JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 11 Number 1 | June 2026: 27-37
DOI: 10.15575/join.v11i1.1781

Maleo-Short: An "In-the-Wild" Indonesian Dataset for Speaker Diarization

Ardi Mardiana[1], Dinda Desmonda Muslimah[2], Ade Bastian[3], Eka Tresna Irawan[4]
[1,2,3]Informatika, Universitas Majalengka, Majalengka, Indonesia
[4]Mayar International Pte.Ltd, Singapore

Article Info

Article history:

Received July 29, 2025
Revised October 19, 2025
Accepted October 31, 2025
Published April 25, 2026

Keywords:

Indonesian
In-the-wild Speech
Speech Dataset
Speaker Diarization

ABSTRACT

Speaker diarization (SD), the task of partitioning an audio stream into speaker-homogenous segments, is fundamental for analyzing multi-speaker recordings. Its application to "in-the-wild" data, such as content from the YouTube platform, poses significant challenges, including overlapped speech, ambient noise, and rapid speaker turns, thereby constituting an active research area. While numerous SD datasets are available, they predominantly focus on English and other high-resource languages. A notable scarcity of publicly accessible datasets exists for the Indonesian language, as extant corpora are primarily engineered for Automatic Speech Recognition (ASR). To address this resource deficit, this research introduces Maleo-Short, a new Indonesian multi-speaker dataset derived from YouTube. The dataset comprises 110 short conversational clips, with a total duration of 1 hours 32 minutes. A reliable ground truth was established through a meticulous manual annotation process using ELAN to generate precise speaker segmentation and transcription files. To validate its utility and assess its complexity, the dataset was evaluated using pre-trained baseline models. The empirical results confirm its status as a challenging benchmark, with the most effective models achieving a Diarization Error Rate (DER) of 32.64% and a Word Error Rate (WER) of 33.78%. Maleo-Short is presented as a valuable, publicly accessible resource intended to catalyze advancements in Indonesian speaker diarization research by facilitating the development and rigorous evaluation of SD systems on acoustically complex and realistic conversational data. Maleo-Short is available at https://doi.org/10.57967/hf/7944.

Corresponding Author:

Andi Mardiana
Informatika, Universitas Majalengka, Majalengka, Indonesia
Email: aim@unma.ac.id

1. INTRODUCTION

Speaker diarization (SD) is the task of determining "who spoke when" in multi-speaker recordings [1]. This technology is crucial for analyzing conversations in various applications, such as meetings and interviews, as it enables speech segmentation based on speaker identity [2]. SD has applications in speech recognition, audio retrieval, and meeting assistance systems [1], [3]. These systems can help overcome challenges in manual note-taking during meetings, such as missed information due to typing speed limitations [3]. However, applying SD to real-world scenarios, particularly "in-the-wild" sources like YouTube, presents numerous challenges [4].

Recent research highlights the challenges in speech recognition and speaker diarization, particularly in complex acoustic environments. These challenges include overlapped speech, which can

p-ISSN: 2528-1682
e-ISSN: 2527-9165

create ambiguity in the audio signal, diverse background noise, acoustic variations, and rapid speaker turns, which can significantly reduce diarization accuracy [2], [5], [6], [7], [8], [9]. Studies have shown that speaker overlap and background noise are major causes of diarization errors. Addressing multi-speaker recordings in realistic conditions remains an active research challenge in the speech processing community [5].

The availability of datasets is a critical prerequisite for advancing SD research. A number of benchmark and diarization datasets have been referenced in speaker diarization research, such as ICSI [10], CHIL [11] and AMI [12], such as meeting speech datasets. The indoor conversation corpus CHiME-6 [13]. Audio-visual datasets, such as VoxConverse [4], AVA-AVD [8] and MSDWild [14]. Chinese datasets, with conversation and conference scenarios such as MagicData-RAMC [15], AliMeeting [16] and AISHELL-4 [17]. However, the majority of available resources are concentrated on the English domain and a few other high resource languages, there is a conspicuous lack of Indonesian SD datasets, specifically designed for speaker diarization research.

Most of the existing Indonesian datasets such as TITML-IDN [18], Indonesian Speech Recognition Corpus (Incar), ASR-IndoCSC, Common Voice Indonesian, LibriVox Indonesia, Nexdata 359-Hours Indonesian Speech, do contain voice recordings of speakers in Indonesian, but their main focus is for Automatic Speech Recognition (ASR) model development. The absence of a representative SD dataset for Indonesian, especially one that captures "in-the-wild" speech variations, is an important gap for this research. Recent research efforts are focusing on creating new datasets, particularly for low-resource languages. The SiTa dataset has been created for Sinhala and Tamil [19]. The recent research on dataset creation for low-resource languages shows the urgency and relevance of developing similar resources for Indonesian.

This paper seeks to address this critical deficit by introducing Maleo-Short, a new Indonesian multi-speaker dataset designed explicitly for the task of speaker diarization. YouTube was selected as the data source for its extensive collection of spontaneous, naturalistic conversational content, which encapsulates a wide range of acoustic variability. This paper's main contribution is the dataset itself, which was meticulously constructed, manually annotated, and evaluated with baseline models. Maleo-Short is intended to function as a public benchmark to stimulate further research and foster the creation of more robust and accurate SD technologies for the Indonesian language. The dataset is available at https://huggingface.co/datasets/maleo-ai/maleo-short.

