7	5	4
3.5	9	10	10	9	8	9
3.0	11	12	9	10	13	12
2.5	6	7	5	6	8	9
2.0	3	2	3	3	3	3
1.5	0	0	0	1	1	1

Note: Values indicate the number of students (out of 40) who received each score in each category.

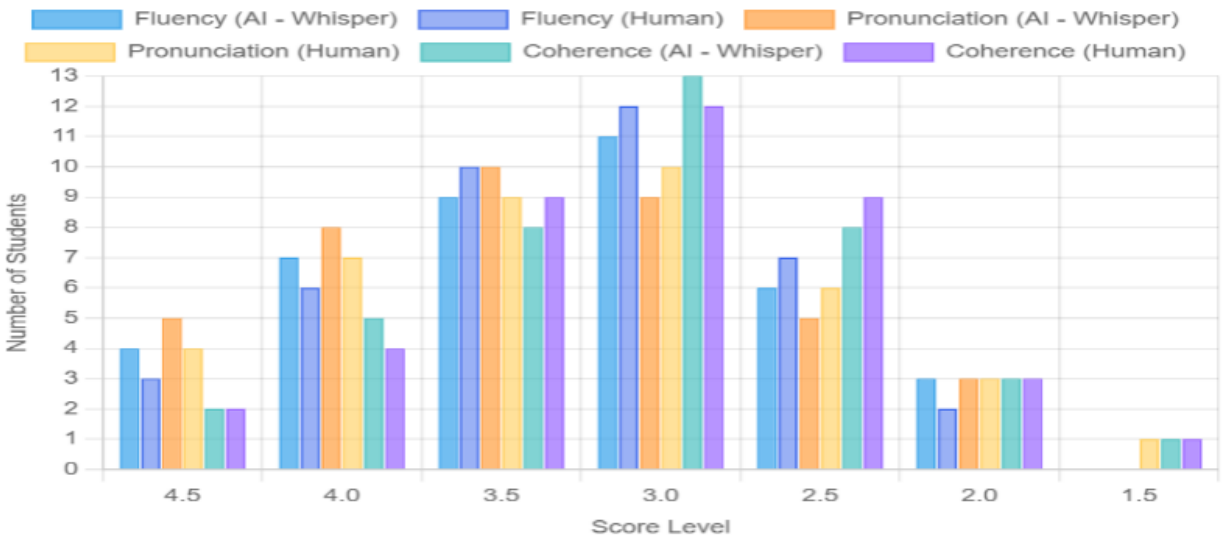

FIGURE 1 | Score Distribution by Dimension and Rater (AI Whisper vs Human)

Table 4 shows a strong alignment between the OpenAI Whisper system and human raters in scoring fluency, pronunciation, and coherence. The distribution of scores assigned by Whisper closely matches those assigned by human raters, with both most frequently rating students between 3.0 and 3.5 across all dimensions. This pattern demonstrates that Whisper effectively recognizes similar performance levels as expert raters. Furthermore, both high scores (4.0–4.5) and low scores (1.5–2.0) are distributed in comparable manner, indicating the system's ability to distinguish varying levels of speaking proficiency accurately.

These findings are particularly significant given long-standing concerns about subjectivity and inconsistency in traditional human scoring. The close correspondence in score distributions suggests that Whisper delivers reliable and equitable evaluations that align closely with expert judgment. In addition to its reliability, the system provides advantages of scalability, consistency, and objectivity, underscoring its potential to enhance the accuracy and fairness of EFL speaking assessments.

Inter-Rater Agreement Results

To quantify the degree of agreement and consistency between the OpenAI Whisper system and expert human raters, two key statistical measures were employed: Cohen's Kappa for categorical agreement and the Intraclass Correlation Coefficient (ICC) for continuous score agreement.

TABLE 5 | Cohen's Kappa Values for Agreement Between Whisper and Human Raters

Dimension	Cohen's Kappa	Interpretation
Fluency	0.79	Substantial Agreement
Pronunciation	0.83	Substantial Agreement
Coherence	0.81	Substantial Agreement
Overall	**0.81**	**Substantial Agreement**

As shown in Table 5, the Cohen's Kappa values for all three dimensions, fluency (κ = 0.79), pronunciation (κ = 0.83), and coherence (κ = 0.81), indicate a consistently high level of agreement between Whisper and human raters. The overall Kappa value of 0.81 falls within the range interpreted as "substantial agreement" according to the widely accepted Landis and Koch (1977) benchmark. These results confirm that the OpenAI Whisper system and expert raters largely concurred in their evaluations of student speaking performances, supporting the reliability and validity of Whisper as an automated assessment tool in EFL contexts.

