Journal of Artificial Intelligence and Software Engineering
Vol. 6, No. 2, Juni 2026, pp. 203~211
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v6i2.9201

Evaluating Whisper Model on Indonesian Educational Videos Transcription with Varying Audio Conditions

Fathia Sabrina[1*], Fitria Nilamsari[2], Rinny Asasunnaja[3]
[1,2,3] Departemen Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Syiah Kuala, Banda Aceh, 23111, Indonesia

Informasi Artikel

Diterima : 25 Mei 2026
Revisi : 23 Juni 2026
Publikasi : 30 Juni 2026

Kata Kunci:

Automatic Speech Recognition
Whisper
Indonesian Language
Educational Video
WER
CER

ABSTRAK

Meningkatnya penggunaan video pembelajaran dalam pendidikan digital menunjukkan pentingnya sistem *Automatic Speech Recognition* (ASR) yang akurat untuk mendukung aksesibilitas, pembuatan subtitle otomatis, dan lingkungan pembelajaran yang inklusif. Penelitian ini mengevaluasi performa model ASR Whisper base pada video pembelajaran berbahasa Indonesia dengan karakteristik audio dan kondisi produksi yang beragam. Beberapa kategori video pendidikan dianalisis, meliputi rekaman kuliah, podcast, wawancara, dan video edukasi animasi. Rekaman audio dikonversi ke format WAV dan dievaluasi menggunakan Word Error Rate (WER) dan Character Error Rate (CER). Rekaman berdurasi panjang juga dibagi menjadi beberapa segmen berdurasi sekitar 30 menit untuk menganalisis konsistensi transkripsi sepanjang waktu. Hasil penelitian menunjukkan bahwa performa transkripsi bervariasi berdasarkan kondisi rekaman, dengan nilai WER berkisar antara 18% hingga 47% dan CER antara 5% hingga 23%. Analisis juga menemukan pola substitusi berulang yang dipengaruhi oleh kemiripan fonetik, ekspresi percakapan, dan frasa kontekstual budaya. Secara keseluruhan, Whisper base menunjukkan kemampuan transkripsi yang cukup efektif untuk multimedia pendidikan berbahasa Indonesia dalam kondisi rekaman realistis.

ABSTRACT

The increasing use of video-based learning in digital education highlights the importance of accurate automatic speech recognition (ASR) systems to support accessibility, subtitle generation, and inclusive learning environments. This study evaluates the performance of the Whisper base ASR model on Indonesian educational videos with diverse audio characteristics and production conditions. Several categories of educational videos were analyzed, including classroom lectures, podcasts, interviews, and animated educational content. Audio recordings were converted into WAV format and evaluated using Word Error Rate (WER) and Character Error Rate (CER). Long-duration recordings were additionally segmented into approximately 30-minute chunks to analyze transcription consistency over time. The results showed that transcription performance varied across recording conditions, with WER values ranging from 18% to 47% and CER values ranging from 5% to 23%. The analysis also identified recurring substitution patterns influenced by phonetic similarity, conversational expressions, and culturally contextual phrases. The findings indicate that Whisper base provides reasonably effective transcription capability for Indonesian educational multimedia under realistic recording conditions.

Penulis Koresponden

Email: fathia.sabrina@usk.ac.id

Cara sitasi IEEE::

[1] F. Sabrina, F. Nilamsari, R. Asasunnaja, "Evaluating Whisper Model on Indonesian Educational Videos Transcription with Varying Audio Conditions" *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 6, no. 2, p. 203-211, Juni 2026. doi:10.30811/jaise.v6i2.9201

1. INTRODUCTION

Modern education increasingly relies on multimedia learning approaches that combine visual and auditory information to improve student engagement, retention, and overall learning effectiveness. Video-based learning, in particular, has been shown to positively influence learning outcomes, motivation, engagement, and instructional effectiveness across various educational settings[1], [2]. As educational videos become an essential component of both formal and informal learning environments, technologies that improve the accessibility, organization, and usability of video content are becoming increasingly important. Along with the rapid implementation of Artificial Intelligence in education, Automatic Speech Recognition (ASR), or Speech-to-Text technology, has emerged as an important supporting technology for video-based learning through applications such as lecture transcription [3], subtitle generation[4], and content summarization [5]. Beyond convenience, ASR technologies also contribute to inclusive education by supporting students with disabilities. Prior studies have shown that ASR systems can facilitate real-time transcription, support note-taking activities, and improve classroom participation for students with hearing impairments and other special needs [6]. In the Indonesian context, ASR-based learning support has also been explored to assist deaf students in classroom activities [7].