2. METHOD

The process that occurs in this research consists of several stages, including Data Collection, Data Preparation, Annotation, Baseline Implementation and Evaluation. The flowchart of the stages in this research can be seen in Figure 1.

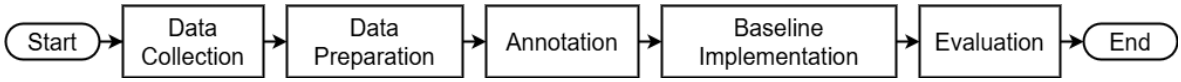

Figure 1. Methodology for creating the Maleo-Short dataset, outlining the key stages from Data Collection to Evaluation

2.1. Data Collection

Data was collected from the YouTube platform, with selection criteria to ensure quality and relevance. The selected videos should feature multi-speaker conversations in the context of predefined topic categories, the topic categories in this study being podcast, talk show, entertainment, movie & drama. In addition, the videos should have at least two speakers involved with a minimum duration of 20 seconds, to provide sufficient conversational context. Audio quality was an important consideration; this research selected videos with acceptable background noise levels, subjectively determined by human annotators to ensure the primary speakers' speech was clearly intelligible.

To provide a qualitative illustration of the acoustic challenges inherent in the dataset, Figure 2 presents a multi-panel visualization of a representative 19-second segment from the Entertainment (ENT) category. This visualization serves to ground the subsequent quantitative analysis by showcasing the "in-the-wild" characteristics that define Maleo-Short. The top panel displays the raw audio waveform, while the middle panel shows the mel spectrogram, which reveals persistent, low-frequency

ambient noise across the entire segment, even during non-speech periods. The bottom panel provides the ground truth annotation, illustrating the complex conversational dynamics present. This timeline highlights several phenomena known to challenge diarization systems. For instance, a clear speech overlap occurs between SPEAKER_1 and SPEAKER_2 (approx. 1.9s to 2.5s), and multiple instances of rapid speaker turns are visible, such as the immediate succession from SPEAKER_3 to SPEAKER_2 at the 11.177s mark. The presence of ambient noise, speaker overlap, and rapid turn-taking exemplifies the core difficulties that state-of-the-art models face when processing unconstrained, spontaneous speech.

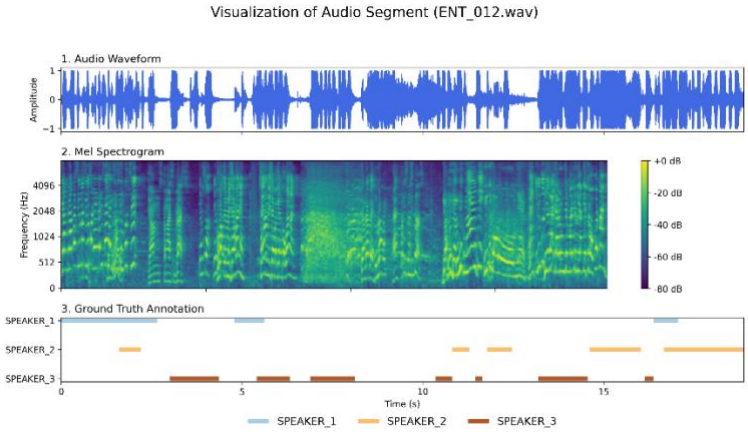

Figure 2. Visualization of a representative audio segment from the Maleo-Short dataset. The panels display (1) waveform, (2) mel spectrogram, and (3) ground truth annotation

In spite of these findings, this study has several limitations that provide a foundation for future research. The current dataset size of 1 hour and 32 minutes limits its suitability for training complex deep learning models from scratch. Acknowledging this, our primary direction for future work is a significant expansion of the corpus. This research is actively working to increase the dataset to a target of 10 hours, focusing on including more diverse genres and challenging scenarios with a higher degree of speech overlap. This expansion will enhance its utility as a more comprehensive resource for both training and robust evaluation.

The creation of the Maleo-Short dataset was guided by established ethical principles for using publicly available data, aligning with guidelines from research bodies such as the ACM and IEEE, which emphasize transparency and the mitigation of potential harm. All data comprises videos that were publicly available on the YouTube platform at the time of collection, and collection methods aimed to comply with YouTube's Terms of Service.

In line with data privacy principles, such as those found in Indonesia's Personal Data Protection (PDP) Law, we implemented measures to protect personal data. While voice is considered biometric data, all speaker labels in the dataset are fully anonymized (e.g., 'SPEAKER_01') and are not linked to any personally identifiable information. The dataset is intended solely for non-commercial research purposes to advance speaker diarization technology for the Indonesian language. Across the entire dataset of 110 clips, there are a total of 351 unique speakers. A detailed recap of the collected datasets is presented in Table 1.

Table 1. Statistics Of The Maleo-Short Dataset. Entries #Sample Duration And #Speaker Are Presented With Three Values Representing The Minimum, Average, And Maximum Values

Topic	#Sample	Total Duration (min)	#Sample Duration (s)	#Speakers	#Total Speaker
Podcast (POD)	45	41	21/55/73	2/4/5	159
Talkshow (TSH)	20	17	34/52/60	2/3/5	50
Entertainment (ENT)	15	11	34/48/60	2/4/6	53
Movie & Drama (MVD)	30	23	23/46/59	2/3/7	89
Total	**110**	**92**	**21/50/88**	**3/2/8**	**351**

p-ISSN: 2528-1682
e-ISSN: 2527-9165

2.2. Data Preparation

Following the collection, each MP4 video file was processed using FFmpeg to extract the audio stream. The audio was then converted to a standardized format to ensure consistency across the dataset, a 16 kHz sampling rate, 16-bit depth, single-channel mono WAV file.