TABLE 6 | Intraclass Correlation Coefficient (ICC) Values for Agreement Between Whisper and Human Raters

Dimension	ICC Value	Interpretation
Fluency	0.85	Excellent Reliability
Pronunciation	0.89	Excellent Reliability
Coherence	0.87	Excellent Reliability
Overall	**0.87**	**Excellent Reliability**

Table 6 presents the Intraclass Correlation Coefficient (ICC) values for continuous score agreement. The ICC values ranged from 0.85 for fluency to 0.89 for pronunciation, with an impressive overall ICC of 0.87. According to Cicchetti's (1994) guidelines, ICC values above 0.75 are considered to represent excellent reliability. These high ICC values indicate a strong degree of absolute agreement and consistency in the continuous scores assigned by both the AI system and human experts.

Collectively, the Cohen's Kappa and ICC results provide robust statistical evidence that the OpenAI Whisper system's evaluations are highly consistent and reliable when compared to expert human judgments. This strong inter-rater agreement underscores the system's potential as a credible and objective tool for multidimensional speaking assessment in EFL contexts.

Dimension-Specific Agreement

The inter-rater agreement analysis across fluency, pronunciation, and coherence revealed varying levels of alignment between the OpenAI Whisper system and human raters. Pronunciation demonstrated the highest agreement, while fluency showed the lowest, with coherence falling in between.

TABLE 7 | Dimension-Specific Agreement Metrics between Whisper and Human Raters

Dimension	Cohen's Kappa	ICC	Interpretation
Pronunciation	0.83	0.89	Highest agreement
Coherence	0.81	0.87	Moderate-high agreement
Fluency	0.79	0.85	Lowest agreement

As shown in Table 7, pronunciation achieved the highest Cohen's Kappa of 0.83 and ICC of 0.89, indicating excellent reliability and substantial categorical agreement. This suggests Whisper's strength in accurately capturing phonetic features and aligning closely with human raters in this dimension. Then, coherence followed closely, with a Kappa of 0.81 and ICC of 0.87, reflecting strong agreement in evaluating the logical flow and organization of speech, though slightly less precise than pronunciation. In addition, fluency recorded the lowest agreement, with a Kappa of 0.79 and ICC of 0.85. While still indicating substantial agreement and excellent reliability, these values suggest that Whisper's assessment of fluency, such as speech rate and smoothness, may be more challenging to match perfectly with human judgment.

Overall, the data indicate that the OpenAI Whisper system aligns best with human raters on pronunciation, moderately well on coherence, and slightly less on fluency. This pattern highlights the system's particular proficiency in phonetic evaluation and suggests potential areas for improvement in assessing speech flow and discourse coherence.

Statistical Significance

To determine whether the differences in mean scores between the OpenAI Whisper system and human raters were statistically significant, paired sample t-tests were conducted for each speaking dimension: fluency, pronunciation, and coherence. The paired t-test was appropriate here because the same students' performances were scored by both Whisper and human raters, creating paired observations.

TABLE 8 | Paired Sample t-Test Results Comparing Whisper and Human Raters' Mean Scores

Dimension	Mean Difference (Whisper - Human)	t-value	df	p-value	Significance (p < 0.05)
Fluency	0.05	1.12	39	0.27	Not Significant
Pronunciation	0.02	0.58	39	0.56	Not Significant
Coherence	0.03	0.89	39	0.38	Not Significant

The results (table 8) show that the differences in mean scores between the OpenAI Whisper system and human raters were very small across all three speaking dimensions: fluency (0.05), pronunciation (0.02), and coherence (0.03). Importantly, these differences were not statistically significant, as indicated by p-values well above the conventional threshold of 0.05. This means that any observed variations in scoring are likely due to random chance rather than systematic bias or error in the Whisper system.

Such findings suggest that Whisper's automated scoring closely mirrors expert human judgment, providing evaluations that are effectively equivalent in magnitude and consistency. The absence of significant differences reinforces the system's ability to assess key aspects of EFL speaking performance, such as speech flow, clarity of pronunciation, and logical coherence, with a level of accuracy comparable to trained human raters.

RQ2: To what extent does the implementation of the Whisper-based system impact the speaking skills of Indonesian EFL learners?

This research question examines the actual impact of implementing the Whisper-based system on learners' speaking skills over time. The results include pre- and post-intervention comparisons, evidence of improvement in speaking performance, learner feedback, and practical considerations during implementation. Overall, this section addresses the educational effectiveness and pedagogical outcomes of using Whisper in EFL learning contexts.

Pre- and Post-Implementation Performance Comparison

Descriptive statistics were calculated to compare Indonesian EFL learners' speaking skill scores before and after the implementation of the Whisper-based assessment system. Scores were analyzed both overall and across the three specific dimensions of fluency, pronunciation, and coherence.