Recent advancements in large language model-based ASR systems, such as Whisper, have significantly improved multilingual transcription performance. For example, Adila et al. [8] evaluated multilingual ASR models using benchmark speech corpora containing diverse speech variabilities, while William and Zahra [9] assessed Whisper performance using structured Indonesian speech datasets. Although these studies provide valuable performance benchmarks, the evaluated recordings remain substantially different from educational multimedia content that often contains spontaneous interactions, environmental noise, multiple speakers, and varying production quality. Prior research has emphasized the importance of evaluating ASR systems under realistic acoustic conditions, as speaking style, speech context, and background noise can significantly influence transcription performance [9], [10]. Although recent studies have demonstrated the effectiveness of Whisper for Indonesian speech recognition [9], most evaluations have been conducted using benchmark datasets, curated speech corpora, or recordings collected under relatively controlled conditions. Such datasets are valuable for measuring baseline transcription accuracy but may not fully represent the acoustic variability commonly encountered in educational multimedia environments as mentioned before. This limitation is particularly relevant in educational contexts, where caption quality directly influences learning comprehension, accessibility, and user experience in video-based learning environments [4]. Consequently, there remains a need to evaluate ASR performance using realistic educational recordings that reflect the diversity of audio conditions encountered in actual learning settings.

Preceding analyses in audio classification commonly categorize audio into speech, music, noise, and combinations of these elements, particularly in multimedia audio analysis and speech processing research [11], [12]. Recent research in educational video analysis has also moved beyond basic metadata to focus on the interactive patterns of the content such as distinguishing between monologue-based (single-speaker) and dialogue-based (multi-speaker) presentations [13], as well as considering the production value and style in its evaluation [14]. Inspired by these foundations, this study aims to evaluate Whisper Model across multiple educational video conditions characterized by differences in noise level, speech interaction, and production style. By adopting this method, this study emphasizes realistic educational audio conditions commonly encountered in Indonesian learning environments which may differ from the previous evaluation that primarily focus on clean or standardized datasets [9].

Through a systematic examination of transcription performance across varied educational audio conditions, this study is designed to yield critical insights into Whisper's robustness within realistic learning environments. The findings are expected to contribute to the Indonesian ASR evaluation research in several ways. First, the study evaluates Whisper-based transcription performance across diverse educational video categories, including lectures, podcasts, interviews, and animated educational contents. Second, the research investigates transcription consistency across long-duration recordings through temporal segmentation analysis. Third, the study performs qualitative word-level error analysis to identify recurring Indonesian transcription patterns beyond aggregate WER and CER metrics. By clarifying the specific challenges ASR systems

encounter when processing diverse Indonesian educational videos, the objective is to provide a foundational yet granular understanding of model behavior under realistic acoustic conditions.

2. METHODOLOGY

To systematically evaluate the transcription performance of Whisper on Indonesian educational videos, this study employed a multi-stage experimental workflow consisting of data collection, audio preprocessing, ASR transcription, quantitative evaluation, and qualitative error analysis. The complete research methodology is presented in Figure 1.

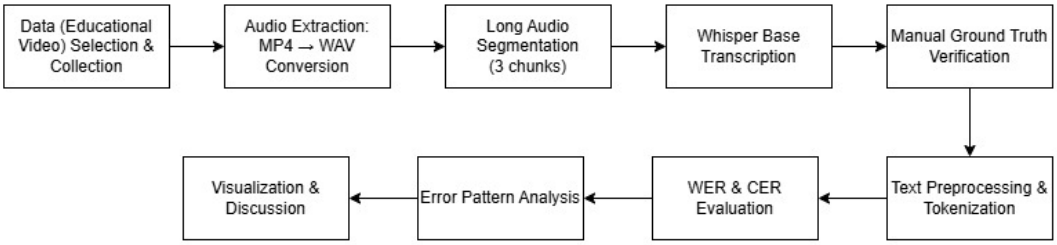

Figure 1. Research methodology workflow for Indonesian educational video transcription evaluation using Whisper ASR