2.3. Annotation

Ground truth was generated through a meticulous manual annotation process using ELAN (EUDICO Linguistic Annotator) [20]. ELAN was chosen for its precision in time-aligned annotations and its widespread use in linguistic research. This process was fundamental to establishing the dataset's reliability. For each audio file, two distinct layers of annotation were created:

1. Speaker Segmentation: The precise start and end timestamps of each speaker's utterance were marked on a dedicated tier. Each segment was assigned an anonymized label (e.g., SPEAKER_1, SPEAKER_2). To maintain consistency, specific guidelines were followed: non-speech sounds like laughter and music were not labeled, and short, ambiguous speaker overlaps were deliberately excluded to ensure the reliability of the ground truth.
2. Transcription: The corresponding orthographic transcription was produced for each annotated speech segment.

The manual annotation workflow in ELAN is depicted in Figure 3. The final annotations were exported into two standard formats, Rich Transcription Time Marked (RTTM) for diarization evaluation, adhering to conventions established in NIST evaluations, and plain text files for ASR evaluation.

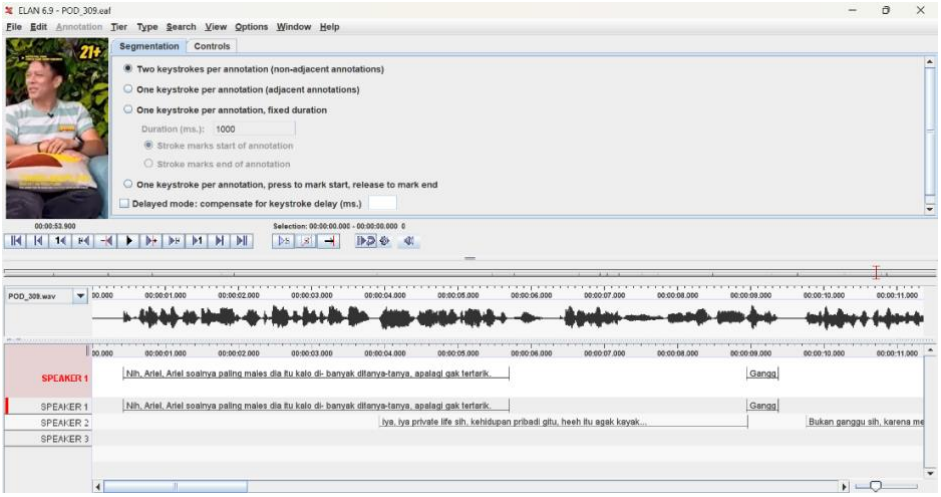

Figure 3. The user interface of the ELAN

2.4. Baseline Implementation

To validate the dataset's utility and establish an initial performance benchmark, several pre-trained models were evaluated off-the-shelf. For Speaker Diarization, the models used were pyannote.audio-3.1, a toolkit that provides neural building blocks for diarization [21], and DiaPer, an end-to-end model with a Perceiver-based architecture [22]. For ASR, the models used were Whisper from OpenAI (medium model), trained via large-scale weak supervision [23], and Wav2Vec, a framework for self-supervised learning [24].

2.5. Evaluation

The evaluation was conducted by calculating standard metrics for diarization and ASR. For diarization, the Diarization Error Rate (DER) is used. As shown in equation (1), DER is calculated as the sum of false alarm (FA), missed speech (MS), and speaker confusion (SC) errors, divided by the total duration of the reference speech.

$$DER = \frac{FA+MS+SC}{Total\ Speech} \quad (1)$$

DER evaluation is performed with a collar parameter of 0.25 seconds.This collar is a period of time around the speaker change boundary that is ignored during the evaluation to provide tolerance for slight timing inaccuracies. Studies on "in-the-wild" datasets show varied DERs. Chung et al. [4] reported a DER of 7.7% on the VoxConverse development set using their audio-visual method, significantly outperforming audio-only baselines (20.2-23.8% DER).

The complexity of "in-the-wild" data is further illustrated by the MSDWild dataset [14], where an audio-only baseline yielded a DER of 21.96% on its few-talker validation set, which was reduced to 12.2% with a fused audio-visual approach; these figures rose to 43.15% and 25.86% respectively on the more challenging many-talker set. The AVA-AVD dataset [8], featuring diverse daily activities and off-screen speakers, saw an audio-only VBx system achieve 21.37% DER (oracle VAD), while the proposed AVR-Net reached 20.57% DER under similar ideal conditions, though performance degraded with predicted upstream modules.

For low-resource languages, the SiTa dataset [19] reported DERs for Sinhala in the range of 6.4-14.3% for modular systems like End-to-End Segmentation, Powerset Cross-Entropy Diarization, and DiaPer, with similar performance on the Tamil subset (6.0-11.5% DER). These benchmarks highlight the progress made but also the remaining challenges in accurately diarizing complex, real-world audio across different conditions and languages, motivating the need for diverse datasets like the Maleo-Short corpus presented here.