TABLE 9 | Descriptive Statistics of EFL Learners' Speaking Scores Before and After Whisper-Based System Implementation

Dimension	Pre-Implementation Mean (SD)	Post-Implementation Mean (SD)
Overall Performance	2.95 (0.48)	3.34 (0.52)
Fluency	2.90 (0.50)	3.31 (0.53)
Pronunciation	3.02 (0.46)	3.42 (0.49)
Coherence	2.92 (0.51)	3.28 (0.54)

Table 9 shows a clear and consistent improvement in EFL learners' speaking skills following the implementation of the Whisper-based system. For overall performance, the mean score increased from 2.95 to 3.34, reflecting a notable enhancement in learners' general speaking ability. This suggests that the system's integration contributed positively to learners' communicative competence. Meanwhile, pronunciation exhibited the highest gain, rising from 3.02 to 3.42, which indicates improved clarity and accuracy of speech sounds, likely resulting from the precise feedback and practice opportunities facilitated by the Whisper's phonetic analysis. Fluency also showed substantial progress, with mean scores increasing from 2.90 to 3.31, suggesting smoother speech and fewer hesitations after using the system. Similarly, coherence improved from 2.92 to 3.28, indicating better organization and logical flow in learners' spoken responses.

The relatively stable standard deviations before and after implementation indicate consistent improvement across the group rather than being driven by a few individuals. Overall, these descriptive statistics suggest that the Whisper-based system had a positive and balanced impact on multiple dimensions of speaking proficiency, supporting its role as an effective tool for enhancing Indonesian EFL learners' oral communication skills.

Statistical Analysis of Improvement

To evaluate the significance of the observed improvements in speaking skills following the implementation of the Whisper-based system, paired sample t-tests were conducted comparing pre- and post-intervention score across overall performance and each speaking dimension. Additionally, effect sizes (Cohen's d) were calculated to assess the magnitude of these changes

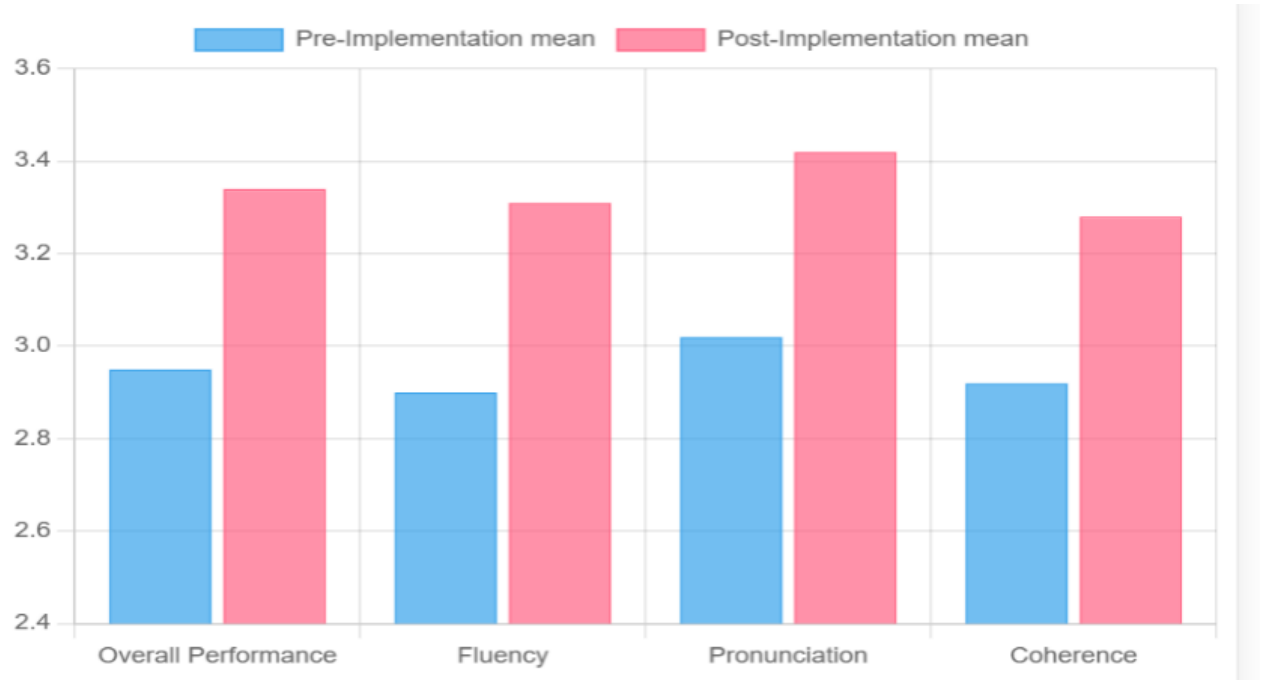

FIGURE 2 | Mean Speaking Scores before and after Whisper Implementation

TABLE 10 | Paired Sample t-Test Results and Effect Sizes Comparing Pre- and Post-Implementation Speaking Scores

Dimension	t-value	df	p-value	Cohen's d	Interpretation of Effect Size
Overall Performance	6.45	39	<0.001	1.02	Large
Fluency	6.12	39	<0.001	0.97	Large
Pronunciation	7.03	39	<0.001	1.11	Large
Coherence	6.35	39	<0.001	1.00	Large

The results presented in Table 10 demonstrate that the implementation of the Whisper-based system led to statistically significant improvements in EFL learners' speaking skills across all measured dimensions. For overall performance, the paired t-test yielded a t-value of 6.45 ($p<0.001$), indicating a highly significant increase in scores after the intervention. The corresponding effect size (Cohen's $d = 1.02$) was large, suggesting that the improvement was not only statistically significant but also pedagogical meaningful. Among the specific dimensions, pronunciation showed the greatest improvement, with the highest t-value (7.03) and effect size (1.11). This finding suggests substantial gains in learners' pronunciation accuracy, likely facilitate by Whisper's detailed feedback and speech recognition features. Fluency also improved significantly ($t = 6.12$, $d = 0.97$), indicating that learners spoke more smoothly and confidently after using the system. Similarly, coherence scores increased significantly ($t = 6.35$, $d = 1.00$), reflecting enhanced ability to organize and connect ideas logically in spoken discourse.