2.1. Data Selection & Collection

Seven types of educational videos were selected to represent distinct audio conditions. The selection adopted in this study is inspired by prior research in audio classification and instructional video analysis, where audio signals are commonly grouped into categories (speech, music, noise and combinations of these elements), speaker interaction patterns (monologue and dialogue) and production style (low production and high production) [11] - [14]. Videos were collected from publicly available YouTube sources with keywords related to the categorization. The dataset consisted of one long lecture recording, one short lecture recording, one long podcast-style educational discussion, one short interview, and three animated educational clips. These videos were intentionally selected to represent diverse educational audio conditions across different levels of noise, speech interaction, and production style.

Noise characterization was conducted using a combination of acoustic indicators and perceptual assessment, as real-world multimedia recordings may exhibit acoustic properties that are not fully represented by numerical measurements alone. The signal-to-noise ratio (SNR) was calculated as the acoustic indicator for each baseline audio sample, following the thresholds established in recent ASR robustness benchmarks [15]. The Root Mean Square (RMS) energy of the speech signal and detected silent segments representing background noise which is used in finding SNR are calculated using Pydub library in Python. The equation of Root Mean Square (RMS) is described as follows:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {x_i}^2} \quad (1)$$

where:

- x_i represent the audio amplitude,
- N represent the total number of samples.

The RMS values were then converted into decibels relative to full scale (dBFS) using:

$$dBFS = 20\ log_{10}(RMS) \quad (2)$$

and the Signal-to-Noise Ratio (SNR) was subsequently calculated as:

$$SNR_{dB} = dBFS_{speech} - dBFS_{noise} \quad (3)$$

where:

- $dBFS_{speech}$ represent the average speech signal level,
- $dBFS_{noise}$ represent the estimated background noise level obtained from automatically detected silent audio segments.

In addition to SNR measurements, perceptual assessment was conducted to characterize recording conditions. Each recording was manually reviewed by the researcher, considering factors such as background noise, speech overlap, recording artifacts, microphone consistency, and overall speech clarity. Based on these observations, descriptive noise condition labels were assigned to each recording to provide contextual information regarding the audio environment. The perceptual assessment was not intended to replace quantitative SNR measurements, but rather to complement them by capturing characteristics that may influence ASR performance in real-world educational multimedia recordings. The detailed summary of each video used in this study is provided in table 1.

Table 1. Detailed summary of educational video used in Whisper ASR review

Video ID	Description	Duration (hour:minute: second)	Signal-to-Noise Ratio (dB)	Perceived Noise Condition	Speech Interaction	Production Style
Lecture_Long	Formal lecture recording provided by university	1:28:54	24.52	Relatively clean speech with environmental noise and student interruptions	Monologue	No post production process (raw)
Lecture_Short	Short lecture recording with controlled environment	11:56	25.91	Cleaner recording with low environmental noise	Monologue	With post production process
Podcast	Professionally recorded dialogue-based discussion	1:23:23	5.9	Low background noise but compressed audio dynamics	Dialogue	High production
Interview	Self-recorded interview with common Indonesian religious expression	6:55	21.75	Moderate environmental and conversational interference	Dialogue	With post production but limited gear used
Animation_1	Foreign animated educational content dubbed into Indonesian	19:04	18.58	Medium noise from background music and sound effects	Mixed	High production
Animation_2	Animated educational content in narrative form with audio effects	11:25	29.9	Very clean speech with controlled background music and sound effects	Monologue	High production
Animation_3	Local animated educational content with realistic kids' speech	10:59	24.68	Relatively clean audio but inconsistent speech and sound effect volume	Dialogue	High production

A notable observation from the acoustic and perceptual assessments is that the podcast recording produced lower computed SNR values than several other recordings, despite being perceived as relatively clean due to its professional production quality and consistent microphone usage. This discrepancy may be attributed to the characteristics of highly produced audio, which often includes dynamic compression, mastering effects, and limited silent segments that can influence RMS-based noise estimation. This finding highlights the importance of complementing quantitative acoustic indicators with perceptual assessment when characterizing real-world educational multimedia recordings.

