JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 3, June 2026, hlm. 429 – 436 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i3.4581
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

PERFORMANCE EVALUATION OF AUTOMATED MEETING SUMMARIZATION BASED ON OPENAI WHISPER AND INDOT5 FINE-TUNING

I Gusti Lanang Oka Wiyana[1*], Putu Indah Ciptayani [1], Ida Bagus Adisimakrisna Peling [1]
[1]Teknologi Rekayasa Perangkat Lunak, Politeknik Negeri Bali
email: *gungwiyana@gmail.com

Abstract: Manual meeting documentation risks losing important information due to cognitive fatigue. Although automated summarization models have evolved, integrated end-to-end systems for Indonesian spoken language remain highly limited. This study aims to design and evaluate an end-to-end automated meeting summarization architecture that directly integrates Automatic Speech Recognition (ASR) via OpenAI Whisper for transcription and the IndoT5 language model for abstractive summarization. IndoT5 was fine-tuned using a dataset of 486 Indonesian spoken language transcript pairs. Testing was conducted on a CPU infrastructure using MP4, MP3, and WAV formats. Results show the optimal fine-tuning configuration significantly improved accuracy, achieving ROUGE-1 (0.4167), ROUGE-2 (0.1973), and ROUGE-L (0.2701) scores. Computationally, the system achieved a Real-Time Factor below 1, processing data faster than the actual recording duration. Conclusively, integrating Whisper and IndoT5 shows potential in producing coherent meeting summaries with lightweight computational overhead, making it viable for local infrastructure implementation to ensure data privacy.

Keywords: abstractive summarization; ASR; end-to-end pipeline; IndoT5; real-time factor

Abstrak: Dokumentasi rapat manual rentan menghilangkan informasi penting akibat keterbatasan kognitif. Meskipun model peringkas otomatis telah berkembang, implementasi sistem terintegrasi (*end-to-end*) khusus percakapan lisan berbahasa Indonesia masih sangat terbatas. Penelitian ini bertujuan merancang dan mengevaluasi arsitektur peringkas rapat otomatis *end-to-end* yang mengintegrasikan langsung *Automatic Speech Recognition* (ASR) melalui OpenAI Whisper untuk transkripsi dan model bahasa IndoT5 untuk peringkasan abstraktif. Adaptasi domain dilakukan melalui *fine-tuning* IndoT5 menggunakan 486 pasang dataset transkrip lisan berbahasa Indonesia. Pengujian pada infrastruktur CPU menggunakan format MP4, MP3, dan WAV. Hasil pengujian menunjukkan konfigurasi *fine-tuning* optimal berhasil meningkatkan akurasi, dengan skor ROUGE-1 (0,4167), ROUGE-2 (0,1973), dan ROUGE-L (0,2701). Sistem mendemonstrasikan efisiensi komputasi dengan nilai *Real-Time Factor* di bawah 1, mengindikasikan waktu pemrosesan lebih cepat dari durasi rekaman asli. Kesimpulannya, integrasi Whisper dan IndoT5 menunjukkan potensi dalam menghasilkan ringkasan yang koheren dengan beban komputasi ringan, sehingga layak diimplementasikan pada infrastruktur lokal organisasi untuk menjaga privasi data.

Kata kunci: ASR; end-to-end pipeline; IndoT5; peringkasan abstraktif; real-time factor

INTRODUCTION

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 3, June 2026, hlm. 429 – 436 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i3.4581
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Meeting activities in modern organizations show a 12.9% increase in virtual meeting frequency per person, along with a 13.5% surge in the number of meeting participants, driven by the transition to post-pandemic work styles [1]. This increased utilization of video conferencing demands more efficient management of post-event deliverables. The importance of this documentation is evident from the high organizational need to archive meeting minutes as a reference for strategic decision-making and proof of performance accountability [2]. However, the manual minuting process remains a primary constraint due to human cognitive limitations in capturing important points in real-time amidst speakers' talking speeds and lengthy discussion durations, which ultimately increases the risk of errors and the loss of critical information [3].

As a solution, developing an automated meeting summarization system based on the integration of Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) is essential. In the NLP domain, abstractive approaches, such as those implemented through the IndoT5 deep learning model, are considered superior compared to multilingual summarization models like mBART or conventional extractive approaches [4], [5], as they are capable of constructing new sentences that retain the core meaning with a more natural linguistic structure. A state-of-the-art transformer model that does not rely on sequential hidden states is required to ensure the coherence of the summary [6]. However, abstractive summarization models generally require pure text input, making ASR technology necessary to map sound waves into text sequences [7]. The OpenAI Whisper model has proven to be robust as an ASR because it is trained using massive weakly supervised methods, allowing it to accurately predict text from various speaker accents and background noise conditions without requiring rigid data standardization [8], [9].