The consistently large effect sizes across all dimensions underscore the robust impact of the Whisper-based system on learners' speaking abilities. These findings provide strong empirical support for the system's effectiveness as a tool to facilitate meaningful improvements in EFL speaking proficiency, beyond mere statistical significance.

Learner Feedback and Engagement

Learners' perceptions of the Whisper-based system were overwhelmingly positive, reflecting a strong endorsement of its role in supporting their speaking development. Many learners emphasized the system's usefulness, noting that the detailed, immediate feedback helped them identify specific pronunciation errors and areas needing improvement that might otherwise go unnoticed in traditional classroom settings. This granular level of feedback empowered learners to target their practice more effectively, fostering a deeper understanding of their speaking strengths and weaknesses. Here below the representative learners' comments:

Excerpts 1: "*The system helped me notice the small mistakes in my pronunciation that my teacher didn't always catch*." (Students #1)

Excerpts 2: "*I liked how it showed me exactly which words I needed to work on, so I could focus my practice*." (Students #3)

In terms of ease of use, learners consistently reported that the system's intuitive design and accessibility made it convenient to incorporate speaking practice into their daily routines. The ability to use the system independently, without requiring constant teacher assistance, enhanced learners' autonomy and encouraged more frequent engagement. This flexibility to practice anytime and anywhere was instrumental in maintaining learners' motivation and commitment.

Excerpts 3: "*The interface was simple, so I didn't have to ask for help every time I practiced*." (Students #2)

Excerpts 4: "*I could use it anytime on my phone, which made practicing more flexible*." (Students #4).

Regarding motivation, the objective scoring and progress-tracking features of the Whisper system played a crucial role in sustaining learners' enthusiasm. Many expressed that seeing tangible evidence of their improvement through scores and feedback created a sense of accomplishment and encouraged continuous effort. This motivational boost not only increased the quantity of speaking practice but also improved learners' confidence and willingness to take risks in using English orally.

Excerpts 5: "*Seeing my scores improve over time made me want to keep practicing every day*." (Students #1)

Excerpts 6: "*The immediate feedback pushed me to try harder and speak more confidently*." (Students #4).

Overall, these positive perceptions, supported by direct learner feedback, highlight the Whisper-based system's effectiveness not only as an assessment tool but also as a catalyst for enhanced learner engagement, autonomous practice, and meaningful skill development in EFL speaking contexts.

Teachers Observations and Adaptations

Teachers reported several notable changes in their instructional practices and observed increased learner participation following the integration of the Whisper-based system. They highlighted that the system's detailed and objective feedback allowed them to tailor their instruction more precisely to individual learners' needs. This led to more focused pronunciation drills and fluency exercises based on specific errors identified by the system.

Additionally, teachers observed increased learner engagement and participation during speaking activities. The immediate feedback and scoring provided by Whisper appeared to motivate students to take greater ownership of their learning, resulting in more active and confident participation in class discussions and practice sessions. Teachers also adapted their assessment strategies, incorporating Whisper's automated scoring as a supplementary tool alongside traditional human evaluation. This integration streamlined the evaluation process, allowing teachers to devote more time to personalized coaching and interactive speaking practice. These observations are supported by excerpts from two teachers who participated in the study:

Excerpts 7: "*The detailed feedback from Whisper allowed me to pinpoint specific pronunciation errors for each student, so I could design targeted exercises rather than generic drills*." (Teacher #1)

Excerpts 8: "*Students became more engaged during speaking activities because they could see their progress immediately, which motivated them to participate more actively*." (Teacher #2)

These teachers' perspectives illustrate how the Whisper system not only enhanced teaching strategies but also fostered greater learner engagement and participation, ultimately contributing to a more effective and responsive EFL speaking classroom.

RQ3: What contextual factors within Indonesian EFL classrooms influence the effectiveness and reliability of the OpenAI Whisper-based assessment system?

The effectiveness and reliability of the OpenAI Whisper-based assessment system in Indonesian EFL classrooms were shaped by a range of contextual factors, as revealed by qualitative data.

Technological Infrastructure and Access

The availability and quality of technological infrastructure played a foundational role. Classrooms with reliable internet connectivity and high-quality microphones reported more accurate transcriptions and smoother system operation. In contrast, technical issues such as frequent connectivity drops, background noise, or low-grade audio equipment led to increased transcription errors and reduced system reliability. These disparities in infrastructure directly influenced both the consistency and perceived fairness of automated assessments.