2.2. Audio Processing and Transcription

Audio preprocessing, extraction, and format conversion were implemented using MoviePy library. Audio tracks were converted into WAV format to preserve spectral speech information and reduce compression artifacts that may negatively affect ASR performance [16]. This choice is also consistent with practical recommendations from modern speech transcription platforms, which commonly process audio internally in uncompressed waveform formats. For videos with a duration exceeding one hour, a segmentation strategy was applied by dividing the recordings into three approximately 30-minute segments in order to analyze potential variations in transcription performance over time and to compare segment-level performance with overall evaluation metrics.

This study utilized the Whisper base model as the primary ASR model for all transcription tasks due to its balance between transcription capability and computational efficiency [17]. The evaluation focused on transcription accuracy rather than real-time inference performance, hence experiments were executed using standard CPU-based processing without GPU acceleration. Transcription output was stored as plain text for further evaluation.

2.3. Ground Truth Preparation and Evaluation

To evaluate transcription accuracy, reference transcripts were manually reviewed and corrected to construct the ground truth dataset. The transcription outputs were carefully checked against the original audio recordings to ensure consistency and minimize transcription errors unrelated to the ASR model performance. This manually verified transcription was subsequently used as the reference text during evaluation.

The primary evaluation metric used in this study is Word Error Rate (WER), which is widely recognized as the standard evaluation metric for ASR systems due to its ability to measure transcription errors at the word level [18]. WER is commonly used to assess how accurately an ASR system reproduces intelligible speech content in comparison to human-verified reference transcripts. The metric is calculated as:

$$WER = \frac{S+D+I}{N} \quad (4)$$

where:

- S means Substitution, represent total words replaced by incorrect words,
- D means Deletion, represent total words unable to be transcribed by ASR system,
- I means Insertion, represent total unnecessary words added by ASR system,
- N represent total correct words in the ground truth.

In addition to WER, Character Error Rate (CER) was also calculated to provide a more detailed evaluation at the character level. CER is particularly useful for multilingual ASR evaluation and for analyzing partial transcription correctness that may not be fully captured by word-level metrics [19]. While following the same algorithmic structure as WER, the CER focuses on character-level edit operations (substitutions, deletions, and insertions) to determine accuracy. Both WER and CER calculation are performed using jiwer library.

This study also performed qualitative and quantitative transcription error analysis to investigate recurring ASR behavior patterns across different educational video conditions. Error components generated during the evaluation process, including substitution, deletion, insertion, and correctly recognized words (hits), were extracted and analyzed to identify dominant transcription error characteristics. The analysis focused particularly on recurring Indonesian lexical substitutions, phonetic similarity patterns, and conversational speech behavior observed in the generated transcriptions.

To further support interpretation of the findings, the evaluation results were visualized using comparative grouped bar charts and normalized stacked bar charts. The grouped visualization was used to compare WER and CER performance across all evaluated recordings, while the stacked visualization illustrated the proportional contribution of substitution, deletion, and insertion errors within each transcription result. These visualizations were intended to provide clearer comparative insight into transcription behavior under different educational multimedia conditions.

3. RESULT AND DISCUSSION

3.1. Overall Transcription Performance

The evaluation results demonstrate that Whisper base model was capable of generating reasonably accurate Indonesian transcriptions across various educational video conditions. The obtained Word Error Rate (WER) values ranged from 18% to 47%, while Character Error Rate (CER) values ranged from 5% to 23% depending on the audio characteristics of each recording.

Figure 2 illustrates substantial variation in transcription performance across different educational video environments. The interview recording produced the highest WER and CER values among all evaluated recordings, indicating that spontaneous dialogue, inconsistent recording quality, environmental interference, and local dialect significantly affected transcription accuracy. In contrast, the animated educational recordings and short lecture recordings achieved lower error rates due to their clearer narration and more controlled audio conditions.

The podcast recording achieved relatively stable transcription performance despite involving multiple speakers. This finding suggests that audio production quality and microphone consistency may have a stronger influence on ASR performance than speaker count alone. The long classroom lecture recording also produced a relatively high WER value of approximately 29%, despite consisting primarily of monologue speech. This number indicates that background classroom interaction, environmental interruptions, and uncontrolled

recording conditions may substantially affect ASR transcription quality even in single-speaker scenarios. Additionally, the consistent gap between WER and CER across all recordings indicates that many transcription errors preserved partial phonetic similarity rather than producing completely unrelated outputs.