In addition to DER, transcription quality is evaluated with *Word Error Rate* (WER). As formulated in equation (2), WER calculates transcription errors by considering substitutions, deletions, and word insertions [25].

$$WER = \frac{Substitutions+Deletions+Insertions}{Number\ of\ Words\ Spoken} \quad (2)$$

With this metric, we measure the diarization performance and transcript accuracy of the Maleo pipeline on the Maleo-Short dataset.

3. RESULT AND DISCUSSION

The empirical evaluation of baseline models on Maleo-Short highlights the inherent complexity of the dataset.

3.1. Diarization Error Rate (DER)

The evaluation of the diarization models performance, as measured by DER, quantifies the inherent difficulty of the Maleo-Short dataset. Table 2 summarizes the average performance of the two models across all samples. The Pyannote model demonstrated superior performance with a DER of 32.64%, compared to the DiaPer model, which recorded a DER of 47.73%. Further analysis of the error components indicates distinct failure modes for each architecture. The primary source of error for Pyannote was Speaker Confusion (SC), accounting for 18.18% of the total DER. This suggests that while its modular approach, particularly its Voice Activity Detection (VAD) component, is effective at identifying speech segments, the model struggles to generate sufficiently discriminative speaker embeddings from the limited acoustic evidence present in short clips. This difficulty in creating robust speaker profiles from brief utterances leads to a higher rate of misattributing speech segments to the wrong speaker.

Conversely, the predominant failure mode for the DiaPer model was Missed Detection (MS), which contributed 25.19% to its DER. This indicates that the end-to-end architecture, which jointly optimizes segmentation and clustering, tends to be more conservative. It often fails to detect speech segments altogether, particularly those that might be of short duration, have lower energy, or occur in noisy backgrounds. This suggests a potential trade-off in end-to-end systems, where the integrated optimization might lead to lower sensitivity in speech detection compared to a specialized VAD module.

Table 2 and Table 3 provide a detailed breakdown of each model's performance across the four content categories. Table 2 details the Pyannote model's results, showing that while it performs best on

p-ISSN: 2528-1682
e-ISSN: 2527-9165

the more structured Talkshow (TSH) category (17.10% DER), its performance degrades significantly on the highly dynamic Entertainment (ENT) category (47.69% DER), where Speaker Confusion is the dominant error. Table 3 presents the results for the DiaPer model, which similarly struggles most with the ENT category (65.71% DER). However, unlike Pyannote, DiaPer's primary failure mode is Missed Detection, which is particularly high in the Podcast (POD) category (31.63%). This categorical analysis highlights that different acoustic environments exacerbate the specific architectural weaknesses of each model.

Table 2. Performance Metrics Of Speaker Diarization Pyannote Model On The Maleo-Short Dataset. FA: False Alarms; MS: Missed Detection; SC: Speaker Confusion; DER: Diarization Error Rate

Category	MS(%)	FA(%)	SC(%)	DER(%)
Podcast (POD)	7.67	5.83	15.33	28.83
Talkshow (TSH)	3.81	2.85	10.45	17.10
Entertainment (ENT)	7.13	7.93	32.64	47.69
Movie & Drama (MVD)	17.35	3.45	20.38	41.19
Average (%)	9.53	4.93	18.18	32.64

Table 3. Performance Metrics Of Speaker Diarization DiaPer Model On The Maleo-Short Dataset. FA: False Alarms; MS: Missed Detection; SC: Speaker Confusion; DER: Diarization Error Rate

Category	MS(%)	FA(%)	SC(%)	DER(%)
Podcast (POD)	31.63	5.86	11.33	48.83
Talkshow (TSH)	21.24	6.23	7.60	35.06
Entertainment (ENT)	27.37	12.63	25.71	65.71
Movie & Drama (MVD)	17.05	14.62	13.87	45.53
Average (%)	25.19	9.24	13.31	47.73

To further dissect these results, Figure 4 and Figure 5 disaggregate the average error components for the Pyannote and DiaPer models, respectively. Figure 4 clearly illustrates that Speaker Confusion (SC) is the principal error source for the Pyannote model in three of the four categories, peaking dramatically in the Entertainment (ENT) domain. This visual evidence reinforces the hypothesis that the model's speaker embedding module is not robust enough for the rapid, overlapping dialogue common in such content.

In contrast, Figure 5 shows a different error profile for the DiaPer model. Missed Detection (MS) is the most significant or second-most significant error component across all categories. The high MS rates in the Podcast (POD) and Entertainment (ENT) categories suggest that the model's end-to-end architecture may be less sensitive to speech segments with lower signal-to-noise ratios or unconventional turn-taking patterns, which are prevalent in these less-structured conversational settings.