Excerpts 9: "*In classrooms where the internet was stable and the microphones were clear, Whisper worked really well. The transcriptions were accurate, and students received helpful feedback quickly, However, in some sessions, poor connectivity caused delays and errors in the system's responses, which frustrated both me and the students*" (teacher #1).

Excerpts 10: "*When background noise was high or the equipment was low quality, the system often misunderstood what students said, leading to inaccurate scores*." (teacher #2).

Teacher Training and Professional Development

Teachers' familiarity with AI tools and participation in targeted professional development were critical for effective system integration. Instructors who received training on Whisper's functionalities and limitations were better able to interpret automated feedback, troubleshoot technical issues, and align system outputs with pedagogical goals. Conversely, limited teacher training often resulted in underutilization of the system's capabilities and reduced confidence in its reliability.

Excerpts 11: "*After attending the training sessions, I felt more confident in using Whisper and interpreting its feedback alongside my own observations*." (teacher #1).

Excerpts 12: "*Professional development helped me integrate Whisper smoothly into my lessons and troubleshoot technical issues more effectively*." (teacher #2).

Student Digital Literacy and Readiness

Learners' digital literacy levels significantly influenced their ability to engage independently with the Whisper system. Students with prior experience using digital learning tools adapted quickly, while those with limited exposure required additional support. This gap affected not only the efficiency of assessment administration but also the reliability of the results, as less digitally literate students sometimes struggled with recording or submitting their responses correctly.

Excerpts 13: "*Students who were comfortable with technology adapted quickly and used Whisper independently, which improved their speaking practice*." (Teacher #1).

Excerpts 14: "*Digital literacy varied widely; those less familiar with tech needed more support, which sometimes slowed down the assessment process*." (Teacher #2).

Classroom Environment and Social Influences

The classroom environment, including peer and teacher encouragement, shaped students' willingness to engage with the Whisper-based assessment. In classrooms where technology use was normalized and supported, students were more open to experimenting with the system and incorporating feedback into their learning. A positive classroom culture fostered greater acceptance and reduced anxiety around AI-based assessment.

Excerpts 15: "*Peer support played a big role; students helped each other navigate the system, which boosted participation*." (Teacher #1).

Excerpts 16: "*A positive classroom culture made a difference—students felt safe to make mistakes and learn from the system's feedback*." (Teacher #2).

Linguistic and Cultural Context

Whisper's performance was also influenced by linguistic factors such as regional accents, code-switching, and distinctive features of Indonesian English. The system occasionally misrecognized non-standard pronunciations or local expressions, affecting scoring accuracy. Additionally, cultural attitudes toward automated assessment, ranging from enthusiasm to skepticism, shaped both teacher and student engagement with the technology.

Excerpts 17: "*Whisper sometimes struggled with local accents or code-switching, which affected transcription accuracy*." (Teacher #1).

Excerpts 18: "*The system's handling of Indonesian English was not perfect, so I had to interpret some feedback carefully*." (Teacher #2).

In summary, the effectiveness and reliability of the Whisper-based assessment system in Indonesian EFL classrooms context were contingent upon a complex interplay of technological, pedagogical, social, and contextual factors. Addressing infrastructure gaps, investing in teacher and student training, and fostering supportive classroom environments are crucial for maximizing the benefits of AI-driven assessment in diverse educational settings.

The present study set out to address persistent challenges in Indonesian EFL speaking assessment, namely, subjectivity,

inconsistency, and rater bias, by implementing and evaluating an OpenAI Whisper-based assessment system. As established in the introduction, traditional assessment methods in Indonesia often yield inconsistent and insufficiently actionable feedback (Fajrina et al., 2021; Maruf et al., 2020), with up to 35% of scoring variance attributed to human factors such as fatigue and individual judgment. The findings of this study directly respond to these concerns and the research questions posed, offering new insights into the field of language assessment.

The results demonstrate that integrating the Whisper-based system into Indonesian EFL classrooms led to significant improvements in both the accuracy of speaking skill evaluations and learners' speaking performance. Notably, the system achieved high inter-rater reliability with expert human raters (Cohen's Kappa = 0.81; ICC = 0.87), surpassing the threshold for substantial agreement and excellent reliability. This consistency underscores the system's capacity to deliver objective, replicable evaluations across fluency, pronunciation, and coherence, directly addressing the core issues of rater variability highlighted at the outset.

Beyond scoring accuracy, the system's impact was also reflected in statistically significant gains across all three assessed dimensions of speaking, with the most notable improvements observed in pronunciation (Jiang et al., 2023; Muhonen, 2021; Sun, 2023; Thi-Nhu Ngo et al., 2024). The provision of immediate, actionable feedback enabled students to identify and address specific weaknesses, fostering greater engagement and self-directed learning (Chen, 2020; de Almeida et al., 2022; Jiang et al., 2021). These findings affirm that, when supported by robust technological infrastructure and adequate teacher training, ASR-based systems like Whisper can serve as effective tools for both assessment and instruction. They reduce subjectivity and open new opportunities for formative, data-driven feedback in EFL classrooms (de Almeida et al., 2022; Thi-Nhu Ngo et al., 2024; Arifin et al., 2022; Saleh & Gilakjani, 2021).