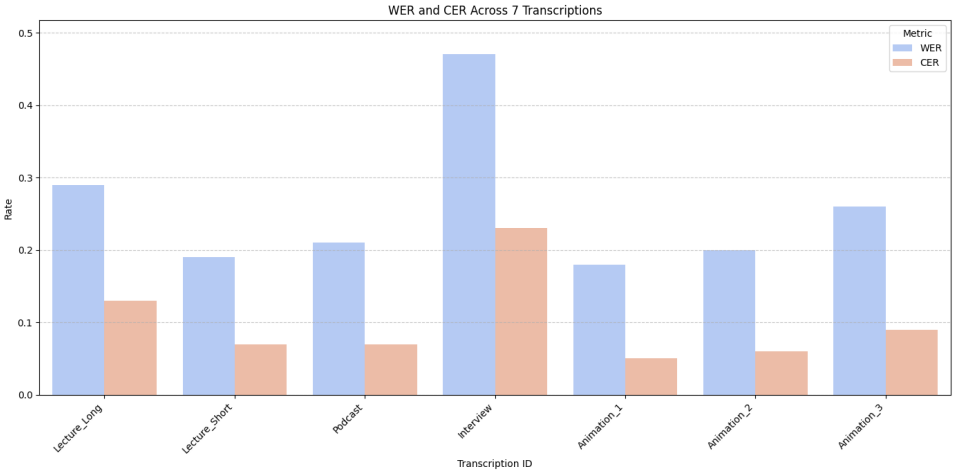

Figure 2. Comparison of Word Error Rate (WER) and Character Error Rate (CER) across all evaluated educational video recordings

3.2. Temporal Segmentation Analysis

Additional analysis was conducted on recordings exceeding one hour in duration by evaluating each approximately 30-minute segment individually. The lecture recording produced WER values of 51%, 54%, and 44% across the three temporal segments, while the podcast recording produced more stable WER values ranging between 26% and 28%. The higher fluctuation observed in the classroom lecture recording may indicate that environmental noise, speaker fatigue, inconsistent microphone distance, and classroom interaction influence transcription consistency over time. In contrast, the relatively stable performance observed in the podcast recording suggests that professionally controlled audio conditions contribute to more consistent ASR behavior throughout long recordings.

A similar pattern was also observed at the character level. The full-video transcriptions consistently produced lower CER values compared to individual chunk transcriptions in both lecture and podcast recordings. For instance, the lecture recording achieved a full-video CER of 13%, while the individual chunks produced CER values ranging from 18% to 21%. Similarly, the podcast recording produced a full-video CER of approximately 7%, whereas chunk-level CER values ranged between 8% and 10%. These findings suggest that longer continuous transcription contexts may help Whisper maintain more stable phonetic and lexical predictions, thereby reducing both word-level and character-level transcription errors. The results also indicate that segment-based processing may introduce local transcription inconsistencies that become less visible during full-context transcription. Detailed number of chunk-level evaluation results for long-duration recordings is presented in table 2.

Table 2. Comparison of chunk-level and full-video WER and CER results for long-duration videos

Video ID	Chunk 1 WER	Chunk 2 WER	Chunk 3 WER	Full Video WER	Chunk 1 CER	Chunk 2 CER	Chunk 3 CER	Full Video CER
Lecture_Long	51%	54%	44%	29%	18%	21%	20%	13%
Podcast	28%	26%	28%	21%	10%	8%	9%	7%

3.3. Word-Level Error Analysis

Beyond quantitative WER and CER evaluation, qualitative error analysis revealed several recurring transcription patterns. Most transcription errors consisted of substitution operations, followed by deletions and insertions. Across nearly all recordings, substitution errors contributed the largest proportion of total transcription errors.

As illustrated in Figure 3, substitution errors consistently represented the largest component of total transcription errors across nearly all recordings. This indicates that Whisper generally preserved sentence structure and word positioning but frequently produced acoustically similar lexical substitutions. The interview recording exhibited the highest substitution and deletion rates, further supporting the observation that uncontrolled conversational audio environments introduce substantial transcription difficulty. In contrast, insertion errors remained comparatively lower across all recordings, suggesting that the model was more likely

to misrecognize existing speech segments than to generate completely unrelated additional words. The educational animation recordings produced lower overall error component rates, demonstrating that clearer narration and reduced environmental interference contribute to more stable ASR performance.