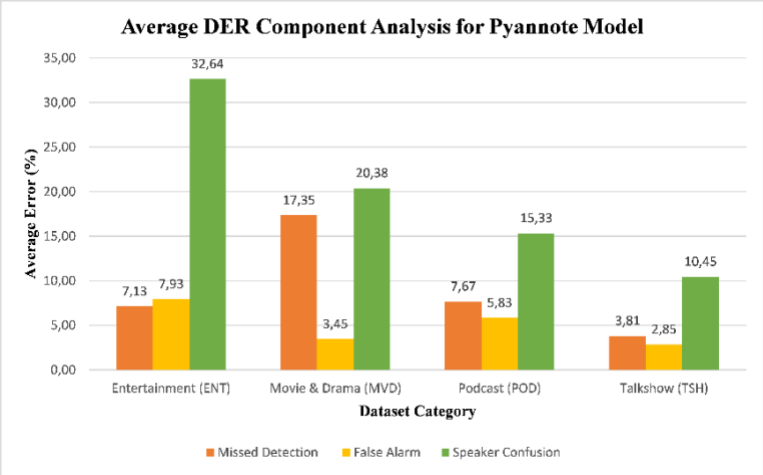

Figure 4. Analysis of average DER error components for Pyannote model by category

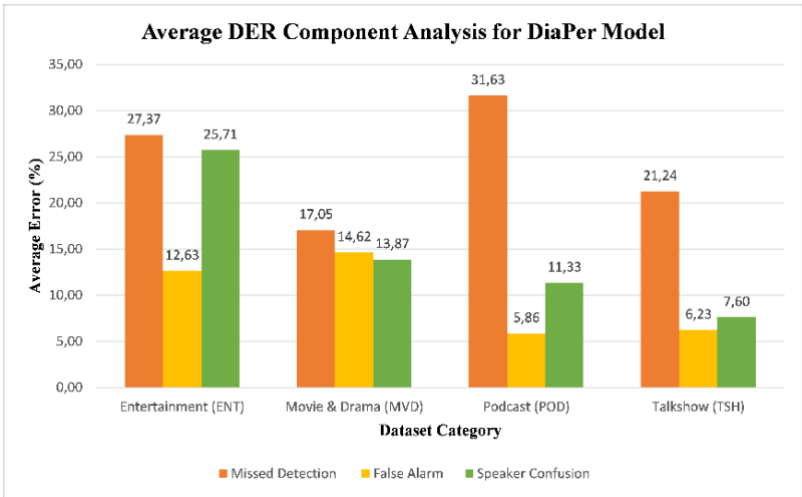

Figure 5. Analysis of average DER error components for DiaPer model by category

The high DER scores on Maleo-Short, when benchmarked against other established datasets as shown in Table 4, validate its status as a challenging testbed. The performance on Maleo-Short shows a significantly higher error rate compared to more controlled datasets like AMI (headset) or even other in-the-wild datasets such as MSDWild. This difficulty stems from two core characteristics of the dataset, its "in-the-wild" nature, which includes background noise and acoustic variability, and the short duration of the clips. Short segments provide very limited data per speaker, making it difficult for models to build robust and discriminative speaker embeddings, which in turn leads to higher confusion rates.

Table 4. DER Benchmark Comparison

Dataset	DER (%)	
	Pyannote	DiaPer
Maleo-Short	32.64	47.73
AMI (*headset*)	18.8	32.94
DIHARD III (*full*)	21.7	24.12
MSDWild	25.3	15.46

3.2. Word Error Rate (WER)

The ASR evaluation further underscores the complexity of the Maleo-Short dataset. As shown in Table 4, the Whisper model, with a WER of 33.78%, significantly outperformed the Wav2Vec model (76.79%). This performance gap can be attributed to several key factors.

Whisper's robustness stems from its training on a massive and diverse 680,000-hour dataset via large-scale weak supervision. This varied, multi-domain training makes it inherently resilient to the acoustic and linguistic variability found in "in-the-wild" data like Maleo-Short.

In contrast, the Wav2Vec model suffers from a significant domain mismatch. While fine-tuned for Indonesian, it was likely trained on cleaner, more formal corpora (such as read-speech from Common Voice) which do not reflect the spontaneous, colloquial nature of the speech in our dataset. This highlights the limitations of transfer learning in this context; fine-tuning on a clean domain is not sufficient to overcome the challenges—such as rapid speaker turns, background noise, and informal language—presented by the target domain. Whisper's training paradigm, which learns from vast amounts of noisy, unlabeled data, prepares it for this domain shift from the outset.

Table 5 and Table 6 present the detailed Word Error Rate (WER) metrics for the Whisper and Wav2Vec models, respectively, broken down by category. Table 5 shows that the Whisper model achieves its best performance on the Talkshow (TSH) category (21.48% WER), likely due to the clearer and more structured speech. Its highest error rate occurs in the Podcast (POD) category (42.37% WER), where a higher number of deletions and substitutions reflects the challenges of transcribing more natural, unscripted conversation.

Table 6 highlights the severe performance degradation of the Wav2Vec model across all categories, confirming a significant domain mismatch. The WER is exceptionally high in the Entertainment (ENT) (87.63%) and Podcast (POD) (79.33%) categories. The error profile is dominated

p-ISSN: 2528-1682
e-ISSN: 2527-9165

by substitution errors, indicating that while the model often detects words, it consistently misidentifies them due to the acoustic and linguistic disparity between its training data and the "in-the-wild" nature of Maleo-Short.