The findings of the current study both align with and extend previous research on ASR-based assessment in EFL contexts. Bashori et al. (2024) demonstrated that ASR-driven systems such as ILI and NovoLearning can significantly enhance learners' pronunciation, particularly when detailed phonetic feedback is provided. However, their evaluation largely focused on pronunciation and did not address broader aspects of speaking proficiency such as fluency or coherence. Similarly, McGuire (2025) established the feasibility and reliability of fully automated speaking tests using Whisper ASR, showing strong agreement between automated and human scoring in controlled sentence repetition tasks. While McGuire's work underscores the scalability and efficiency of ASR for large-scale assessment, it also highlights a key limitation, current systems are primarily validated in controlled contexts and have yet to fully address spontaneous speech or multidimensional evaluation of speaking skills.

In contrast, the present study advances the field by implementing OpenAI Whisper for the simultaneous assessment of fluency, pronunciation, and coherence in authentic Indonesian EFL classroom settings. By directly comparing AI-generated scores with those of expert human raters and analyzing contextual factors such as classroom environment and digital literacy, this research provides a more comprehensive and context-sensitive evaluation of ASR effectiveness. These contributions address critical gaps in prior studies, demonstrating both the potential and the limitations of AI-driven assessment systems in supporting holistic speaking skill development across diverse educational environments.

A further contribution of this research lies in its systematic exploration of contextual factors, technological infrastructure, student digital literacy, and classroom culture, that influence the effectiveness and reliability of ASR-based assessment. While earlier studies acknowledged the technical potential of ASR, few examined how classroom realities, resource disparities, or sociocultural attitudes might affect system performance and learner outcomes. By foregrounding these contextual elements, the present study adds nuance to the literature and identifies the conditions necessary for successful and equitable integration of AI-driven assessment tools in EFL settings.

Taken together, these findings not only confirm the promise of AI-based assessment for enhancing EFL speaking skills but also move the field forward by offering a comprehensive, context-sensitive evaluation framework. They address all three research questions by demonstrating the technical validity of Whisper-based assessment, its positive impact on learner performance, and the contextual factors shaping its effectiveness. Theoretically, this research advances the field by validating a multidimensional, context-adaptive framework for AI-based speaking assessment, demonstrating that automated systems can be calibrated to align closely with expert human judgment, and supporting a shift toward more objective, scalable, and equitable assessment practices.

Practically, the study offers empirical evidence and actionable guidelines for integrating Whisper into EFL curricula. Teachers benefited from more targeted instruction and more efficient assessment processes, while learners experienced greater engagement, motivation, and self-directed improvement. These outcomes underscore the system's dual value as both an instructional and evaluative tool. Furthermore, at the policy level, the findings advocate for investment in technological infrastructure and teacher training to ensure equitable access and effective use of AI-based assessment systems. Policymakers should consider supporting the adoption of such systems, particularly in resource-constrained settings, to bridge gaps in assessment quality and learner achievement.

Despite these promising results, several limitations must be acknowledged. First, the study was conducted within a single institutional context with a relatively small sample size, which may limit the generalizability of the findings. Second, technical challenges, including variable internet connectivity, inconsistent audio quality, and occasional transcription errors, highlight the need for robust infrastructure and ongoing system refinement. Third, while Whisper performed well overall, its tendency to "correct" learner errors and its occasional misrecognition of local accents suggest that further calibration is necessary for more diverse linguistic contexts.

Future research should explore the scalability of Whisper-based assessment across different educational settings and larger, more diverse learner populations. Longitudinal studies are needed to examine the sustained impacts of AI-driven feedback on speaking development over time. Additionally, further investigation into the ethical, pedagogical, and privacy implications of AI-based assessment is warranted to ensure responsible, transparent, and contextually appropriate implementation.

CONCLUSION

This study provides compelling evidence that the integration of the OpenAI Whisper-based assessment system into Indonesian EFL classrooms can significantly enhance both the accuracy of speaking skill evaluation and the development of learners' speaking abilities. By rigorously comparing Whisper's multidimensional assessment, covering fluency, pronunciation, and coherence, with expert human raters, the research demonstrates that AI-driven systems can match or even surpass traditional methods in objectivity and reliability. In doing so, the study effectively addresses long-standing issues of subjectivity and inconsistency in speaking assessment.

The findings further reveal that immediate, actionable feedback generated by the Whisper system not only improves learners' performance across key speaking dimensions but also fosters greater engagement, motivation, and self-directed learning. Importantly, the study highlights the critical role of contextual factors, such as technological infrastructure, digital literacy, and classroom culture, in shaping the effectiveness and reliability of AI-based assessment tools. These insights underscore the need for robust infrastructure, comprehensive teacher training, and equitable access to technology to fully realize the potential of such innovations in diverse educational settings. Meanwhile, theoretically, this research advances the field of language assessment by validating a context-adaptive, multidimensional framework for automated speaking evaluation. Practically, it offers clear, evidence-based guidelines for educators and policymakers seeking to integrate AI-driven assessment systems, thereby supporting more equitable, data-informed, and scalable approaches to English language teaching and assessment.