Several previous studies have explored the use of pre-trained T5 family models for summarizing news documents; however, their effectiveness on non-standard spoken structures has not been widely tested [10]. Other research has attempted to apply ASR technology for audio transcription but stopped at the text extraction stage without further summarization processes [11]. Furthermore, evaluations of transformer models on dialogue transcript datasets are often implemented specifically in English, which has fundamentally different linguistic characteristics from Indonesian [12], [13]. Based on this literature review, the integration of the OpenAI Whisper ASR model and IndoT5 within an End-to-End pipeline architecture for Indonesian meeting summarization, particularly analyzing computational efficiency across audio formats, remains underexplored and has not been comprehensively evaluated. Therefore, this study aims to design and evaluate the performance of this architecture by measuring computational time (inference time) and validating text feasibility using the ROUGE metric [14]. The final results of this study are expected to serve as an automation solution that supports post-meeting knowledge management in organizations.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 3, June 2026, hlm. 429 – 436 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i3.4581
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

METHOD

This study employs a quantitative experimental approach to develop and evaluate the End-to-End (E2E) pipeline architecture of an automated meeting summarization system. The entire computational process was built using the Python 3.10 programming language and the PyTorch deep learning framework [15], [16]. Generally, the research process was carried out systematically following the stages illustrated in Image 1.

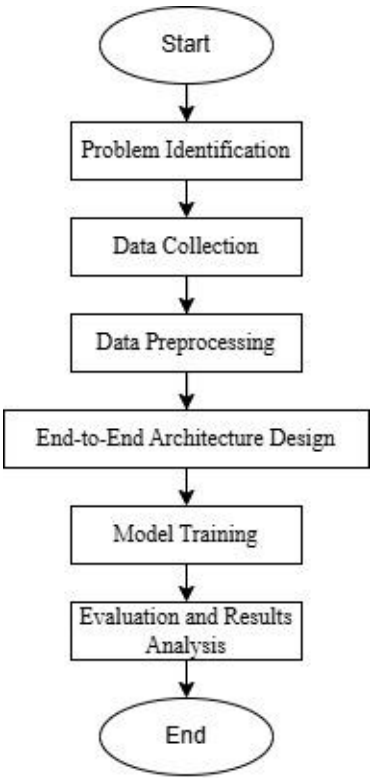

Image 1. Research Flowchart

The workflow began with the collection of 30 video recordings of Plenary Meetings of the House of Representatives of the Republic of Indonesia (DPR RI). The data was processed through chunking to extract 486 text pairs with an 80:20 split ratio to prevent data leakage, and extracted into MP3 and WAV formats using FFmpeg. After the data was prepared, the research proceeded to the system architecture design stage, model training experiments (fine-tuning), and the final stage comprising system performance testing and evaluation analysis before drawing a conclusion.

More specifically, in the system design stage, the constructed architecture integrates two primary artificial intelligence modules executed entirely in a local environment. The first module is speech recognition using the Medium variant of the OpenAI Whisper ASR model to achieve a balance of high transcription accuracy while avoiding network latency issues. The text output from this module is automatically forwarded as input to the second module, an abstractive text summarizer using the pre-trained IndoT5-base model from the Wikidepia repository, which has undergone a domain adaptation process [17]. The technical data flow within this integrated system from the uploading of audio documents by the user to the presentation of the final summary through a Streamlit-based web interface is thoroughly illustrated in Image 2.