Table 5. Performance Metrics Of ASR Whisper Model On The Maleo-Short Dataset. S: Substitutions; D: Deletions; I: Insertions; WER: Word Error Rate

Category	S	D	I	WER(%)
Podcast (POD)	32	40	3	42.37
Talkshow (TSH)	13	16	2	21.48
Entertainment (ENT)	28	20	4	41.06
Movie & Drama (MVD)	14	4	3	25.44
Average	23	22	2	33.78

Table 6. Performance Metrics Of ASR Wav2Vec Model On The Maleo-Short Dataset. S: Substitutions; D: Deletions; I: Insertions; WER: Word Error Rate

Category	S	D	I	WER(%)
Podcast (POD)	71	68	1	79.33
Talkshow (TSH)	55	32	2	66.02
Entertainment (ENT)	65	46	0	87.63
Movie & Drama (MVD)	43	21	1	74.73
Average	59	45	0	76.79

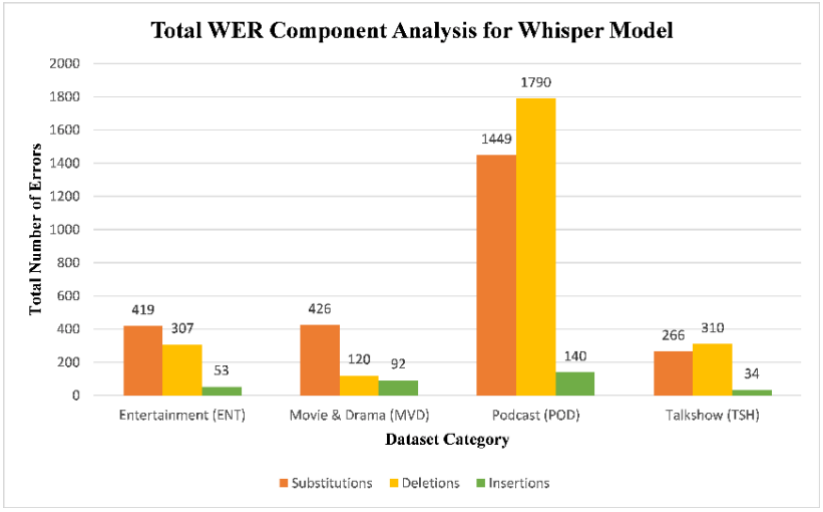

Figure 6. Analysis of total WER error components for Whisper model by category

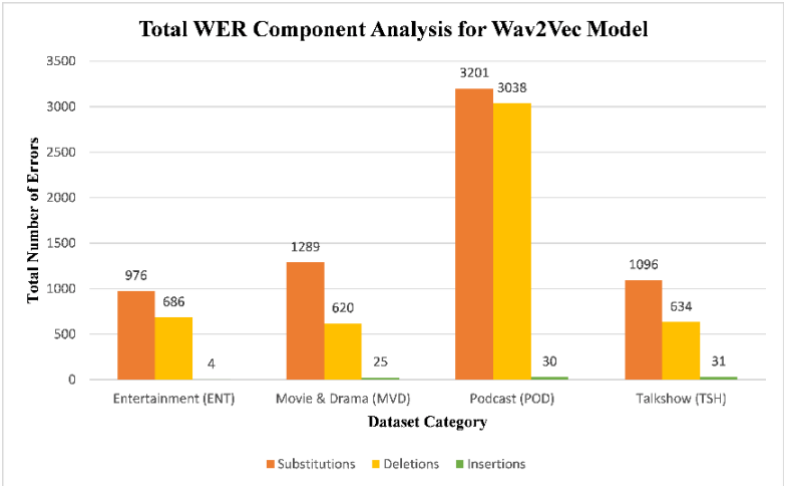

Figure 7. Analysis of total WER error components for Wav2Vec model by categor

The decomposition of total errors by category for each model, as visualized in Figure 6 and Figure 7, providing a clearer picture of the error distribution. Figure 6 demonstrates that for the superior Whisper model, the Podcast (POD) category is the primary source of errors, contributing a total of 1,790 deletions and 1,449 substitutions. This large volume of errors underscores the difficulty of transcribing the informal language, varied speaking rates, and non-standard syntax common in podcasts.

Figure 7 reinforces this observation for the Wav2Vec model but at a much larger scale. The Podcast (POD) category again dominates the error count, with substitution errors (3,201) being the most frequent issue. This visual comparison makes it evident that while both models are most challenged by the podcast domain, Wav2Vec's failure is more pronounced, particularly in its inability to correctly map acoustic features to the correct lexical units, resulting in a cascade of substitution errors.

The benchmark comparison in Table 5 further contextualizes these results. The substantial performance gap between Maleo-Short and cleaner, read-speech datasets like Common Voice or Fleurs is evident. The high WER on Maleo-Short is a direct consequence not only of spontaneous speech filled with colloquialisms and disfluencies but also of the short clip duration. For an ASR model, short segments severely limit the available linguistic context needed to accurately predict words and resolve ambiguity. Each clip effectively forces the model into a "cold start" without the opportunity to adapt to a speaker's specific accent or style, thus rigorously testing its zero-shot capabilities.

Table 7. WER Benchmark Comparison

Dataset	WER (%)	
	Whisper	Wav2Vec
Maleo-Short	33.78	76.79
Common Voice 11	12.62	-
Fleurs	10.2	-
Indonesian Common Voice dataset	-	14.29

Beyond the acoustic challenges common to "in-the-wild" corpora, such as background noise and variable recording quality, Maleo-Short presents unique linguistic complexities specific to spontaneous Indonesian conversation. The most prominent of these is the prevalent phenomenon of code-switching. The data frequently features speakers switching between Indonesian, English, and sometimes regional dialects within the same utterance. This behavior, common in Indonesian media like podcasts and talk shows, poses a severe challenge for ASR systems not explicitly trained to handle such fluid language mixing.