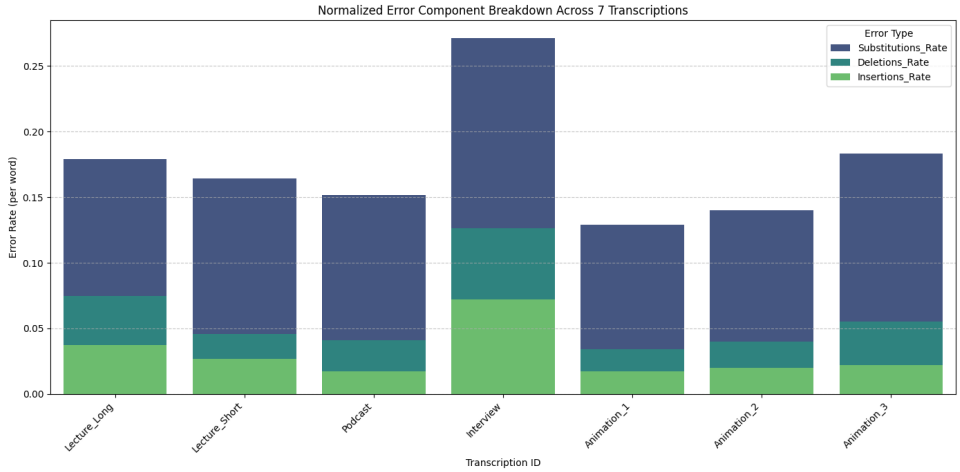

Figure 3. Normalized WER substitution, deletion, and insertion error rates across all evaluated recordings

Several recurring substitution patterns were observed involving acoustically similar Indonesian words or informal conversational expressions. Similar phonetic confusion patterns appeared in several other recordings as well, involving words with comparable pronunciation structures. Whisper also occasionally generated spelling variations that remained phonetically close to the original words, such as “fisika” becoming “visika”, “semesta” becoming “semenesta”, and “psikologi” becoming “pesikologi”. These findings suggest that the model often preserved partial phonetic information despite generating orthographic inaccuracies. This phenomenon may also explain why several recordings produced relatively higher WER values while maintaining comparatively lower CER values, as many transcription errors involved partial word modifications rather than completely unrelated outputs.

Interestingly, some substitutions observed in the low-production interview recording involved culturally or contextually specific expressions. For example, the Islamic greeting “assalamualaikum” was transcribed as “selamat malam” under noisy conversational conditions. This finding suggests that the ASR model may occasionally replace unfamiliar or acoustically ambiguous expressions with more statistically common conversational phrases in Indonesian speech. Similar patterns were also observed in several omitted conversational expressions during spontaneous dialogue segments.

Several deletion patterns also frequently appeared in conversational recordings, particularly involving filler words or informal expressions such as “ya”, “kan”, and “itu”. This may indicate that Whisper tends to omit low-information conversational markers when processing spontaneous Indonesian speech. Insertion errors were generally less frequent than substitutions and deletions. However, several recurring inserted words such as “di”, “yang”, and “ke” appeared repeatedly across multiple recordings. This behavior may indicate that the language model component occasionally predicts common Indonesian functional words during uncertain acoustic conditions. Table 3 summarizes several recurring transcription error patterns identified during word-level analysis.

Table 3. Examples of recurring transcription error patterns identified in Indonesian educational video transcriptions

Ground Truth	Whisper Output	Error Type	Observation
ia	iya	Substitution	Phonetically similar conversational expression
ia	ya	Substitution	Informal Indonesian pronunciation ambiguity
fisika	visika	Substitution	Consonant confusion while preserving phonetic structure
psikologi	pesikologi	Substitution	Partial syllable insertion preserving pronunciation
semesta	semenesta	Substitution	Additional phoneme insertion
Assalamualaikum warahmatullahi wabarakatuh	Selamat malam	Substitution	Contextual substitution toward a more common Indonesian greeting
ya / kan / itu	omitted	Deletion	Conversational filler omission
di / yang / ke	inserted	Insertion	Frequent prediction of common functional words