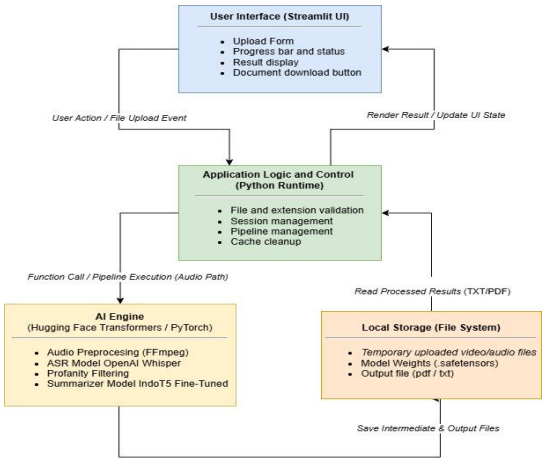

Image 2. End-to-End Pipeline Architecture of the Meeting Summarization System

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 3, June 2026, hlm. 429 – 436 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i3.4581
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

The system testing stage focused on computational efficiency and linguistic quality. Efficiency was measured through inference latency, which is the total execution time from the upload process to the output presentation. Summary quality was evaluated by comparing the system's output text against reference documents (ground truth) using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric [14]. The systematic calculation of the ROUGE metric is formulated in Equation (1).

$$ROUGE - N = \frac{\sum_{gram_n \in Ref} min(Count_{sys}, Count_{ref})}{\sum_{gram_n \in Ref} Count_{ref}} \quad (1)$$

The description of Equation (1) includes N as the length of the n-gram used in the evaluation, $gram_n$ as the n-th n-gram in the summary text, Ref as the human reference summary, $Count_{sys}$ as the number of matching n-grams in the system output, and $Count_{ref}$ as the number of n-grams in the reference. A higher ROUGE percentage value directly represents a higher level of accuracy and a more representative informational similarity between the computational system and the reference [14].

RESULT AND DISCUSSION

The implementation of artificial intelligence in this system began with determining the model configuration through a series of fine-tuning experiments on the IndoT5 model. This stage is essential to ensure that the end-to-end testing utilizes the most optimal parameters. The domain adaptation process utilized a secondary dataset comprising 486 pairs of spoken transcriptions and manual summaries in Indonesian, with a total file size of 1.37 MB. For the training process, the dataset was partitioned using an 80:20 ratio, where 80% (390 data pairs) was allocated as training data and 20% (96 data pairs) as testing data. The data separation was conducted programmatically using the train_test_split module from the scikit-learn library in Python. It was grouped based on the original file identity so that segments from the same recording were not distributed into two different sets, aiming to prevent data leakage. This file identity-based splitting strategy is recommended in natural language processing literature to ensure that model validation is free from repetitive context bias [18]. The comparison results of the hyperparameter combinations are presented in Table 1.

The data in Table 1 shows that the EXP-013 configuration serves as the baseline for the system architecture as it produced the highest linguistic accuracy. This optimized model was consistently used in the system performance testing stages. To validate computational stability and avoid evaluation bias, inference latency was executed on 11 test data samples of varying durations. The computational scenarios were tested on a CPU infrastructure with MP4, MP3, and WAV input formats. The detailed test results are presented in Table 2.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 3, June 2026, hlm. 429 – 436 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i3.4581
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 1. IndoT5 Model Fine-Tuning Hyperparameter Experiment Results

Experiment ID	Epoch	Batch Size	LR	ROUGE		
				1	2	L
EXP-BASE	-	-	-	0.2484	0.0908	0.1637
EXP-001	5	4	1e-4	0.3141	0.1244	0.1864
EXP-002	5	4	3e-5	0.2994	0.112	0.1697
EXP-003	5	4	5e-5	0.3056	0.1132	0.1769
EXP-004	5	8	1e-4	0.3078	0.1195	0.1824
EXP-005	5	8	3e-5	0.3182	0.1282	0.1838
EXP-006	5	8	5e-5	0.3119	0.1193	0.179
EXP-007	10	4	1e-4	0.3847	0.1686	0.2434
EXP-008	10	4	3e-5	0.2999	0.111	0.1748
EXP-009	10	4	5e-5	0.3147	0.1203	0.1815
EXP-010	10	8	1e-4	0.3228	0.1345	0.1945
EXP-011	10	8	3e-5	0.3015	0.1135	0.1735
EXP-012	10	8	5e-5	0.3092	0.1235	0.1816
EXP-013	15	4	1e-4	0.4167	0.1973	0.2701
EXP-014	15	4	3e-5	0.307	0.114	0.1785
EXP-015	15	4	5e-5	0.3584	0.1468	0.222
EXP-016	15	8	1e-4	0.3846	0.1667	0.2389
EXP-017	15	8	3e-5	0.3069	0.1129	0.1758
EXP-018	15	8	5e-5	0.3164	0.1217	0.1838