Furthermore, the dataset is rich in colloquialisms, slang, and discourse particles (e.g., sih, dong, kok) that are underrepresented in formal text and read-speech corpora used to train most ASR language models. The high WER scores observed are therefore not only a product of acoustic difficulty but also a direct consequence of these linguistic phenomena, which test the zero-shot and adaptation capabilities of state-of-the-art models in a realistic, low-resource language context.

3.3. Qualitative Error Analysis

To move from quantitative metrics to a concrete illustration of the dataset's challenges, we present a qualitative analysis of a characteristic error case from file ENT_012. Figure 8 displays the full 57-second segment from the Entertainment (ENT) category, which is characterized by rapid speaker exchanges and significant speech overlap. The figure provides a direct comparison between the manually annotated Ground Truth and the Model Hypothesis generated by the Pyannote model.

The analysis reveals a classic and severe Speaker Confusion (SC) error. The Ground Truth (Panel 2) clearly shows a complex, overlapping conversation between four distinct speakers (SPEAKER_1, SPEAKER_2, SPEAKER_3, and SPEAKER_4). In contrast, the Model Hypothesis (Panel 3) fails to differentiate this complexity. The model only identifies two speakers and incorrectly 'collapses' the speech segments from all four speakers into these two erroneous labels.

For example, the utterances from SPEAKER_3 (e.g., at 3.3s, 5.7s) and SPEAKER_4 (e.g., at 29.4s, 37.2s) are consistently mislabeled as SPEAKER_1 or SPEAKER_2. This failure is emblematic of the primary failure mode for the Pyannote model on our dataset. As established in the quantitative analysis (Table 2 and Figure 4), Speaker Confusion is the model's largest error component, peaking dramatically in this specific (ENT) category. This case study visually confirms how the complex, multi-speaker dynamics in Maleo-Short create a difficult testbed, validating its utility as a challenging benchmark.

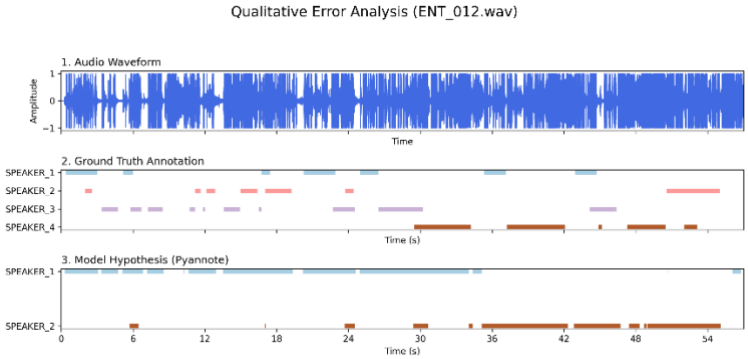

Figure 8. Example of a severe Speaker Confusion (SC) error from the Pyannote model on clip ENT_012. The panels compare the Ground Truth, against the Model Hypothesis (bottom, 2 speakers), highlighting a characteristic failure mode

4. CONCLUSION

This research has successfully constructed, annotated, and validated Maleo-Short, a new 1 hours 32 minutes dataset for Indonesian speaker diarization derived from “in-the-wild” conversational data. Through a meticulous manual annotation protocol, a reliable ground truth for both speaker segmentation and transcription has been established. Empirical evaluation using state-of-the-art baseline models confirms that Maleo-Short constitutes a challenging new benchmark, the high DER and WER scores reflect the inherent complexities of processing spontaneous, unconstrained Indonesian speech. The dataset is released publicly to address a critical resource deficit and is intended to foster the development of more robust speech processing systems for the Indonesian language, contributing to the broader effort of advancing technology for low-resource languages.

In spite of these findings, the limitations of this study provide a clear foundation for a broader research agenda. The current dataset's size limits its use for training models from scratch, and ambiguous overlaps were deliberately excluded. However, Maleo-Short serves as a critical foundation for future Indonesian speech research beyond its primary use. It enables the benchmarking of speech enhancement models in realistic noisy environments and provides a rich resource for linguistic analysis of the code-switching and colloquialisms prevalent in modern Indonesian speech.

Accordingly, future work will expand upon this foundation with a clear roadmap. First, we plan to develop "Maleo-in-the-Wild," a significantly larger version of the dataset with a target of over 10 hours of annotated audio. This expanded corpus will include more diverse genres and scenarios with high degrees of speech overlap, creating a more comprehensive resource for both training and evaluation. Second, future work will involve creating finer-grained annotations to explicitly mark language transitions within the dataset. This will support the development of ASR and diarization models specifically architected to be robust against the linguistic complexities of spontaneous Indonesian conversations.

ACKNOWLEDGEMENTS

The authors would like to express their profound gratitude to Mayar.id for the generous funding and comprehensive support that made this research possible. Their contribution was instrumental to the projects success. Sincere appreciation is also extended to Universitas Majalengka, particularly the Faculty of Engineering, for providing the necessary facilities and fostering a supportive academic environment throughout the research process. Finally, the authors are grateful to their colleagues for the insightful discussions and valuable feedback that greatly improved this manuscript.