Overall, the qualitative findings demonstrate that ASR errors in Indonesian educational videos are not entirely random, but are influenced by phonetic similarity, conversational speech structure, recording quality, and environmental noise conditions. These observations provide additional insight beyond aggregate WER and CER values and help explain how different educational audio environments affect transcription behavior. Despite the promising findings, this study focused primarily on Indonesian educational multimedia recordings with limited domain variation and utilized a single ASR model configuration based on Whisper base. Therefore, the obtained results mainly reflect transcription behavior within educational audio environments rather than representing general Indonesian speech recognition performance across all domains and speaking conditions. Future work may involve evaluating additional Whisper model variants, larger educational datasets, and multilingual or regional Indonesian speech variations to further investigate ASR robustness under more diverse linguistic conditions.

4. CONCLUSION

This study evaluated the performance of the Whisper base ASR model on various Indonesian educational video conditions, including lecture recordings, podcasts, interviews, and animated educational content. The results demonstrated that transcription performance varied substantially depending on audio quality, production style, environmental noise, and speech interaction patterns. Recordings with cleaner audio and more controlled narration generally produced lower WER and CER values, while spontaneous dialogue and noisy recording conditions significantly increased transcription errors. The analysis also showed that many transcription errors involved acoustically similar Indonesian expressions, conversational filler words, and contextual substitutions involving culturally specific phrases. Several transcription outputs preserved partial phonetic similarity with the original words, which may explain why some recordings produced relatively higher WER values while maintaining comparatively lower CER values.

The temporal segmentation analysis further demonstrated that transcription consistency may fluctuate across long-duration recordings, particularly in uncontrolled educational environments. Full-video transcription evaluations generally produced lower WER and CER values than chunk-level evaluations, suggesting that longer transcription contexts may help Whisper maintain more stable semantic and phonetic predictions during continuous speech processing.

In general, the study indicates that Whisper base model provides reasonably effective transcription capability for Indonesian educational multimedia while remaining computationally accessible for low-resource research settings. The findings may support future development of automatic subtitle generation, educational accessibility systems, and Indonesian ASR evaluation research under realistic multimedia learning conditions.

This research was conducted using a limited number of educational multimedia recordings and a single Whisper model configuration. Consequently, the findings primarily describe ASR behavior within the evaluated educational scenarios and may not fully represent transcription performance across broader Indonesian speech domains and acoustic environments. Future work may involve evaluating additional Whisper model variants, exploring larger educational datasets, and incorporating more advanced linguistic error analysis techniques to further investigate ASR behavior in Indonesian regional dialect variations, multilingual educational recordings, and spontaneous conversational environments.

REFERENSI

[1] M. Noetel *et al.*, “Video Improves Learning in Higher Education: A Systematic Review,” *Rev. Educ. Res.*, vol. 91, no. 2, pp. 204–236, Apr. 2021, doi: 10.3102/0034654321990713.

[2] E. Navarrete, A. Nehring, S. Schanze, R. Ewerth, and A. Hoppe, “A Closer Look into Recent Video-based Learning Research: A Comprehensive Review of Video Characteristics, Tools, Technologies, and Learning Effectiveness,” *Int. J. Artif. Intell. Educ.*, vol. 35, no. 4, pp. 1631–1694, Dec. 2025, doi: 10.1007/s40593-025-00481-x.

[3] P. Khong, D. Holmes, B. Masoudian, G. C. Lund, and S. Garwood, “Lecture Capture, Transcripts, and Captioning in US Colleges of Osteopathic Medicine: Descriptive Cross-Sectional Survey,” *Med. Sci. Educ.*, vol. 35, no. 2, pp. 625–628, Nov. 2024, doi: 10.1007/s40670-024-02224-4.

[4] S. Malakul and I. Park, “The effects of using an auto-subtitle system in educational videos to facilitate learning for secondary school students: learning comprehension, cognitive load, and satisfaction,” *Smart Learn. Environ.*, vol. 10, no. 1, p. 4, Jan. 2023, doi: 10.1186/s40561-023-00224-2.

[5] H. Gonzalez *et al.*, “Automatically Generated Summaries of Video Lectures May Enhance Students’ Learning Experience,” in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 382–393. doi: 10.18653/v1/2023.bea-1.31.