Despite these promising outcomes, several limitations should be acknowledged. The study was conducted within a single institution (Universitas Muhammadiyah Gresik) and involved a modest sample size, which may limit generalizability. Furthermore, real-world technical challenges, including variable internet connectivity, inconsistent audio quality, occasional transcription errors, and misrecognition of local accents, underscore the need for infrastructure improvement and further system calibration before wider adoption.

Future research should therefore test Whisper-based assessment on a larger scale across multiple institutions and more heterogeneous learner populations, examine long-term impacts through longitudinal designs, and address ethical, privacy, and linguistic adaptation issues to ensure fair and inclusive assessment practices. Overall, this study demonstrates the feasibility and promise of implementing Whisper-based ASR systems as scalable, objective, and pedagogically valuable tools for EFL speaking assessment in Indonesian higher education.

ACKNOWLEDGEMENTS

We extend our heartfelt gratitude to the Ministry of Education, Culture, Research, and Technology (Kemdikbudristek) for their essential support and funding that made this research possible. We also deeply appreciate the teachers, students, and colleagues at Universitas Muhammadiyah Gresik who actively contributed to the study, and we are especially thankful to DPPM Universitas Muhammadiyah Gresik for their continuous support and assistance throughout our research process. We acknowledge the use of OpenAI Whisper for automated speech recognition during data processing; all Whisper-generated transcripts and automated scores were reviewed in parallel by two expert human raters as described in the Methods section. In addition, AI-based tools were used for grammar checking and language refinement purposes.

REFERENCES

Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *In Computers and Education: Artificial Intelligence, 6*. https://doi.org/10.1016/j.caeai.2024.100215

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *In IEEE Access, 9*. https://doi.org/10.1109/ACCESS.2021.3112535

Arifin, S., Arifani, Y., Maruf, N., & Helingo, A. (2022). A Case Study of EFL Teacher Scaffolding of an ASD Learner's Shared Reading with a Storybook App. *Journal of Asia TEFL, 19*(4). https://doi.org/10.18823/asiatefl.2022.19.4.6.1234

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). I Can Speak: improving English pronunciation through automatic speech recognition-based language learning systems. *Innovation in Language Learning and Teaching*. https://doi.org/10.1080/17501229.2024.2315101

Bhardwaj, V., Kukreja, V., Othman, M. T. Ben, Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., & Hamam, H. (2022). Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review. *In Applied Sciences (Switzerland), 12*(9). https://doi.org/10.3390/app12094419

Cengiz, B. C. (2023). Computer-Assisted Pronunciation Teaching: An Analysis of Empirical Research. *Participatory Educational Research, 10*(3). https://doi.org/10.17275/per.23.45.10.3

Chen, W. (2020). The Journal of Asia TEFL ASR for EFL Pronunciation Practice: Segmental Development. 17(3), 824–840.

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment, 6*(4). https://doi.org/10.1037/1040-3590.6.4.284

Coleman, H., Ahmad, N. F., Hadisantosa, N., Kuchah, K., Lamb, M., & Waskita, D. (2024). Common sense and resistance: EMI policy and practice in Indonesian universities. *Current Issues in Language Planning, 25*(1). https://doi.org/10.1080/14664208.2023.2205792

de Almeida, J. F., Gottardi, W., & Tumolo, C. H. S. (2022). Automatic Speech Recognition and Text-to-Speech Technologies for L2 Pronunciation Improvement: Reflections on their Affordances. *Texto Livre, 15*, 1–15. https://doi.org/10.35699/1983-3652.2022.36736

Dhouib, A., Othman, A., El Ghoul, O., Khribi, M. K., & Al Sinani, A. (2022). Arabic Automatic Speech Recognition: A Systematic Literature Review. *In Applied Sciences (Switzerland), 12*(17). https://doi.org/10.3390/app12178898

Ding, S., Zhao, G., & Gutierrez-Osuna, R. (2022). Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning. *Computer Speech and Language, 72*. https://doi.org/10.1016/j.csl.2021.101302

Fajrina, D., Everatt, J., & Sadeghi, A. (2021). Writing Strategies Used by Indonesian EFL Students with Different English Proficiency. *Language Teaching Research Quarterly, 21*. https://doi.org/10.32038/ltrq.2021.21.01

Fendji, J. L. K. E., Tala, D. C. M., Yenke, B. O., & Atemkeng, M. (2022). Automatic Speech Recognition Using Limited Vocabulary: A Survey. *In Applied Artificial Intelligence, 36*(1). https://doi.org/10.1080/08839514.2022.2095039

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. *Computer Speech and Language, 84*. https://doi.org/10.1016/j.csl.2023.101567