REFERENCES

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, “A review of speaker diarization: Recent advances with deep learning,” *Comput Speech Lang*, vol. 72, p. 101317, Mar. 2022, doi: 10.1016/j.csl.2021.101317.

[2] K. Kumar, "Speaker Diarization: A Review," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, no. 06, Jun. 2023, doi: 10.55041/IJSREM24075.
[3] D. Soesanto, B. Hartanto, and Melisa, "Meeting Assistant System Berbasis Teknologi Speech-to-Text," *Teknika*, vol. 10, no. 1, pp. 1–7, Jan. 2021, doi: 10.34148/teknika.v10i1.307.
[4] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the Conversation: Speaker Diarisation in the Wild," in *Interspeech 2020*, ISCA: ISCA, Oct. 2020, pp. 299–303. doi: 10.21437/Interspeech.2020-2337.
[5] H. Wijaya, "Teknologi Pengenalan Suara tentang Metode, Bahasa dan Tantangan: Systematic Literature Review," *bit-Tech*, vol. 7, no. 2, pp. 533–544, Dec. 2024, doi: 10.32877/bt.v7i2.1888.
[6] J. Tian *et al.*, "The Royalflush Automatic Speech Diarization and Recognition System for In-Car Multi-Channel Automatic Speech Recognition Challenge," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, IEEE, Apr. 2024, pp. 1–2. doi: 10.1109/ICASSPW62465.2024.10626136.
[7] N. Ryanta *et al.*, "Enhancement and Analysis of Conversational Speech: JSALT 2017," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2018, pp. 5154–5158. doi: 10.1109/ICASSP.2018.8462468.
[8] E. Z. Xu, Z. Song, S. Tsutsui, C. Feng, M. Ye, and M. Z. Shou, "AVA-AVD: Audio-visual Speaker Diarization in the Wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, New York, NY, USA: ACM, Oct. 2022, pp. 3838–3847. doi: 10.1145/3503161.3548027.
[9] K. Kinoshita, M. Delcroix, and N. Tawara, "Advances in Integration of End-to-End Neural and Clustering-Based Diarization for Real Conversational Speech," in *Interspeech 2021*, ISCA: ISCA, Aug. 2021, pp. 3565–3569. doi: 10.21437/Interspeech.2021-1004.
[10] A. Janin *et al.*, "The ICSI meeting project: Resources and research," in *Proceedings of the 2004 ICASSP NIST Meeting Recognition Workshop*, 2004.
[11] D. Mostefa *et al.*, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Lang Resour Eval*, vol. 41, no. 3–4, pp. 389–407, Dec. 2007, doi: 10.1007/s10579-007-9054-4.
[12] W. Kraaij, T. Hain, M. Lincoln, and W. Post, "The AMI meeting corpus," in *Proc. International Conference on Methods and Techniques in Behavioral Research*, 2005, pp. 1–4.
[13] S. Watanabe *et al.*, "CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings," in *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, ISCA: ISCA, May 2020, pp. 1–7. doi: 10.21437/CHiME.2020-1.
[14] T. Liu *et al.*, "MSDWild: Multi-modal Speaker Diarization Dataset in the Wild," in *Interspeech 2022*, ISCA: ISCA, Sep. 2022, pp. 1476–1480. doi: 10.21437/Interspeech.2022-10466.
[15] Z. Yang *et al.*, "Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational(RAMC) Speech Dataset," in *Interspeech 2022*, ISCA: ISCA, Sep. 2022, pp. 1736–1740. doi: 10.21437/Interspeech.2022-729.
[16] F. Yu *et al.*, "M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2022, pp. 6167–6171. doi: 10.1109/ICASSP43922.2022.9746465.
[17] Y. Fu *et al.*, "AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario," in *Interspeech 2021*, ISCA: ISCA, Aug. 2021, pp. 3665–3669. doi: 10.21437/Interspeech.2021-1397.
[18] D. P. Lestari, K. Iwano, and S. Furui, "A Large Vocabulary Continuous Speech Recognition System for Indonesian Language," in *Book name 15th Indonesian Scientific Conference in Japan, Vol., No.*, 2006, pp. 17–22.
[19] U. Thayasivam, T. Gnanenthiram, S. Jeewantha, and U. Jayawickrama, "SiTa - Sinhala and Tamil Speaker Diarization Dataset in the Wild," in *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, K. Sarveswaran, A. Vaidya, B. Krishna Bal, S. Shams, and S. Thapa, Eds., Abu Dhabi, UAE: International Committee on Computational Linguistics, Jan. 2025, pp. 83–92. [Online]. Available: https://aclanthology.org/2025.chipsal-1.8/
[20] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a Professional Framework for Multimodality Research," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, Eds., Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: https://aclanthology.org/L06-1082/
[21] H. Bredin *et al.*, "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020, pp. 7124–7128. doi: 10.1109/ICASSP40776.2020.9052974.
[22] F. Landini, M. Diez, T. Stafylakis, and L. Burget, "DiaPer: End-to-End Neural Diarization With Perceiver-Based Attractors," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 32, pp. 3450–3465, 2024, doi: 10.1109/TASLP.2024.3422818.
[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, in ICML'23. JMLR.org, 2023.
[24] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.
[25] A. C. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Interspeech 2004*, ISCA: ISCA, Oct. 2004, pp. 2765–2768. doi: 10.21437/Interspeech.2004-668.