[6] A. Kotevski, "Using Automatic Speech Recognition To Support Students With Disabilities," *UKLO Proc.*, vol. 1, no. 1, pp. 187–193, May 2025, doi: 10.20544/AISC.1.1.25.P18.
[7] Hasbullah Azis, E. P. Indreswari, and R. Wisudawanto, "Pemanfaatan Teknologi Automatic Speech Recognition Dalam Menciptakan Pembelajaran Inklusif: Implementasi, Efektifitas Dan Tantangan," *GANESHA J. Pengabdi. Masy.*, vol. 5, no. 1, pp. 27–33, Jan. 2025, doi: 10.36728/ganesha.v5i1.4171.
[8] A. Adila, D. Lestari, A. Purwarianti, D. Tanaya, K. Azizah, and S. Sakti, "Enhancing Indonesian Automatic Speech Recognition: Evaluating Multilingual Models with Diverse Speech Variabilities," in *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, Hsinchu City, Taiwan: IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/O-COCOSDA64382.2024.10800336.
[9] E. William and A. Zahra, "Speech Recognition Dengan Whisper Dalam Bahasa Indonesia," *Action Res. Lit.*, vol. 9, no. 2, pp. 386–397, Feb. 2025, doi: 10.46799/arl.v9i2.2573.
[10] K. Kuhn, V. Kersken, B. Reuter, N. Egger, and G. Zimmermann, "Measuring the Accuracy of Automatic Speech Recognition Solutions," *ACM Trans. Access. Comput.*, vol. 16, no. 4, pp. 1–23, Dec. 2023, doi: 10.1145/3636513.
[11] K. Zaman, M. Sah, C. Direkoglu, and M. Unoki, "A Survey of Audio Classification Using Deep Learning," *IEEE Access*, vol. 11, pp. 106620–106649, 2023, doi: 10.1109/ACCESS.2023.3318015.
[12] L. Fredianelli, F. Artuso, G. Pompei, G. Licitra, G. Iannace, and A. Akbaba, "Environmental Noise Dataset for Sound Event Classification and Detection," *Sci. Data*, vol. 12, no. 1, p. 1712, Oct. 2025, doi: 10.1038/s41597-025-05991-w.
[13] O. E. Tumewu, S. Saprudin, U. Sambiri, R. Achmad, M. H. Rahman, and F. Hamid, "The implementation of gamification with video media variations to improve students' concept mastery in science learning," *JPPI J. Penelit. Pendidik. Indones.*, vol. 11, no. 4, pp. 72–81, Dec. 2025, doi: 10.29210/020256620.
[14] S. Alley, "Do Higher Production Value Videos Lead to Improved Engagement and Learning Outcomes? A Field Experiment," *J. Educ. Online*, vol. 22, no. 1, Jan. 2025, doi: 10.9743/JEO.2025.22.1.15.
[15] S. Z. Pranida, R. A. Genadi, M. C. Airlangga, and S. Shehata, "ASR Under Noise: Exploring Robustness for Sundanese and Javanese," in *Proceedings of the 9th Widening NLP Workshop*, Suzhou, China: Association for Computational Linguistics, 2025, pp. 87–99. doi: 10.18653/v1/2025.winlp-main.16.
[16] M. Borsky, P. Mizera, P. Pollak, and J. Nouza, "Dithering techniques in automatic recognition of speech corrupted by MP3 compression: Analysis, solutions and experiments," *Speech Commun.*, vol. 86, pp. 75–84, Feb. 2017, doi: 10.1016/j.specom.2016.11.007.
[17] OpenWhispr, "Whisper Model Sizes Explained," OpenWhispr Blog. Accessed: May 24, 2026. [Online]. Available: https://openwhispr.com/blog/whisper-model-sizes-explained
[18] D. Moser, N. Stanic, and M. Sariyar, "Benchmarking speech-to-text robustness in noisy emergency medical dialogues: an evaluation of models under realistic acoustic conditions," *JAMIA Open*, vol. 8, no. 6, p. ooaf147, Nov. 2025, doi: 10.1093/jamiaopen/ooaf147.
[19] T. D. K, J. James, D. P. Gopinath, and M. A. K, "Advocating Character Error Rate for Multilingual ASR Evaluation," in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico: Association for Computational Linguistics, 2025, pp. 4926–4935. doi: 10.18653/v1/2025.findings-naacl.277.