Table 2. Inference Latency Testing on 11 Test Data Samples

Data Sample	Actual Duration (seconds)	MP4 Time (seconds)	MP3 Time (seconds)	WAV Time (seconds)
Data 1	246.99	204.07	182.13	184.53
Data 2	226.00	172.86	173.45	172.11
Data 3	228.00	173.88	172.37	165.96
Data 4	197.02	164.17	190.32	155.88
Data 5	200.02	163.49	164.21	169.37
Data 6	194.00	152.54	157.17	156.50
Data 7	199.02	163.00	163.29	163.68
Data 8	175.99	154.43	154.43	154.61
Data 9	200.04	164.79	156.70	165.94
Data 10	203.04	174.38	168.71	173.82
Data 11	249.02	198.06	205.42	197.84
Average	210.83	171.42	171.65	169.11

The data in Table 2 demonstrates the system's performance stability across various recording durations. On average, computing the MP4 format required 171.42 seconds, and the MP3 format required 171.65 seconds. Meanwhile, the WAV format recorded the most efficient processing time, averaging 169.11

seconds. All format variations achieved a Real-Time Factor index below 1, indicating that the system is capable of processing the minute transcripts faster than the original audio duration. The achievement of this computational time efficiency is quite significant, considering previous studies emphasized that the primary obstacle in implementing artificial intelligence models in real-world scenarios is often constrained by high computational loads and processing latency [19]. The efficiency in the WAV format occurs because the acoustic data is lossless, allowing the model to extract spectral representations without the bottleneck of heavy audio decompression processes.

In addition to time analysis, the integrity of the machine's text reconstruction quality was measured against 11 reference document samples using the ROUGE metric. The performance comparison of average linguistic accuracy across all test data is presented in Table 3.

Table 3. Average ROUGE Metric Evaluation Across Formats

Format	ROUGE		
	1	2	L
MP4	0.4310	0.2020	0.2776
MP3	0.4102	0.1963	0.2782
WAV	0.4310	0.2020	0.2776

The evaluation results in Table 3 validate that the input representations from the MP4 and WAV formats successfully achieved the highest identical unigram similarity level of 0.4310. On the other hand, the compressed MP3 format experienced a slight decrease in the ROUGE-1 score to 0.4102. This phenomenon is closely related to the lossy compression nature of MP3 files, which reduces certain frequency spectrum ranges to minimize memory size. This condition produces microscopic acoustic distortions that cause the model to predict phonemes with slight deviations. The accuracy degradation due to signal compression aligns with previous research showing that maintaining audio signal integrity (lossless) is crucial for preserving feature representation quality, thereby improving Transformer-based model performance in the information extraction process [20]. The stable average ROUGE-2 acquisition at around 0.20 reflects the characteristics of abstractive summarization, where the IndoT5 model performs high-level semantic reconstruction by paraphrasing the main ideas using new lexical structures rather than merely performing literal sentence extraction. This abstractive capability is highly relevant for reducing narrative redundancy in unstructured spoken conversations into professional minute documents.

CONCLUSION

The integration of the OpenAI Whisper Medium variant model and the IndoT5 language model, which has undergone a domain adaptation process, successfully transformed spoken conversations into coherent text summaries. Based on the evaluation results, the fine-tuning process was proven to significantly improve linguistic quality compared to the baseline model, with an increase in ROUGE-1 from 0.2484 to 0.4167, ROUGE-2 from 0.0908 to 0.1973, and ROUGE-L from 0.1637 to 0.2701. The system's efficiency, demonstrated by an average Real-Time Factor value below one, indicates that the system has the potential to be

implemented for professional documentation needs that demand processing speed and data privacy. As a future development effort, the integration of a Speaker Diarization feature can be applied to enhance the system's functionality in automatically identifying speaker transitions in multi-participant meeting scenarios to produce more informative meeting minutes.

BIBLIOGRAPHY

[1] E. DeFilippis, S. M. Impink, M. Singell, J. T. Polzer, and R. Sadun, "The impact of COVID-19 on digital communication patterns," *Humanit. Soc. Sci. Commun.*, vol. 9, no. 1, Dec. 2022, doi: 10.1057/s41599-022-01190-9.

[2] M. Setyorini, Y. Kartika Sari, and M. K. Ansor, "Pengembangan Sistem Manajemen Rapat Dengan Notifikasi Whatsapp Di Polres Tulungagung Menggunakan Kerangka Kerja Scrum," 2025. [Online]. Available: https://www.jurnal.stkippgritulungagung.ac.id/index.php/joincos

[3] D. M. Hilty *et al.*, "Findings and Guidelines on Provider Technology, Fatigue, and Well-being: Scoping Review," May 01, 2022, *JMIR Publications Inc.* doi: 10.2196/34451.