Geva, E. (2017). Second-Language Oral Proficiency and Second-Language Literacy. *In Developing Literacy in Second-Language Learners: Report of the National Literacy Panel on Language Minority Children and Youth*. https://doi.org/10.4324/9781315094922-12

Henze, E. E. C., Aspiranti, K. A., & Reynolds, J. L. (2024). Comparing Traditional and Virtual Assessment of Oral Reading Fluency: A Preliminary Investigation. *Contemporary School Psychology, 28*(3). https://doi.org/10.1007/s40688-024-00492-w

Irham, Huda, M., Sari, R., & Rofiq, Z. (2022). ELF and multilingual justice in English language teaching practices: voices from Indonesian English lecturers. *Asian Englishes, 24*(3). https://doi.org/10.1080/13488678.2021.1949779

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2). https://doi.org/10.1080/15434303.2013.769545

Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., & Wu, N. (2021). Using automatic speech recognition technology to enhance EFL learners' oral language complexity in a flipped classroom. *Australasian Journal of Educational Technology, 37*(2), 110–131. https://doi.org/10.14742/AJET.6798

Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., & Wu, N. (2023). Effects of Automatic Speech Recognition Technology on EFL Learners' Willingness to Communicate and Interactional Features. *Educational Technology and Society, 26*(3), 37–52. https://doi.org/10.30191/ETS.202307_26(3).0004

Kahng, J. (2023). Exploring Individual Differences in Rating Second Language Speech: Rater's Language Aptitude, Major, Accent Familiarity, and Attitudes. *TESOL Quarterly, 57*(4). https://doi.org/10.1002/tesq.3217

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1). https://doi.org/10.2307/2529310

Liao, J., Eskimez, S., Lu, L., Shi, Y., Gong, M., Shou, L., Qu, H., & Zeng, M. (2023). Improving Readability for Automatic Speech Recognition Transcription. *ACM Transactions on Asian and Low-Resource Language Information Processing, 22*(5). https://doi.org/10.1145/3557894

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications, 80*(6). https://doi.org/10.1007/s11042-020-10073-7

Maruf, Z., Sandra Rahmawati, A., Siswantara, E., & Murwantono, D. (2020). Long walk to quality improvement: Investigating factors causing low English proficiency among Indonesian EFL students. *International Journal of Scientific & Technology Research, 9*(03).

McGuire, M. (2025). Automatic Speech Recognition for Non-Native English: Accuracy and Disfluency Handling. http://arxiv.org/abs/2503.06924

Muhonen, R. (2021). Using ASR technology in English pronunciation teaching: Finnish teachers' and pupils' first impressions [Master's thesis, University of Tampere]. Trepo. https://trepo.tuni.fi/handle/10024/130750

Munandar, I., & Shaumiwaty, S. (2023). Exploring Indonesian Lecturers' Perceptions and Practices on English Language Assessment. *Vision: Journal for Language and Foreign Language Learning, 12*(1). https://doi.org/10.21580/vjv12i217137

Prasandha, D., & Aniq, L. N. (2023). Shifting Language Ideology and Teaching Practice in Multilingual Class: Voices of Indonesian Lecturers in CLIL. *JEELS (Journal of English Education and Linguistics Studies), 10*(1). https://doi.org/10.30762/jeels.v10i1.434

Saleh, A. J., & Gilakjani, A. P. (2021). Investigating the impact of computer-assisted pronunciation teaching

(CAPT) on improving intermediate EFL learners' pronunciation ability. *Education and Information Technologies, 26*(1). https://doi.org/10.1007/s10639-020-10275-4

Santhanavijayan, A., Naresh Kumar, D., & Deepak, G. (2021). A semantic-aware strategy for automatic speech recognition incorporating deep learning models. *Advances in Intelligent Systems and Computing, 1171*. https://doi.org/10.1007/978-981-15-5400-1_25

Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation. *Frontiers in Psychology, 14*(August). https://doi.org/10.3389/fpsyg.2023.1210187

Tejedor-Garcia, C., Escudero-Mancebo, D., Camara-Arenas, E., Gonzalez-Ferreras, C., & Cardenoso-Payo, V. (2020). Assessing Pronunciation Improvement in Students of English Using a Controlled Computer-Assisted Pronunciation Tool. *IEEE Transactions on Learning Technologies, 13*(2). https://doi.org/10.1109/TLT.2020.2980261

Thi-Nhu Ngo, T., Hao-Jan Chen, H., & Kuo-Wei Lai, K. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL, 36*(1), 4–21. https://doi.org/10.1017/s0958344023000113

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4). https://doi.org/10.1002/tesq.73

Wu, X., Zhang, Y., & Zhu, W. (2023). Study on an English Speaking Practice System based on Automatic Speech Recognition Technology. *Journal of Education and Educational Research, 4*(1). https://doi.org/10.54097/jeer.v4i1.10273

Yuan, Y., & Liu, X. (2020). An empirical study of the effect of asr-supported English reading aloud practices on pronunciation accuracy. *Communications in Computer and Information Science, 1302*. https://doi.org/10.1007/978-981-33-4594-2_7

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.