[4] A. P. Widyassari *et al.*, "Review of automatic text summarization techniques & methods," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1029–1046, Apr. 2022, doi: 10.1016/J.JKSUCI.2020.05.006.

[5] A. Sakti Wiradinata and C. M. Viny, "Abstractive Text Summarization Berita Bahasa Indonesia Menggunakan Retrieval-Augmented Generation," *Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 13, no. 1, 2025, doi: https://doi.org/10.24912/jiksi.v13i1.32861.

[6] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," Jul. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app13137620.

[7] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd (Draft). Stanford University, 2025. Accessed: Jan. 06, 2026. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *International Conference on Machine Learning*, Dec. 2022. doi: https://doi.org/10.48550/arXiv.2212.04356.

[9] M. Labied, A. Belangour, M. Banane, and A. Erraissi, "An overview of Automatic Speech Recognition Preprocessing Techniques," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, IEEE, Mar. 2022, pp. 804–809. doi: 10.1109/DASA54658.2022.9765043.

[10] A. Bahari and K. E. Dewi, "PERINGKASAN TEKS OTOMATIS

ABSTRAKTIF MENGGUNAKAN TRANSFORMER PADA TEKS BAHASA INDONESIA," *Jurnal Ilmiah Komputer dan Informatika*, vol. 13, no. 1, 2024.

[11] M. Maurya, M. Zaheer, N. Mohammad, S. siddiqui, M. Z. Khan, and M. A. Akram, "Speech Recognition Technologies: Design, Challenges, and Real-World Applications," *International Journal of Innovative Research in Computer Science and Technology*, vol. 13, no. 3, pp. 55–61, May 2025, doi: 10.55524/ijircst.2025.13.3.9.

[12] E. Daraghmi, L. Atwe, and A. Jaber, "A Comparative Study of PEGASUS, BART, and T5 for Text Summarization Across Diverse Datasets," *Future Internet*, vol. 17, no. 9, Sep. 2025, doi: 10.3390/fi17090389.

[13] I. G. A. I. U. Putri, I. N. P. Trisna, and N. K. D. Rusjayanthi, "Abstractive Text Summarization to Generate Indonesian News Highlight Using Transformers Model," *Journal of Information Systems and Informatics*, vol. 7, no. 2, pp. 1248–1263, Jun. 2025, doi: 10.51519/journalisi.v7i2.1082.

[14] M. Wahyu Bagus Dwi Satya *et al.*, "Comparative Analysis of T5 Model Performance for Indonesian Abstractive Text Summarization," *Sistemasi: Jurnal Sistem Informasi*, vol. 14, no. 3, pp. 2540–9719, 2025, doi: https://doi.org/10.32520/stmsi.v14i3.4884.

[15] S. Lynch, *Python for Scientific Computing and Artificial Intelligence*. Boca Raton: Chapman and Hall/CRC, 2023. doi: 10.1201/9781003285816.

[16] Iswahyudi, D. Hindarto, and H. Santoso, "PyTorch Deep Learning for Food Image Classification with Food Dataset," *sinkron*, vol. 7, no. 4, pp. 2651–2661, Oct. 2023, doi: 10.33395/sinkron.v8i4.12987.

[17] M. Fuadi, A. Dharma Wibawa, and S. Sumpeno, "idT5: Indonesian Version of Multilingual T5 Transformer," 2023. doi: https://doi.org/10.48550/arXiv.2302.00856.

[18] K. John, D. D. Saurette, and B. Heung, "The problematic case of data leakage: A case for leave-profile-out cross-validation in 3-dimensional digital soil mapping," *Geoderma*, vol. 455, p. 117223, Mar. 2025, doi: 10.1016/J.GEODERMA.2025.117223.

[19] M. Ali *et al.*, "A Machine Learning Approach to Reduce Latency in Edge Computing for IoT Devices," *Engineering, Technology and Applied Science Research*, vol. 14, no. 5, pp. 16751–16756, Oct. 2024, doi: 10.48084/etasr.8365.

[20] H. K. Hameed, "AI-Driven Near-Lossless Audio Compression Modeling via Autoencoders," *Al Rafidain Journal of Engineering Sciences*, vol. 3, no. 2, Sep. 2025, doi: 10.61268/c23c6z11.